Using Emotions to Help Detect Fake Tourism Reviews

Mansour Saleh Almansour, BSc, MSc
School of Computing and Communications
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

February, 2026

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography.
Mansour Saleh Almansour

Using Emotions to Help Detect Fake Tourism Reviews

Mansour Saleh Almansour, BSc, MSc.
School of Computing and Communications, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy*. February, 2026

Abstract

Detecting fake reviews has emerged as a urgent critical challenge in NLP, with significant implications for practical applications across various industries. In the tourism sector, fake reviews pose a serious threat by misleading potential customers, damaging the credibility of genuine businesses, and eroding trust in online platforms. The ability to effectively identify and handle deceptive content is essential for safeguarding tourism consumers and maintaining the integrity of related review platforms. While the NLP community has made notable advancements in fake tourism reviews detection, there remains limited understanding of how emotion information can be leveraged to enhance the detection of tourism fake reviews.

This research fills this gap by exploring the integration of emotional information into LLM-based approaches for tourism fake review detection. Specifically, the study evaluates the performance of three prominent large language models (LLMs), including BERT, DistilBERT, and RoBERTa on a tourism-specific dataset, named GeFaRe, both with and without the inclusion of emotion-based information. The dataset used in this research was annotated with emotional labels derived from Plutchik's eight primary emotions, allowing for the incorporation of rich emotional information into the fake reviews classification process. Various experimental configurations were designed to assess the impact of emotional information on the performance of each of the models.

My study provides compelling evidence of the value of integrating emotion information into LLM-based fake tourism review detection models. Across all of my configurations, the inclusion of emotional information led to an improvement of most models' performance, with DistilBERT achieving the highest levels of accuracy up to 0.956. This underscores the potential of emotional information as key information in the context of fake tourism review detection. Furthermore, my findings highlight the importance fine-tuning the LLMs for domain-specific datasets for emotion and fake tourism reviews detection.

Acknowledgements

First and foremost, I am profoundly grateful to Allah the Almighty for granting me the strength, patience, and wisdom to complete this thesis. His countless blessings, and mercy have been my source of perseverance throughout this journey. Without His will and support, none of this would have been possible.

I would like to express my sincere gratitude to my supervisors, Dr. Scott Piao and Prof. Paul Rayson, for their invaluable guidance, continuous encouragement, and constructive feedback throughout the development of this thesis. Their mentorship and expertise have played a crucial role in shaping this research. Their insightful suggestions and continuous support have been instrumental in overcoming challenges and refining all aspects of this work. Their encouragement and belief in my capabilities have motivated me to push the boundaries of my research and strive for excellence. Their patience and dedication have contributed to the quality of this thesis and enhanced my academic and professional growth.

I am deeply thankful to my family, whose continuous support and encouragement have been the foundation of my journey. Their patience, understanding, and complete belief in me have given me the strength to navigate the challenges of this research. From their countless words of encouragement to their unconditional support during moments of doubt, my family has been my greatest source of motivation. This thesis would not have been possible without their sacrifices and steadfast presence in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis investigates the issue of how to improve the detection of fake tourism reviews using emotion information. It explores the integration of emotion-based features into the Large Language Models (LLMs) with the aim of enhancing the identification of fake reviews. At the beginning of this project, a novel, tourism-related dataset was created and annotated with emotion information by adopting a systematic crowdsourcing approach. This annotated dataset serves as a foundation for developing an emotion classification model that is trained on tourism-related reviews. The model is then employed to extract emotion-based features, which are subsequently integrated into fake review detection systems. By systematically evaluating various configurations and architectures, this thesis establishes that incorporating emotion information improves the performance of LLM-based classifiers in terms of distinguishing between truthful reviews and fake ones.

1.1 Research Background

1.1.1 The Importance of Detecting Fake Tourism Reviews

The widespread availability of online tourism reviews has transformed tourists' consumer decision-making, making these reviews a central component when evaluating products and services. Platforms like TripAdvisor and Yelp host millions of reviews, which tourists rely on in order to make informed choices about products and services. However, the increasing prevalence of deceptive, or fake, reviews presents a significant challenge, threatening the reliability of these platforms and, consequently, consumer confidence. In this thesis, all of the review-related data under discussion were drawn from the tourism domain.

Several studies have examined the extent of fake reviews, with Ott et al. (2013) estimating that 1-6% of the reviews provided by the major platforms may be deceptive.

While this might seem like a small percentage, even a 1% deception rate could mean tens of thousands of fraudulent reviews on a platform that hosts millions of user-generated reviews. For example, if a platform hosts 10 million reviews, even a modest 1% rate of deception would equate to 100,000 fake reviews. A further study provides a higher estimate, as Salehi-Esfahani and Ozturk (2018) suggest that about one-third of online reviews could be fake, indicating that this phenomenon represents a far more widespread problem across various platforms. A study on Yelp reviews indicates that over 80% of accounts may be unreliable, with more than 80% of highly-rated businesses subject to review manipulation (Trinh et al., 2020). Another study focusing on e-commerce platforms reveals that approximately 4% of all internet reviews are estimated to be fraudulent, impacting worldwide online purchases by $152 billion annually (Richards et al., 2023). Another study on the prevalence of fake reviews in dietary supplements and healthcare products found that 42.8% of shared content on social media platforms contains inaccurate or fake information (Kannal et al., 2024). Such volumes of deceptive reviews can mislead consumers, distort the perceived product value, and diminish the reliability of genuine feedback.

The landscape of fake review detection has become significantly more challenging with the emergence of generative Artificial Intelligence (AI) technologies. Some studies demonstrated that AI-generated fake reviews are becoming increasingly sophisticated and harder to distinguish from authentic content (Ignat et al., 2024; Huang and Sun, 2024). Gambetti and Han (2023) have shown that ChatGPT can generate fake restaurant reviews that are hard to distinguish from real ones when assessed by both automated detection systems and human evaluators. Notably, AI-generated fake reviews often exhibit better grammar, structure, and even clinical detail than authentic reviews, making some detection methods less effective (Shajalal et al., 2024). Recent investigation reveal that between 6.5% and 16.9% of text in some online reviews platforms may be substantially modified or generated by LLM's, indicating widespread adoption of AI assistance in content creation (Liang et al., 2024).

The impact of fake reviews extends beyond individual purchasing choices. For consumers, a reliance on deceptive reviews can lead to poor buying decisions, resulting in dissatisfaction, financial loss, and decreased trust in the digital platforms. For businesses, fake positive reviews offer an unearned competitive edge, while fake negative reviews can unfairly harm reputations and reduce revenue (Martínez Otero, 2021). Consequently, unchecked fake reviews can destroy consumer confidence, diminishing the credibility of online review platforms over time and disrupting the competitive balance in the marketplace (Paul and Nikolaev, 2021).

Addressing the issue of fake reviews is essential for upholding the integrity of the online feedback systems. In regard to Natural Language Processing (NLP), this task has received significant research attention, as detecting deceptive reviews requires

robust linguistic analysis and classification methods. As online reviews remain a key factor in consumer behaviour, developing effective methods for detecting and filtering fake reviews is critical for maintaining the reliability and utility of these platforms. Advancements in the NLP-based detection techniques, such as machine learning and deep learning models, are an essential components of these efforts. The findings of the current research not only advance our understanding of deceptive language patterns but also enhance the tools available for detecting fake reviews or developing new models for domain-specific text, maintaining the integrity of the online platforms for consumers and businesses alike (Maurya et al., 2023).

1.1.2 Impact of Reviews on Consumer Decisions

Fake reviews have emerged as a significant factor in consumer decision-making, and are capable of affecting individuals' choices as well as impacting the overall market dynamics. These reviews, whether positive or negative, disrupt the authentic consumer feedback cycle and can mislead buyers about the true quality of products and services. It has been found that 95% of customers check the online reviews of previous users in order to gain a clear brand image or review the features of a product. Of these customers, 82% are estimated to be reading and basing their decisions on fake reviews, which reduces the overall transparency of the system (globenewswire, 2022). A study conducted by Nam et al. (2020), found that 84% of the participants trusted online reviews as much as personal recommendations, underscoring the high degree of trust consumers place in these reviews. Additionally, the study found that 74% of consumers are more inclined to trust companies that attract positive reviews, highlighting the role that favourable feedback plays in shaping brand perception. This reliance on reviews as a trusted information source demonstrates the urgent need for robust detection methods to maintain the integrity of the online review platforms and ensure that consumers' decisions are based on genuine feedback. This has prompted an increasing focus within the NLP field on the detection and mitigation of fake reviews, particularly given the high reliance on review platforms associated with sectors such as travel, dining, and retail.

Positive fake reviews are strategically constructed to inflate the reputation of specific products or services, giving them an undeserved competitive advantage. For instance, a recently-launched restaurant might post a series of enthusiastic reviews to project an image of high-quality service and ambiance, drawing in customers under false pretences. These exaggerated reviews may lead consumers to develop unrealistic expectations, resulting in disappointment when the experience fails to match the advertised standards. This breach of trust not only impacts individual consumers but can also lead to a loss of confidence in review platforms as reliable sources of information. Conversely, negative fake reviews aim to undermine competitors by

introducing misleading information that damages their reputation. Such reviews may be deployed to turn potential customers away from a competing business, thus securing an unfair market advantage. For example, a business might engage third-party services to flood the review platforms with negative comments about a specific competitor's products or services, organising a coordinated distort campaign. These actions can significantly harm a business's revenue and reputation, even if the critical comments are entirely baseless.

The impact of fake reviews goes beyond misleading individual consumers, and has broader implications for market stability and competition. For consumers, decisions based on fraudulent reviews may lead to dissatisfaction, financial losses, and decreased trust in digital platforms. For businesses, especially those that rely on authentic customer feedback, fake reviews can distort the public's perception, impacting these firms' market share, profitability, and long-term brand integrity. The importance of reviews is particularly evident in the travel industry, where platforms such as TripAdvisor report that consumer feedback exerts a huge influence on individuals' decision-making. According to data published by TripAdvisor in the power of reviews report [1], 82% of consumers consider reviews highly important for accommodation, 77% for attractions, and 70% for restaurants. This extensive dependence illustrates the significant influence that reviews exert on consumer choices within the travel sector.

This reliance on reviews highlights the need for robust, NLP-driven solutions to detect and mitigate the effects of fake reviews. To address this, previous research has explored various approaches to fake review detection, including sentiment analysis, textual analysis, behavioural analysis, and machine learning models that have been trained on large datasets of both genuine and deceptive reviews. These methods analyse the linguistic patterns, contextual markers, and sentiment dynamics in order to identify discrepancies, that are indicative of deception. The research presented in this thesis attempts to identify fake reviews from a different angle, however, by incorporating emotion information with fine-tuned LLMs which could enhance the accuracy of fake review detection. Implementing effective mechanisms for detecting fake reviews is essential for preserving consumer trust in the online review platforms and promoting fair competition across various industries.

1.1.3 Research Motivation

Detecting fake reviews, however, presents considerable challenges. Fake reviews are deliberately created in such a way that they simulate the language and structure of genuine reviews, often by using subtle linguistic cues that make them difficult to distinguish from authentic reviews. The traditional approaches to fake review

[1] https://www.tripadvisor.com/powerofreviews.pdf, (accessed 6 January 2025)

detection have primarily relied on analysing features such as review length, sentiment, and lexical choices. These features, while informative, are often insufficient in isolation, due to the increasing sophistication of the fake review generators (Wang et al., 2022a). This limitation has encouraged the research for this thesis to embrace more complex features, which is the emotional content expressed in the reviews.

Emotion information is a promising dimension for enhancing the effectiveness of fake review detection systems. Genuine reviews are typically driven by authentic emotional experiences, whether positive, negative, or neutral. In contrast, fake reviews, even when linguistically convincing, often lack genuine emotional coherence and display patterns that may reveal their inauthenticity. By analysing the emotional information contained within reviews, it becomes possible to gain deeper insights into the authenticity of the emotional tone, potentially revealing any deceptive patterns. Emotion detection models, like those developed for this research, can be used to capture the complex emotional expressions in reviews.

Integrating emotion information within LLM based-classifiers for fake tourism review detection represents a new approach. To the best of my knowledge, this thesis represents the first systematic exploration of this approach. Emotion-aware classifiers can combine traditional textual features with emotion-based information, potentially improving the classification accuracy by providing a multi-dimensional view of a review's authenticity. This integration can help to capture the depth of the emotional response conveyed in the text, thereby enhancing the classifier's ability to detect deception. The use of LLM-based classifiers that incorporate emotion information represents a novel contribution to the field of fake review detection. The integration of emotion-based features into the classifiers provides a promising direction for addressing the limitations of the traditional methods, that rely primarily on behavioural and textual features. Therefore, this research aims to explore the efficacy of emotion-informed LLM-based classifiers in relation to fake review detection, contributing to the development of more reliable fake review detectors.

1.2 Research Questions

This study addresses the challenge of detecting fake reviews by examining how emotional content can improve the level of accuracy by applying machine learning methods. The following research questions (RQs) guide the investigation, each targeting a specific aspect of this research.

1. RQ1: How can crowdsourcing annotation help to produce a reliable tourism emotion dataset?

2. RQ2: How are the existing emotion detection tools effective for tourism reviews?

3. RQ3: How can fine-tuned LLMs improve emotion detection?

4. RQ4: How can emotion information help to detect fake tourism reviews?

In this thesis, RQ1 is answered in Chapter 3, which focuses on creating a tourism-related dataset that is annotated with emotion information, leveraging a crowdsourcing technique. RQ2 is examined in Section 4.2, where the performance of the existing emotion detection tools is evaluated using the tourism dataset to assess their applicability and limitations in this domain. RQ3 is addressed in Section 4.3, which focuses on developing and fine-tuning the LLMs to make them a more effective tool for emotion detection in tourism-related reviews. Finally, RQ4 is discussed in Chapter 5, where the integration of emotion-based features into the fake review detection models is designed, implemented and analysed. Through attempting to answer these questions, this research seeks to develop a comprehensive understanding of how emotional information can help to build an accurate model for identifying deceptive content using Plutchik's eight primary emotions, based on tourism-related data.

The contributions of this research address several gaps in the field of fake review classification by enhancing the detection of emotional expressions in tourism-related texts, thereby improving the reliability of the fake review classifiers. By focusing exclusively on reviews authored by real users, the study ensures that all analysed data reflect authentic human emotion and linguistic patterns, avoiding confounding effects from AI-generated content. Consequently, the findings and developed models are grounded in genuine behaviour, offering reliable and contextually relevant insights for detecting deceptive reviews.

1.3 Organisation of the Thesis

This thesis comprises six chapters and is organised as follows:

- **Chapter 1: Introduction**
  The first chapter introduces the research problem and establishes the significance of fake review detection in maintaining the integrity of the online review platforms. It highlights the research motivation, and presents the guiding research questions for this thesis. Additionally, it provides an overview of the structure of this thesis.

- **Chapter 2: Literature Review**
  This chapter reviews the key literature on emotion detection and fake review identification, focusing on the theoretical frameworks, methodologies, and datasets. It explores foundational emotion theories, like Plutchik's Wheel of

Emotions, and surveys the use of crowdsourcing for annotating datasets in NLP. Additionally, it investigates the emotion detection approaches and provides an overview of the current methodologies for identifying fake reviews.

- **Chapter 3: Tourism Review Corpus Construction with Emotion Annotation**
  Annotation quality is pivotal for developing reliable models in NLP tasks, particularly for detecting complex emotion categories in tourism-related reviews. This chapter details the crowdsourcing process that was followed in order to create a labelled dataset of emotion information, based on Plutchik's primary emotions, describing the procedures implemented to ensure a high annotation quality. This chapter discusses the design of the annotation task and provides an analysis of the annotated data.

- **Chapter 4: Emotion Detection**
  This chapter presents the methodologies and techniques that were employed in order to identify the emotional categories within the tourism-related dataset. It evaluates the performance and limitations of the existing emotion detection tools and explores the fine-tuning of LLMs so that they better accommodate emotion classification, focusing on learning rates and data augmentation techniques. The chapter concludes by analysing the classification accuracy of these models with regard to detecting emotions in tourism reviews.

- **Chapter 5: Impact of Emotion Information on Fake Review Detection**
  This chapter presents the core experimental work of this thesis related to fake review detection, focusing on the integration of emotion-based features with LLM-based classifiers for fake tourism review detection. It details the feature engineering process and methodologies for incorporating emotional information into the classification models, highlighting the advantages of emotion-aware models in terms of improving accuracy. Comparative analyses validate the effectiveness of the proposed approach, providing evidence for its role in enhancing fake review detection.

- **Chapter 6: Conclusions and Future Research Directions**
  The final chapter summarises the thesis, revisits the research questions, and discusses the potential future research directions in this area.

Chapter 2

Literature Review

2.1 Introduction

Detecting fake tourism reviews is a highly challenging task; however, analysing the emotions expressed in reviews might prove an effective method for spotting fakes ones. Emotions influence the way in which individuals express themselves, so finding patterns in sentimental speech might assist the determination of the authenticity of a review (Paul and Nikolaev, 2021). Fake reviewers are becoming more skilled, and may purposefully duplicate the emotional patterns that they have observed in genuine reviews. Accordingly, to enhance the trustworthiness of research findings, a comprehensive method that combines various detection approaches is recommended. This chapter provides a comprehensive review of the literature on emotion detection and fake review identification, with particular attention to theoretical models, methodological frameworks, and commonly employed datasets. It examines main theories of emotion, including Plutchik's Wheel of Emotions, and discusses the application of crowdsourcing as a strategy for dataset annotation in NLP. In addition, it analyses existing approaches to emotion detection and presents an overview of current methodologies developed for the identification of fake reviews.

2.2 Emotion Models in Psychology

2.2.1 Plutchik's Wheel of Emotion

The Plutchik Wheel of Emotion (Plutchik, 1980) is a psychological paradigm that depicts the emotions of people and their interactions visually. Plutchik proposed that eight fundamental emotions exist: "joy", "sadness", "anger", "fear", "trust", "disgust", "anticipation", and "surprise", that he defines as being physiologically grounded and as performing an adaptive purpose with regard to human behaviour.

Semeraro et al. (2021) described the Plutchik Wheel of Emotion as a circular graphic, with the eight basic emotions placed equidistantly around the outer circle. Each basic emotion is assigned a colour and linked to the other emotions by lines. Troiano et al. (2023) argued that, according to the paradigm of Robert Plutchik, the emotions on the adjacent lines are more closely connected, whereas contrary emotions are perceived as contrasting or clashing. Plutchik hypothesised that eight secondary emotions exist alongside the eight main ones, which are generated by mixing two adjacent primary emotions. Lee (2019) suggested that secondary emotions, commonly referred to as "*dyads*", are more complicated emotional states, since mixing "joy" with "trust", for example, leads to "love", whereas mixing "fear" with "surprise" results in "awe". The secondary emotions are located on the wheel's outer ring, and are not linked to the fundamental emotions. Figure 2.1 illustrates the Plutchik emotion model, with its axis and intensity, marked by different colours in focused cycles. It also shows the secondary emotions that lie interspersed between the basic emotions.

Plutchik's approach considers the intensity of emotions as well, in order to evaluate the variation of degree and its effects. Each emotion's range of colours on the wheel signify the various degrees of intensity of that emotion. According to Tang et al. (2023), Plutchik contended that mixing fundamental emotions of varying intensities might result in a broader spectrum of feelings. For example, different amounts of delight might produce sensations such as rapture, peace or contentment. According to Sleim (2022), the Plutchik Wheel of Emotion offers a framework for comprehending the complexities of human emotions and their interdependence. It implies that emotions are interrelated and dynamic rather than separate, isolated states. On the contrary, Pak (2020) argued that the paradigm emphasises the concept of emotions blending, transitioning, and transforming into one another, to produce a complex tapestry that includes mental experiences.

The Plutchik Wheel of Emotion has had an impact on many research disciplines and applications. Hoemann et al. (2021) noted that it has enhanced the knowledge of psychological processes and emotional growth, including psychopathology in psychotherapy. The framework has also been used in sociology, anthropology, and information technology studies.Vuijk et al. (2023) stated that the investigation of emotional representation in many different kinds of communication (e.g., literature, art and music) is one real-world use of the Plutchik Wheel of Emotion. The approach may be used by investigators and analysts to detect and categorise the emotions elicited by different media, in order of their psychological influence on humans. Moreover, the previous studies of this framework indicate that the primary emotions of an individual might be combined with other reactions to form a more complex reaction representation. Therefore, a complex relationship exists between human emotions and their very interrelationships.

Although, as Scherer (2022) states, the Plutchik Wheel of Emotion provides a

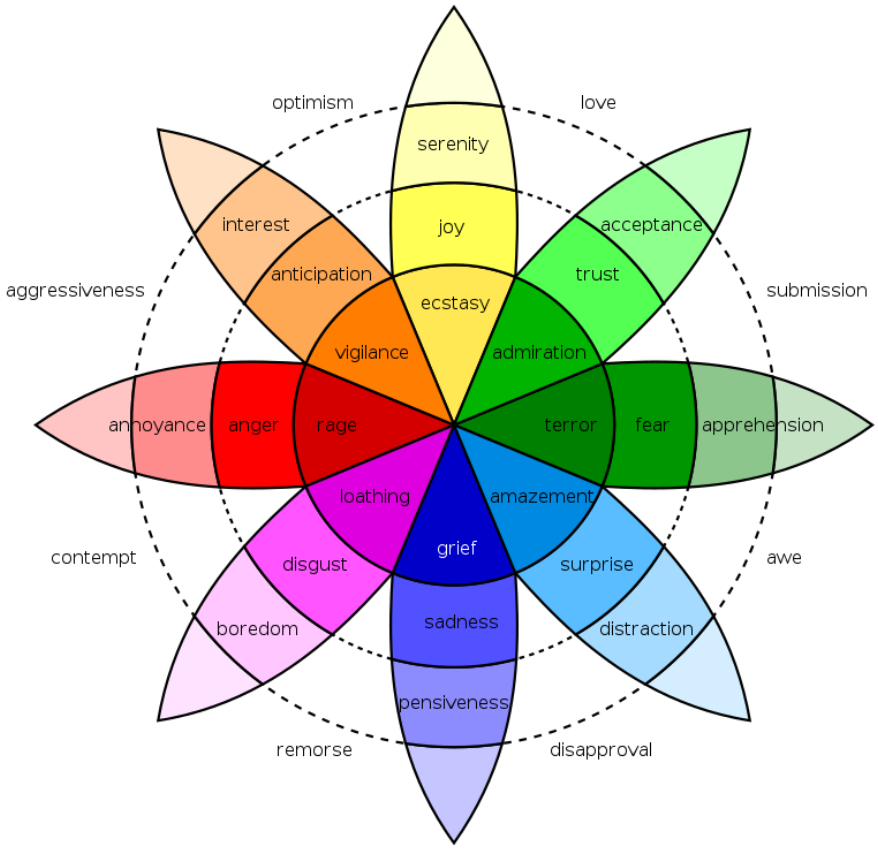


Figure 2.1: Plutchik's Wheel of Emotions (Plutchik, 1980)

useful framework, it is only one of several models that are available in the discipline of emotion psychology. Emotions are complicated, multidimensional phenomena, and many theories emphasise different elements or categories of emotions. In this thesis, the Plutchik Wheel of Emotion was employed as the framework for classifying tourism-related reviews in Chapter 3. This model, with its eight primary emotion categories, was selected due to its well-established psychological foundation and ability to capture a wide range of human emotional experiences. Using this model made it possible to adopt a systematic approach to annotating and analysing the emotions expressed within tourism reviews, enabling the identification of both simple and complex emotional patterns.

2.2.2 Ekman's Basic Emotion

The fundamental emotion model of Ekman (Ekman, 1992) is a psychological theory that was proposed by Paul Ekman, a well-known psychologist noted for his research on feelings and facial expressions. According to Batbaatar et al. (2019), in this paradigm, there are six primary universal emotions, which are expressed and recognised across cultures: "anger", "contempt", "fear", "happiness", "sorrow", and "surprise". These basic feelings, according to Ekman, are intrinsic and physiologically-defined. Sato et al. (2019) describes them as universal, since they are built into all human brains in order to allow constant expression and identification, independent of cultural or societal influences. According to Datz et al. (2019), each fundamental emotion is connected with a unique facial expression, that Ekman refers to as a *"microexpressions"*. Such microexpressions are natural, brief facial movements, that occur whenever an emotion is felt. "Anger", for instance, is symbolised by narrowed eyes, a clenched jaw, and arched brows, while "contentment" is characterised by a grin and elevated cheeks.

According to Yang et al. (2019), Ekman's gestures and emotion study included cross-cultural experiments, during which he showed images of people displaying these fundamental emotions to individuals from other cultures. It was subsequently discovered that people from many cultures could correctly identify the emotions being portrayed, which implies that these emotional manifestations are universal. Wang et al. (2022b) indicated that, although Ekman's approach focuses largely on fundamental emotional facial expressions and emphasises the importance of additional indicators, it also includes body language, the sound of the voice and information, for the exchange of information and detection of emotions. It is worth noting that Ekman's basic emotional model does not claim that these six feelings are the only emotions that humans feel. Ekman subsequently maintained that these are the fundamental, universal emotions, that serve as the foundation for more sophisticated emotional experiences. Moreover, Hayes et al. (2021) noted that, by identifying and categorising fundamental emotions, Ekman's approach indicates that feelings have adaptive purposes. Cowen et al. (2021) pointed out that each basic emotion has a distinct role to play in human behaviour and preservation. "Fear", for instance, activates the fight-or-flight response, which prepares the individual to respond to potential danger, "happiness" encourages connection with others and well-being, whilst 'sorrow" may cause retreat and introspection.

Although the fundamental emotion model has garnered widespread endorsement and acknowledgement in the field of psychological research, Zengilowski et al. (2021) underlined that it has attracted considerable criticism and so argued that emotions are more complicated and nuanced than the model predicts and that culture, including individual variables, might impact emotional experiences and displays. Furthermore, Hayes et al. (2021) noted the ongoing debate over the question of whether emotions are

genuinely universal or whether there may exist cultural differences in their conveyance and perception.

Nonetheless, Cowen et al. (2021) indicated that Ekman's basic emotion paradigm has had a substantial impact on a variety of study and practice domains. It has been employed to construct algorithms for emotion identification in multiple domains, including sociology, anthropology and psychology as well as computer science. Additionally, Nalis and Neidhardt (2023) noted that the framework has been used in the field of clinical psychology to evaluate and address a variety of emotional problems. In outline, it could be predicted that Ekman's basic emotion paradigm offers a useful foundation for comprehending the underlying emotions that underpin human experience. Although, as Ezzameli and Mahersia (2023) argue, it has limits and is still debated, it has made major contributions to the discipline of emotion study and has practical implications for a variety of disciplines.

2.3 Crowdsourcing for Emotion Classification in NLP

Crowdsourcing has become an essential method for emotion classification tasks within NLP, addressing the need for large, diverse datasets that are essential for training robust machine learning (ML) models (Mohammad and Turney, 2013; Buechel and Hahn, 2017). The demand for data that reflect the complex nature of human emotions has surged alongside the advancements in the emotion-driven NLP applications, such as sentiment analysis, social media monitoring, and customer reviews analysis (Kusal et al., 2023; Stajner, 2021). Unlike traditional datasets, which often focus on narrow linguistic interpretations, crowdsourced datasets introduce variability, that mirrors the real-world range of emotional expressions and interpretations, making them particularly valuable for enhancing model generalisation (Öhman et al., 2018). This section discusses the benefits, quality control methods, and challenges associated with crowdsourcing for emotion detection.

2.3.1 Advantages of Crowdsourcing

One of the primary motivations for crowdsourcing in emotion classification is its ability to capture subjective perspectives on emotional content, reflecting the diversity inherent in human emotional expression. The traditional methods of emotion annotation typically involve domain experts who, while often highly accurate, tend to provide interpretations that may not fully encompass the complex emotional variability seen across broader demographics (Snow et al., 2008; Mohammad and Bravo-Marquez, 2017a). Expert-driven annotations are not only costly and time

intensive but also risk imposing a uniform perspective that may not generalise fully across heterogeneous user bases or different languages and cultures (Stajner, 2021). Crowdsourcing, in contrast, makes it possible to leverage *the wisdom of the crowd*, utilising a broad range of annotators with diverse backgrounds to enable the creation of data that are more representative of varied emotional interpretations (Mohammad and Turney, 2013). This diversity can be especially valuable when developing NLP applications intended for multilingual or cross-cultural contexts, as it introduces a richer variety of emotional perceptions and biases into the training datasets (Stoev et al., 2023).

Moreover, crowdsourcing offers a highly scalable, cost-effective solution, allowing the rapid collection of vast amounts of labelled data. This scalability is crucial, given that emotion detection tasks require extensive datasets in order to train models that are capable of accurately identifying subtle or overlapping emotional states (Snow et al., 2008; Kusal et al., 2023). Platforms such as Amazon Mechanical Turk (MTurk) have made it possible to access thousands of annotators within a short time frame, significantly reducing the resource constraints associated with the traditional annotation methods (Hettiachchi et al., 2023). For instance, multiple studies have demonstrated that crowdsourcing can yield high quality data even for complex annotation tasks, provided that strict quality control mechanisms, such as gold standard questions, are in place (Stajner, 2021; Mohammad and Bravo-Marquez, 2017a).

The benefit of crowdsourcing in emotion detection also extends to its ability to capture evolving, context-sensitive expressions of emotions. Emotions conveyed through text can be complex and context-dependent, often requiring a degree of interpretative flexibility that fixed annotation guidelines struggle to capture (Buechel and Hahn, 2017). Crowdsourced annotators, with their diverse individual perspectives, contribute to datasets that more accurately capture these variations, reflecting emotions as they are perceived in real-world scenarios (Öhman et al., 2018). This is particularly useful for applications in dynamic fields such as social media, where the language usage and emotional expressions can shift rapidly over time and may vary widely, based on demographic factors (Stoev et al., 2023).

Furthermore, various studies have also highlighted the importance of crowdsourcing in creating multi label emotion datasets, where annotators can assign more than one emotional label to a single text instance, thereby capturing the complex interaction between different emotions (Anand et al., 2023; Suyal and Singh, 2024). This is especially relevant for real-world applications where users often experience mixed emotions (Buechel and Hahn, 2017; Mohammad and Bravo-Marquez, 2017a). By enabling annotators to label multiple emotions, the crowdsourcing frameworks support the development of NLP models that are better equipped to handle the complex nature of human emotions (Stajner, 2021).

2.3.2 Reliability and Quality Control in Crowdsourcing

Ensuring the quality of crowdsourced data for emotion analysis is a critical concern due to the inherently subjective nature of emotions and the variability between annotators' interpretations. To maintain data reliability, various quality control strategies have been adopted, such as gold standard questions, inter-annotator agreement (IAA) metrics, consensus algorithms, and filtering methods to mitigate low-quality contributions (Snow et al., 2008; Mohammad and Turney, 2013; Mohammad and Bravo-Marquez, 2017a). Gold standard questions are a widely-used technique, where pre-annotated questions with known answers are inserted within the task to evaluate annotator accuracy and consistency. Annotators who fail to meet the accuracy thresholds on these questions are excluded from the annotation task, ensuring a baseline level of reliability (Mohammad and Bravo-Marquez, 2017a).

Another widely-used technique for enhancing data quality is the aggregation of multiple annotations per data item, commonly by applying methods such as majority voting or weighted consensus. Majority voting is an approach that averages out multiple annotators' answers in order to mitigate individual biases and filter out noise, leading to a more representative label for ambiguous emotional expressions (Buechel and Hahn, 2017; Hettiachchi et al., 2023). Other studies also explore consensus algorithms that weigh annotator contributions based on their past performance or consistency with other annotators, thereby refining the aggregation process and improving the dataset's overall reliability (Öhman et al., 2018). This approach aligns well with the subjective nature of emotions, as individual interpretations may vary, but an aggregated label can capture the collective understanding of complex annotation tasks.

Furthermore, IAA metrics, such as Cohen's kappa, Krippendorff's alpha (Passonneau, 2006), and the Plutchik Emotion Agreement (PEA) metric (Desai et al., 2020), are employed to assess the consistency of emotion annotations across different annotators. High levels of agreement indicate reliable annotations, although some level of disagreement is expected during subjective tasks like emotion detection. Lower agreement may suggest the need for task reformulation, additional annotator training, or more refined annotation guidelines (Stajner, 2021).

Studies show that combining multiple quality control strategies can significantly enhance annotation consistency and accuracy, particularly with regard to distinguishing subtle or complex emotions (Buechel and Hahn, 2017; Stoev et al., 2023). For instance, combining gold-standard checks with majority voting and inter annotator agreement metrics provides a multi-layered approach to quality control, filtering out low quality data while allowing for a more robust and reliable dataset. Such methods are particularly effective in crowdsourced emotion annotation, where the inherent subjectivity of emotions necessitates the application of a comprehensive quality assurance framework in order to capture reliable emotional representations

in annotated text (Öhman et al., 2018; Hettiachchi et al., 2023).

In this thesis, a combination of gold standard questions, majority voting, and IAA metrics were utilised to evaluate and refine crowdsourcing workers' annotations, ultimately supporting the development of the tourism-related dataset that is presented in Chapter 3. Gold standard questions served as a quality control mechanism, ensuring that the contributors demonstrated the necessary understanding of the annotation task both before and during the process. Majority voting was employed as a strategy for consolidating multiple annotations for each data instance, providing a reliable consensus by resolving conflicts and minimising individual biases. Additionally, IAA metrics were applied to measure the consistency and reliability of the annotations, highlighting the overall agreement among the workers.

2.3.3 Challenges related to Crowdsourcing

Despite the advantages of employing crowdsourcing for emotion detection, several notable challenges arise due to the subjective, complexity, and context-dependent nature of emotions. A primary issue is the inherent subjectivity of emotional interpretation, which varies widely across individuals, due to personal, cultural, and linguistic influences. These factors affect how annotators recognise and label emotions, leading to potential inconsistencies in the data. For example, different cultural contexts may interpret the same emotional expression in divergent ways, impacting the labelling process and reducing dataset uniformity (Kusal et al., 2023). Emotions are also frequently expressed in complex, overlapping states, where a single text can convey multiple emotions, such as "anger" mixed with "sadness". This overlap makes it challenging for annotators to distinguish and accurately label multiple emotions within a single instance (Buechel and Hahn, 2017; Suyal and Singh, 2024). These complex emotional differences are challenging to capture, especially in crowdsourcing environments, where the annotators may lack specialised training in emotional analysis (Stoev et al., 2023).

Another significant challenge related to crowdsourced emotion annotation is the context-dependent nature of emotional expressions. Emotional content often relies on specific background information that may be unknown to the annotators, particularly when they are working with texts associated with different languages or cultures. With insufficient contextual knowledge, annotators may miss subtleties in the emotional expressions, leading to inaccuracies in the labelled data (Öhman et al., 2018). For instance, sarcasm or irony, which can alter the identified emotion of a text, may be difficult for annotators to recognise, especially if they are unfamiliar with the linguistic or cultural context in which the emotion is conveyed (Stoev et al., 2023). Studies have shown that annotators working outside their native language or cultural context can struggle with accuracy, underscoring the need for a careful consideration

of any cultural differences in emotion annotation tasks (Stajner, 2021).

To address these issues, various strategies for enhancing annotation quality in crowdsourced emotion datasets have been proposed. One approach involves improving the task instructions and providing illustrative examples, which has been shown significantly to improve annotation consistency and accuracy across diverse annotator groups (Stajner, 2021). By clearly defining the emotion categories, specifying the annotation criteria, and offering example emotion categories, the degree of subjectivity in annotations can be reduced. Another effective strategy is to use iterative feedback and training mechanisms, which allow annotators to refine their understanding of emotional distinctions over time, improving the overall annotation reliability.

Furthermore, multi-label annotation, a method that allows annotators to assign multiple emotion labels to a single text, is helpful for capturing the complex, often overlapping emotional states conveyed in natural language (Anand et al., 2023). This approach enhances the ability of emotion-annotated datasets to reflect complex emotional expressions that cannot be represented adequately by a single label. For example, annotators can tag combinations of emotions, such as 'joy" and "surprise" or "anger" and "disgust", within a single sentence, reflecting the complex emotional experiences found in real-world language. Multi-label annotations are particularly valuable for capturing the subtle relations between emotions, which are essential for accurate emotion recognition in more complex texts, such as social media posts. In Chapter 3, the crowdsourcing annotation task was carefully designed to ensure clarity and reliability by providing workers with detailed instructions and illustrative examples of each emotion class. Additionally, to address the inherent complexity of annotating the emotions expressed in text, workers were allowed to assign up to two emotion labels per sentence. This multi-label annotation approach accommodated the possibility of mixed or overlapping emotions featuring in tourism-related reviews.

2.4 Emotion Detection in NLP

Emotion detection in NLP is a growing field of interest and research (Das et al., 2022). This subject focuses on how computational techniques might identify and interpret human emotions as expressed in text. This field is interdisciplinary in nature, as it combines elements of linguistics, computer science and ML to solve complex problems. The ML techniques, mainly linked to deep learning, are frequently employed in NLP to detect emotion (Arbieu et al., 2021). Systems built for emotion detection are often trained on large datasets, where the written content has been labelled according to the emotions that it may convey (Kusal et al., 2023). Sentiment analysis is a well-established application of emotion detection, that seeks to clarify the sentiment, whether "positive", "negative", or "neutral", underlying a piece of text. However, emotion detection can delve more deeply and attempt to discern more

specific emotions such as "joy", "anger", "surprise", and "sadness" (Pradhan et al., 2023), which emotion classes, among others, were discussed above in Section 2.2.

Emotion detection has significant potential applications. Businesses might use this technology to gauge customer sentiment as expressed in reviews or social media comments, which would help to inform their customer service and marketing strategies (Liu, 2020). Text written by humans is ingrained with emotion. Which is expressed through both explicit language and more subtle indicators, such as word choice and syntax. This makes emotion detection an inherently challenging task (Guo, 2022). Sarcasm, for example, would be especially difficult for an algorithm to discern, as it involves understanding the context and often entail expressing something that is contrary to one's intended meaning (Poria et al., 2019b). Furthermore, cultural differences may also affect how emotions are expressed and perceived in text (Schuller and Batliner, 2013). What is considered a positive expression in one culture might be neutral or even negative in another. However, advances in deep learning and the availability of larger, more diverse training datasets are improving the performance of the emotion detection systems (Gosai et al., 2018). The following two subsections will examine the Lexicon-based approaches, and Machine-learning-based approaches as methods for emotion detection in NLP.

2.4.1 Lexicon-Based Approaches

Lexicon-based approaches are one of the primary techniques used in emotion detection in NLP. In this context, a lexicon is defined as a list or dictionary of words and phrases, each of which is associated with their corresponding emotional values (Rabeya et al., 2017). These values may represent different types of emotions or sentiments, like "joy", "sadness", "anger", or more generally "positive", "negative", and "neutral" sentiments (Pradhan et al., 2023). One strength of lexicon-based approaches is their simplicity. Unlike ML methods, lexicon-based techniques do not need training on large-labelled datasets. This means that lexicon-based systems can be implemented relatively quickly and might be particularly useful in scenarios where training data are scarce (Bharti et al., 2022). Additionally, lexicon-based approaches could give explicit explanations for their predictions. This contrasts with deep learning models which are often criticised as *"black boxes"* because their decision-making processes can be opaque, as indicted by Rudin (2019). By using a lexicon, one can trace back which words in a text contributed to the identified emotion, which in turn could be valuable in applications where interpretability is essential.

Lexicon-based approaches also suffer from significant limitations (Khanpour and Caragea, 2018), the most prominent of which is their reliance on predefined emotional values for words or phrases. Language is fluid and the emotional connotation of a word might change based on the context in which it is used. For instance, the

word *"bright"* might evoke a sense of "joy" in the sentence *"It is a bright sunny day"* but could might "sadness" in the phrase *"He was the bright, lost star of his generation"*. Lexicon-based approaches typically fail to account for such context-dependent emotional connotations. Furthermore, building a comprehensive lexicon may be a time-consuming, laborious task, considering the diversity of human language (Zad et al., 2021). The lexicon might also need to be updated regularly both to keep pace with the changes in language use over time and also to incorporate new words or phrases that arise.

Moreover, lexicon-based methods could struggle with linguistic phenomena like sarcasm or irony, where the emotional connotation of a sentence is not determined by the emotions associated with individual words (Schuller and Batliner, 2013). In these cases, understanding the broader context and the underlying intention of the speaker is critical, yet beyond the scope of standard lexicon-based approaches. In addition, lexicon-based methods might have difficulty dealing with cultural differences in language use (Kamal et al., 2019). Words and phrases may carry different emotional connotations in different cultures, so a lexicon developed for one cultural context may not be applicable in another. In contrast, the ML models, when given sufficient data, have the potential to learn this cultural aspect implicitly. Thus, it can be deduced that, while lexicon-based approaches for emotion detection in NLP offer simplicity and interpretability, their limitations are considerable. To overcome these issues, lexicon-based methods might be combined with other techniques, such as ML, to leverage the respective strengths of each approach. Despite the challenges, lexicon-based approaches remain a valuable tool in the NLP toolkit for emotion detection.

2.4.2 Machine-Learning-Based Approaches

Machine-learning-based approaches represent an alternative to lexicon-based techniques for emotion detection in NLP (Pugliese et al., 2021). By learning from large datasets of labelled examples, ML algorithms have the potential to discern complex patterns in data and make accurate predictions. In the field of emotion detection, these algorithms are trained to associate specific textual features with different emotional states (Pugliese et al., 2021). Different ML techniques are utilised within the field of emotion detection and each one has its unique strengths and weaknesses (Khan et al., 2022). On the one hand, the traditional ML methods, like Support Vector Machines (SVMs) and Naive Bayes classifiers, might be used to identify emotions from textual data (Afshari et al., 2022). These methods typically require feature engineering, which is a process whereby experts identify and extract informative features from text, although designing and extracting these features requires significant expertise and can be time-consuming.

On the other hand, deep learning methods, such as Recurrent Neural Networks

(RNNs), and LLMS, like BERT, would automatically learn features from the data, which in turn would remove the need for feature engineering (Dhruv and Naskar, 2020; Spoon et al., 2021). These models have achieved state-of-the-art results for many NLP tasks, including emotion detection. They can effectively capture the context from textual data, which allows them to understand how the meaning of words may change based on the surrounding text (Dhruv and Naskar, 2020; Spoon et al., 2021). However, while these methods are powerful, they also suffer from limitations. Deep learning models, for example, are notorious for requiring large amounts of labelled data (Yang et al., 2021). In scenarios where data are scarce, these models may perform poorly compared to their simpler counterparts. Additionally, while they may be capable of capturing complex patterns, they are often considered difficult to interpret in terms of their decision-making processes (Chan et al., 2020). ML techniques also struggle to cope with imbalances in data (Tarekegn et al., 2021). If certain emotions are underrepresented in the training data, the model might struggle to recognise these emotions in new, previously unseen data. Techniques like over-sampling, under-sampling or synthetic minority over-sampling could be used to address this issue (Rodríguez et al., 2021).

Furthermore, ML models are sensitive to the quality of the training data (Nguyen et al., 2021). If the labelled data contain errors or biases, the model could learn and perpetuate these mistakes. Moreover, similar to the lexicon-based methods, the ML approaches face challenges related to dealing with linguistic phenomena linked to connotative meaning as well as cultural variations in language use (Xu et al., 2021). The deep learning models could potentially learn to understand these complexities, given sufficient data (Xu et al., 2021). BERT, alongside other LLMs including DistilBERT and RoBERTa, was utilised for emotion detection for tourism related-text, as discussed in Chapter 4. Furthermore, given the challenge of data scarcity, multiple oversampling techniques were employed to address the class imbalances and improve the model performance. By augmenting the dataset, these oversampling methods ensured that the models had sufficient data to allow them to detect emotion categories, enhancing the robustness of the emotion detection model.

2.4.3 Hybrid-Based Approaches

Hybrid-based approaches in emotion detection in NLP aim to combine the strengths of different techniques to improve the accuracy and reliability of the emotion detection. They integrate ML techniques, such as supervised and unsupervised learning, with lexicon-based methods, in order to benefit from the unique advantages of each approach. Hybrid-based approaches could offset the respective limitations of the individual methods. As such, as lexicon-based methods might struggle with context-dependent meanings, ML methods are designed to understand these complexities.

Likewise, the interpretability issue associated with machine learning models might be mitigated by the transparency of the lexicon-based methods. Furthermore, hybrid methods may enhance the robustness of the emotion detection systems. By relying on more than one technique, these systems may be able to maintain reasonable performance even if one method fails or underperforms in a particular scenario. However, the hybrid-based approaches are masked by the complexity associated with integrating different techniques. Each method has its own unique requirements and principles, so balancing these could be a complex task.

Further, maintaining and updating such systems could be more resource-intensive compared to using simpler, single-method systems (Anuja et al., 2022). It has been argued that hybrid methods might inherit the weaknesses of their constituent techniques. That is to say, the need for labelled data, a challenge for machine learning, and the requirement of comprehensive emotion lexicons, a concern for lexicon-based methods, might still be present in hybrid-based systems (Wang et al., 2020). Moreover, interpreting the results of hybrid systems could be more difficult compared to when using single-method systems. Understanding how different methods contribute to the final decision and analysing these contributions might require considerable expertise and effort (Graterol et al., 2021). Consequently, hybrid-based approaches represent a promising direction in the field of emotion detection based on NLP. They offer the potential to leverage the strengths and offset the weaknesses of the individual techniques. However, the complexity of integrating different methods, the potential for inherited weaknesses, and the challenges of interpretation represent significant obstacles, that need to be addressed in order to realise the full potential of these methods (Saffar et al., 2023).

2.4.4 Publicly Available Datasets for Emotion Detection

There are a few structured annotated datasets that have been created for emotion detection in NLP, that provide researchers with structured, diverse data which they may use to train and test their models. There follows an overview of five major datasets that are commonly used during emotion detection tasks:

ISEAR: The International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Wallbott, 1994) is recognised as one of the pioneering datasets that are dedicated to the study and detection of emotions (Balahur et al., 2011). This extensive dataset includes seven distinct emotion categories, including "joy", "sadness", "fear", "anger", "guilt", "disgust", and "shame". These categories were identified through comprehensive, multi-cultural questionnaire surveys that were conducted across 37 different countries. In these studies, 3,000 participants were asked to complete detailed questionnaires about their personal experiences and emotional reactions to various events. The outcome of these surveys was a rich dataset,

consisting of 7,665 sentences, each labelled according to the corresponding emotion.

The ISEAR dataset has become a pioneer dataset that researchers have used to compare the efficiency of several NLP approaches. It presents a general description of the entire range of emotions and the considerable variety of reactions, making it an essential tool for studying various aspects of the human emotional sphere (Bharti et al., 2022). This dataset has been instrumental in advancing the research in the field of emotion detection and analysis, providing valuable insights into how people from diverse cultural backgrounds experience and express their emotions.

SemEval: The SemEval (Semantic Evaluations) dataset, designed specifically for emotion detection, comprises a diverse collection of news headlines in both Arabic and English (Rosenthal et al., 2017). These headlines were sourced from reputable platforms, such as the BBC, CNN, and Google News, as well as major newspapers and other trusted sites. This diversity ensures a broad coverage of topics and linguistic styles, enhancing the robustness and generalisability of the models that are trained on it. The SemEval dataset contains 1,250 instances, each labelled with one of the six basic emotional categories proposed by Ekman (happiness, sadness, anger, fear, surprise and disgust).

The dataset is also noted for its rich annotations. Each entry is annotated with regard to various semantic attributes, supporting tasks such as emotion detection. It was assembled as part of the annual SemEval competition, which is a prominent event in the field of NLP. This competition aims to advance the state-of-the-art of semantic analysis by fostering the development and evaluation of new techniques and methodologies. The inclusion of this dataset in the competition provides a valuable resource for improving and refining the emotion detection systems, thereby contributing to the overall progress of semantic understanding and analysis within NLP.

EmoInt: The WASSA-2017 (Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis) dataset, known as EmoInt, was created to address the absence of a resource dedicated to predicting emotion intensity in social media text, particularly in tweets (Mohammad and Bravo-Marquez, 2017b). This dataset was developed for the shared task on emotion intensity that was organised as part of the WASSA-2017 workshop, and was specifically designed to advance the research on understanding and quantifying the intensity of emotions that are expressed in textual data. Unlike previous datasets, EmoInt provides labels for the emotions and their intensity scores, reflecting the strength of the emotions felt, as indicated by the tweets.

The WASSA-2017 EmoInt dataset provides fine-grained annotations, indicating the intensity of emotions. These emotions include "anger", "fear", "joy", and "sadness", with intensity scores ranging from 0 (*no emotion*) to 1 (*maximum emotion*). The intensity scores in the dataset were obtained through crowdsourcing,

ensuring a diverse, representative set of judgments. Multiple annotators evaluated each tweet, and the final intensity score for each emotion was computed as the average of these individual ratings.

Daily Dialog: The Daily Dialog dataset was specifically designed to support research on dialogue systems (Li et al., 2017). It is used for various tasks, including dialogue generation and emotion detection. Daily Dialog comprises conversational data, with dialogues that reflect everyday communication. The conversations are manually drafted, ensuring coherence and naturalness, which is crucial for training realistic dialogue systems. The dataset includes dialogues related to various domains, such as daily life, education, relationships, and work. It was built by creating and annotating dialogues, and the resulting models directly reflect how people talk to one another.

The Daily Dialog dataset contains over 13,000 sentences, each of which is linked to one of the seven emotional categories. Positive emotions include "happiness", whereas negative emotions include "anger", "disgust", "fear", and "sadness", with two further categories: "surprise" and "neutral". The focus on dialogues in this dataset makes it ideal for studying emotions within a conversational context, where emotions can shift between one message and the next, and are often interconnected with the speakers' turn-taking. Due to the large body of realistic, conversation data, the Daily Dialog dataset is used to train and test the emotion detection models. The dataset's wide range of emotions and authentic conversational style makes it particularly valuable for developing systems that can accurately identify and respond to the emotions of respondents (Li et al., 2017).

MELD: The MELD (Multimodal EmotionLines Dataset) dataset is a noteworthy resource in the field of multimodal emotion analysis (Poria et al., 2019a). It was designed to aid the development and evaluation of models that are capable of understanding and predicting emotions based on multimodal data, including text, audio, and visual information. MELD extends the EmotionLines dataset (Hsu et al., 2018) by incorporating multimodal features, making it particularly useful when building models to draw emotions from textual, visual and audio data streams, which is essential for conversational models.

MELD includes text, audio and visual data for each dialogue, providing a comprehensive dataset for studying emotion recognition across multiple modalities. The dataset consists of multiparty conversations that have been extracted from the US TV show *"Friends"*. Each conversation is segmented into utterances, with each utterance annotated for emotion and sentiment. This conversational context is essential for understanding how emotions evolve and interact over the course of a dialogue, providing valuable insights for dialogue systems and emotion recognition models.

The dataset consists of more than 1,400 dialogues and approximately 13,000

utterances, that are annotated with seven emotion categories: "anger", "disgust", "sadness", "joy", "surprise", "fear", and "neutral". Additionally, it includes sentiment labels: "positive", "negative", and "neutral", enabling a detailed analysis of both the emotional and sentiment dynamics of conversations. This fine-grained annotation makes it a rich resource for studying complex emotional interactions. The dialogues in MELD cover a wide range of everyday scenarios and interactions, making this dataset highly realistic and diverse. This diversity is crucial for training models that can generalise well to various NLP applications.

2.4.5 Existing studies for Emotion Detection

The reported accuracy for text-based emotion detection shows remarkable variability, influenced by the methodological approach, dataset characteristics, and evaluation metrics employed. Traditional machine learning methods, particularly Support Vector Machines (SVM), generally demonstrate moderate performance, with results varying across studies. For instance, the emotion classification system for English tweets at the SemEval-2018 shared task on Affect in Tweets achieved an F-score of 0.493 using an SVM-based model (De Bruyne et al., 2018). More recent applications of SVM have reported improved results, with F-scores reaching up to 0.75 (del Arco et al., 2020). These advancements can be attributed to enhanced feature engineering strategies, such as the use of word embeddings and linguistic features, as well as more refined preprocessing techniques that contribute to better handling of noisy social media text.

The adoption of deep learning architectures has led to notable improvements in emotion detection accuracy, with Long Short-Term Memory (LSTM) networks consistently achieving F-scores in the range of 0.72 to 0.75 across different datasets. For example, contextual emotion detection on Twitter data using deep learning methods reported an F-score of 0.7185 (Rashid et al., 2020). Similarly, the EmoDet2 system, which integrates BERT with a BiLSTM classifier, attained an F-score of 0.748 on shared task datasets (Al-Omari et al., 2020). These results highlight the capacity of deep learning approaches to capture contextual dependencies and semantic nuances in text.

The development of transformer-based models marks the recent advancement in emotion detection research. For instance, a BERT model combined with a Broad Learning System (BLS) achieved an F-score of 0.7332 on the SemEval-2019 and SMP2020-EWECT tasks (Peng et al., 2021). Other transformer architectures, such as ALBERT, DistilBERT, and RoBERTa, have also demonstrated strong performance, with DistilBERT reaching an accuracy of 0.792 in comparative evaluations across multiple transformer models on the IEMOCAP dataset (Ali et al., 2024). These results underscore the capacity of transformer-based models to leverage contextualised embeddings and attention mechanisms for improved classification outcomes. Table 2.1

provides an overview of some of the existing studies for text-based emotion recognition, outlining the datasets employed, the classification methods applied, and the reported evaluation metrics.

Reference	**Dataset**	**Method**	**Performance Measure**
De Bruyne et al. (2018)	Twitter	SVM	F-score 0.493
del Arco et al. (2020)	Twitter	SVM	F-score 0.75
Al-Omari et al. (2020)	EmoContext	BiLSTM	F-score 0.748
Rashid et al. (2020)	EmoContext	LSTM	F-score 0.7185
Peng et al. (2021)	EmoContext and SMP2020-EWECT tasks	BERT	F-score 0.7332
Ilyas et al. (2023)	YouTube and Twitter	CNN	F-score 0.747
Ali et al. (2024)	IEMOCAP	DistilBERT	Accuracy 0.792
Reddy and Palaniswamy (2024)	MELD	ResNet, LSTM and RoBERTa	Accuracy 0.5869
Vora and Mehta (2024)	EmoInt and ISEAR	ANN	F-scores 0.7514 for EmoInt and 0.5175 for ISEAR
Pithava et al. (2024)	Facebook, Instagram and Twitter	KNN	Accuracy 0.60
Reviriego and Raynova (2024)	Twitter	NAÏVE BAYES	F-score 0.748
Ibnath et al. (2025)	Hugging Face	KNN	F-score 0.7693

Table 2.1: Overview of the studies for emotion detection

2.5 Fake Review Detection

The growing influence of online reviews on consumer behaviour has made the detection of fake reviews a significant research focus. Fake reviews are intentionally deceptive

opinions that are posted online with the goal of misleading consumers. Therefore, it has become important to identify fake reviews in order to safeguard the integrity of the numerous online sites. Detecting these fake reviews is challenging, due to their similarity to genuine reviews. Many approaches have been suggested for detecting fake reviews, that fall into two main categories: the feature-based approach and the machine learning approach. There follows an overview of these approaches to fake review detection.

2.5.1 Feature-Based Approach

The feature-based approach to fake review detection involves analysing specific features of reviews to distinguish between genuine and deceptive content. This approach typically encompasses both behavioural features and textual features. Behavioural features focus on the actions and patterns associated with the reviewer, such as the timing and frequency of the reviews, the consistency of the ratings, and the historical behaviour of the reviewer. In contrast, textual features analyse the content of the review itself, examining aspects like word choice, sentence structure, sentiment, and overall coherence. Therefore, it is important to identify and compare the features associated with fake and real reviews, respectively, in order to clarify the nature of the reviews and help to develop appropriate models for fake review detection (Anoop et al., 2019).

2.5.1.1 Behavioural Features

The key features associated with employing the behavioural characteristic to detect fake reviews are statistics-driven, focusing on the activity and behaviour of the reviewer rather than the content of the review itself. These features are premised on the assumption that the behaviour exhibited by fake reviewers differs significantly from that displayed by genuine ones. By analysing this behaviour, certain patterns can be identified that are often characterised by an abundance of fraudulent activities. Common behavioural features that assist fake review detection include:

Review Frequency and Timing: One of the most revealing behavioural features is the frequency and timing of the reviews. Genuine reviewers typically post reviews infrequently, at intervals that reflect normal purchasing and usage patterns. In contrast, fake reviewers often exhibit abnormal posting behaviour, such as submitting multiple reviews in quick succession or within a short time interval (Zhang et al., 2023c). Therefore, it would appear that the frequency at which a reviewer submits comments facilitates the detection of fake reviews. Fake reviewers often operate rapidly, possibly as part of a coordinated campaign to influence product ratings quickly. This behaviour is particularly visible in situations where reviews are time-sensitive, such as during product launches or promotions (Barbado et al., 2019).

Time stamping can also help to identify fake reviews. For instance, if many positive reviews might flood in at a particular time, such as the period when a particular product is being promoted, then it may be assumed that these reviews are fake (Xie et al., 2012). When several reviewers post comments simultaneously, frequently using the same language and phrases, this phenomenon is again indicated. In contrast, genuine reviewers' comments tend to be posted at longer intervals (Zhang et al., 2023c) and also during regular hours, reflecting typical daily routines. Conversely, fake reviews might be posted at unusual times, possibly due to automated systems or reviewers working across different time zones. Finally, while frequency alone cannot be used to determine whether a reviewer is genuine or not, it can be used as one factor in conjunction with other indicators to help to identify potential fake reviews.

Reviewer History: The history of a reviewer, including their past behaviour and the consistency of their reviews, is another crucial behavioural feature. A user is a valuable customer, and their reviewing history can provide valuable insights into their personality (Duma et al., 2024). Genuine reviewers often focus on specific product categories, that are aligned with their interests or needs. In contrast, fake reviewers may review a wide range of unrelated products, which may signal a lack of genuine interest. Mukherjee et al. (2016) highlight that fake reviewers might be paid to post positive reviews across various categories, which happens less frequently among genuine users. Conversely, genuine reviewers normally leave feedback using balanced language, which includes a mixture of both positive and negative comments.

The reputation of the reviewer, often indicated by metrics such as helpfulness votes or badges, can also be an important feature. Genuine reviewers tend to accumulate positive feedback over a long period of time, while fake reviewers might have a lower reputation or even receive no feedback at all (Mukherjee et al., 2013b). Incorporating reviewer reputation into the detection models can provide additional context for assessing the credibility of a review. Original reviewers can also be identified by their long-standing accounts on the review platform. These users tend to exhibit an orderly pattern of reviewing behaviour over time, whereas fake reviewers often create new accounts specifically for the purpose of posting reviews (Salehan and Kim, 2016). The evaluation of the review distribution patterns, account longevity, and reviewing consistency provides valuable insights into the degree of authenticity of user feedback on online platforms. This nuanced understanding can inform strategic initiatives aimed at detecting fake reviews.

Reviewer Network: The reviewer network plays an important role in the detection of fake reviews by analysing the relationships and interactions among the reviewers on a review platform. Unlike the independent evaluation of individual reviews, reviewer networks focus on the connections between reviewers, such as shared behaviour, mutual reviews, and co-reviewing patterns. This approach leverages the idea that fake reviewers rarely act alone, but are frequently part of a coordinated group

or network that collectively works to manipulate the perception of a specific product or service (Mukherjee et al., 2012). By mapping out these connections, it is possible to identify clusters of reviewers who are more likely to be engaged in fraudulent activities, such as posting reviews of the same set of items within a similar timeframe or repeatedly supporting each other's reviews through offering likes, comments, or up-votes.

In contrast, Paul and Nikolaev (2021) have shown that authentic reviewers tend to be more communicative, engaging with other users through offering comments on or ratings of others' reviews. In contrast, fake reviewers often fail to demonstrate this level of interaction, suggesting a lack of genuine engagement with the online community. Furthermore, analysing the relational connections between reviewers can provide valuable insights into their degree of authenticity. Studies have demonstrated that reviewers who have strong social ties within an online platform are more likely to be genuine (Barbado et al., 2019). Overall, reviewer networks offer a powerful tool for understanding the social dynamics of the online review ecosystems, and provide valuable insights into the ways in which fake reviewers organise themselves and collaborate in order to achieve their goals. When combined with other detection techniques, reviewer network analysis can improve the detection of fake reviews, helping to maintain the integrity and trustworthiness of the online review platforms.

Review Patterns: The similar format of different reviews can indicate fraudulent activity. Fake reviewers often employ templates or phrases that are repeated across several reviews, making it possible to detect such patterns (Saleh Nagi Alsubari, 2022). Analysing repetitive wording in reviews makes it possible to develop methods for identifying fake reviews and distinguishing them from genuine ones. Additionally, genuine reviewers often focus on specific product categories, that are aligned with their interests or needs. In contrast, fake reviewers may review a wide range of unrelated products, which may signal a lack of genuine interest. Hajek et al. (2023) highlight that fake reviewers might be paid to post positive reviews across various categories, which is a less common practice among genuine users.

Rating Patterns: Rating patterns offer another layer of behavioural insight. Genuine users usually give ratings that are aligned with their actual experiences, resulting in a more varied distribution of scores. In contrast, fake reviewers might give consistently extreme ratings (either very high or very low), depending on whether they are trying to promote or disparage a product (Salminen et al., 2022; Manaskasemsak et al., 2023). These extreme rating patterns are often coupled with short, generic comments, which fail to provide any substantive feedback but work to manipulate the product's overall rating (Wang et al., 2011). The detection of such patterns can be enhanced by examining the divergence between a user's ratings and the average rating of a product, where considerable inconsistencies might signal the presence of fraudulent behaviour.

2.5.1.2 Textual Features

In the realm of NLP, the analysis of review texts has emerged as a crucial approach to detecting and mitigating online deception. By examining the textual features of reviews, it is possible to identify patterns or characteristics that are indicative of fake content. These features focus on various aspects of the textual content, including the style, sentiment, and punctuation. Each of these aspects provides a unique perspective and helps to identify different aspects of fake reviews, enabling a more comprehensive approach to detection. The following textual features are commonly used in fake review detection:

Linguistic Style: The linguistic style of a review can provide valuable insights into its authenticity, making it a crucial aspect to consider for detecting fake reviews. Authentic reviews are often characterised by variations in vocabulary and sentence structure, whereas fake reviews tend to exhibit more uniformity in their choice of words (Le et al., 2022). This is because real reviewers typically employ a range of linguistic phrases to convey their opinions and experiences, resulting in a richer, more nuanced text, which reflects the diverse backgrounds and unique expressions of real users. In contrast, fake reviewers often rely on simplistic language and repetitive phrases, which can be identified through quantitative analysis. Another measure that has been shown to be effective in determining the textual richness of reviews is the type token ratio (Le et al., 2022). This metric calculates the proportion of distinct words in the total words of a text, providing an indication of its vocabulary density. By applying this measure, fake reviews that contain overly simplistic language and limited vocabulary can be identified.

A key indicator of deception in linguistic style is the use of extreme language without offering any supporting evidence or specific details. Deceptive reviews often employ superlative adjectives (e.g., "the best", "incredible", and "unbelievable") or adverbs (e.g., "absolutely" and "extremely") to amplify their emotional appeal, as they aim to manipulate the reader's perception through making exaggerated claims (Ott et al., 2011). Genuine reviews, on the other hand, tend to offer more balanced language, providing both advantages and disadvantages and specific narratives that provide credibility to the reviewer's experience. Additionally, fake reviews may display an inconsistent or unnatural tone, where the style is misaligned with the claimed identity of the reviewer. In addition, Ott et al. (2013) focus on detecting deceptive opinion spam reviews, that have been deliberately written to appear authentic but are intended to manipulate the readers' perceptions. By employing various analytical approaches, including text categorisation, psycho-linguistic deception detection, and genre identification, they explore the linguistic features that distinguish fake reviews from genuine ones. These features may include sentiment analysis, and lexical choices that are unique to deceptive content, which the NLP models can exploit in order to improve the detection accuracy.

Furthermore, Qiao and Rui (2023) found that fake reviews are often easier to read due to their simpler syntax and lexicon, since fake reviewers tend to avoid complex sentence structures and nuanced vocabulary, opting instead for a more straightforward writing style. A quantitative assessment of text difficulty can be performed using metrics, such as the Flesch-Kincaid score (Shetty, 2019), which provides an estimate of a text's readability based on factors like its sentence length and word complexity. These findings highlight the importance of considering linguistic features for fake review detection. By examining the style, vocabulary, and syntax of reviews, effective methods for detecting fake content can be developed.

Sentiment Analysis: Sentiment analysis is an important factor for detecting fake reviews, as it involves the computational identification and categorisation of the emotions that are being expressed in the text. Fake reviews often exhibit extreme sentiment (either excessively positive or excessively negative), designed to manipulate the public's perception (Jindal and Liu, 2008; Ott et al., 2011). By detecting these extremes of sentiment, sentiment analysis can help to identify reviews that are misaligned with typical user experiences, suggesting the possibility of deception.

Furthermore, sentiment consistency is another aspect of sentiment analysis that is used for fake review detection. Genuine reviews typically exhibit a coherent sentiment throughout, where the overall emotional language is aligned with the specific points discussed within the text. In contrast, fake reviews may display inconsistencies, such as an overall positive rating paired with mainly negative sentences, or *vice versa*. The presence of contradictions in the sentiment of a review may signify that it is fake (Kaur and Malik, 2021). By analysing these inconsistent sentiments, detection systems can identify reviews that exhibit unnatural emotional shifts or inconsistent sentiment patterns, which are characteristic of deceptive content.

Punctuation: Punctuation is an informative element in the detection of deceptive reviews. The use of punctuation can reveal important clues about the authenticity of a review, as fake reviews frequently display abnormal punctuation patterns compared to genuine ones. For instance, fake reviews may include an excessive use of exclamation marks to amplify the emotional phrases and make the content appear more convincing or exciting (Ott et al., 2011). Genuine reviews, on the other hand, tend to use punctuation in a more balanced, contextually appropriate manner, reflecting natural speech patterns. Similarly, the inconsistent capitalisation of words or phrases, such as "BEST PRODUCT EVER!", is another technique that is used by fake reviewers to attract attention (Hajek and Sahut, 2022). By examining these punctuation-related anomalies, detection systems can effectively differentiate between authentic and fake reviews.

Repetitiveness and Redundancy: Repetitiveness and redundancy are common indicators of fake reviews and play a significant role in their detection. Fake reviews often contain repeated phrases, words, or entire sentences, reflecting a lack of genuine

content or specific experiences to share. This repetition can take several forms, such as the frequent use of particular keywords related to the reviewed item, which may be intended to manipulate the readers (Mukherjee et al., 2013a). Furthermore, redundancy in fake reviews often extends to content that repeats the same point in different ways, adding little to no new information. This redundancy can be manifested as multiple sentences conveying identical ideas, which may indicate an effort to meet a certain word count or create an impression of depth and thoroughness (Abri et al., 2020). In contrast, genuine reviews are generally more concise and to the point, reflecting the real experiences of users who have no need to add filler content. Redundant information in fake reviews can provide a good sign for the detection algorithms, highlighting reviews that lack the diversity and depth that characterise authentic user feedback.

2.5.2 Machine-Learning-Based Approaches

Machine learning based models make use of the progressive learning technique to identify fake reviews, since it has the ability to learn features and patterns from the data. These methods can be classified under two main categories of learning: namely, supervised learning and unsupervised learning.

2.5.2.1 Supervised Learning

Supervised learning provides the foundation for many ML approaches to detecting fake reviews in the field of NLP. This method involves training classifiers on a labelled dataset, where each review is marked as either truthful or fake. Various classifiers have been utilised in this context, including Support Vector Machines (SVM), Random Forests, Naive Bayes, and Gradient Boosting Machines. According to Khalif and Mane (2024), these models rely heavily on features that are extracted from the text, such as n-grams, part-of-speech tags, syntactic patterns, and sentiment indicators, to differentiate between genuine and deceptive reviews. The primary advantage of the supervised learning models is their relatively high degree of accuracy when the training data are well-represented and of sufficient size.

Recent studies have focused on improving the performance of these models by utilising more advanced feature engineering methods. For instance, Zhang et al. (2023b) emphasised the use of psycho linguistic features, such as those derived from the Linguistic Inquiry and Word Count (LIWC) tool, to identify deceptive content based on psychological markers. Additionally, incorporating semantic features using word embeddings, like Word2Vec and GloVe, has improved model performance by capturing the contextual meaning of words (Hájek et al., 2020). However, while these techniques enhance model accuracy, they often lack robustness when applied to different domains or languages, which limits their generalisability.

With the advent of deep learning, more complex models. such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have been increasingly used when employing supervised learning for fake review detection. CNNs have proven effective at capturing local patterns and dependencies in text, such as word collocations and syntax structures, which are often indicative of fake reviews (Hájek et al., 2020). On the other hand, LSTM networks, known for their ability to capture long-range dependencies in sequential data, have been successful in identifying temporal patterns and inconsistencies in writing style that are characteristic of fake reviews (Mohawesh et al., 2024).

Moreover, transformer-based models, such as BERT, have emerged as state-of-the-art methods for supervised fake review detection. As Gupta et al. (2021) note, BERT's ability to understand the context of words within sentences, using a bidirectional approach, allows it to capture the subtle linguistic nuances that are associated with deceptive writing. This has enhanced the detection accuracy, especially in the case of large-scale datasets. Nonetheless, the effectiveness of the supervised models heavily depends on the availability of high-quality labelled data, which is often a limiting factor in real NLP applications (Khalif and Mane, 2024).

2.5.2.2 Unsupervised Learning

Unsupervised learning methods aim to detect fake reviews without relying on prelabelled data, which are often scarce or expensive to obtain. These methods typically involve clustering, anomaly detection, or the use of generative models to identify patterns that deviate from the norm. Recent studies have explored various unsupervised techniques to support fake review detection. For instance, Latent Dirichlet Allocation (LDA) and other topic modelling techniques have been employed to discover hidden themes in reviews and identify those that fail to match the general topic distribution (Şule Öztürk Birim et al., 2022). This approach assumes that fake reviews often contain content that diverges significantly from the usual topics that are discussed in genuine reviews, thereby allowing them to be flagged as suspicious.

Autoencoders, a type of neural network used for unsupervised learning, have also demonstrated good results in this field. Autoencoders can be trained to reconstruct genuine reviews and then measure the reconstruction error when attempting to reconstruct new, potentially fake reviews. High reconstruction errors can indicate fake reviews, as they deviate from the normal patterns learned during training. A study by Khattar et al. (2019) showed that multimodal variational autoencoder (MVAE), which extend the basic autoencoder framework by incorporating probabilistic elements, can effectively capture the underlying structure of genuine reviews and differentiate them from fake ones.

Another prominent approach in unsupervised fake review detection involves the use of generative adversarial networks (GANs), which consist of two neural networks, a generator and a discriminator, that are trained simultaneously. The generator creates synthetic reviews, while the discriminator attempts to distinguish between real and generated reviews. Over time, the discriminator learns to identify subtle features that differentiate real from fake reviews, thereby improving its ability to detect deceptive content in an unsupervised manner. Zhang et al. (2022) found that GANs could outperform the traditional unsupervised learning methods due to their ability to learn more sophisticated representations of deceptive content.

2.5.3 Publicly Available Datasets for Fake Review Detection

Detecting fake reviews is important for maintaining the credibility of online review platforms. The presence of fake reviews can significantly undermine the trust of users, as they rely on these reviews when making their decisions about products or services. Several datasets have emerged, that aim to improve the development of fake review detection, primarily sourced from e-commerce platforms and review sites. These datasets are helpful for training and evaluating the models that are designed to detect fraudulent reviews. The publicly available datasets used for fake review detection include the following:

Yelp Dataset: The Yelp Dataset [2] is a rich, extensive dataset that is provided by Yelp as part of its data science competition: the Yelp Dataset Challenge. This dataset was designed to advance the research and development in multiple fields, such as sentiment analysis and fake review detection. It is considered one of the primary datasets in the field of fake review detection. This dataset encompasses a vast selection of reviews, detailed business attributes, and rich user interactions, making its coverage comprehensive. This dataset serves as a cornerstone resource in the study of fake reviews.

Containing over six million user reviews related to more than 150,000 businesses, the dataset spans a diverse collection of businesses, including restaurants, retail stores, and service providers. Each review includes detailed text, star ratings, timestamps, and user information, offering a rich source of data for analysing consumer emotions and behaviour. Beyond the reviews' content, the dataset provides extensive metadata about the businesses, including their names, categories, locations, and operating hours, which allows researchers to explore how different business characteristics influence reviews' patterns and authenticity. The breadth and depth of the Yelp dataset make it an invaluable tool for developing sophisticated models to detect fake reviews and understand user sentiment in greater depth.

[2]https://www.yelp.com/dataset, (accessed 12 January 2025)

Multiple datasets have been derived from the Yelp dataset, including Yelp CHI (Mukherjee et al., 2013a) and Yelp ZIP (Rayana and Akoglu, 2015). Yelp CHI is specifically structured to analyse and detect fake reviews. This dataset is notable for its comprehensive nature, including a wealth of user reviews and detailed metadata, that makes it ideal for studying deceitful review behaviour. It includes more than 67,000 user reviews of 201 restaurants and hotels in the Chicago area of the US. The dataset includes comprehensive metadata about each review, including the text, star ratings, timestamp, user information and business information. The reviews are labelled as either genuine or suspicious, based on the Yelp anti-fraud filtering algorithm. In this dataset, 13.23% of the reviews were classified as fake, originating from 20.33% of the users who were identified as spammers.

Similar to the Yelp CHI dataset, the Yelp ZIP dataset is designed to advance research on the detection of opinion spam. The dataset begins by selecting a ZIP code in New York State and gather reviews for restaurants located within that area. It then increments the ZIP code number systematically and repeats the process. Because ZIP codes are organised geographically, this method results in the collection of restaurant reviews from a contiguous region of the US, which extends beyond New York, encompassing parts of New Jersey, Vermont, Connecticut, and Pennsylvania. Furthermore, it includes over 600,000 Yelp reviews of 5,044 restaurants by over 260,000 reviewers, complete with detailed metadata about the reviews, users, and businesses. Each review features a full text, star ratings, and timestamp, while the user profiles provide data such as the number of reviews written, their average ratings, and account creation date. Additionally, the business metadata includes the business name, category, location, and operating hours, facilitating a thorough analysis of both individual and collective behaviour that is indicative of opinion spam.

Amazon Dataset: The Amazon Reviews dataset is a major resource, that was specifically designed to analyse and detect fake reviews on online platforms (Jindal and Liu, 2008). This dataset encompasses 5.8 million reviews, each accompanied by detailed textual content and metadata, including a product ID, reviewer ID, rating, date, title, text, amount of helpful feedback and amount of feedback in general. The dataset incorporates reviews for over six million different products, making it exceptionally comprehensive and varied. These products have been extracted from four categories (books, music, DVDs and industry manufactured products). This extensive coverage ensures a diverse range of product reviews, providing a valuable resource for in-depth analysis.

The primary focus of this dataset is to aid in the detection of opinion spam by examining review text and user behaviour. Additionally, a significant portion of the dataset is manually labelled to distinguish between genuine and fraudulent reviews, providing a reliable benchmark for training the ML models. This labelling process involves the application of expert judgment and various heuristics in order

to distinguish between genuine and fraudulent reviews. Linguistic analysis of the review text was also performed to identify the characteristics of fake reviews, including the presence of repetitive or formulaic language. By offering a rich set of labelled data and extensive metadata, this dataset serves as a crucial foundation for developing sophisticated detection algorithms and enhancing our understanding of spam behaviour in online reviews.

TripAdvisor Dataset: The TripAdvisor dataset contains a collection of hotel reviews of 20 popular hotels in Chicago (Ott et al., 2011, 2013). This dataset provides a valuable resource for researchers aiming to understand and detect deceptive opinion spam. It comprises a total of 1,600 reviews, which are evenly divided between truthful and deceptive opinions. This balanced composition includes 400 truthful positive reviews, 400 truthful negative reviews, 400 deceptive positive reviews, and 400 deceptive negative reviews, allowing a comprehensive analysis across different sentiment categories.

Each hotel in the dataset has 20 reviews, offering a balanced representation of opinions. The reviews in this dataset are carefully annotated to distinguish between truthful and deceptive content. Truthful reviews are sourced from actual TripAdvisor reviews and some other sources, ensuring authenticity and real-world applicability, while deceptive reviews are generated through Amazon Mechanical Turk, where crowdsourced workers are instructed to write convincing fake reviews. This dual sourcing methodology enhances the dataset's robustness, making it a valuable resource for training and testing machine learning models aimed at detecting opinion spam. In this thesis, the fake reviews contained in this dataset were used as the primary source for developing and evaluating the fake review detection methods that are discussed in Chapter 5.

2.5.4 Existing studies for Fake Review Detection

The reported accuracy for fake review detection demonstrates considerable variation across methodological approaches, ranging from traditional machine learning to more recent deep learning and transformer-based models. Among traditional machine learning techniques, Support Vector Machines (SVM) remain widely applied and perform competitively, with different accuracies reported across different datasets. For instance, an SVM baseline achieved 0.83 accuracy on TripAdvisor reviews (Ahmed et al., 2018), while combining SVM with BERT embeddings improved performance to 0.8781 on the same dataset (Mir et al., 2023). The results confirm the adaptability of SVM when integrated with contextual embeddings. Tree-based classifiers such as Random Forest have also shown promising results, reaching 0.8723 accuracy on Yelp reviews (WANG et al., 2023). These findings indicate that while traditional classifiers remain effective, their performance varies significantly depending on dataset

Reference	**Dataset**	**Method**	**Accuracy**
Periasamy et al. (2024)	E-commerce platforms	BERT, SVM	0.787
Shahariar et al. (2019)	TripAdvisor	CNN	0.9158
Singh et al. (2023)	TripAdvisor	CNN	0.9183
Gupta et al. (2022)	Yelp	DistilBERT	0.68
Ennaouri et al. (2024)	TripAdvisor	K-means	0.91
Ferdinan et al. (2024)	Tokopedia	LSTM	0.8132
Saxena (2025)	Kaggle	BERT	0.9434
WANG et al. (2023)	Yelp	Random Forest	0.8723
Mewada et al. (2025)	Amazon	multiple Stacking models (MLP, CNN, LSTM, and BiLSTM)	0.9004
Mir et al. (2023)	TripAdvisor	BERT, SVM	0.8781
Hájek et al. (2020)	Amazon	DFFNN	0.8280
Ahmed et al. (2018)	TripAdvisor	SVM	0.83

Table 2.2: Overview of the studies for fake review detection

characteristics and feature representation.

Deep learning methods generally report higher accuracy in fake review detection. Convolutional Neural Networks (CNN) trained on TripAdvisor reviews achieved accuracies of 0.9158 and 0.9183 in separate studies (Shahariar et al., 2019; Singh et al., 2023), demonstrating the strong ability of CNNs to capture linguistic patterns in review texts. More advanced architectures such as stacking-based ensembles, which combine multiple models including CNN, LSTM, and BiLSTM, achieved 0.9004 accuracy on Amazon reviews (Mewada et al., 2025), highlighting the benefit of leveraging multiple deep learning classifiers together. These results demonstrate that deep learning methods can outperform traditional classifiers and adapt effectively across platforms and domains.

In the progression of fake review detection methods, transformer-based models have emerged as the leading approach. BERT-based approaches in particular shows robust results across multiple datasets, with a BERT framework for fake

web recommendations on the Kaggle dataset achieving 0.9434 accuracy (Saxena, 2025). On TripAdvisor reviews, BERT combined with SVM reached 0.8781 accuracy, confirming its good performance across domains (Mir et al., 2023). However, performance varies depending on the specific transformer variant and dataset: for example, DistilBERT achieved only 0.68 accuracy on Yelp reviews (Gupta et al., 2022). These findings suggest that while transformers generally outperform both traditional and deep learning methods, their effectiveness depends on model choice, dataset characteristics, and domain-specific training. Table 2.2 provides a unified summary of some of the existing studies for fake review detection, presenting the datasets, classification methods, and reported accuracy.

2.6 Chapter Summary

The reviewed literature in this chapter highlights the substantial progress that has been made in the domains of emotion detection, fake review detection, and the creation of annotated datasets for NLP tasks. However, several critical gaps remain, particularly related to the intersection of these fields within the context of tourism-related reviews. Despite the availability of emotion annotated datasets, none of the existing datasets were specifically designed for application in the tourism domain, as discussed in Section 2.4.4. This limitation undermines the generalisability of the current methods for emotion detection when applied to reviews in this context. Tourism reviews are distinct due to their emotional and subjective nature, as users often express complex feelings regarding experiences like travel, hospitality, and leisure. The lack of a domain-specific dataset severely restricts the ability to fine-tune models to suit the unique linguistic and emotional patterns that are present in this domain, thereby underscoring the necessity of creating a dedicated tourism review dataset that is annotated with emotion information.

The development of such a dataset represents a significant contribution to the field. By addressing the limitations of the existing resources, this new dataset will provide a foundation for more accurate, contextually-relevant emotion detection models. Furthermore, this research also aims to fine-tune LLMs specifically for emotion detection in tourism reviews. By fine-tuning a domain-specific LLM on the newly-created dataset, this research will establish a model that is capable of achieving good performance in emotion detection related to tourism reviews.

The importance of integrating emotion detection with fake review classification cannot be overstated. Emotions play a pivotal role in deceptive reviews, as fake reviewers often attempt to manipulate readers by exaggerating positive or negative emotions in order to influence their perceptions. However, the existing fake review detection methods predominantly rely on textual features or sentiment analysis, as discussed in Section 2.5.1, neglecting the emotional dimension. Incorporating

emotion information into fake review classification presents an opportunity to enhance the accuracy and robustness of the existing detection models. By examining the emotional categories in the text, it becomes possible to identify deceptive reviews more effectively. The necessity of exploring the integration of emotion detection into fake review classification becomes even more apparent when considering the challenges posed by the increasingly sophisticated nature of fake reviews.

Finally, this research seeks to address the gaps in the literature by developing a novel dataset of tourism-related reviews that is annotated with emotion information, using the crowdsourcing method. This dataset will be utilised to fine-tune LLMs for detecting emotions in tourism-related text. Furthermore, this research will evaluate the impact of incorporating emotion information into an LLM classifier to enhance the detection of fake reviews, demonstrating the value of emotion analysis in improving the reliability of review authenticity detection.

Chapter 3

Tourism Review Corpus Construction with Emotion Annotation

3.1 Introduction

One of the contributions of this thesis is the creation of a tourism-specific corpus, annotated with emotion information, named TORCE (the Tourism Corpus of Emotion). This addresses a gap in the existing emotion annotated datasets, which lack the domain specific text that is necessary to analyse emotional categories in tourism-related reviews, as discussed in Section 2.4.4. By constructing this dataset, this work aligns the data annotation efforts with the unique linguistic and emotional characteristics of the tourism domain, thereby facilitating the better targeting of emotion detection tasks.

There are two versions of the TORCE dataset, each designed to meet different requirements related to emotion annotation in the tourism domain. TORCEv1 is annotated with eight emotion categories based on the Plutchik Wheel of Emotions. This version provides a precise, detailed representation of emotional expressions, making it suitable for fine grained emotion detection and analysis. TORCEv2, on the other hand, is a streamlined version of the dataset, annotated with five primary emotion categories. This simplification is designed to make it more practical to use for the subsequent emotion detection work.

This chapter explores the potential of using the crowdsourcing platform to address the data annotation challenges that are related to tourism reviews in particular. It offers a comprehensive examination of both the opportunities and challenges associated with using this methodology, including quality control techniques, inter-annotator agreement, and their implications for data reliability. Moving beyond the

theoretical discussion, this chapter also emphasises the practical complexities involved in using crowdsourcing to generate a labelled dataset, highlighting its transformative role in generating labelled data for domain-specific NLP applications.

3.2 Tourism Review Data Collection

3.2.1 Data Sources Survey and Selection

To construct a dataset of user reviews related to tourism and hospitality venues, particularly attractions located in the United Kingdom, reviews were gathered from the TripAdvisor platform [3]. This platform was chosen due to its extensive repository of user-generated content, which was expected to encompass a wide range of expressions of emotion. Initially, several data sources were considered, including Tourpedia [4], TripAdvisor, and tourism-related subreddits on the Reddit platform [5], such as "travel", "longtermtravel", "solotravel", and "wanderlust". However, a manual inspection and comparison of the sample data revealed that the comments published through the Reddit platform often took the form of complaints and negative conversations about tourism services or lacked explicit expressions of emotions. Additionally, the Tourpedia data were outdated, due to its closure in 2014, and related solely to attractions in London. Consequently, TripAdvisor was selected as the primary data source for the tourism-related reviews.

The selection of an appropriate information source is critical for the effective execution of every annotation effort. TripAdvisor surfaced as the most appropriate source because of its massive collection of tourism-related recommendations and user-generated material. As a result of its popularity, it proved to be an excellent choice for obtaining a diverse variety of phrases to be annotated. TripAdvisor is a widely-used website that offers tourism information, providing details about approximately 8.7 million tourism attractions, hotels, and restaurants. To date, over 860 million people have reviewed and commented on TripAdvisor, in 28 different languages [6]. TripAdvisor categorises tourism attractions into various types, including landmarks, parks, zoos, aquariums, spas, nightlife, museums, and shopping. Furthermore, users can rate tourism services using this platform. The destinations on the platform are rated using a scale from 1 (terrible) to 5 (excellent). Reviews must be a minimum of 100 characters in length, with no specified maximum length. Additionally, users can express their approval by pressing the *like* button in order to add a "like" tag to someone else's review.

[3] https://www.tripadvisor.co.uk, (accessed 13 January 2025)
[4] http://tour-pedia.org/about, (accessed 13 January 2025)
[5] https://www.reddit.com, (accessed 13 January 2025)
[6] https://ir.tripadvisor.com, (accessed 16 January 2025)

3.2.2 Techniques and Tools for Online Data Collection

Online reviews have become a cornerstone for understanding consumer experiences, offering organisations and researchers a valuable source of data. Reviews encompass a wide range of emotions and comments, addressing aspects ranging from product quality to service satisfaction. Collecting and analysing these data requires systematic approaches and specialised tools (vanden Broucke and Baesens, 2018). A key method for gathering reviews is web scraping, a technique that enables the automated extraction of data from websites. Popular Python libraries, such as BeautifulSoup [7], Scrapy [8], and Selenium [9], facilitate this process. While the first two of these are effective for parsing static web content, Selenium stands out due to its ability to interact with dynamic, JavaScript-driven websites. This capability makes Selenium particularly useful for scraping websites that employ AJAX or other techniques to load content dynamically.

Crawling tools facilitate automated browsing, communication with components, and data extraction, reducing the need for administrative interaction. The Selenium tool not only enables data collection but also mimics real-world user actions, such as scrolling, filling in forms, and clicking, and can also wait for dynamic content to load before scraping, allowing it to interact with complex web environments seamlessly. When integrated with Python's robust data processing libraries like Pandas and Numpy, as well as storage solutions, these tools enable an end-to-end workflow for extracting, transforming, and storing data efficiently.

The process of collecting data via web scraping must be approached responsibly, however, adherence to the ethical guidelines and compliance with the website terms of service are critical, in order to avoid legal and reputational risks. Unauthorized or excessive scraping may not only breach the legal boundaries but is also frowned upon by the web community. Therefore, prior to initiating any data collection activity, a website's terms must be thoroughly evaluated and the scraping processes must be designed to fall within acceptable limits. For this research, the TripAdvisor platform granted approval for the scraping process, and further approval was obtained from the ethics committee of the Computer Science Department Reference Number FST-2021-0656-RECR-2.

The user-generated reviews were gathered between the 1^{st} of February and the 25^{th} of May 2022. These reviews provide comprehensive coverage of every attraction in the United Kingdom. In order to collect these reviews from TripAdvisor, a dedicated crawling tool was crafted using Python, leveraging the Selenium package. This package played an essential role in automating the navigation and interaction

[7] https://www.crummy.com/software/BeautifulSoup, (accessed 20 January 2025)
[8] https://scrapy.org, (accessed 20 January 2025)
[9] https://pypi.org/project/selenium, (accessed 20 January 2025)

with the web. Using this tool, the TripAdvisor platform systematically browsed and extracted information about every attraction in the UK. The extraction process involved parsing the HTML source code, resulting in the acquisition of a substantial dataset, comprising 82,805 attraction URLs.

Upon obtaining these URLs, the next step was to navigate through each one, extracting detailed information about the respective attraction, and retrieving every associated English language reviews. It is noteworthy that the default setting on the platform ensured the display of exclusively English language reviews, as a result of navigating within the TripAdvisor UK domain. The dataset encompasses diverse information, including the attraction's name, rating, the text of each review, the date of each review, the number of likes received, the reviewer's rating of the attraction, and the reviewer's total number of contributions to the TripAdvisor platform. In total, the online crawling process yielded a raw dataset comprising over 7,146,000 user-generated reviews.

Once collected, raw data typically require extensive pre-processing to make them usable (Lu, 2023). The data cleaning phase involves removing irrelevant fields and standardising the text format, such as converting text to lowercase and removing punctuation. Reviews are often rich in textual content, necessitating further pre-processing, such as tokenisation, which entails splitting text into individual words or phrases and removing stopwords like "and", "the", or "is". These pre-processing steps were used solely for statistical analysis and sentence selection. For the crowdsourcing annotation task itself, workers were shown the original, unprocessed sentences. Following these steps ensures that the dataset is ready for meaningful analysis.

3.2.3 Data Structure and Organisation

At the beginning of structuring and organising the collected data, following a careful manual examination of the reviews, a key observation was that many of the reviews comprised multiple sentences, each potentially reflecting a different emotional category. This variability made it challenging to assign a single emotion class to a lengthy review, as the diverse emotions mentioned within it might have introduced noise into the annotation process. To address this challenge, the data was subjected to a segmentation process, where complex reviews were broken down into individual sentences. The objective was to achieve a more granular segmentation of the text, providing annotators with a better opportunity to assign the correct emotion class. Given the time-consuming nature of manual segmentation, however, additional filtering mechanisms were applied to streamline the dataset preparation. Ultimately, the application of these filtering steps resulted in the selection of 2,500 sentences for the subsequent task.

1) Split a review into sentences

A key step involved splitting each review into its constituent sentences. This segmentation process was accomplished using a sentence tokenizer that was built on pre-trained algorithms within the spaCy English language model [10]. The tokenizer, relying on a dependency parser developed by Honnibal and Johnson (2015), determined the sentence boundaries, expanding the dataset to incorporate a comprehensive 37,646,428 sentences. This expanded dataset formed the basis for our subsequent in-depth analysis and annotation efforts.

Evaluation Type	**Score**
Precision	0.92
Recall	0.89
F-score	0.90

Table 3.1: Evaluation scores for the tokeniser

2) Stopwords and punctuation removal

Stopwords, such as "the", "and", and "have", are frequently occurring words that contribute minimal semantic content. For this reason, SpaCy's predefined stopword list was employed to filter out such terms, with the objective of retaining only the most informative words for analysis. In addition, punctuation marks were removed, as they do not contribute to lexical meaning. After applying these preprocessing steps, the average sentence length decreased from 18.17 to 9.29 words per sentence, indicating a considerable reduction in textual redundancy. All subsequent computational steps were carried out on this cleaned version of the data to reduce noise and focus on meaningful lexical features.

It is important to note, however, that these modifications were applied exclusively for computational processing and not for the crowdsourcing annotation task. The sentences presented to annotators were kept in their original form, including punctuation and stopwords, because these elements contribute to sentence readability, grammatical structure, and contextual clarity. Removing such elements could have resulted in fragmented or ambiguous sentences, thereby reducing annotation quality and consistency. The distinction between the two text versions was therefore intentional: one optimised for computational analysis, and the other preserved for crowdsourcing annotation.

3) Exclude sentences without any likes

[10]https://spacy.io/models/en, (accessed 20 January 2025)

The platform enables users to add a "like" tag to a review. Reviews that lack any "like" tags are more likely to be deemed uninformative, as they have failed to garner any positive feedback. Consequently, all of the sentences that were extracted from reviews without any "like" tags were excluded from the subsequent steps of the process. This led to a reduction in the number of sentences from 37,646,428 to a total of 14,544,512 sentences.

4) Minimum of three word sentence lengths

Short sentences are more prone to be insufficient for providing descriptive details due to their limited use of words. Hence, a minimum threshold of three words was established in order for a sentence to be considered during the subsequent steps. This adjustment led to a decrease in the number of sentences from 14,544,512 sentences to 12,397,713 sentences.

5) Sentences containing emotion words

Two emotion lexicons were employed in this step. The first was the NRC emotion lexicon (Mohammad and Turney, 2013), and the second the USAS semantic lexicon (Rayson et al., 2004). The former lexicon comprises 14,182 words, classified into eight basic emotions, namely "anger", "joy", "sadness", "disgust", "anticipation", "surprise", "trust", and "fear", based on the Plutchik Wheel of Emotion. Of these words, 4,463 are related to at least one emotion category, while the rest are emotionally neutral. Only words that were labelled with at least one emotion class were considered in the subsequent analysis, while the rest were excluded.

On the other hand, the USAS semantic lexicon provides a framework for automatically performing the semantic analysis of text. It includes 55,662 lemmas, of which 3,312 words have been classified into at least one of six emotion classes, namely, "general", "liking", "calm/violent/angry", "happy/sad", "fear/bravery/shock", and "worry/concern/confident". Additionally, 703 words are included in both lexicons. Table 3.2 provides further details.

Lexicon Name	**Total Words**	**Emotion Words**
NRC	14,182	4,463
USAS	55,662	3,312
Combined	-	7,072

Table 3.2: An overview of the lexicons employed

In this crucial step, lemmatisation was applied to all of the words within the corpus, with the aim of transforming inflected words into their base forms.

The SpaCy lemmatiser, a fundamental component of the SpaCy English language pipeline, played a crucial role in this process. Additionally, the NRC emotion words underwent lemmatisation, mirroring the treatment of corpus words, as certain words in the lexicon were observed to be inflected. In contrast, emotion words that were derived from the USAS semantic lexicon were already in their lemmatised form. The underlying rationale for this lemmatisation step was to ensure the uniform representation of all words, from both the collected data and the two emotion lexicons mentioned earlier, in their base forms.

This standardised representation was then utilised in the subsequent comparison step. Furthermore, each sentence underwent scrutiny with reference to the two lexicons mentioned earlier to guarantee the presence of at least one emotion word. A simple matching algorithm was employed for this purpose, calculating the number of emotion words within each sentence. Subsequently, sentences that lacked any emotion words were systematically excluded. This detailed process resulted in the reduction in the total number of sentences from an initial count of 12,397,713 to a refined total of 9,268,491.

6) Dataset division based on review ratings

Following the comprehensive pre-processing steps, the dataset was strategically divided into five subsets. based on the overall rating assigned to the original review containing each sentence. TripAdvisor gives users the ability to rate the experience reviewed on a scale from 1 (low) to 5 (high). This deliberate division was implemented to ensure that all rating classes were equally considered, with the aim of creating a final dataset containing an equal number of sentences per rating class. A detailed list of the distribution of the sentences across each rating class is presented in Table 3.3.

Rating Type	**Number of sentences**
Rating 1	868,785
Rating 2	593,554
Rating 3	836,865
Rating 4	1,943,187
Rating 5	5,026,100
Total	9,268,491

Table 3.3: An overview of the subsets, divided per rating class

7) Ignore sentences outside the range of the subset mean

To ensure that all of the sentences contained a similar number of words, the average

length of the sentences was calculated for each subset. It was found that sentences associated with an overall review rating of 1 were the longest, with an average length of 9.99 words, while sentences associated with an overall review rating of 5 were the shortest, with an average length of 9 words. The average length for each subset was then used as a filtering threshold, where sentences that either exceeded or fell short of the average length by 25% were excluded. This filtering helped to exclude excessively long sentences, as these can contain several emotion types within a single sentence, which would make them more complicated for crowdsourcing workers to annotate. Similarly, excessively short sentences were also removed, because these can be less informative. In addition, such filtering also helped to ensure that all of the annotators were given a similar text length per annotation task and spent roughly the same time and effort on each sentence annotation. This process reduced the number of sentences for all subsets combined from 9,268,491 to 3,925,842 sentences. Table 3.4 shows the mean, the threshold range for each subset, and the included number of sentences.

Rating Type	**The Threshold**	**Sentence Range**	**Number of sentences**
Rating 1	9.99	8 - 12	375,945
Rating 2	9.91	8 - 12	263,894
Rating 3	9.69	8 - 12	380,190
Rating 4	9.29	7 - 11	891,226
Rating 5	9.00	7 - 11	2,014,587
Total	-	-	3,925,842

Table 3.4: An overview of the averaged subsets

8) Select sentences for annotation

Each subset was sorted in descending order, based on the number of emotion words per sentence. A word is considered an "emotion word" if it features in the two previously mentioned lexicons, the NRC emotion and USAS semantic lexicons. The sorting process ensured that sentences that were hypothetically rich in emotion were included in the crowdsourcing annotation task. In contrast, sentences that contained fewer emotion words were less likely to be chosen. Following the sorting process, the first 500 sentences from each subset were selected to form the final dataset for the crowdsourcing annotation task. Several sentences that contained a high number of emotion words were excluded, however, for various reasons. For example, the sentence "Awful, Awful, Awful, Awful, Awful, Awful, Awful, Awful, Awful." was

been excluded due to containing excessive repetition. Also, certain sentences were related to other sentences and so, in isolation, would have appeared ambiguous to the annotators, such as, "This was done all at the same time and I felt like they were very rushed because just as I was starting to feel relaxed the lady said okay". Finally, a total of 2,500 sentences were included in the final dataset for the crowdsourcing task.

The procedure for collecting and assembling the data for the emotion labelling was a precise, methodical activity. As mentioned earlier, this step entailed identifying multiple key phases, each of which was designed to enhance and improve the information that would form the basis for the eventual crowdsourced annotation. The filtration and refinement processes comprised multiple, critical sub-steps. The use of powerful tokenisation techniques to split comments into specific phrases decreased the interference in the annotation method. Furthermore, removing sentences that had failed to attract any likes and establishing a minimum sentence length criterion guaranteed that the collected data contained only significant, pertinent text. The addition of emotion lexicons added an emotional component to the text, harmonising it with the goals of the annotation task. This phase ensured that the dataset that was utilised for the crowdsourced annotation was enhanced by sentences that were likely to elicit particular emotions. Excluding statements with uncertain connotations or repeated phrases improved the dataset's performance even further. Figure 3.1 shows a sample TripAdvisor review together with its rating, from which the highlighted sentence was selected for annotation. Table 3.5 provides a statistical overview of the collected data, which was filtered until it formed the final dataset.

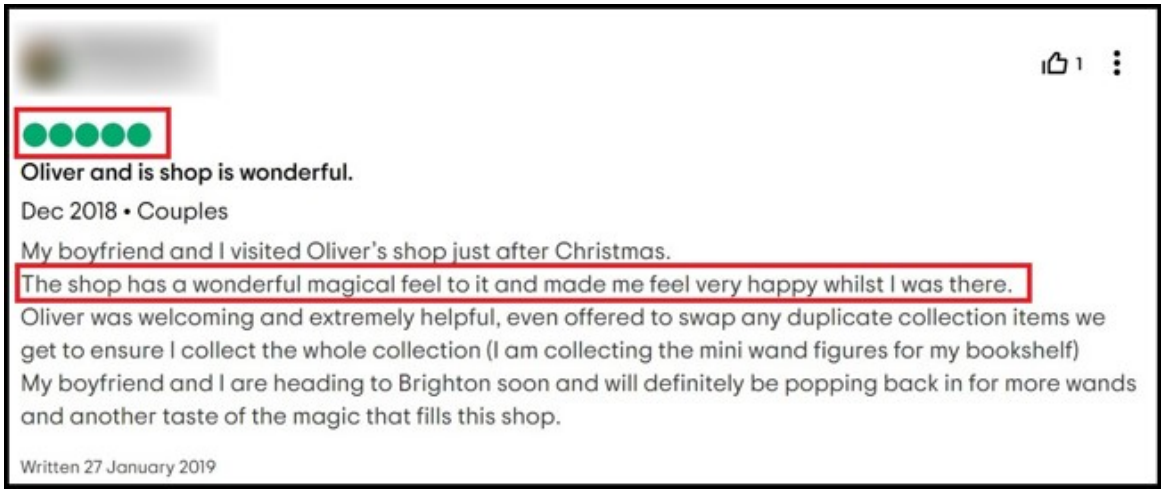


Figure 3.1: An example TripAdvisor review

Number of attractions	82,805 attractions
Number of reviews	7,146,099 reviews
Number of tokens	543,620,839 tokens
Number of types	1,288,090 words
Number of sentences with more than one like	14,544,512 sentences
Number of sentences with at least three words length	12,397,713 sentences
number of sentences with at least one emotion word	9,268,491 sentences
number of sentences with 25% around the mean length	3,925,842 sentences
number of selected sentences	2,500 sentences

Table 3.5: A statistical overview of the collected data

3.3 Crowdsourcing Method for Corpus Annotation

Crowdsourcing is a method that entails hiring multiple people to perform a specific human intelligent task (HIT). One of the key challenges in data science is how to label collected data, especially if they are vast in size. Crowdsourcing solved this challenge by making it possible to split the process into small tasks so that, by applying a decent quality control technique, reliable, high-quality data can be obtained through outsourcing tasks to multiple people. Also, by offering a suitable payment, that is usually a low amount in return for performing relatively small tasks, crowdsourcing offers a viable method for employing and training people to perform these small tasks. Crowdsourcing has been used in NLP for many different tasks, including dataset annotation (Mohammad and Turney, 2013). Hence, crowdsourcing offers a flexible, adaptable strategy, that has transformed how businesses handle their data categorisation and other intelligence-related jobs.

3.3.1 Crowdsourcing Methods and Tools

One of the longest-standing challenges associated with data analysis is the need for the precise, effective labelling of large datasets (Zhen et al., 2021). By breaking down the labelling tasks into smaller, more manageable chunks, crowdsourcing provides a helpful solution to this challenge. One of the primary benefits of crowdsourcing is the capacity to tap into the collective wisdom of a varied group of people (Huiqi et al., 2021). Organisations may glean an exceptionally broad range of information by gathering comments from a worldwide pool of participants, drawing on a huge

number of unique viewpoints and areas of expertise. This variety can be very useful in NLP, where comprehending the difficulties of syntax and context is critical.

Furthermore, the use of crowdsourcing in NLP demonstrates its adaptability and efficacy in a variety of NLP assignments, notably dataset annotation. This method has not only sped up NLP study and enhancement, but has also cleared the way for new applications that depend on well-annotated information, including sentiment analysis, machine interpretation, and chatbot creation (Zhu et al., 2019). Ultimately, crowdsourcing has become a game-changing paradigm in the data sciences, providing a scalable, economic, and varied method for data categorisation. Its importance in NLP alongside various fields is growing, and academics and businesses acknowledge the possibility of using crowd knowledge to improve the accuracy of their information and project outcomes (Deichmann et al., 2021).

Platforms such as Amazon Mechanical Turk (MTurk) [11], Prolific [12], CloudFactory [13], Toloka [14], and Appen [15], provide accessible and scalable solutions for conducting crowdsourcing tasks. These platforms offer a range of tools for designing annotation tasks, recruiting annotators, and managing quality control through applying mechanisms like redundancy and inter-annotator agreement measures. By integrating these tools, a balance between efficiency and accuracy can be achieved. The use of crowdsourcing might expedite the annotation process and foster diversity within annotation, which is crucial for handling complex tasks like emotion detection in domain-specific contexts.

3.3.2 Mechanic Turk for Emotion Annotation

In order to implement the annotation process, the first task was to select a crowdsourcing platform. All of the crowdsourcing platforms mentioned above were considered and compared. After careful consideration and comparison, the MTurk platform was chosen for the annotation work, for the following reasons:

1. **Widened Usage of NLP Research**: Arguably, the key motivation for choosing MTurk as a tool for emotion annotation stems from the widespread acceptance of the service within the NLP community. MTurk has earned the trust of researchers worldwide, emerging as a reliable, effective resource for obtaining annotation data across various NLP applications (Oppenlaender et al., 2020). MTurk annotations have played a key role in constructing numerous NLP benchmark datasets, solidifying its reputation as a trustworthy tool for

[11]https://www.mturk.com, (accessed 8 January 2025)
[12]https://www.prolific.co, (accessed 8 January 2025)
[13]https://www.cloudfactory.com, (accessed 8 January 2025)
[14]https://toloka.ai, (accessed 8 January 2025)
[15]https://appen.com, (accessed 8 January 2025)

NLP tasks. Experts routinely evaluate their algorithms and models using these datasets to ensure their compliance with the industry standards.

2. **User Interface Flexibility for Task Creation**: Another compelling aspect supporting the choice of MTurk is its user-friendly nature and highly customisable interface for task creation. When dealing with complex annotated tasks like emotional labelling, the platform offers significant flexibility in designing essential workflow annotations (Deichmann et al., 2021). MTurk allows task creators to design tasks using templates that can be adjusted to meet the unique requirements of the text classification activity. Requesters can formulate tasks with text requests, answer alternatives, and supplementary instructions to ensure that the workers comprehend the complexities of the assignment. The platform incorporates a continuous tracking capability, enabling requesters to remain informed about the progress of their annotated projects. This ensures that tasks are completed on time and also provides an opportunity for prompt intervention should issues arise during the annotation process.

3. **API Optimization for Task Customisation and Quality control**: The Application Programming Interface (API) of MTurk offers distinctive features that empower requesters to control the task customisation and quality control at a granular level. The API enables requesters to implement customised quality assurance processes to ensure the accuracy of the annotations. Examples of this include verification checks, examinations, and the automatic rejection of low-quality work. By leveraging the API, requesters can dynamically administer jobs, allowing them to modify the task settings, clarify the instructions, and add additional workers as needed throughout the annotation process.

4. **Abundance of Workers Despite Their Proficiency**: It has been observed that one of the most critical factors that contributes to the success of annotation projects is the availability of a qualified, motivated workforce. MTurk offers a vast, diversified base of employees, with an estimated over 100,000 active workers at any given time (Difallah et al., 2018), who might be recruited for the task of emotion annotation, enriching it with their expertise (Schemmann et al., 2016). The platform boasts a substantial network of workers with expertise in NLP annotation assignments. Over time, this community has grown, facilitating the identification of skilled, experienced workers who might participate in data annotation projects. Whether the annotation task involves a small dataset or a large-scale project, the extensive workforce of MTurk is thought to provide easy scalability in order to meet the requirements of the project. The platform is assessed to offer expertise, with a range of working skills and backgrounds. In the context of emotion annotation, diversity is often seen as advantageous,

since different cultural workforces may provide varying levels of linguistic and psychological proficiency (Simaei et al., 2023). This may further involve the incorporation of diverse psychological expertise with different ways of expressing emotions, enriching the annotation process.

5. **Cost-Effective Annotation**: When choosing a crowdsourcing platform for information annotation, affordability is critical, especially when dealing with vast databases. MTurk operates on a pay-per-task model, allowing requesters to set their own budget for each annotation job. This flexibility enables effective cost control, particularly when funds are limited. The competitive nature of MTurk motivates the workers to execute jobs swiftly and precisely (Zhu et al., 2020). As employees seek to earn more by completing tasks faster, this competition generally leads to lower total costs for the task creators. Thus, MTurk allows requesters to set qualifications based on specific criteria for completing a particular task. It also facilitates the selection of workers based on their prior ratings and performance history, ensuring that the work is of high quality (Zhen et al., 2021). MTurk enables requesters to avoid the administrative costs associated with attracting and overseeing a specialised in-house staff for data analysis. The resulting reduction in administration expenses can have a considerable impact on the overall value for money of annotation initiatives.

Crowdsourcing is considered valuable and one of the most effective tools for data annotation in NLP due to its ability to tackle annotation challenges. By leveraging a large, diverse pool of contributors, crowdsourcing enables requesters to annotate vast amounts of data efficiently and at scale, which would otherwise be impractical using in-house staff. This approach is particularly advantageous in the case of complex tasks, such as emotion annotation, where varied perspectives help to capture complex interpretations of the text.

3.4 Annotating a Tourism Review Corpus

Emotion classification is inherently subjective, as it relies on individual judgments that are formed by unique experiences, cultural backgrounds, and cognitive interpretations. This subjectivity often results in varying opinions among annotators when categorising the same text. Leveraging crowdsourcing and the collective knowledge of diverse annotators, however, offers a solution to this challenge, enabling the generation of reliable, high quality annotations, that mitigate individual biases through collective intelligence (Mohammad and Turney, 2013). For the annotation of tourism-related reviews, the classification of emotions draws on the Plutchik Wheel of Emotion, which categorises emotions into eight primary groups: “anger”, “anticipation”,

"disgust", "fear", "joy", "sadness", "surprise", and "trust". These categories provide a comprehensive representation of humans' emotional responses (Toby Lea and Jungaberle, 2020).

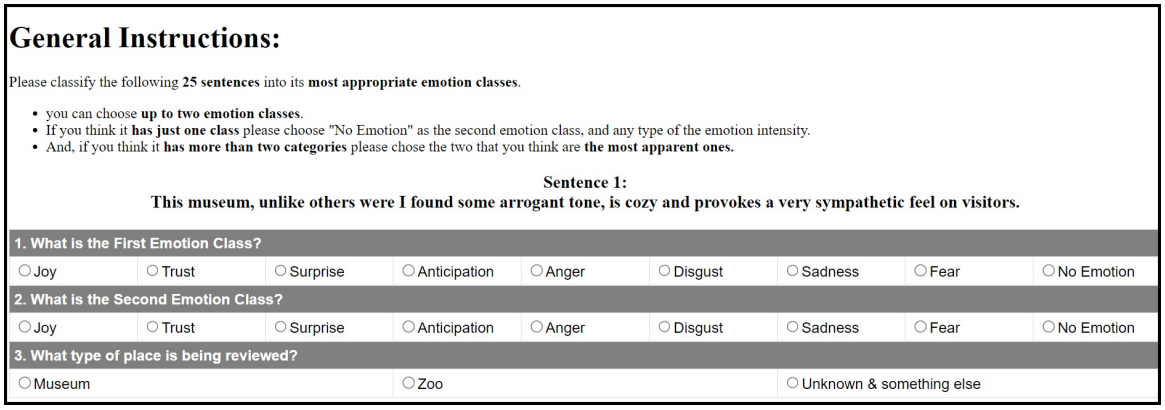
The selection of these emotion categories provides a systematic, widely-accepted foundation for emotion annotation. This alignment guarantees that this study captured the range of emotions expressed in the text, facilitating multidisciplinary cooperation and the creation of comprehensible emotion identification algorithms. Employing this framework allows the creation of metadata rich datasets, which are essential for training robust models and clarifying textual emotional expressions (Uban et al., 2021). Additionally, the consistency offered by this structured approach aids crowdsourced workers on the MTurk platform to perform annotations with uniformity, which enhances the credibility and utility of the annotated corpus.

During the annotation process, a crowdsourcing task was designed using the MTurk API to annotate each of the 2,500 sentences in the dataset with emotion categories, as determined by the MTurk workers. The workers were instructed to place each sentence into the most appropriate emotion category or categories from Plutchik's eight emotion classes, using radio buttons. They were allowed to select up to two emotion categories per sentence, from "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", and an additional class of "No emotion" for cases where no emotion was detected in the text. Each sentence in the corpus was annotated by five different workers. This approach was implemented to capture the complexity and nuance of human emotions, as certain sentences might evoke multiple emotions simultaneously. Allowing multiple classifications ensured a more comprehensive and accurate representation of the emotional content, reflecting the practical scenarios in which people sometimes express multiple emotions in their texts. This method also aimed to reduce the cognitive load on workers by acknowledging and accommodating the natural overlap between different emotional states.

Clear instructions on how to annotate the Human Intelligence Task (HIT) were provided to the workers to increase the quality of their annotations. These instructions included a definition of each emotion category along with an example sentence from the Cambridge Dictionary [16] and a previous experiment. For instance, the emotion class "joy" was defined as "great happiness", accompanied by the example sentence: "Lots to see outside, and we had fun pretending we were about to set sail on an adventure!" Figure 3.2 presents an example of the task of annotating a sentence and the instructions given to the workers. A screenshot of the definitions provided to the workers is shown in Figure 3.3, while Figure 3.4 displays a screenshot of the examples given to the workers.

Additionally, the MTurk API allows the task requester to set multiple requirements for workers before they accept the task. These requirements aimed to reduce the

[16] https://dictionary.cambridge.org, (accessed 28 January 2025)



General Instructions:

Please classify the following **25 sentences** into its **most appropriate emotion classes**.

- you can choose **up to two emotion classes**.
- If you think it **has just one class** please choose "No Emotion" as the second emotion class, and any type of the emotion intensity.
- And, if you think it **has more than two categories** please chose the two that you think are **the most apparent ones.**

Sentence 1:
This museum, unlike others were I found some arrogant tone, is cozy and provokes a very sympathetic feel on visitors.

1. What is the First Emotion Class?

○ Joy	○ Trust	○ Surprise	○ Anticipation	○ Anger	○ Disgust	○ Sadness	○ Fear	○ No Emotion

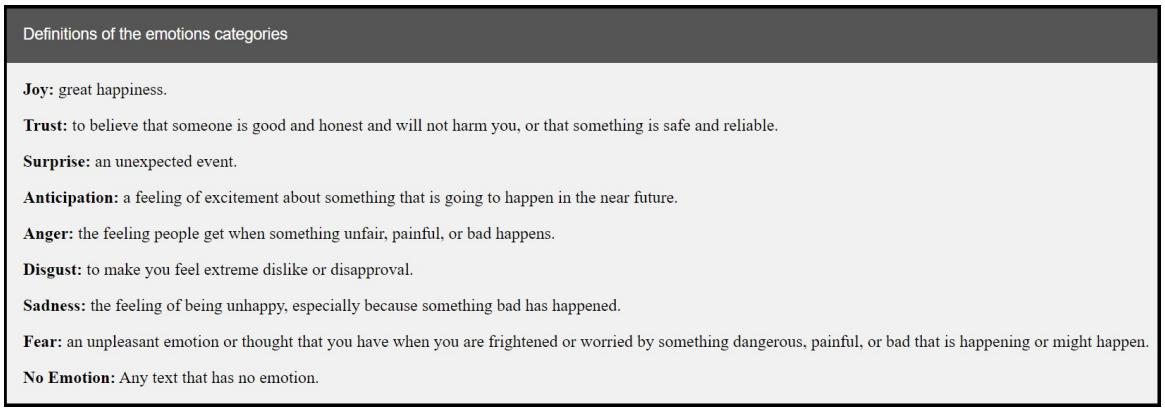
2. What is the Second Emotion Class?

○ Joy	○ Trust	○ Surprise	○ Anticipation	○ Anger	○ Disgust	○ Sadness	○ Fear	○ No Emotion

3. What type of place is being reviewed?

○ Museum	○ Zoo	○ Unknown & something else

Figure 3.2: An example annotation task



Definitions of the emotions categories

Joy: great happiness.

Trust: to believe that someone is good and honest and will not harm you, or that something is safe and reliable.

Surprise: an unexpected event.

Anticipation: a feeling of excitement about something that is going to happen in the near future.

Anger: the feeling people get when something unfair, painful, or bad happens.

Disgust: to make you feel extreme dislike or disapproval.

Sadness: the feeling of being unhappy, especially because something bad has happened.

Fear: an unpleasant emotion or thought that you have when you are frightened or worried by something dangerous, painful, or bad that is happening or might happen.

No Emotion: Any text that has no emotion.

Figure 3.3: Definitions of the emotions' categories

number of spammers submitting annotations. To ensure the reliability of the workers, the annotation task was restricted to workers who met the following minimum criteria:

1. **Number of Previous Tasks**: Workers needed to have at least 10,000 approved HITs, based on their previous submissions.

2. **Approval Rate**: Workers were required to have a minimum HIT approval rate of 95% for all previously completed HITs.

Moreover, considering the language of the data, it was expected that all of the workers who participated in the annotation process were either native or fluent English speakers. When setting up the crowdsourcing task, the dataset was divided into 100 chunks, each containing 25 sentences to be annotated by five workers. Each chunk

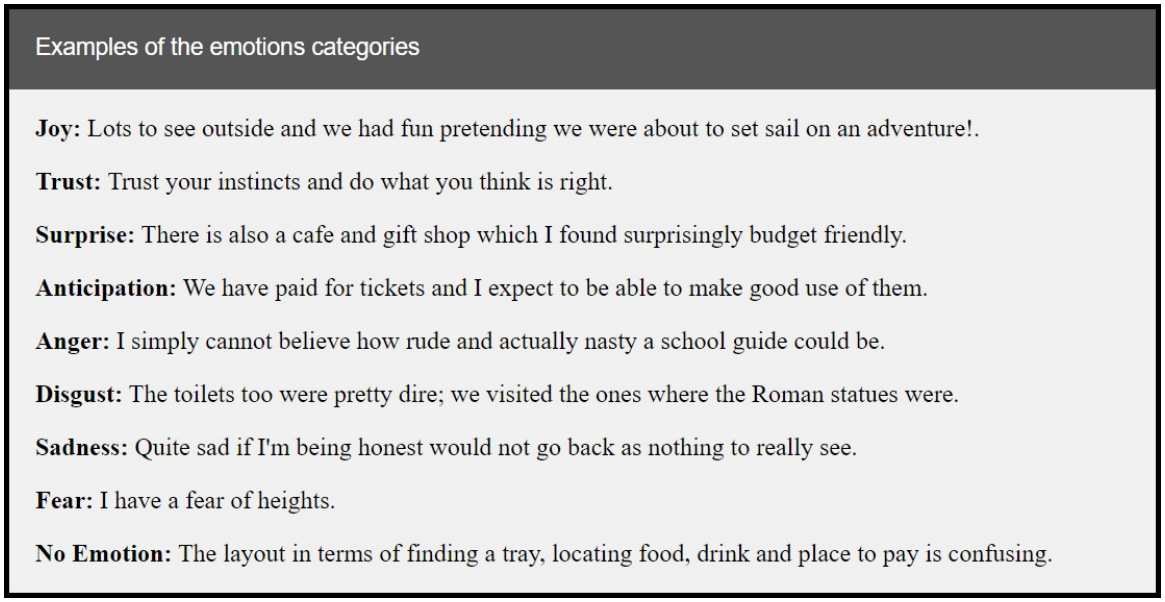


Figure 3.4: Examples of the emotions' categories

was processed as a single HIT. For each sentence, the workers were asked to answer three questions: two to classify a sentence into the appropriate emotion classes and an additional question to determine the type of place the sentence was about, as shown in Figure 3.2.

To manage the quality of the HITs that were submitted by the workers, gold standard questions were injected into each HIT. These questions had pre-validated, known answers, and provided a method for assessing worker performance and identifying spammers, based on their handling of the gold standard questions (Liu et al., 2012). The gold standard questions were carefully selected to have clear, definite answers, without any ambiguity. This ensured a fair evaluation of the workers' performance. Given the subjective nature of emotion classification, different types of questions were chosen as the gold standard questions. Prior to initiating the crowdsourcing task, standard answers to a set of 300 questions, related to determining the type of place mentioned in the sentence, were manually created.

These questions focus on specific types of places. For instance, the predetermined standard answer to the sentence: *"What a waste of money absolutely rubbish a poor excuse for a museum save your money and don't bother"* was internally defined as "museum". The expectation was that the MTurk worker would choose this exact answer. Within the crowdsourcing task, each chunk comprising 25 sentences included three, pre-set gold standard answers (12%). Consequently, 24% of the tasks that were submitted by the workers were subjected to quality assurance filtering, leading to

rejection due to an incorrect answer to at least one gold standard question.

The internal estimation indicates that each chunk took the workers approximately eight minutes to complete. This estimation was derived from a pilot study that was conducted prior to the main annotation task. During the pilot phase, a group of workers was asked to annotate sample chunks, and the time that they spent on this task was recorded. The average time taken by these workers provided a benchmark for estimating the duration of the entire annotation process. The workers received compensation of $1.44 per chunk, resulting in an hourly pay rate of $10.80, which surpasses the minimum wage in the United States [17]. Table 3.6 furnishes an overview of the crowdsourcing task.

Number of sentences in the dataset	2,500 sentences
Number of dataset chunks	100 chunks
Number of sentences per chunk	25 sentences
Number of questions per sentences	3 questions
Number of Gold standard questions per chunk	3 questions
Number of submitted tasks by the workers	662 tasks
Number of accepted tasks	500 tasks
Number of rejected tasks	162 tasks
Number of annotations per sentence	5 annotations

Table 3.6: An overview of the crowdsourcing task

3.5 Analysis of Emotion Annotation and Quality

3.5.1 Analysis of Emotion Annotation Samples

This section describes the analysis of selected samples of the outcomes of a crowdsourcing emotion annotation task, focusing on the comparative annotations of workers across sample-distinct tasks. The consistency and reliability of these annotations was explored by identifying any significant discrepancies among the annotators' interpretations. In undertaking this comparative analysis, the aim was to uncover any patterns and insights that reveal the strengths and challenges of using crowdsourcing for emotion annotation. The following sentence samples will be discussed in detail. Each sentence was annotated by five different annotators, who

[17]https://www.dol.gov/general/topic/wages/minimumwage, (accessed 21 June 2024)

provided their interpretations of the emotions conveyed, with each annotator assigning up to two emotion categories per sentence.

Sample Review 1: "This museum, unlike others were I found some arrogant tone, is cozy and provokes a very sympathetic feel on visitors."

As shown in Table 3.7, the most frequently identified emotion is "Joy", which was chosen by four of the five annotators, either as emotion 1 or emotion 2. This strong consensus suggests that the description of the museum as "cozy" and as evoking a "sympathetic feel" predominantly elicits feelings of happiness and contentment among the annotators. The positive, welcoming tone of the sentence probably contributes to this widespread perception of joy.

In addition to joy, the emotions of "Trust" and "Surprise" are also notable, each of which were selected by two annotators. The selection of "Trust" indicates that the museum's atmosphere was perceived as reliable and comforting, fostering a sense of confidence in the visitors' experience. On the other hand, the emotion of "Surprise" reflects the unexpected pleasantness of the museum, particularly when contrasted with others that have an "arrogant tone". This element of surprise adds a layer of unexpected delight to the annotators' interpretations.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Surprise	Joy	Joy	Trust
Emotion 2	Trust	Joy	Surprise	-	-

Table 3.7: Emotion distribution of the first sentence

Sample Review 2: "Never in my life have I ever been made to feel so bad, and paying fortune for tickets to feel like that."

As shown in Table 3.8, the major emotion that was identified by the annotators is "Anger", which appeared in every annotator's selections. This high frequency of "Anger" suggests that the sentiment of being made to feel bad, especially after paying a significant amount for tickets, predominantly evokes feelings of frustration and resentment among the annotators. This significant monetary investment, coupled with a negative experience, clearly leads to feelings of anger and dissatisfaction.

"Surprise" was selected by two annotators, indicating that the negative experience was also unexpected. This emotion reflects the shock and dismay felt by the individual due to the stark contrast between the high cost and the poor experience. The presence of "Surprise" alongside "Anger" underscores the unexpected nature of the disappointing experience, adding a layer of shock to the emotional response.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Surprise	Anger	Surprise	Anger	Anger
Emotion 2	Anger	Sadness	Anger	-	-

Table 3.8: Emotion distribution of the second sentence

Sample Review 3: "Just a pity that the person serving was rather brusque in manner and lacking charm."

As shown in Table 3.9, "Anger" is the most frequently identified emotion, appearing four times across the annotations. This suggests that the brusque manner and lack of charm of the respective servers predominantly evoked strong feelings of frustration and resentment among the annotators. The negative interaction clearly led to anger, reflecting significant dissatisfaction with the service provided.

"Disgust" was identified twice, indicating that the behaviour of the server not only angered but also repelled the reviewers. The choice of "Disgust" suggests that the service was perceived as fundamentally unpleasant, highlighting a strong aversion to the brusque, discourteous manner described.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anger	Anger	Disgust	Surprise	Disgust
Emotion 2	-	-	Anger	Anger	-

Table 3.9: Emotion distribution of the third sentence

Sample Review 4: "A brilliant experience, value for money a true love for the club was present throughout the tour."

As shown in Table 3.10, the main emotion identified by the annotators is "Joy", which was selected by four of the five annotators. This strong consensus suggests that the experience described in the sentence evokes a significant feeling of happiness and satisfaction. The use of phrases like "brilliant experience" and "value for money", along with the sentiment of a "true love for the club", contributes to this overall positive emotional response, reflecting contentment and delight. The occurrence of "Trust" adds another layer to the emotional response, indicating that the experience also fostered a sense of confidence and trust.

Sample Review 5: "I can't understand why they can't provide good padding for a seat which isn't cheap so people can enjoy the shows."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Joy	Joy	Trust	Joy
Emotion 2	-	-	-	-	-

Table 3.10: Emotion distribution of the fourth sentence

As shown in Table 3.11, the uppermost emotion identified is "Anger", with all five annotators selecting this emotion category. This strong consensus indicates that the sentence evokes a significant feeling of frustration and resentment. The complaint about the lack of good padding for expensive seats, which impacts the ability to enjoy the shows, is a clear source of dissatisfaction. The repeated identification of "Anger" emphasises the intensity of the negative emotional response to the perceived inadequacy and the failure to meet expectations, especially given the high cost of the seats.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anger	Anger	Anger	Anger	Anger
Emotion 2	-	-	-	-	-

Table 3.11: Emotion distribution of the fifth sentence

Sample Review 6: "The entrance to the spa also had a strong smell of cooked food, not what you would expect from a 5 star spa."

As illustrated in Table 3.12, the primary emotion identified is "Disgust", which was selected by all five annotators. This unanimous selection indicates that the smell of cooked food at the entrance of a 5-star spa evoked strong feelings of repulsion and aversion. The reviews appeared to find this situation highly unpleasant and inappropriate for the setting, which is expected to be pristine and relaxing. The consistent recognition of "Disgust" underlines the intensity of the negative reaction to the unexpected, inappropriate smell in a supposedly luxurious environment.

Additionally, "Surprise" was identified by two annotators. The presence of "Surprise" indicates that the smell appeared to be incongruous with the ambience expected at a 5-star spa. The shock may have stemmed from the contrast between the expectations of a high-end spa experience and the reality of encountering an unpleasant odour.

Sample Review 7: "A disappointing expensive experience, going to spas regularly worldwide I was expecting something an experience at a

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Disgust	Disgust	Disgust	Disgust	Disgust
Emotion 2	Anger	Surprise	-	-	Surprise

Table 3.12: Emotion distribution of the sixth sentence

different level."

As shown in Table 3.13, the predominant emotion identified is "Sadness", which was selected by four of the five annotators. This indicates that the disappointing, costly experience predominantly evoked feelings of unhappiness and disappointment. The reviewers report a sense of loss and dissatisfaction, as the high expectations that they had, based on their previous experiences at other spas worldwide, were not met. The consistent identification of "Sadness" underscores the emotional impact of these unfulfilled expectations and the perceived failure of the spa to deliver a high-quality experience.

In addition, "Anticipation" was identified twice. This suggests that there was a significant level of expectation and eagerness leading up to the experience. The reviewers anticipated a superior experience due to their familiarity with high-standard spas globally. The presence of "Anticipation" highlights the gap between their expectations and the reality of the experience, which probably intensified their feelings of sadness.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Sadness	Anticipation	Sadness	Sadness	Sadness
Emotion 2	-	Anger	Anticipation	-	-

Table 3.13: Emotion distribution of the seventh sentence

Sample Review 8: 'After hearing good things about this place and it being recommended by a colleague I brought my mother for a rare mother and daughter treat."

As shown in Table 3.14, the foremost emotion identified is "Anticipation", selected by all five annotators. This indicates that the reviewer was looking for a significant level of eagerness and excitement about the visit. The positive recommendations and the special occasion of treating their mother likely heightened their expectations. The consistent recognition of "Anticipation" underscores the emotional investment and the sense of looking forward to the experience.

Additionally, "Trust" was identified by two annotators. This suggests that the reviewer not only felt anticipation but also a sense of confidence and reliability based on the colleague's recommendation. The occurrence of "Trust" highlights the reviewer's belief in the credibility of the positive feedback they received and their expectation that the experience would meet those high standards.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anticipation	Anticipation	Anticipation	Anticipation	Anticipation
Emotion 2	-	-	Trust	-	Trust

Table 3.14: Emotion distribution of the eighth sentence

Sample Review 9: "Good point - The food when it finally arrived was nice and better than expected given level of service."

As demonstrated in Table 3.15, the major emotion identified was "Surprise", which was selected by three of the five annotators. This indicates that the food quality exceeded the reviewer's expectations, especially considering the poor service experienced. The reviewer probably felt an unexpectedly positive reaction when the food proved to be better than anticipated, highlighting the contrast between their expectations and the actual outcome. The continuous identification of "Surprise" highlights the impact of this unexpected positive experience.

"Joy" was identified by two annotators, indicating a sense of happiness and satisfaction with the food. Although the level of service was below standard, the enjoyment of the food contributed to a positive emotional response. The appearance of "Joy" emphasises that the final outcome of the meal brought pleasure, which somewhat compensated for the earlier dissatisfaction with the service.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Surprise	Joy	Joy	Surprise	Surprise
Emotion 2	-	-	-	-	-

Table 3.15: Emotion distribution of the ninth sentence

Sample Review 10: "It was the perfect celebration for a birthday and we're immensely grateful to all the staff involved, both paid and volunteer, as they are the bedrock of the operation."

As shown in Table 3.16, the predominant emotion recognised is "Joy", which was selected by all five annotators. This indicates that the birthday celebration evoked a

significant feeling of happiness and satisfaction. The phrase "the perfect celebration" reflects the high level of enjoyment and the positive experience that were shared by those involved. The consistent identification of "Joy" highlights the overwhelming sense of contentment and delight with the event.

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Joy	Joy	Joy	Joy
Emotion 2	-	Trust	-	-	-

Table 3.16: Emotion distribution of the tenth sentence

In conclusion, these samples highlight the varied emotional responses expressed regarding different scenarios. The consistent identification of emotions across the annotators demonstrates the reliability of the crowdsourcing approach in capturing the nuanced emotional impact of specific situations. Whether the unanimous sense of anger toward inadequate seating, the prevalent feeling of disgust at an unpleasant spa experience, the shared anticipation and trust in a highly recommended venue, the mixed emotions of surprise and joy at an unexpectedly good meal, or the overwhelming joy and trust in a well-organised birthday celebration, the annotations provided valuable insights into the annotators' collective emotional perceptions. These findings underscore the importance of considering multiple emotion classes per sentence and treating them with equal significance, as they collectively enrich our understanding of how various experiences resonate on an emotional level. By treating both selection of emotion classes with equal importance, a comprehensive understanding of the varied emotional impact of the described scenario is gained, emphasising the multifaceted nature of the emotional responses. This comprehensive analysis validates the efficacy of employing crowdsourcing for emotion annotation and emphasises the diverse emotional landscapes that different contexts can represent.

3.5.2 Evaluation of Annotation Quality

The evaluation of fine-grained emotion annotation raises numerous challenges, particularly in regard to assessing Inter-Annotator Agreement (IAA). A contributing factor is the potential for reviews to have multiple annotations, a characteristic that is pertinent to this work, where workers can assign a sentence to two emotion classes concurrently. Consequently, the agreement metric needed to accommodate the values of multicategory sets. To measure the quality of the experiment's results, the agreement metric was vital. It is important to note that emotion classification

is subjective by nature, and different individuals may have varying opinions when classifying the same text.

In order to assess the quality of the annotation obtained through MTurk, however, two metrics of the IAA were calculated; namely, the Krippendorff's alpha with MASI (Measuring Agreement on Set-valued Items) (Passonneau, 2006), and the Plutchik Emotion Agreement (PEA) metric (Desai et al., 2020).

The Krippendorff's alpha with MASI is adept at measuring the agreement between a set of valued items. It serves as a distance metric between sets, with a value of 1 indicating identical sets and 0 indicating disjoint sets. Formula 3.1 represents MASI, where A and B denote two sets of annotations.

$$Jacc(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$Pass(A, B) = \begin{cases} 0 \text{ if } A \cap B = \emptyset \\ \frac{1}{3} \text{ if } A \cap B \neq \emptyset \\ \frac{2}{3} \text{ if } A \subset B \text{ or } B \subset A \\ 1 \text{ if } A = B \end{cases} .$$

$$MASI(A, B) = Jacc(A, B) \times Pass(A, B) \tag{3.1}$$

This metric treats disagreement regarding emotion classes equally, which is not ideal for the Plutchik emotion classes, so another metric method needs to be tested for this unique kind of annotation agreement.

The PEA metric (Desai et al., 2020) was employed to calculate the IAA for the annotations based on the categories of the Plutchik Wheel of Emotion. The PEA tackles the penalisation issue of disagreement by considering relatively similar emotion classes. For instance, "Joy" is deemed more similar to "Trust" than to "Sadness". Figure 3.5 illustrates an example of the PEA metric, where the distance score for the selected emotion pair is 0.25.

Moreover, Desai et al. (2020) introduced Formula 3.2 for calculating the agreement among workers $d(w_x, w_y)$ as follows:

$$\frac{1}{n} \sum_{i=1}^{n} \underset{j}{max} \left(\left| 1 - \frac{1}{\pi} \left| f\left(e_x^{(i)}\right) - f\left(e_y^{(j)}\right) \right| \right| \right) \tag{3.2}$$

where, w_x and w_y represent the workers, $f\left(e_x^{(i)}\right)$ and $f\left(e_y^{(j)}\right)$ indicate the annotation sets by the workers, and $\frac{1}{\pi}$ is a normalising constant.

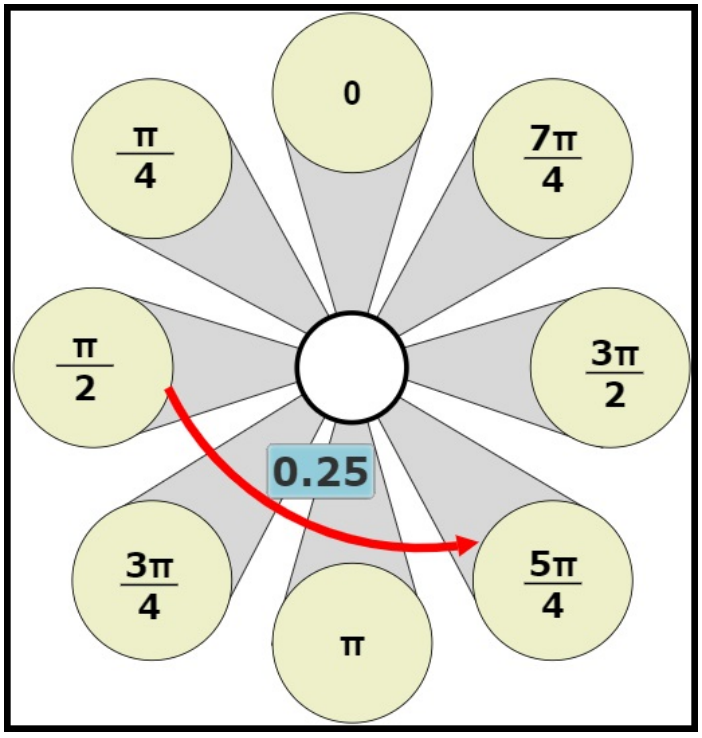


Figure 3.5: Illustration of the PEA metric, where each emotion is given a radian value

Additionally, the interpretation of the agreement score values can have diverse implications for different tasks. For instance, Landis and Koch (1977) proposed six levels of IAA, including "Poor Agreement", "Slight Agreement", "Fair Agreement", "Moderate Agreement", "Substantial Agreement", and "Almost Perfect Agreement". This interpretation framework is widely adopted in the field of NLP in relation to crowdsourcing annotation tasks (Mohammad and Turney, 2013; Ahmad et al., 2019; Stajner, 2021). Table 3.17 provides the ranges of the agreement score for the six levels of inter-annotator agreement.

For this experiment, the Krippendorff's alpha agreement score was 0.37, as shown in Table 3.18, which indicates fair agreement. On the other hand, the PEA, which is more suitable for the Plutchik Wheel of Emotions produced better agreement scores. Because the workers were allowed to annotate each sentence with more than one emotion category, the PEA agreement score was calculated using three different methods, as follows:

1. **PEA High**: This method takes into account the best set of annotations.

2. **PEA Low**: This method takes into account the worst set of annotations.

Agreement Score	**Interpretation**
<0	Poor Agreement
0.00 – 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 - 0.60	Moderate Agreement
0.61 - 0.80	Substantial Agreement
0.81 - 1.00	Almost Perfect Agreement

Table 3.17: Landis and Koch's six ranges of IAA

3. **PEA**: This method averages the **PEA High** and **PEA Low** scores.

For example, if the first worker chooses ("Joy" and "Anticipation") and the second worker chooses ("Joy" and "Trust"), then the **PEA High** will calculate the agreement using the best set of annotations ("Joy" and "Joy"), while the **PEA Low** will be based on the worst set of annotations ("Anticipation" and "Trust").

As shown in Table 3.18, the obtained agreement score using the **PEA High** method was 0.86, that obtained using the **PEA Low** method was 0.65, and that obtained with the **PEA** method was 0.75. The interpretation of these values, using Landis and Koch's agreement levels, indicates that the **PEA High** value reflects almost perfect agreement, while the **PEA Low** and **PEA** values indicate substantial agreement degrees.

Agreement method	**Agreement Score**	**Interpretation**
Krippendorff	0.37	Fair Agreement
PEA high	0.86	Almost Perfect Agreement
PEA low	0.65	Substantial Agreement
PEA	0.75	Substantial Agreement

Table 3.18: The agreement score for the MTurk crowdsourcing annotation

3.5.3 Distribution of Data for the Emotion Categories

In order to produce the TORCEv1 dataset, the majority voting of emotion class was implemented. Majority voting is a method for choosing the most selected emotion class by the Mturk workers to represent the emotion class for a given sentence. Using

majority votes is a valuable method for consolidating annotations from multiple annotators, mitigating personal biases and inaccuracies, and achieving agreement on the most suitable emotion class for each sentence. This approach enhances the dataset's reliability. In the crowdsourcing annotation, the workers were instructed to annotate each sentence with a maximum of two emotion classes, resulting in a maximum of ten choices per sentence. With five workers annotating each sentence, majority voting required at least three votes. Additionally, due to the possibility of multiple choices, a double majority vote scenario was addressed as follows:

1. **Equal Annotations**: If a sentence received equal votes for two emotion classes, both classes were considered valid. For instance, if "joy" and "trust" each received four votes, the sentence would be annotated with both the "joy" and "trust" emotion classes.

2. **Non-equal Annotations**: If one emotion class received more votes than the other, the sentence was assigned to the majority emotion class only. For example, if "joy" received five votes and "trust" received three, the sentence would be annotated with the "joy" emotion class only.

Table 3.19 illustrates the number of sentences with equal majority voting classes, revealing the patterns in conflicting annotations, especially for the pairs ("Joy", "Trust") and ("Anger", "Disgust"). Notably, the "Sadness" emotion class exhibited no conflicting annotations. Conversely, Table 3.20 showcases the number of sentences with a unique majority vote. The number of sentences with at least one emotion class is displayed in Table 3.21. In the dataset, the "Joy" emotion class predominated, while the "Fear" emotion class was infrequent. Moreover, 378 sentences achieved complete agreement on the emotion class among the workers. However, due to the subjective nature of emotion classification tasks, most of the sentences received a majority vote of three.

To summarize, out of 2,500 sentences in the TORCEv1 dataset, 2,357 were annotated with one emotion class, 116 with two emotion classes, and only 27 lacked a decisive vote by the workers, emphasising the effectiveness of the majority voting approach in emotion annotation tasks. Figure 3.6 illustrates the distribution of these emotion categories across the TORCEv1 dataset, highlighting the prevalence of each emotion class among the annotated sentences. Notably, "joy" emerges as the most frequently represented category, accounting for a substantial proportion of the dataset. Conversely, "fear" is the least represented category, suggesting that this particular emotion is rarely expressed in tourism reviews.

Emotion Pair	**Votes of 4**	**Votes of 3**	**Total**
Anger - Anticipation	0	2	2
Anger - Disgust	10	26	36
Anger - Surprise	0	2	2
Anticipation - Disgust	0	2	2
Anticipation - Joy	0	3	3
Anticipation - Trust	0	6	6
Disgust - Fear	0	1	1
Disgust - Surprise	0	2	2
Joy - Surprise	0	11	11
Joy - Trust	12	31	43
Surprise - Trust	0	8	8
Total	22	94	116

Table 3.19: The counting of sentences that have double majority votes in TORCEv1

Emotion Class	**Votes of 5**	**Votes of 4**	**Votes of 3**	**Total**
Anger	51	74	197	322
Anticipation	28	37	103	168
Disgust	45	66	165	276
Fear	0	8	23	31
Joy	119	95	446	660
Sadness	51	62	162	275
Surprise	29	41	181	251
Trust	55	51	268	374
Total	378	434	1545	2357

Table 3.20: The counting of sentences that have a unique majority vote in TORCEv1

3.5.4 Adjusting the Emotion Annotation Categories for Practical Application

To address the need for a more focused benchmark dataset that is designed for tourism-related emotional analysis, TORCEv2 was subsequently developed.

Emotion Class	Votes of 5	Votes of 4	Votes of 3	Total
Anger	51	84	227	362
Anticipation	28	37	116	181
Disgust	45	76	196	317
Fear	0	8	24	32
Joy	119	107	491	717
Sadness	51	62	162	275
Surprise	29	41	204	274
Trust	55	63	313	431
Total	378	478	1733	2589

Table 3.21: The counting of sentences that have majority votes in TORCEv1

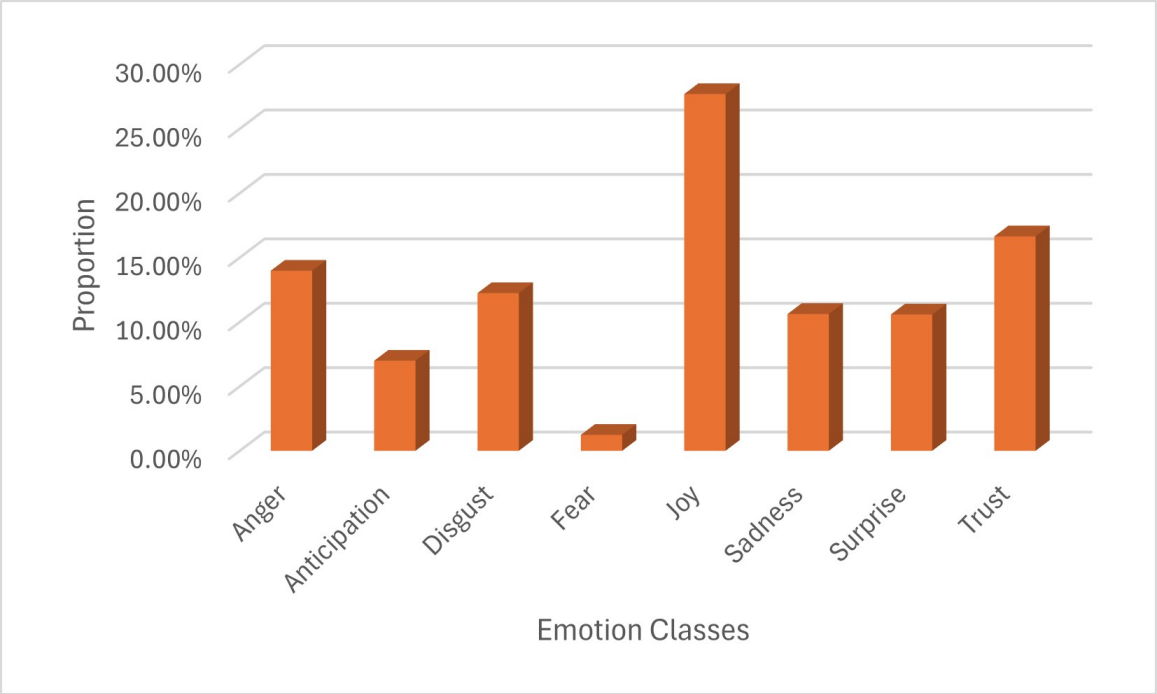


Figure 3.6: Distribution of the emotion classes across the TORCEv1 dataset

TORCEv2 inherits the data and annotations from the TORCEv1 dataset but aims to provide a more refined, targeted resource for emotion detection experiments within

the tourism domain. This dataset was created to fill a gap in the available resources for emotion detection based on tourism-related text, offering a domain-specific dataset that accurately reflects the unique linguistic characteristics and emotional range found in tourism reviews. This version, TORCEv2, will be used in the emotion classification experiment that will be discussed in Chapter 4.

As part of the dataset refinement process, an additional analysis was conducted to identify the correlations among the different emotion categories present in the TORCEv1 dataset. Specifically, the Mutual Information (MI) association metric was calculated for emotion pairs, allowing the examination of how frequently certain emotions were co-assigned by the MTurk annotators to the same sentence. This metric provides a quantitative measure of association, highlighting how frequently particular emotion categories tend to occur together within the tourism review dataset.

Mutual Information (MI) is an information-theoretic measure that quantifies the degree of association between two random variables by comparing their joint distribution with the product of their marginal distributions. In the context of emotion co-occurrence, MI evaluates whether two emotion categories appear together in sentences more frequently than would be expected by chance. Mathematically, MI between two categories X and Y is defined as:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) \tag{3.3}$$

where $p(x, y)$ represents the joint probability of the two categories being assigned together, and $p(x)$ and $p(y)$ represent their individual marginal probabilities (Cover and Thomas, 2006). A higher MI value indicates a stronger association, meaning that the two emotion categories co-occur more often than expected. Conversely, a low value suggests little to no meaningful association.

Furthermore, a value of one was selected as the threshold for distinguishing between weak and strong associations when interpreting MI scores. Values below one were considered to indicate limited or weaker co-occurrence between emotion categories, while values above one were taken to represent stronger associations that occur more frequently than expected. It is important to note that MI has no fixed upper bound, and its interpretation is therefore relative to the distribution of values observed within the dataset. Consequently, the threshold of one was adopted as a practical criterion to separate weaker associations from those considered to be more substantial.

The results of this analysis, as detailed in Table 3.22, demonstrate that strong associations exist between specific emotion pairs. Specifically strong associations were observed between both “anger” and “disgust” (with an MI score of 2.221), and “joy” and “trust” (with an MI score of 1.415). These associations suggest that these pairs of emotions are often regarded as overlapping or interconnected

within the context of tourism reviews, potentially due to the varied expressions of satisfaction, disappointment, or trust found in user reviews. For example, "anger" and "disgust" often co-occur, possibly reflecting negative emotions that arise from unpleasant experiences or dissatisfaction within tourism-related reviews. Similarly, the frequent co-occurrence of "joy" and "trust" may reflect positive emotional responses associated with satisfaction, reliability, and positive experiences in tourism text. The patterns of association align with Plutchik's model, where emotions that demonstrate complementary or adjacent characteristics are positioned close to each other (Troiano et al., 2023).

Emotion Class	**Anticipation**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Trust**
Anger	0.32	2.221	0.871	0	0.734	0.749	0
Anticipation	-	0.218	0	0.511	0.339	0.765	0.873
Disgust	-	-	0.752	0	0.882	0.696	0
Fear	-	-	-	0	0.629	0.565	0
Joy	-	-	-	-	0	0.508	1.415
Sadness	-	-	-	-	-	0.496	0
Surprise	-	-	-	-	-	-	0.892

Table 3.22: The mutual information association scores among the emotion classes

Therefore, based on these MI scores, the emotion pairs with high association scores were merged into single categories for the purpose of the experiments. Specifically, "anger" was merged with "disgust", and "joy" with "trust", creating composite categories that more accurately reflect the co-occurrence patterns observed in the dataset. This adjustment step enhances the dataset's suitability for the intended experiments by improving the accuracy and interpretability of the emotion classification. By combining strongly-associated emotion categories, the dataset better reflects the emotional landscape expressed in tourism reviews, where certain emotional classes are commonly correlated. This adjustment provides a more reliable

foundation for evaluating the performance of machine learning models that analyse expressions of emotions in tourism-related content.

Following these adjustments, the TORCEv2 dataset composition was finalised, with updated statistics displayed in Table 3.23. This table details the distribution of the majority vote counts across the emotion categories in the dataset, after the classification structure was refined. Notably, the "fear" category was excluded from the TORCEv2 dataset due to its low frequency, with only 14 sentences receiving majority votes for "fear" across all annotations. This category represented less than 1% of the total annotations, making it insufficiently represented for reliable training and evaluation purposes.

Table 3.23 also shows the counts for each emotion class based on the majority votes, with the annotations categorised into groups that received 3, 4, or 5 votes out of a possible 5. For example, "joy" is the most frequently occurring category, with 1,214 sentences receiving majority votes for this category, including 187 sentences with complete agreement among the workers. This suggests that "joy" may reflect a wide spread of positive emotions in tourism reviews. In contrast, "surprise", "anticipation" and "sadness" received fewer majority votes, indicating that they occur less frequently. By eliminating the rare "fear" category and combining the associated categories, the TORCEv2 dataset provides a more balanced, representative sample of expressions of emotions in tourism reviews. This final structure enhances the dataset's utility with regard to training and evaluating machine learning models. Despite these efforts to create a more balanced dataset, however, an imbalance remains. This arises primarily from the challenge of predicting variations in workers' outputs. Factors such as differences between individual judgments and varying levels of expertise exacerbate this issue.

Emotion Class	**Votes of 5**	**Votes of 4**	**Votes of 3**	**Total**
Anger	99	180	334	613
Anticipation	28	37	79	144
Fear	0	8	6	14
Joy	187	213	814	1214
Sadness	51	62	145	258
Surprise	29	41	176	246
Total	394	541	1554	2489

Table 3.23: The counting of sentences that have majority votes in TORCEv2

The final distribution of the emotion categories within the TORCEv2 dataset is

illustrated in Figure 3.7, which demonstrates the frequency of each emotion class following the dataset's refinement. Approximately 49% of the sentences falls under the "joy" category, with the remaining 51% distributed across the four other emotion classes. This distribution reflects the significance of positive emotions in tourism reviews, where satisfaction and enjoyment are commonly expressed.

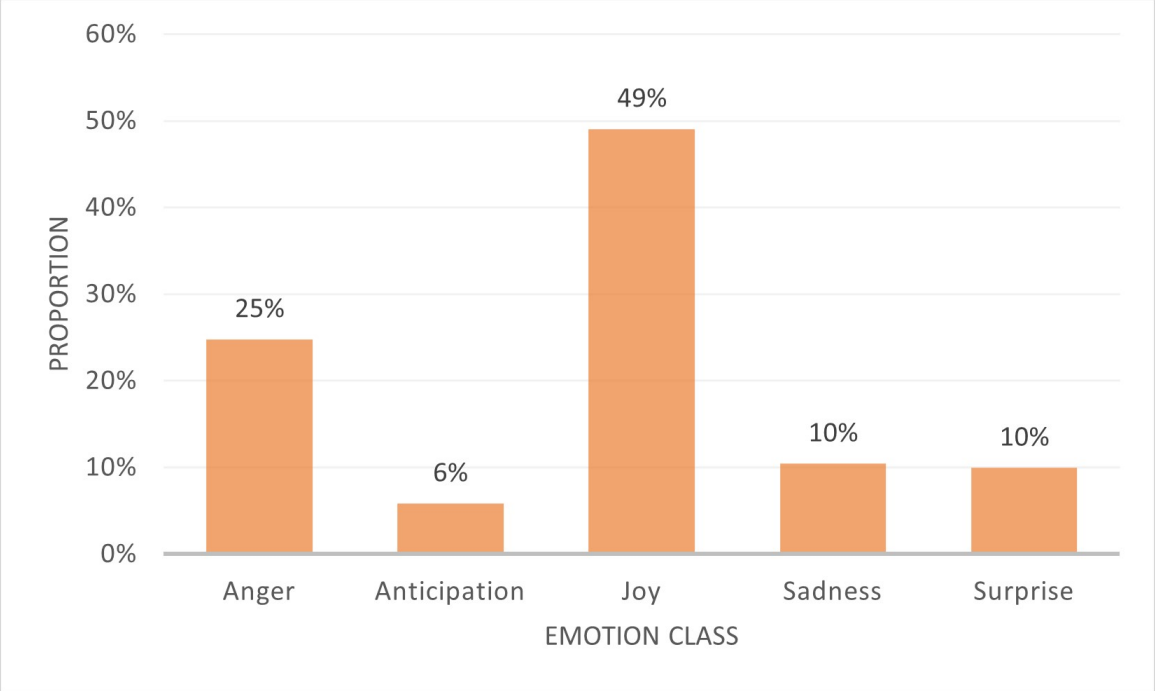


Figure 3.7: Distribution of the emotion categories across the TORCEv2 dataset

Within the dataset, a small number of sentences, 11 in total, did not achieve a clear majority vote among the annotators regarding the assigned emotion category. These sentences were subsequently classified under the label "no emotion". The presence of such cases is expected in crowdsourced emotion annotation tasks, as some sentences may be inherently ambiguous, neutral in tone, or open to multiple interpretations depending on the annotator's perspective. For instance, sentences that describe factual information or practical aspects of tourism, such as transport details, directions, or simple statements of service, may lack explicit affective content and thus do not produce agreement on an emotional label. Furthermore, disagreement among annotators can arise when a sentence contains subtle or mixed indications, where no single emotion is sufficiently dominant to achieve majority agreement. By assigning these sentences to the "no emotion" category, the dataset preserves annotation reliability by avoiding forced classification into categories that are not

clearly supported by the annotators' judgments. At the same time, the inclusion of this category reflects the reality that not all textual content in tourism reviews carries identifiable emotional categories.

3.6 Chapter Summary

In summary, the tourism-related reviews were collected using the TripAdvisor platform, which was selected due to its extensive database of user-generated content. The data collection process was guided by the aim to create a high-quality corpus, suitable for emotion annotation. The collected data were filtered using specific criteria to select 2,500 sentences, ensuring that the chosen subset of tourism reviews remaining suitable for the crowdsourcing task.

The crowdsourcing process for annotating these sentences with emotion labels was performed on the MTurk platform, providing a structured approach to obtain high-quality annotations. The workers were provided with clear instructions and asked to assign up to two emotions per sentence from the eight categories: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", and an extra option "No Emotion" (indicating no emotion in a given sentence). The MTurk API facilitated the creation of a user-friendly annotation task with clear instructions to guide the workers, ensuring consistent, high-quality responses.

The quality of the annotations was evaluated using IAA metrics, specifically the Krippendorff's alpha with MASI and the PEA metric. The PEA metric was particularly effective, as it incorporated the relationships and similarities between Plutchik's emotion classes, addressing potential penalisation issues for disagreements. Among the methods, PEA High achieved almost perfect agreement, emphasising the reliability of the annotations. The agreement level was interpreted using Landis and Koch's agreement scale, providing a comprehensive evaluation of worker consistency.

Two versions of the TORCE dataset annotation were developed. TORCEv1 includes all eight emotion categories from Plutchik's model, capturing nuanced expressions of emotions within tourism reviews. In contrast, TORCEv2 simplifies the TORCEv1 dataset by focusing on five core emotion categories, offering a more streamlined approach, that was suitable for the subsequent emotion detection task. To ensure the reliability of the annotations, a majority voting method was employed with the two versions, to resolve any conflicting votes effectively.

Finally, through this part of the thesis, RQ1 was addressed by demonstrating the efficacy of crowdsourcing annotation in producing a reliable tourism emotion dataset. The experimental results showed that the high agreement levels among the annotators, as demonstrated by using the PEA high metric, validated the reliability of the crowdsourcing process. This agreement score validated the effectiveness of crowdsourcing as a robust methodology for emotion annotation tasks.

Chapter 4

Emotion Detection

4.1 Introduction

This chapter presents a comprehensive analysis of emotion detection within the context of tourism reviews, evaluating the effectiveness of the existing mainstream tools and exploring the adaptation of LLMs to provide better classification results for the tourism-related data. Emotion detection is the process of identifying and categorising emotions in textual data, that has gained importance across various fields, including tourism, where customer feedback plays an imperative role in determining the consumer perceptions and business strategies. This chapter addresses the need for carefully designed approaches to capture these emotional classes accurately, keeping in mind the uniqueness of the language used in tourism reviews.

The primary objective of this chapter is to assess the limitations of the mainstream emotion detection tools when applied to tourism reviews and, secondly, to investigate the impact of fine-tuning LLMs to achieve greater accuracy in this context. Generic language models are indeed very powerful for general tasks but might lack the specificity required to perform optimally in the case of domain-specific applications. Thus, adapting these models through fine-tuning and data augmentation for the rarer emotion classes in the dataset is essential to unlock their potential for detecting nuanced expressions of emotions within tourism-related content. Overall, this chapter aims to contribute to broaden our understanding of emotion detection in tourism-related text, highlighting the practical applications and challenges associated with deploying LLM-based tools in this field.

4.2 Testing the Mainstream Emotion Detection Tools

The domain of emotion detection encompasses a diverse, evolving range of tools that are designed to interpret and classify the expressions of emotions embedded within textual data, as discussed in 2.4. These tools, ranging from algorithms and frameworks to specialised machine learning models, are applied across different domains to analyse user-generated content and extract the underlying emotions. Central to the emotion detection technology are resources such as annotated emotion corpora and machine learning models that are fine-tuned for emotion recognition. Each tool within this domain employs distinct methodologies, whether based on rule-based systems, lexicon-driven approaches, or advanced machine learning techniques, including deep learning, to identify and categorise the expressions of emotions in text accurately.

Evaluating the existing emotion detection tools, particularly within the context of tourism reviews, poses unique challenges. Tourism-related content is characterised by domain-specific language, cultural references, and subjective expressions, that reflect the highly personal and often emotional experiences of tourists. Unlike general-purpose text, tourism reviews contain language that captures both the enthusiasm associated with positive experiences and the disappointment arising from unmet expectations, often weaving in colloquial terms, idiomatic expressions, and specific cultural tones. These reviews are inherently subjective, with each piece of feedback shaped by individual perspectives, making the detection and accurate classification of these expressions of emotions a complex task.

This section assesses the performance of the mainstream emotion detection tools when applied to tourism reviews, examining their ability to navigate the difficulties of this domain-specific language. A critical objective of this evaluation is to determine whether these tools, which were designed primarily for use with generic text data, can effectively interpret the diverse expressions of emotions that are unique to tourism-related data. Given that these emotion detection tools are typically trained on datasets representing general or non-specialised language, their degree of adaptability to the specialised vocabulary and characteristics of tourism reviews is of particular interest. Ultimately, this investigation aims to clarify the extent to which the mainstream emotion detection tools can meet the specific needs of tourism review analysis.

4.2.1 Selected Tools for Evaluation

To evaluate emotion detection tools, four prominent approaches were analysed, each exemplifying a distinct methodology within the domain of NLP. These tools demonstrate varied strategies and structures for emotion recognition, offering insights

into their suitability for application in tourism-related contexts, where the text is often rich with emotions and subjective experiences.

1. LeXmo [18]: serves as an example of the lexicon-based emotion analysis tools, which are widely utilised for emotion detection. Built upon the NRC Emotion Lexicon (Mohammad and Turney, 2013), LeXmo is a Python-based tool that utilises a comprehensive lexicon of over 14,000 words, each associated with one or more emotions. The list of contained emotion classes includes: "anger", "disgust", "fear", "joy", "sadness", "surprise", "trust", and "anticipation". Developed through a crowdsourcing approach, this lexicon provides associations that enable LeXmo to classify the expressions of emotions within textual data effectively. The tool's primary strength lies in its ability to map words to emotional categories, making it a valuable resource for sentiment analysis, sarcasm detection, and emotion detection, in various contexts. In the context of tourism reviews, LeXmo was tested to evaluate its effectiveness in identifying emotional text associated with user experiences. By leveraging the predefined associations in the NRC Lexicon, LeXmo aids the interpretation of expressions of emotions by categorising the words and phrases that tourists commonly use to express their emotions.

2. EmoNet [19]: represents an advanced neural network-based approach to emotion detection, employing Gated Recurrent Neural Networks (GRNNs) for precise language understanding (Abdul-Mageed and Ungar, 2017). Unlike the lexicon-based methods, which rely on predefined word associations, GRNNs are capable of learning the dependencies across words and sentences, enabling the model to detect expressions of emotions even in contextually complex texts. EmoNet is trained on a diverse dataset, that includes social media posts, news articles, and movie reviews, providing it with a rich contextual foundation. This diversity allows EmoNet to adapt well to the varied expressions of emotions that occur in complex, challenging contexts. GRNNs, in particular, offer the advantage of recognising the long-term dependencies in text, making EmoNet a robust tool for capturing the subtle interplay between emotions in complex narratives. Its performance in detecting and categorising expressions of emotions was evaluated to assess its adaptability to tourism-related data, which often includes subjective reflections and emotional responses.

3. Pysentimiento [20]: is a toolkit that leverages state-of-the-art LLMs that are specifically designed for emotion recognition and social NLP tasks. Proposed by Pérez et al. (2023), Pysentimiento provides access to multilingual datasets and models for

[18]https://github.com/dinbav/LeXmo, (accessed 22 January 2025)
[19]https://github.com/UBC-NLP/EmoNet, (accessed 22 January 2025)
[20]https://github.com/pysentimiento/pysentimiento, (accessed 22 January 2025)

detecting six primary emotions: "anger", "disgust", "fear", "joy", "sadness", and "surprise". By incorporating LLMS, Pysentimiento is equipped to handle complex linguistic expressions across languages and cultures, making it suitable for analysing diverse user-generated content. In this study, Pysentimiento was tested on tourism-related data to evaluate its efficacy in detecting a range of emotional tones as expressed in reviews. Its design allows for a more sophisticated understanding of context and emotion, extending beyond simplistic emotion classification. This toolkit's robust multilingual support enhances its applicability in tourism contexts, where reviews often contain cultural and regional language variations.

4. ETT (Efficient Task Transfer) [21]: was proposed by Poth et al. (2021), and represents a strategic advancement in the NLP methodologies, focusing on task selection to enhance performance across various language tasks, including emotion detection. ETT optimises intermediate task selection, providing a computationally efficient alternative to the traditional, few-shot learning techniques. By training on six core emotions: "anger", "love", "fear", "joy", "sadness", and "surprise", ETT enables the transfer of knowledge from related tasks, thereby improving emotion classification without requiring extensive retraining for each specific application. ETT's efficacy was tested by tourism review analysis, aiming to measure its ability to adapt to the highly subjective, emotional language that is associated with tourist feedback. Its intermediate-task selection capability is particularly relevant in dynamic contexts like tourism, where the complexity of the language and expressions of emotions require adaptable, resource-efficient models.

Each of these tools contributes uniquely to the field of emotion detection, employing different methodologies, ranging from lexicon-based systems to advanced neural networks and task-transfer models, to analyse and interpret the emotional content of text. Their evaluations of tourism-related data highlight their respective strengths and limitations, offering a comprehensive understanding of their applicability in contexts where subjective, emotional language dominates. This section provides a foundation for determining the suitability of these tools for capturing and categorising the emotional tones within tourism reviews, a critical step in refining the emotion detection capabilities for domain-specific applications.

4.2.2 Selection Criteria for the Tools for the Experiment

In establishing an experimental framework for evaluating the emotion detection tools, the selection process for these tools required carefully designed criteria, emphasising the alignment with the unique characteristics of the tourism review data and the selected emotion scheme for this research. This methodical selection ensured that

[21] https://github.com/adapter-hub/efficient-task-transfer, (accessed 22 January 2025)

the tools chosen were not only accessible and academically credible but could also be adapted to handle the complex, subjective nature of user-generated text on social sites. Six primary criteria were applied to guide this process, selecting tools that were robust, relevant, and capable of capturing the complexities of the expressions of emotions that were embedded in tourism-related text.

Academic Citability and Published Papers: A key criterion for tool selection was the presence of formal academic documentation in the form of published papers. Tools associated with peer-reviewed research provide a citable foundation, establishing a methodological precision that strengthens the academic integrity of the experimental evaluation. By selecting tools like LeXmo, EmoNet, Pysentimiento, and ETT, each of which is linked to published studies, this criterion was applied to ensure that every tool utilised had been studied and validated within the academic community.

Public Accessibility: Prioritising tools that are openly available to researchers and practitioners was also an important consideration during the tool selection. Publicly accessible tools not only foster the transparency and reproducibility of research but also encourage broader adoption and facilitate collaborative advancements in emotion detection. The chosen tools (LeXmo, EmoNet, Pysentimiento, and ETT) are all accessible via open-source platforms, enabling a transparent assessment of their functionalities.

Alignment with Plutchik's Emotion Scheme: Given the focus on tourism reviews that were annotated using Plutchik's emotion model, tools that either fully or partially adopt this emotion scheme were prioritised. This alignment with Plutchik's model is essential for accurately interpreting the varied emotion classes expressed within the TORCEv2 dataset. LeXmo, which is built on the NRC Emotion Lexicon, directly aligns with Plutchik's model by offering associations with all of Plutchik's primary emotions. Similarly, EmoNet, Pysentimiento, and ETT were selected for their compatibility with a structured emotion framework, ensuring that these tools were effectively able to capture and classify the range of emotional expressions related to the TORCEv2 dataset. However, it is worth noting that the "anticipation" emotion class is absent from both Pysentimiento and ETT.

Adaptability to Domain-Specific Language and Context: Domain-specific content often contains unique expressions, and cultural references that general-purpose emotion detection tools may struggle to interpret effectively (Alaei et al., 2019), so it was important to select tools with a demonstrated adaptability to specialised lexicons and diverse linguistic contexts. Pysentimiento, for example, was chosen for its flexibility and ability to leverage LLMs that accommodate varied

vocabularies and complex expressions across different domains. By focusing on tools that are capable of adapting to these specific linguistic characteristics, adopting this criterion ensures the tools' ability to interpret emotions within highly-contextualised, culturally-rich content.

Sensitivity to Complex, Layered Expressions of Emotions: Tourism narratives often convey layered emotional expressions that extend beyond basic positive, negative, or neutral sentiments. The tools selected needed to exhibit sensitivity to this complexity, enabling them to detect fine-grained emotions that reflect the multifaceted nature of travel experiences. EmoNet, employing Gated Recurrent Neural Networks (GRNNs), particularly matches this criterion, as it is designed to recognise subtle emotional tones across different contexts. Its advanced neural network structure allows it to detect emotions within complex narratives, capturing not only the obvious emotions but also the underlying emotional tones that characterise tourism reviews.

Alignment with User Sentiment in Tourism Contexts: Understanding user sentiment in tourism contexts involves recognising the unique challenges and subjective experiences that are often embedded in tourism-related content. Tools that can interpret the subjective, varied perspectives of travellers were essential for this evaluation. The Efficient Task Transfer (ETT) methodology aligns well with this criterion, employing a strategy to select intermediate tasks that improve performance in emotion detection in complex, domain specific contexts. ETT's approach allows it to go beyond basic emotion recognition, reaching into the complex emotional layers that characterise tourism experiences.

The application of these criteria yielded a diverse set of tools, including LeXmo, EmoNet, Pysentimiento, and ETT, each of which brought distinct methodological strengths to the study of emotion detection in tourism reviews. LeXmo offers a lexicon-based approach that directly aligns with Plutchik's model, making it effective for basic emotion classification. EmoNet, with its GRNN architecture, is highly effective for detecting complex expressions of emotions in contextualized narratives. Pysentimiento, leveraging LLMs, provides adaptability to multilingual and domain-specific lexicons, enhancing its relevance in tourism-related emotion analysis. Lastly, ETT's methodology broadens its applicability, enabling it to capture complex user emotions.

The careful selection of these tools, based on the criteria of academic citability, accessibility, alignment with Plutchik's emotion schemes, adaptability to domain-specific contexts, sensitivity to complex expressions, and suitability for interpreting tourism-specific emotions, lays a robust foundation for evaluating their performance.

By aligning these tools, this study aims to yield meaningful insights into the effectiveness of each tool with regard to capturing the emotions within the tourism domain.

4.2.3 Testing and Comparing

The mainstream emotion detection tools, including LeXmo, EmoNet, Pysentimiento, and ETT, were evaluated in order to examine their performance with regard to the TORCEv2 dataset. The primary aim was to test the ability of these tools to detect and classify expressions of emotions accurately in the context of tourism-related text, while also establishing a baseline for the subsequent evaluations involving LLMs. The selection criteria for these tools prioritised factors such as published research papers, public availability, and alignment with Plutchik's emotion model. These criteria were applied carefully to ensure that the chosen tools suited the unique structure and emotional composition of the dataset.

The widely used F-score was employed as the primary metric for evaluating tool performance, as it balances precision and recall, providing a holistic view of each tool's classification accuracy. Due to the imbalanced nature of the TORCEv2 dataset, as shown in Table 4.1, it was better to use the weighted calculation of the F-score metric. This choice of metric allowed a fair assessment across the varying levels of class representation within the dataset. Emotion categories with more frequent sentences in the dataset, such as "joy", carried more weight in the overall F-score calculation. This approach helped to provide a realistic evaluation that accounts for the imbalanced distribution of the expressions of emotions within the TORCEv2 dataset. The weighted F-score was calculated as follows:

$$F\_score_{\text{ weighted}} = \frac{\sum_i (\text{support}_i \times F_i)}{\sum_i \text{support}_i} \tag{4.1}$$

where support_i is the number of reviews in class i, and F_i is the F-score for class i. By employing this weighted F-score, the emotion classes with a higher number of reviews in the dataset will have a higher influence on the overall F-score.

Table 4.2 presents the number of sentences in the test dataset that are used in this chapter's experiments. The total count of sentences is 494. Specifically, Anger has 122 sentences, Anticipation has 28, Joy is the most represented with 242 sentences, Sadness has 52, and Surprise has 50 sentences.

As displayed in Table 4.3, the evaluation metrics for the emotion detection tools reveal clear performance differences, with the F-score providing the most balanced indicator of effectiveness. Pysentimiento achieved the highest F-score at 0.68, demonstrating good performance across both precision 0.67 and recall 0.71, and suggesting that it is the most effective tool among those evaluated for capturing

Emotion class	**Weight**
Anger	25%
Sadness	10%
Joy	49%
Surprise	10%
Anticipation	6%

Table 4.1: The weight of all of the emotion classes in the TORCEv2 dataset

Emotion Category	**Number of Sentences**
Anger	122
Anticipation	28
Joy	242
Sadness	52
Surprise	50
Total	494

Table 4.2: Number of sentences in the testing dataset

emotions in tourism-related text. ETT followed with an F-score of 0.59, supported by balanced precision and recall values 0.61 and 0.64, respectively, which highlights its comparatively stable performance. EmoNet achieved a moderate F-score of 0.54, consistent with its precision and recall scores, suggesting a reasonable ability to identify emotions but with reduced accuracy relative to Pysentimiento and ETT. By contrast, LeXmo obtained the lowest F-score at 0.45, alongside lower accuracy and recall, indicating limited capability in handling the emotional content of the tourism-related data. Overall, the F-score results confirm that tools relying on advanced language models, such as Pysentimiento, are more effective in capturing emotional nuances, while lexicon-based tools such as LeXmo are less suited for this task.

To gain deeper insights into the effectiveness of each tool, an emotion class-specific analysis was conducted. The results are presented in Figures 4.1, 4.2, and 4.3, which report the F-score, precision, and recall, respectively. In addition, Figures 4.4, 4.5, 4.6, and 4.7 display the corresponding confusion matrices for LeXmo, EmoNet, Pysentimiento, and ETT. The analysis shows clear differences in performance across emotion classes, indicating that each tool has different strengths and limitations in relation to detecting particular emotions.

Emotion Detection Tool	F-score	Accuracy	Precision	Recall
LeXmo	0.45	0.46	0.52	0.46
EmoNet	0.54	0.59	0.54	0.59
Pysentimiento	0.68	0.71	0.67	0.71
ETT	0.59	0.64	0.61	0.64

Table 4.3: Evaluation metrics for the mainstream emotion detection tools

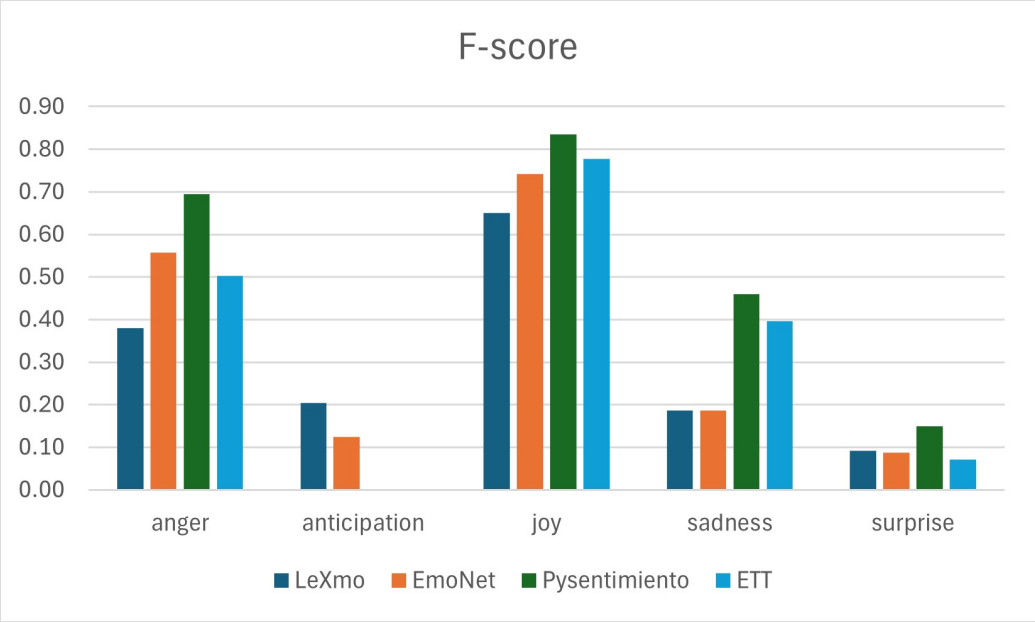


Figure 4.1: F-score for each emotion class

The detection of "joy" consistently achieved the highest F-scores across all tools, confirming it as the best-supported emotion. Pysentimiento performing particularly well in this category with an F-score of 0.83, followed by ETT 0.78 and EmoNet 0.74. This high performance is corroborated by the confusion matrices; for example, ETT's matrix shows a very high true positive count for "joy" 227, which directly contributes to its high recall 0.94, the strongest among all tools for any emotion. These results imply that positive emotions are well supported by the existing emotion detection algorithms, possibly due to the higher representation of the "joy" class in the dataset.

Performance for "anger" detection was more variable. Pysentimiento was the most

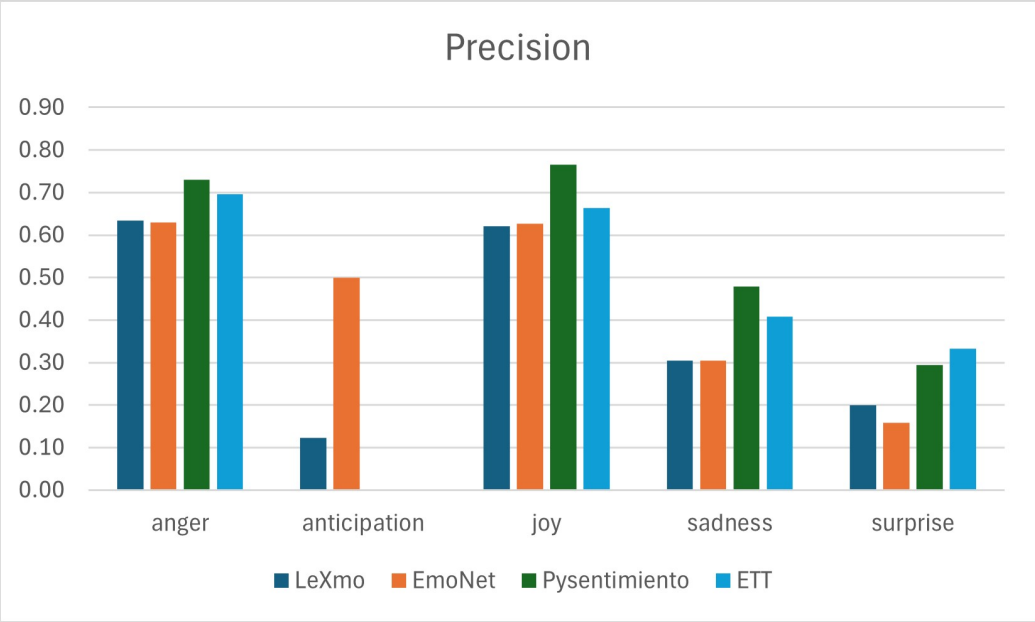


Figure 4.2: Precision for each emotion class

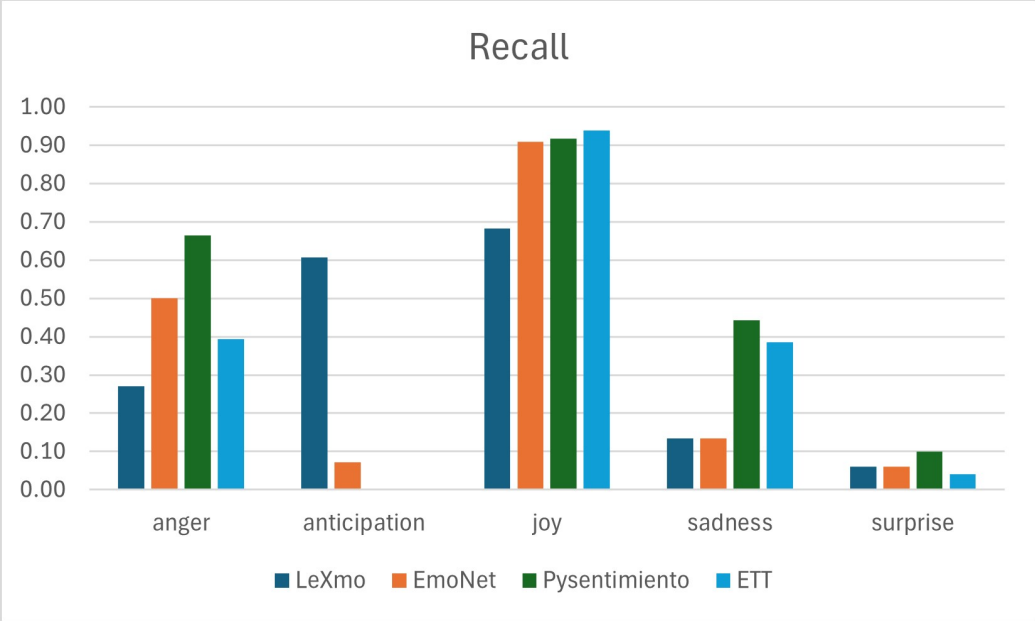


Figure 4.3: Recall for each emotion class

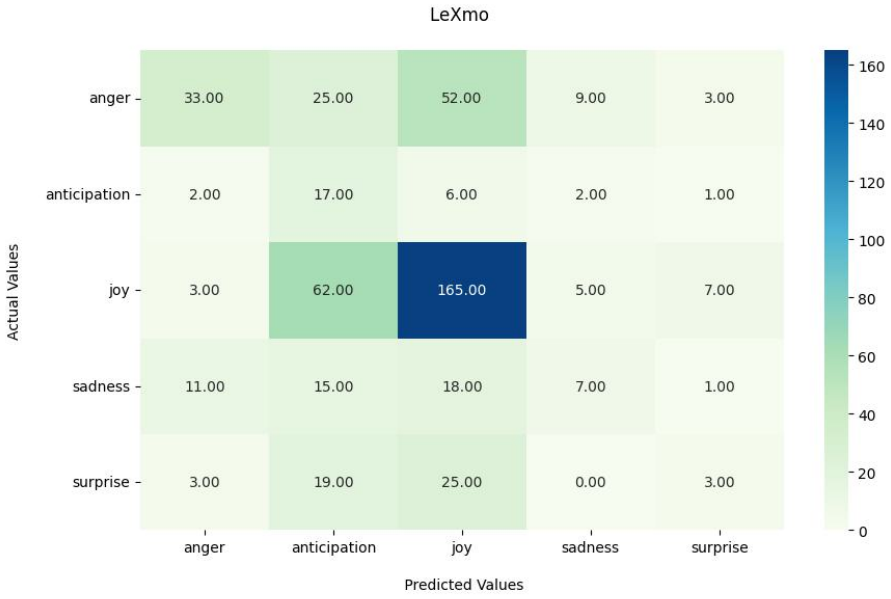


Figure 4.4: The confusion matrix for LeXmo

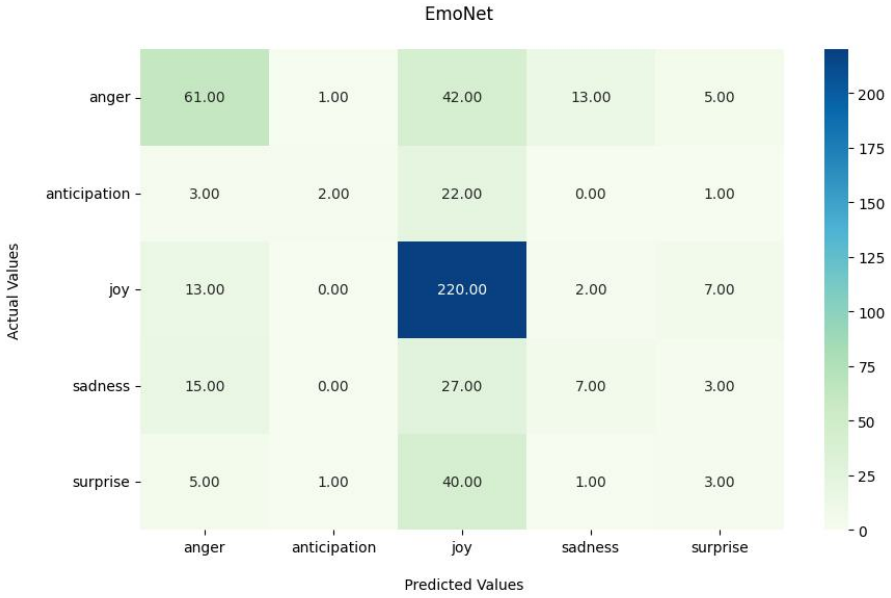


Figure 4.5: The confusion matrix for EmoNet

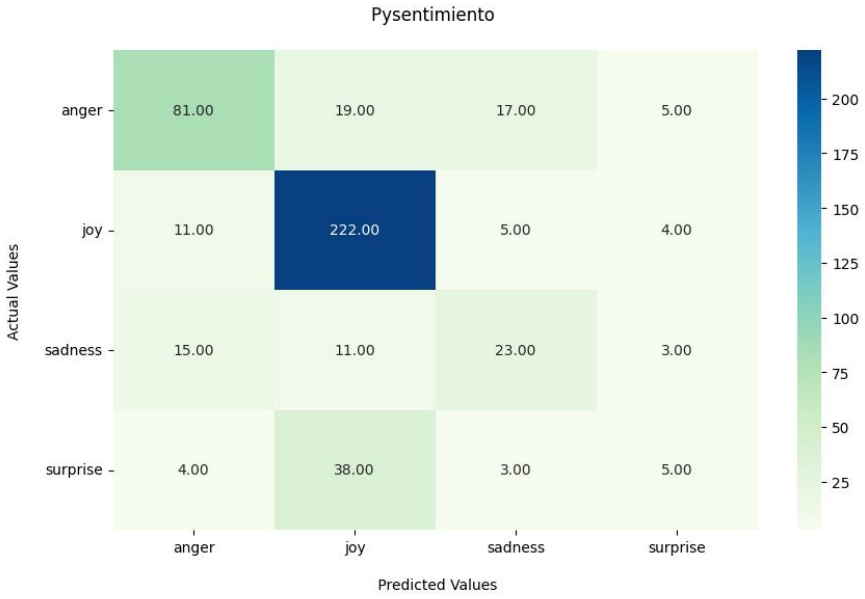


Figure 4.6: The confusion matrix for Pysentimiento

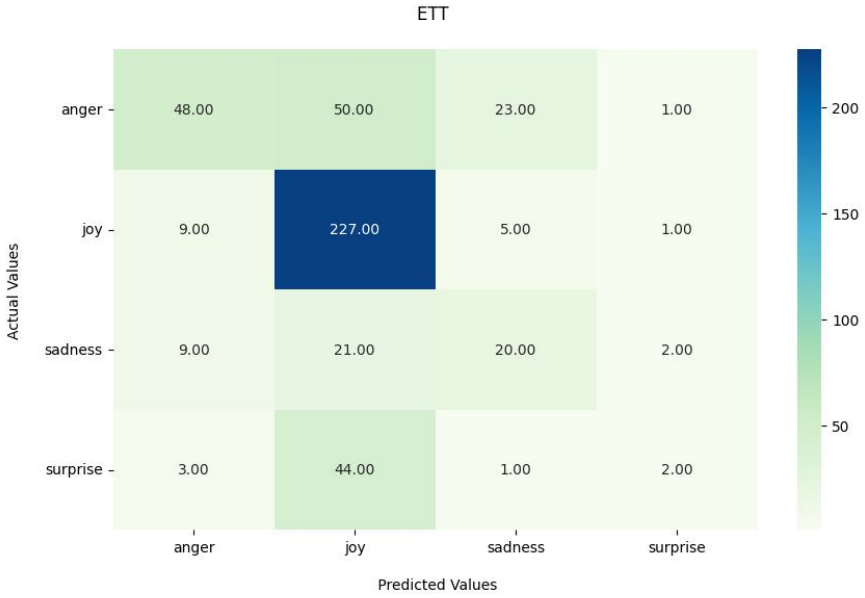


Figure 4.7: The confusion matrix for ETT

effective tool, achieving the highest F-score 0.70, which is supported by relatively high precision 0.73 and the highest recall 0.66 for this emotion. The confusion matrix for Pysentimiento shows a good number of true positives 81. In contrast, LeXmo's low F-score 0.38 is primarily due to very poor recall 0.27, indicating it fails to identify most actual "anger" sentences. This is evident in its matrix, where the true positive count for "anger" is 33 which is low compared to the models.

The "sadness" category presented a considerable challenge for most tools. While Pysentimiento and ETT achieved moderate F-scores 0.46 and 0.40 respectively, the performance of LeXmo and EmoNet was notably poor 0.19 for both of them. The confusion matrices reveal the cause with high number of false negatives. For instance, LeXmo and EmoNet's matrices show 7 true positives for "sadness", but a high number of instances misclassified as other emotions, leading to their very low recall 0.13. Pysentimiento's superior performance is attributed to a better balance, correctly identifying more true positives 23 and suffering from fewer misclassifications. However, The F-scores for "sadness" were generally low across all tools, indicating that this emotion class presents challenges for accurate detection in the TORCEv2 dataset.

The categories of "surprise" and "anticipation" proved to be the most challenging, with all tools scoring extremely low. For "surprise", no tool exceeded an F-score of 0.15. The matrices show that for every tool the number of false negatives is higher than the true positives. ETT, for example, has a very high number of false negatives were 44 sentences classified as "joy", leading to a near-zero recall of 0.04. "anticipation" has similar issues, in particular, LeXmo and EmoNet have F-scores of 0.20 and 0.13 respectively. LeXmo's slightly higher score is due to a higher recall 0.61, meaning it labels many instances as anticipation, but its very low precision 0.12 reveals that most of these predictions are incorrect. Additionally, it is worth noting that both Pysentimiento and ETT completely lack any support for the "anticipation" category, which further restricts their effectiveness in capturing the full emotional spectrum of the TORCEv2 dataset. These findings suggest that emotions involving forward-looking or unexpected elements are particularly difficult for the tools to interpret, possibly due to the complex, context-dependent nature of the expressions of such emotions in reviews.

The varying F-scores across the emotion classes reflect the underlying challenges faced by the mainstream emotion detection tools when applied to domain-specific contexts like the TORCEv2 dataset. Although certain emotions, such as "joy", were efficiently detected, the difficulty of accurately identifying emotions like "surprise" and "anticipation" highlights the limitations of these tools for handling complex or infrequent emotional classes. This analysis highlights the necessity for continued advancements in emotion detection technology, with a focus on developing models that are adjusted to the unique linguistic and contextual characteristics encountered

in domain-specific data. Developing such tools, that are adaptable and sensitive to specific contexts, will be essential for achieving comprehensive, accurate emotion detection, especially in the case of tourism-related text.

The evaluation outcomes help to clarify the performance of the mainstream emotion detection tools and highlight areas where improvements are required. These findings emphasise the importance of developing tools that are aligned well with domain-specific contexts, as the general-purpose emotion detection tools may fail to capture the full emotional range present in specialised datasets, such as the TORCEv2 dataset. This foundational evaluation prepares the ground for the next phase of this research experiment, which will involve adapting and fine-tuning LLMs to address the identified challenges. The observed limitations, especially those related to the accurate detection of complex, less frequently-expressed emotions, like "surprise" and "anticipation", indicate the need for more specialised approaches to emotion detection.

4.3 Fine-tuning LLMs for Emotion Detection

This section presents the experiments that involved fine-tuning the LLMs to enhance emotion detection specifically within the context of tourism reviews. The initial evaluation of the existing emotion detection tools demonstrated their limitations, particularly when applied to tourism-related text, highlighting the need to develop a new model that has been trained on a domain-specific dataset to improve performance when analysing tourism-related reviews. Subsequently, this research explored the fine-tuning of three prominent LLMs: BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), and RoBERTa (Liu et al., 2019), using the benefit of leveraging data augmentation techniques to adapt these models to incorporate the specific differences associated with tourism-related emotional content.

4.3.1 The Performance of Original Untuned Models

Before fine-tuning the above-mentioned models for emotion classification, the original models were tested using TORCEv2 and they produced poor F-scores, as shown in Table 4.4. Such results were expected because the models are trained on generic, large-scale data. They need to be fine-tuned to suit specific tasks, such as text classification (Devlin et al., 2019).

Language Model	F-score	Accuracy	Precision	Recall
BERT	0.18	0.17	0.19	0.17
DistilBERT	0.28	0.29	0.28	0.29
RoBERTa	0.29	0.28	0.31	0.28

Table 4.4: Evaluation metrics for the LLMs without fine-tuning

4.3.2 Data Augmentation

4.3.2.1 The Challenge of Data Scarcity

Machine Learning models are sensitive to the quality and quantity of the training data. If the labelled data contain errors or biases, the model could learn and preserve these mistakes. Also, it is difficult for machine learning to deal with imbalances within data. For example, if a model is trained on an imbalanced emotion dataset, it would be more likely to make mistakes in relation to the minority emotion classes. This is because the model has encountered fewer examples of the minority emotion classes during the training and so may not have learned to identify them effectively. To address this issue, several techniques can be used to mitigate the negative impact of an imbalanced training dataset. Such techniques include oversampling, under sampling and weighted learning. By using these techniques, the performance of machine learning models in the case of imbalanced datasets can be improved.

The TORCEv2 dataset, while vital for tourism emotion analysis, exhibited imbalances in the distribution of the various emotion classes. This imbalance in the dataset occurred due to the inability to anticipate the variations in the workers' outputs. As shown in Table 4.1, certain emotions were underrepresented, posing a significant challenge for the LLMs during the training. In response, this study strategically employed data augmentation techniques to strengthen the training set, ensuring a less biased representation of all of the emotion categories.

4.3.2.2 Methodological Approach to Addressing Data Scarcity Challenges

To enhance the performance of LLMs related to detecting expressions of emotions in tourism reviews, this research recognised the imperative need for targeted data augmentation techniques. As shown in Table 4.1, the tourism review dataset suffers from a scarcity of certain emotion classes. Therefore, testing different techniques to oversample the scarce emotion classes was important in order to bring them closer to the level of the majority classes in the training dataset. The objective was to mitigate the challenges arising from the scarcity of certain emotion classes within the TORCEv2 dataset. To address the limitations of this imbalance, multiple data

augmentation techniques were employed, based on the methodology suggested by Wei and Zou (2019). The new sentences were generated from the original sentences by applying these data augmentation methods, thereby expanding the dataset while maintaining its semantic consistency. The following techniques were applied to the scarce classes:

1. **Random Insertion (RI)**: Leveraging the contextual understanding provided by BERT, up to two words were randomly inserted into each review. This method aimed to introduce subtle variations and differences, enriching the dataset with diverse expressions.

2. **Random Deletion (RD)**: Introducing a stochastic element, RD involved the random removal of up to two words per review. This process simulated a more dynamic linguistic landscape, fostering adaptability in the LLMs.

3. **Random Swapping (RS)**: This technique entailed the random swapping of two neighbouring words within a review, introducing variations in word order. This approach aimed to capture different contextual shifts, especially in tourism reviews.

4. **Synonym Replacement (SR)**: Harnessing the power of BERT's language model, SR randomly replaced up to two words with their synonyms. This technique sought to diversify the vocabulary contained within the dataset, enhancing the models' ability to discern subtle emotional differences.

5. **All data augmentation methods combined**: Recognising the complementary nature of the individual data augmentation methods, we explored the synergy arising from their combined application. This comprehensive approach aimed to enrich the training data holistically, addressing diverse linguistic scenarios.

Table 4.5 illustrates the number of sentences in the training dataset prior to and following the data augmentation process. Initially, there were 1,981 sentences, with each category containing varying numbers of instances: Anger had 491 sentences, Anticipation had 116, Joy had the highest with 972, Sadness had 206, and Surprise had 196. Following the data augmentation, each category was increased to ensure that it contained an equal number of sentences, specifically 972, as the number of sentences for the highest emotion category, resulting in a total of 4,860 sentences across all categories. The application of these data augmentation techniques aimed not only to address the data scarcity but also to implant a heightened robustness within the LLMs. By introducing controlled variations that are reflective of real-world linguistic difficulties, these techniques sought to equip the models with a more

diverse understanding of the expressions of emotions within the unique context of tourism reviews.

Emotion Category	**Before Data Augmentation**	**After Data Augmentation**
Anger	491	972
Anticipation	116	972
Joy	972	972
Sadness	206	972
Surprise	196	972
Total	1981	4860

Table 4.5: Number of sentences in the training dataset prior to and following data augmentation

4.3.3 Experimenting with Learning Rates

With regards to fine-tuning the LLMs, a critical factor is to choose an optimal learning rate value. Learning rates play a key role in the training dynamics of machine learning models. The learning rate is a hyperparameter that controls the speed at which a machine-learning model updates its parameters. If a learning rate that is too high is applied, an LLM may be unable to fully learn the patterns in the training data and may have difficulty generalising new data. On the other hand, an LLM with a too low learning rate may take a very long time to train and may be unable to achieve the desired performance metrics. Therefore, it is important to perform trials with different learning rates to determine the optimal learning rate. In this experiment, various learning rates for each LLM were systematically tested. The selected learning rates included 2e-5, 1e-5, 5e-6, and 1e-6. This exploration aimed to strike a balance, avoiding rates that were too high, impeding model convergence, or too low, resulting in prolonged training periods.

4.3.4 Model Architecture

This section introduces the architecture of the fine-tuned LLMs that have been developed for emotion detection in tourism-related reviews. The architecture was designed with several components, including Input, Tokenisation and Masking, Embedding Layer, CLS Vector, Fully Connected Layer, Activation, and Classification

Label, as demonstrated in Figure 4.8. Each of them enhances the model's ability to process and classify the input text efficiently and accurately. Figure 4.8 illustrates each layer and its role within the overall architecture.

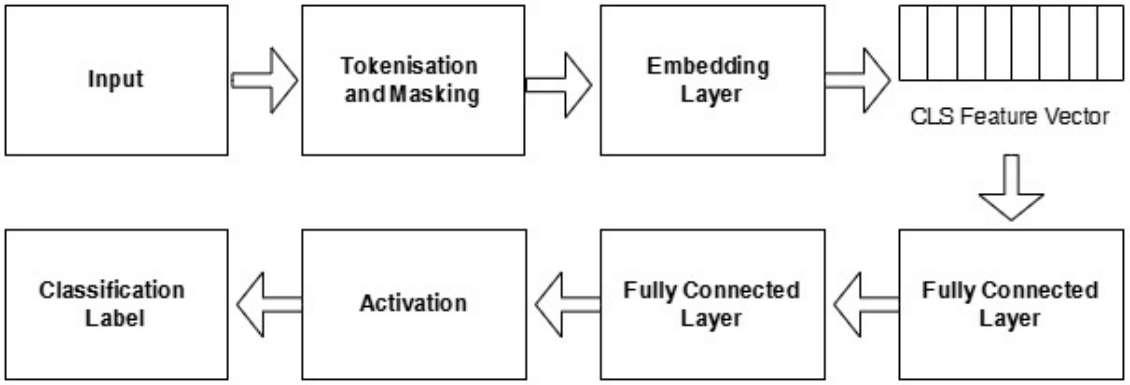


Figure 4.8: Flowchart of the models' architecture for emotion classification

Input Text, Tokenisation and Masking: The input layer serves as an entry point for the text data into the model, taking sentences from the tourism reviews and preparing them in a format that is suitable for the next steps. The tokenisation and masking step involve breaking down the input text into smaller units, typically words or subwords, called tokens. Tokenisation is essential for managing variable-length inputs, allowing the model to process textual data in a structured format. Masking, on the other hand, is another technique that is used in training to randomly hide certain tokens, encouraging the model to learn the contextual relationships among words. This helps the model to understand the dependencies between words and contextual nuances, which are important for emotion classification. Together, tokenisation and masking ensure that the input data are both manageable and contextually rich.

Embedding Layer: In the embedding layer, each token is transformed into a fixed-dimensional vector that represents its meaning in a multidimensional space. The semantic relationships between words are thereby captured by these embeddings by mapping similar words to nearby points in space. The embedding layer is fundamental, since it provides a medium through which the model learns and retain information about words' similarities and relationships, which is crucial for accurately distinguishing between the expressions of emotions in text. By converting tokens into embeddings, this layer reduces the high-dimensional nature of language data into a more tractable form and makes the subsequent computations considerably easier and

more efficient.

CLS Vector: The CLS Vector is a special token that summarises the entire input sequence. It is usually at the beginning of the sequence and is updated during the processing of the model. This CLS vector, after passing through the layers, contains information about the entire input, enabling it to serve as a comprehensive representation of the text content. This vector is optimised during training to capture relevant features in the input text that supports accurate emotion classification. Additionally, the dimension size of the CLS feature vector used in this research is 768 dimensions.

Fully Connected Layers: These are an intermediate layer that processes the output from the CLS vector. These layers consist of neurons that are fully connected to each other, allowing the model to learn complex patterns and interactions within the embedded data. By applying linear transformations and weights, these layers refine the information extracted from the input, preparing it for the final classification. The fully connected layers empower the model to distinguish hidden emotional distinctions within the text, which is particularly valuable in the context of the current emotion classification task.

Activation: The activation function, in this model specifically the Rectified Linear Unit (ReLU), is applied to the output of the fully connected layer to introduce non-linearity. ReLU is a non-linear function that outputs zero for any negative input and the input itself for any positive input. By applying ReLU, the model can capture a wide range of relationships that are inherent in the data, which is important for the proper classification of the emotions. It also mitigates the vanishing gradient problem, enabling a more rapid convergence during training. Additionally, it offers an ideal combination of simplicity and efficiency in terms of performance.

Classification Label: The classification label is the final layer of the architecture, where the model outputs a predicted label that represents one of the five emotion categories. This process classifies the input text into one of the five predefined classes, including “joy”, “anger”, “sadness”, “anticipation”, and “surprise”, based on the features extracted and refined through each of the previous layers. This final output translates the model’s internal processing into a usable prediction, that can be applied in real-world emotion detection tasks.

In general, this LLM architecture for emotion detection is carefully structured to achieve a high level of accuracy with regard to processing and classifying textual data. From the Input Layer to the final Classification Label, every component is important

in making the model capable of recognising emotional classes in tourism-related reviews. The Embedding Layer and CLS Vector effectively capture the semantic and contextual depth of the input, while the Fully Connected Layer and ReLU Activation introduce essential non-linearity, enabling the model to distinguish between subtle emotional variations. This architecture develops a powerful framework for emotion detection, making it a valuable tool for the final step of this research, which is to develop a fake review classifier, using emotion information.

Listing 4.1 provides the Python code for the classifier class used in this experiment. This class is central to the model's functionality, as it processes input data and assigns it to the appropriate emotion category, based on the predefined classification parameters. Specifically, the classifier takes the original 768-dimensional hidden layer output from the LLM and reduces it to a 5-dimensional vector, where each dimension corresponds to one of the target emotion categories. This transformation enables the model to effectively categorise the emotional content of the input text.

```
class classifier(nn.Module):
  def __init__(self, model_name, num_classes):
    super(classifier, self).__init__()
    self.fModel = sModel.from_pretrained(model_name)
    self.pre_classifier = nn.Linear(768, 768)
    self.dropout = nn.Dropout(0.1)
    self.classifier = nn.Linear(768 , num_classes)

  def forward(self, input_ids, attention_mask):
    outputs = self.fModel(input_ids = input_ids,
                          attention_mask = attention_mask)
    pooler = outputs[0][:, 0]
    pooler = self.pre_classifier(pooler)
    pooler = nn.ReLU()(pooler)
    pooler = self.dropout(pooler)
    logits = self.classifier(pooler)
    return logits
```

Listing 4.1: Emotion classification function

Where the model name can take any of the following values: "bert-base-uncased", "roberta-base", and "distilbert-base-uncased", depending on the chosen LLM.

4.3.5 Model Fine-tuning

In the pursuit of refining the emotion detection capabilities of LLMs for tourism reviews, a comprehensive fine-tuning process was implemented. The objective was to identify the optimal learning rates and data augmentation techniques that would enhance the models' performance with regard to capturing the various emotional

differences present in the TORCEv2 dataset. For the purpose of the LLMs fine-tuning, three models were tested including:

1. **BERT**: The base uncased, trained on two datasets - English Wikipedia and BookCorpus (Zhu et al., 2015).
2. **DistilBERT**: The base uncased, trained on the same datasets as BERT, but it is smaller and faster, because it has fewer parameters than the BERT model.
3. **RoBERTa**: The base trained on five datasets, with a combined weight of 160GB.

Those models were selected from the Transformers package [22].

In the fine-tuning process, each of the BERT, DistilBERT and RoBERTa were fine-tuned, using four data augmentation techniques as well as a combination of them, as discussed earlier. In addition, it involved experimenting with four learning rates, that were tested for each model. Moreover, the batch size was equal to 16 and trained for 20 epochs. The TORCEv2 dataset was divided into training and testing sets, with 80% of the data allocated to training and the remaining 20% to testing. The statistics for the training set are presented in Table 4.5, while those for the testing set are shown in Table 4.2.

4.3.5.1 Fine-tuning the BERT Model

In this comprehensive experiment examining BERT-based model performance across multiple evaluation metrics, six methods were evaluated, including no data augmentation, random insertion, random deletion, random swapping, synonym replacement, and a combined method, each demonstrating unique performance characteristics and learning rate sensitivities across F-score, accuracy, precision, and recall metrics. Table 4.6, Table 4.7, Table 4.8, and Table 4.9 show the produced scores for F-score, accuracy, precision, and recall respectively.

The no data augmentation approach exhibits the most performance variability across different learning rates, with F-scores ranging from 0.32 to 0.60, representing the largest performance variation of 0.28 among all methods tested. This approach demonstrates a distinctive performance pattern where it achieves its optimal results at intermediate learning rates of 1e-5 and 5e-6, both yielding F-scores of 0.60, accuracy and recall scores of 0.70, and precision of 0.54. The method shows notable performance degradation at both the highest learning rate of 2e-5, where F-score drops to 0.32 with corresponding accuracy of 0.49, precision of 0.24, and recall of 0.49, and at the lowest learning rate of 1e-6, where performance declines to F-score 0.35, accuracy 0.50,

[22] https://github.com/huggingface/transformers, (accessed 27 January 2025)

precision 0.43, and recall 0.50. This substantial variability indicates that this method is highly sensitive to learning rate optimisation and requires careful hyperparameter tuning to achieve acceptable performance levels.

Random insertion demonstrates high performance across all learning rates, with F-scores ranging from 0.73 to 0.77. This technique achieves its highest performance at learning rates 1e-5 and 5e-6, both producing F-scores of 0.77, accuracy of 0.79, precision of 0.79 and 0.78, and recall of 0.79. Notably, random insertion excels in precision performance, achieving the highest precision score of 0.80 at the learning rate of 2e-5, making it particularly valuable for applications where minimising false positives is critical. The method maintains high performance even at the lowest learning rate of 1e-6, with F-score of 0.73. Random insertion consistently produces high scores across all learning rates, making it as a reliable data augmentation strategy.

Random deletion demonstrates consistent stability in the tested settings, with F-scores varying minimally from 0.72 to 0.74 across all learning rates. This method consistently underperforms relative to other data augmentation techniques, producing the lowest F-score of 0.74 at learning rate 2e-5. The method maintains accuracy scores between 0.75 and 0.77, precision scores from 0.73 to 0.79, and recall scores from 0.75 to 0.77 across all experimental learning rates. Random deletion shows particular strength in precision at intermediate learning rates, achieving 0.79 precision at both 1e-5 and 5e-6, matching the performance of random insertion and random swapping.

Random swapping achieves the highest score among all individual data augmentation methods, reaching F-scores of 0.78 at learning rate 2e-5, accompanied by the highest accuracy and recall scores of 0.80. This method demonstrates robust performance at higher learning rates, maintaining F-scores of 0.77 at 1e-5 and 0.76 at 5e-6, but shows more significant degradation at the lowest learning rate of 1e-6, where score drops to 0.72. The method shows balanced performance across all metrics, with precision scores ranging from 0.71 to 0.79 and recall scores from 0.73 to 0.80.

Synonym replacement demonstrates the most stability among all methods tested, with F-scores varying by only 0.01 across all learning rates. This method achieves F-scores between 0.74 and 0.75. The technique maintains consistent accuracy scores of 0.77 and 0.78, precision scores of 0.75 and 0.78, and recall scores of 0.77 and 0.78 across all experimental learning rates. Uniquely among all methods, synonym replacement performs best at the lowest learning rate of 1e-6, achieving the highest F-score of 0.74.

The all combined method, incorporating all individual data augmentation techniques simultaneously, achieves highest score at learning rate 2e-5, with F-scores of 0.78, accuracy of 0.80, precision of 0.78, and recall of 0.80. This comprehensive approach maintains strong performance at intermediate learning rates, achieving F-scores of 0.77 at both 1e-5 and 5e-6, demonstrating the potential benefits of leveraging multiple data augmentation strategies together. However, the combined method

shows notable performance degradation at the lowest learning rate of 1e-6, dropping to F-score 0.72, suggesting that the complexity introduced by multiple simultaneous data augmentations requires adequate learning capacity for optimal effectiveness.

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.32	0.60	0.60	0.35
Random Insertion	0.76	0.77	0.77	0.73
Random Deletion	0.74	0.74	0.74	0.72
Random Swapping	0.78	0.77	0.76	0.72
Synonym Replacement	0.75	0.75	0.75	0.74
All Combined	0.78	0.77	0.77	0.72

Table 4.6: F-scores for the BERT-based models

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.70	0.70	0.50
Random Insertion	0.78	0.79	0.79	0.75
Random Deletion	0.77	0.77	0.77	0.75
Random Swapping	0.80	0.79	0.78	0.73
Synonym Replacement	0.78	0.78	0.78	0.77
All Combined	0.80	0.78	0.78	0.74

Table 4.7: Accuracy scores for the BERT-based models

In summary, random swapping and the combined data augmentation method yield the best results across all data augmentation techniques. A lower learning rate, especially 1e-6, generally constrains the performance across all methods, indicating that a minimum threshold learning rate is necessary for the data augmentation methods to be effective. Overall, this analysis demonstrates that, while data augmentation can improve performance, it is sensitive to the learning rate, with optimal results achieved at higher rates rather than the lowest tested rate.

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.24	0.54	0.54	0.43
Random Insertion	0.80	0.79	0.78	0.73
Random Deletion	0.75	0.79	0.79	0.73
Random Swapping	0.78	0.79	0.78	0.71
Synonym Replacement	0.75	0.77	0.78	0.76
All Combined	0.78	0.76	0.76	0.73

Table 4.8: Precision scores for the BERT-based models

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.70	0.70	0.50
Random Insertion	0.78	0.79	0.79	0.75
Random Deletion	0.77	0.77	0.77	0.75
Random Swapping	0.80	0.79	0.78	0.73
Synonym Replacement	0.78	0.78	0.78	0.77
All Combined	0.80	0.78	0.78	0.74

Table 4.9: Recall scores for the BERT-based models

4.3.5.2 Fine-tuning the DistilBERT Model

In this comprehensive analysis of DistilBERT-based model performance across multiple evaluation metrics, the DistilBERT model produced the worst F-scores. Compared to the other models, it under-performed with all of the fine-tuning methods. Table 4.10, Table 4.11, Table 4.12, and Table 4.13 show the produced scores for F-score, accuracy, precision, and recall respectively.

The no data augmentation method displays considerable performance variation across different learning rates, with F-scores ranging from 0.32 to 0.60. This approach achieves its optimal performance at intermediate learning rates of 1e-5 and 5e-6, reaching F-scores of 0.60, accuracy scores of 0.70 and 0.71, precision of 0.54 and 0.53, and recall of 0.70 and 0.71. The method shows significant performance decline at the highest learning rate of 2e-5, where F-score drops to 0.32, accompanied by accuracy of 0.49, precision of 0.24, and recall of 0.49. Similar degradation occurs at the lowest

learning rate of 1e-6, with F-score falling to 0.34, accuracy to 0.49, precision to 0.28, and recall to 0.49. This performance pattern indicates that the no data augmentation approach requires careful learning rate selection to achieve acceptable results, with a narrow optimal range around intermediate learning rates.

Random insertion demonstrates good performance across most learning rate conditions, achieving F-scores between 0.65 and 0.74. The method scored its highest performance at learning rate 1e-5, producing F-score 0.74, accuracy 0.75, precision 0.73, and recall 0.75. At the learning rate of 2e-5, it achieves F-score 0.72, accuracy 0.73, precision 0.71, and recall 0.73, while maintaining nearly similar performance at 5e-6 with F-score 0.73, accuracy 0.74, precision 0.73, and recall 0.74. At the lowest learning rate of 1e-6, the method produces modest performance with F-score 0.65, accuracy 0.65, precision 0.67, and recall 0.65.

Random deletion shows moderate performance levels with F-scores ranging from 0.64 to 0.73 across all experimental settings. The method achieves its best results at learning rate 2e-5, reaching F-score 0.73, accuracy 0.75, precision 0.74, and recall 0.75. At learning rates 1e-5 and 5e-6, random deletion produces F-scores of 0.72, with accuracy ranging from 0.73 to 0.74, precision from 0.71 to 0.72, and recall from 0.73 to 0.74. Performance declines at the lowest learning rate of 1e-6, where F-score drops to 0.64, accuracy to 0.63, precision to 0.67, and recall to 0.63. Additionally, the generally underperforms compared to other data augmentation methods.

Random swapping demonstrates highest scores across multiple learning rate settings, achieving F-scores between 0.65 and 0.74. The method performs optimally at learning rates 2e-5 and 1e-5, both producing F-score 0.74, with accuracy of 0.75, precision of 0.73, and recall of 0.75. At learning rate 5e-6, random swapping maintains almost similar performance with 0.74 at all evaluation metrices. Even at the lowest learning rate of 1e-6, the method achieves F-score 0.65, accuracy 0.64, precision 0.68, and recall 0.64.

Synonym replacement displays moderate performance levels with F-scores ranging from 0.66 to 0.73 across all learning rates. The method achieves its highest performance at learning rates 2e-5 and 1e-5, reaching F-score 0.73, accuracy 0.74, precision 0.72, and recall 0.74. While at 5e-6, performance slightly declines to F-score 0.71, accuracy 0.73, precision 0.73, and recall 0.73. Notably, at the lowest learning rate of 1e-6, synonym replacement achieves the highest F-score of 0.66 among all methods, with accuracy 0.66, precision 0.68, and recall 0.66.

The all combined method, incorporating all individual data augmentation techniques, achieves good performance with F-scores ranging from 0.65 to 0.74. The method reaches its highest score at learning rate 2e-5, producing F-score 0.74, accuracy 0.74, precision 0.73, and recall 0.74. At intermediate learning rates of 1e-5 and 5e-6, the method achieves F-scores of 0.72, with accuracy of 0.72, precision of 0.72 and 0.73, and recall of 0.72. At the lowest learning rate of 1e-6,

performance declines to F-score 0.65, accuracy 0.64, precision 0.68, and recall 0.64. The combined approach demonstrates the potential benefits of integrating multiple data augmentation strategies, though it does not consistently outperform the best individual methods.

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.32	0.60	0.60	0.34
Random Insertion	0.72	0.74	0.73	0.65
Random Deletion	0.73	0.72	0.72	0.64
Random Swapping	0.74	0.74	0.74	0.65
Synonym Replacement	0.73	0.73	0.71	0.66
All Combined	0.74	0.72	0.72	0.65

Table 4.10: F-scores for the DistilBert-based models

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.70	0.71	0.49
Random Insertion	0.73	0.75	0.74	0.65
Random Deletion	0.75	0.74	0.73	0.63
Random Swapping	0.75	0.75	0.74	0.64
Synonym Replacement	0.74	0.74	0.73	0.66
All Combined	0.74	0.72	0.72	0.64

Table 4.11: Accuracy scores for the DistilBert-based models

In summary, the F-score results highlight that the random swapping and random insertion data augmentation methods yield the best F-score. As the learning rate decreases, the effectiveness of all of the methods diminishes, with a marked reduction at 1e-6. This suggests that data augmentation is most beneficial at high learning rates, where the model can learn more easily from the augmented data. Lower learning rates, particularly 1e-6, restrict the model's capability, leading to weaker performance across all methods.

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.24	0.54	0.53	0.28
Random Insertion	0.71	0.73	0.73	0.67
Random Deletion	0.74	0.72	0.71	0.67
Random Swapping	0.73	0.73	0.74	0.68
Synonym Replacement	0.72	0.72	0.73	0.68
All Combined	0.73	0.73	0.72	0.68

Table 4.12: Precision scores for the DistilBert-based models

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.70	0.71	0.49
Random Insertion	0.73	0.75	0.74	0.65
Random Deletion	0.75	0.74	0.73	0.63
Random Swapping	0.75	0.75	0.74	0.64
Synonym Replacement	0.74	0.74	0.73	0.66
All Combined	0.74	0.72	0.72	0.64

Table 4.13: Recall scores for the DistilBert-based models

4.3.5.3 Fine-tuning the RoBERTa Model

In this comprehensive analysis of RoBERTa-based model performance across multiple evaluation metrics, RoBERTa model with random insertion at a learning rate of 1e-5 achieved the highest F-score among all tested models for emotion detection of 0.80. Table 4.14, Table 4.15, Table 4.16, and Table 4.17 show the produced scores for F-score, accuracy, precision, and recall respectively.

The no data augmentation method presents lowest performance variation across different learning rates, with F-scores ranging from 0.32 to 0.62. This approach achieves its best performance at the intermediate learning rate of 5e-6, reaching an F-score of 0.62, accuracy of 0.72, precision of 0.55, and recall of 0.72. The method shows consistent poor performance at learning rates 2e-5, 1e-5, and 1e-6, where F-scores remain at 0.32, with accuracy of 0.49, precision of 0.24, and recall of 0.49. This unique pattern where performance peaks at 5e-6 suggests that the no data

augmentation approach for RoBERTa models has a specific learning rate sensitivity that differs from other transformer architectures, requiring precise learning rate tuning to achieve acceptable results.

Random insertion demonstrates strong performance across all learning rate settings, achieving F-scores between 0.76 and 0.80. The method reaches its highest performance at learning rate 1e-5, producing an F-score of 0.80, accuracy of 0.81, precision of 0.80, and recall of 0.81, making it the highest-performing individual method overall. At learning rate 2e-5, random insertion achieves F-score 0.78, accuracy 0.80, precision 0.79, and recall 0.80, while maintaining solid performance at 5e-6 with F-score 0.77, accuracy 0.79, precision 0.77, and recall 0.79. Even at the lowest learning rate of 1e-6, the method sustains good performance with F-score 0.76, accuracy 0.77, precision 0.76, and recall 0.77.

Random deletion shows robust performance with F-scores ranging from 0.75 to 0.79 across all experimental conditions. The method achieves its best results at learning rate 2e-5, reaching F-score 0.79, accuracy 0.81, precision 0.80, and recall 0.81, which the highest scores at this learning rate. At learning rates 1e-5 and 5e-6, random deletion produces F-scores of 0.77, with accuracy of 0.78, precision of 0.77, and recall of 0.78. Performance remains good at the lowest learning rate of 1e-6, where F-score reaches 0.75, accuracy 0.76, precision 0.75, and recall 0.76. Random deletion demonstrates particular effectiveness in precision performance at higher learning rates, achieving 0.80 precision at 2e-5, matching the performance of random insertion.

Random swapping shows consistent performance across multiple learning rate settings, achieving F-scores between 0.77 and 0.79. The method performs optimally at learning rate 1e-5, producing F-score 0.79, accuracy 0.81, precision 0.79, and recall 0.81. At learning rates 2e-5 and 5e-6, random swapping maintains F-scores of 0.77, with accuracy ranging from 0.78 to 0.79, precision of 0.77, and recall from 0.78 to 0.79. Notably, at the lowest learning rate of 1e-6, random swapping achieves the highest F-score of 0.77 among all methods, with accuracy 0.77, precision 0.76, and recall 0.77. This method consistently produces balanced performance across all metrics.

Synonym replacement demonstrates strong performance with F-scores ranging from 0.76 to 0.79 across all learning rates. The method achieves its highest performance at learning rate 1e-5, synonym replacement produces F-score 0.79, accuracy 0.81, precision 0.79, and recall 0.81. At learning rates 2e-5 and 5e-6, both produce F-score 0.79, with accuracy of 0.80, precision ranging from 0.79 to 0.80, and recall of 0.80. At learning rate 1e-6 with it produces F-score of 0.76, accuracy of 0.77, precision of 0.76, and recall of 0.77.

The all combined method, incorporating all individual data augmentation techniques, achieves good performance with F-scores ranging from 0.75 to 0.79. The method reaches its highest performance at learning rate 1e-5, producing F-score 0.79, accuracy 0.81, precision 0.80, and recall 0.81. At learning rates 2e-5 and 5e-6, the

combined method achieves F-scores of 0.77, with accuracy of 0.79, precision ranging from 0.76 to 0.77, and recall of 0.79. At the lowest learning rate of 1e-6, performance declines to F-score 0.75, accuracy 0.76, precision 0.75, and recall 0.76.

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.32	0.32	0.62	0.32
Random Insertion	0.78	0.80	0.77	0.76
Random Deletion	0.79	0.77	0.77	0.75
Random Swapping	0.77	0.79	0.77	0.77
Synonym Replacement	0.79	0.79	0.79	0.76
All Combined	0.77	0.79	0.77	0.75

Table 4.14: F-scores for the RoBERTa-based models

Data Augmentation Method	**Learning Rate**			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.49	0.72	0.49
Random Insertion	0.80	0.81	0.79	0.77
Random Deletion	0.81	0.78	0.78	0.76
Random Swapping	0.78	0.81	0.79	0.77
Synonym Replacement	0.80	0.81	0.80	0.77
All Combined	0.79	0.81	0.79	0.76

Table 4.15: Accuracy scores for the RoBERTa-based models

In conclusion, this experiment demonstrates that the learning rate of 1e-5 is optimal for the majority of the data augmentation methods, particularly random insertion, which reaches the highest score of 0.80. The lower learning rates, especially 1e-6, tend to reduce model performance across all of the methods, while random insertion and synonym replacement are the most effective data augmentation strategies, performing well across a range of learning rates. The combination of all methods provides sound but unspectacular results, indicating that a targeted approach using specific methods like random insertion may be more beneficial than an aggregated approach.

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.24	0.24	0.55	0.24
Random Insertion	0.79	0.80	0.77	0.76
Random Deletion	0.80	0.77	0.77	0.75
Random Swapping	0.77	0.79	0.77	0.76
Synonym Replacement	0.79	0.79	0.80	0.76
All Combined	0.77	0.80	0.76	0.75

Table 4.16: Precision scores for the RoBERTa-based models

Data Augmentation Method	Learning Rate			
	2e-5	**1e-5**	**5e-6**	**1e-6**
No Data Augmentation	0.49	0.49	0.72	0.49
Random Insertion	0.80	0.81	0.79	0.77
Random Deletion	0.81	0.78	0.78	0.76
Random Swapping	0.78	0.81	0.79	0.77
Synonym Replacement	0.80	0.81	0.80	0.77
All Combined	0.79	0.81	0.79	0.76

Table 4.17: Recall scores for the RoBERTa-based models

4.3.5.4 Analysis of the Impact of Fine-tuning

The F-scores presented in Table 4.6, Table 4.10, and Table 4.14 underscore the substantial impact of model fine-tuning on emotion classification. RoBERTa, in particular, demonstrated remarkable consistency in outperforming its counterparts. Table 4.18 illustrates the best F-score obtained for each language model, together with its learning rate and data augmentation method. The BERT model achieved its highest F-score of 0.78 when all of the data augmentation methods were employed, utilising a learning rate of 2e-5. This suggests that the combined approach effectively supports BERT's performance, demonstrating that the model benefits from various data variations. The DistilBERT model reached an F-score of 0.74 with the random swapping data augmentation method, at a learning rate of 5e-6. This learning rate is lower than the one that provides the best F-score for BERT, suggesting that DistilBERT's optimal performance may require more gradual adjustments to be

made to the parameter updates. On the other hand, the RoBERTa model recorded the highest F-score in these experiments, with a score of 0.80. This was achieved using random insertion as the data augmentation method, at a learning rate of 1e-5. RoBERTa's performance, surpassing that of both BERT and DistilBERT, highlights its effectiveness when paired with random insertion and a moderate learning rate.

Overall, the careful integration of learning rates and data augmentation techniques led to the more accurate detection of emotions within tourism reviews. Remarkably, a very low learning rate usually provides the worst performance, which proves that it needs more training epochs to improve its performance, if this is even possible at all. However, the results indicate that the optimal data augmentation method and learning rate vary according to the model architecture. RoBERTa demonstrated the highest performance, particularly when coupled with random insertion and a learning rate of 1e-5. BERT performed well with a combination of all of the data augmentations at a learning rate of 2e-5, while DistilBERT, the lightweight version, required a more conservative learning rate of 5e-6 to perform to its optimum, with random swapping. These findings highlight the need to modify the data augmentation methods and training parameters in a way that best suits each model's architecture to achieve optimal results.

Language Model	**Data Augmentation Method**	**Learning Rate**	**F-score**
BERT	All Combined	2e-5	0.78
DistilBERT	Random Swapping	5e-6	0.74
RoBERTa	Random Insertion	1e-5	0.80

Table 4.18: Best F-score for each LLM

A comparison between the fine-tuned language models and mainstream emotion detection tools reveals a notable improvement in F-scores, as confirmed by the results in Table 4.3 and Table 4.18. This enhancement suggests that model fine-tuning directly improves performance, particularly with regard to the accurate capturing of nuanced aspects of the expressions of emotions in tourism-related texts. Additionally, the analysis of the emotion classification results prior to and following the language model fine-tuning, presented in Table 4.4 and Table 4.18, showcase a substantial increase in the F-scores across all of the language models. This result underscores the positive impact of fine-tuning, which refines the models' ability to distinguish emotional text more effectively, thereby improving the overall F-score for emotion classification. The observed gains in F-scores indicate that model fine-tuning is a critical step in optimising the language models for application to complex emotion detection tasks.

Moreover, a comparison between Figure 4.1 and Figure 4.9, which present the best performance F-score of both the existing tools and the fine-tuned LLMs for the individual emotion categories, highlights the substantial improvements in the emotion classification accuracy. By establishing the best performance of the four existing tools as a baseline, a notable increase in the F-scores is observed across four of the five emotion categories. Specifically, the F-scores improve by 0.15 for "anger", 0.10 for "joy", 0.01 for "sadness", 0.27 for "surprise", and an exceptional 0.47 for "anticipation". These results provide strong evidence that applying appropriate fine-tuning techniques to LLMs can enhance the efficacy of automatic emotion detection and classification. Additionally, in comparison with the studies reported in Section 2.4.5, the RoBERTa model which achieved an F-score of 0.80 has exceeded the performance reported in the reviewed studies. This result demonstrates the effectiveness of the model for emotion detection from text.

4.3.5.5 RoBERTa: the Best Performing Model

As shown in Table 4.14, the RoBERTa model, with a learning rate of 1e-5, that was trained on a dataset with the random insertion data augmentation method, yielded the highest F-score result across all of the tested models. In detail, Figure 4.9 illustrates the F-scores achieved for each emotion class: "anger", "joy", "sadness", "surprise", and "anticipation". Each score reflects the performance produced by this model by classifying the test dataset sentences associated with these emotions. This figure clearly illustrates the variations in the model's performance across the different emotional categories, providing a clear indication of its relative strengths with regard to detecting specific emotions. The differential performance observed emphasises the model's capability to capture certain emotional expressions more effectively and accurately than others. Furthermore, the Figure 4.10 shows the confusion matrix for this model

The emotion category "joy" achieved the highest F-score of 0.93, indicating strong performance in identifying and classifying this emotion. This suggests that the features related to "joy" are well captured and distinct in the dataset, enabling accurate classification. "Anger" recorded an F-score of 0.85, the second highest among the emotion's classes. "Anticipation" obtained a moderate F-score of 0.67, while "sadness" and "surprise" achieved the lowest F-scores, with values of 0.45 and 0.42, respectively. These low scores suggest that distinguishing sentences that express these emotion categories may prove challenging.

Finally, because this model demonstrates the highest F-score among all of the tested models, it was used in the next part of my experiment on fake review detection (discussed in the next chapter) as the primary tool for classifying the emotion information of reviews. This classification process represents a preliminary step for

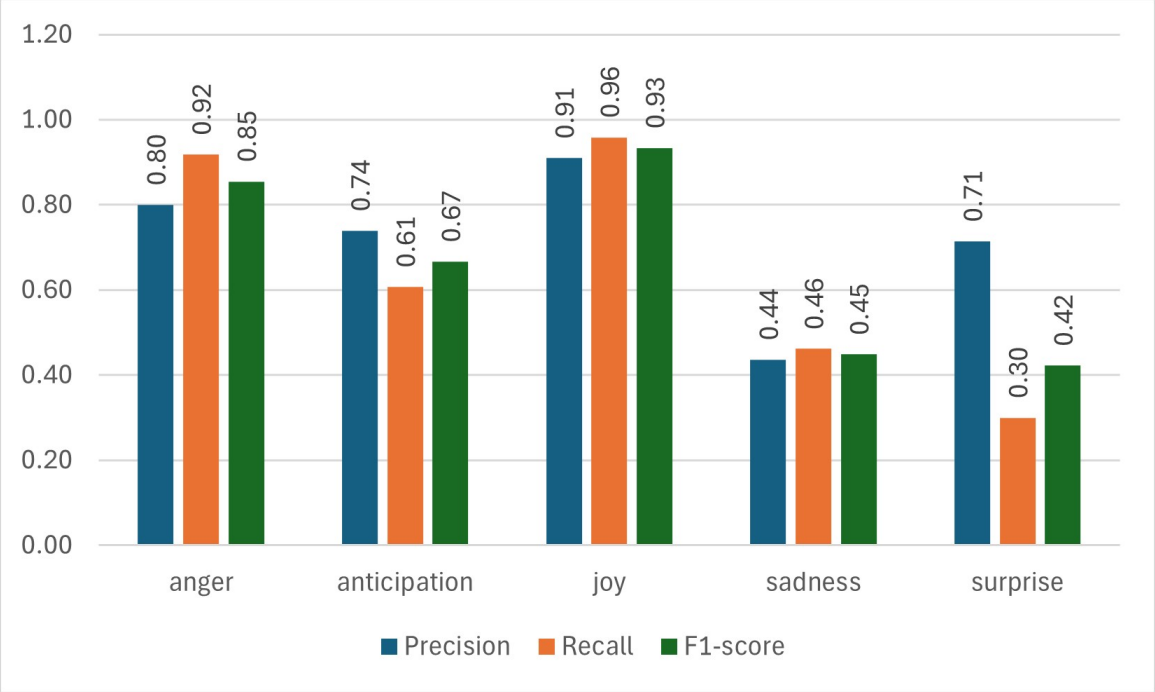


Figure 4.9: The evaluation metrics for each emotion class produced by the best performing model

conducting experiments on fake review detection.

4.3.6 Conclusion on the Model Fine-tuning Experiment

Enhancing Emotion Detection for the Tourism Domain: The comparative analysis of the fine-tuned LLMs against the existing mainstream emotion detection tools, as illustrated in Table 4.3 and Table 4.18 for the existing tools and fine-tuned models respectively, demonstrates the substantial benefits gained from the fine-tuning strategies applied in this research. The improvements in F-scores that were observed across all of the emotion categories are shown in Figure 4.1 and Figure 4.9 for the existing tool and best performance model, respectively. This improvement emphasises the effectiveness of the fine-tuning processes in enhancing the reliability of the emotion classification for tourism-related text.

Enhancing Large Language Models (LLMs): The comparative examination of F-scores prior to and following the fine-tuning process, as presented in Table 4.4 and Table 4.18, reveals a marked enhancement in model performance. The fine-

Figure 4.10: The confusion matrix produced by the best performing model

tuned F-scores consistently and significantly exceed those achieved by the original, untuned models. This considerable improvement highlights the efficacy of the fine-tuning approach that is applied in this research, providing strong evidence that fine-tuning LLMs substantially enhances their capacity to identify and categorise emotions accurately within tourism-related reviews. These findings emphasise the importance of adopting fine-tuning strategies to optimise model accuracy and reliability in emotion classification tasks, particularly in domain-specific applications, such as the emotion classification of tourism-related text.

The notable increase in F-scores observed across all three of the LLMs confirms the effectiveness of the selected data augmentation techniques and optimised learning rates. These experimental results confirm that adapting generic language models to suit the unique characteristics of tourism-related text, via carefully adjusted fine-tuning processes, produces measurable improvements in the models' ability to perform emotion detection. This enhancement highlights the importance of domain-specific

fine-tuning for achieving superior model performance, particularly in applications requiring the accurate detection of emotional tones, such as the detection of emotion within tourism-related reviews.

4.4 Case Study: Analysis of the Classification of the Sample Reviews

This section provides a qualitative analysis of the emotion classification results for the tools tested so far, including LeXmo, EmoNet, Pysentimiento, ETT, and RoBERTa. Specifically, I compare the tools' outputs against the gold standard emotion annotation of ten sample tourism review sentences obtained through crowdsourcing to provide a detailed qualitative analysis of each tool's performance. The gold standard annotation serves as a reference point, derived from the consensus of human annotators from the crowdsource MTurk platform. In the following, for each of the ten sample review sentences, I examine the output of the emotion classification of the above-mentioned tools in detail.

Sample Review 1: "After queueing for aprox 25 minutes we finally got served only to find they had then stopped serving food?". [Gold standard annotation: Surprise].

This sentence reflects a scenario that is filled with frustration and unexpected disappointment. The gold standard annotation for this sentence is surprise, as it conveys a situation where the reviewer is surprised by an unforeseen outcome after a long wait. The sentence presents a clear emotional shift from relief at finally being served to disbelief upon realising that food was no longer available.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Joy	Anger	Surprise	Joy	Anger

Table 4.19: The predicted emotion class of the first sentence produced by the emotion detection classifiers

The predicting tools, as shown in Table 4.19, interpret this sentence quite differently. LeXmo and ETT both classify the sentence as expressing joy, probably due to focusing on the initial part of the sentence, where the reviewer mentions finally being served, interpreting this as a positive event. However, this classification overlooks the crucial twist in the second half of the sentence, where the positive

anticipation turns into disappointment. The joy annotation seems to miss the full context and, as a result, fails to reflect accurately the overall emotional tone.

On the other hand, Pysentimiento correctly identifies the emotion as surprise, in line with the gold standard. This model appears to recognise the unexpected turn of events that defines the reviewer's emotional experience. Meanwhile, EmoNet and RoBERTa classify the sentence as expressing anger, which is a reasonable interpretation but does not align with the gold standard annotation. There is certainly a strong element of frustration and potential anger in the situation, especially after waiting for a significant amount of time only to be let down, which explains why these models interpreted this disappointment as anger, though they overlook the element of shock that characterises the sentence.

Sample Review 2: "Following this we had to go through what felt like airport security which did not feel friendly at all". [Gold standard annotation: Anger].

The sentence conveys a sense of dissatisfaction and frustration with an unpleasant experience. The gold standard annotation for this sentence is anger, which reflects the irritation felt by the reviewer as they describe a process that felt unnecessarily strict and unfriendly, similar to airport security.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anticipation	Joy	Joy	Joy	Anger

Table 4.20: The predicted emotion class of the second sentence produced by the emotion detection classifiers

In detail, as Table 4.20 shows, LeXmo annotates the sentence with anticipation, which appears to be a misclassification. The use of the phrase "had to go through" and "did not feel friendly" indicates an experience that was unwanted and unpleasant, with no sense of anticipation about what is to follow. This suggests that LeXmo misinterpreted the context, possibly focusing on the structured process of going through security rather than the reviewer's negative feelings about it.

EmoNet, Pysentimiento, and ETT all classify the sentence as expressing joy, which is clearly misaligned with the tone of the sentence. These models may have misinterpreted phrases like "airport security" or "had to go through" as neutral or procedural, focusing on the logistics of the experience rather than the negative sentiment attached to it. The presence of the phrase "did not feel friendly" is a clear indicator of dissatisfaction, which was overlooked in these cases. The joy classification

shows how these models can struggle when the emotional text is conveyed through more indirect or subtle phrasing.

In contrast, RoBERTa correctly identified the emotion as anger, in line with the gold standard. This model understood the full context of the sentence, recognising the frustration in the comparison to airport security and the explicit mention of the unfriendly atmosphere. Its accurate classification suggests that this model was able to capture the reviewer's underlying dissatisfaction and resentment.

Sample Review 3: "How can you not get ice cream at the ice cream farm when you've promised your children ice cream?". [Gold standard annotation: Surprise].

This sentence expresses disbelief and frustration at an unexpected, disappointing situation. The gold standard annotation is surprise, which captures the reviewer's devastation at the irony of being unable to obtain the promised ice cream at a place that specialises in it.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anticipation	Anger	Anger	Joy	Surprise

Table 4.21: The predicted emotion class of the third sentence produced by the emotion detection classifiers

As demonstrated in Table 4.21, the tools provide a diverse set of annotations, reflecting the complexity of the sentence. LeXmo classifies the sentence as anticipation, which is inaccurate in this context. While the reviewer does express an expectation by promising ice cream, the dominant emotion is not evoked by looking forward to the event, but rather the surprise and frustration at the fact that the situation fell short of expectations.

EmoNet and Pysentimiento both classify the sentence as anger, which is understandable given the frustration in the tone, especially regarding the unfulfilled promise made to children. However, the classification overlooks the core emotion of Surprise, which derives from the unexpectedness of being unable to obtain ice cream at a dedicated ice cream farm. ETT interestingly labels the emotion as joy, which is a clear misclassification. There is no indication of joy or happiness in the sentence. This may be a result of misinterpreting the focus on ice cream, a typically joyful topic, without considering the full context of the review.

RoBERTa is in line with the gold standard and correctly classifies the emotion as surprise. It captures the disbelief expressed in the metaphorical question, recognising

the irony and shock arising from an inability to obtain ice cream at a farm that specialises in ice cream.

Sample Review 4: "I came to London fully expecting a state of the art transit system that runs like a well-oiled machine". [Gold standard annotation: Anticipation].

This sentence reflects the reviewer's clear expectation of encountering a highly efficient transportation system in London. The emotion conveyed here is anticipation, as the reviewer is looking forward to something that they presume will be impressive.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anticipation	Joy	Surprise	Joy	Anticipation

Table 4.22: The predicted emotion class of the fourth sentence produced by the emotion detection classifiers

As Table 4.22 shows, some of the tools are successfully aligned with this emotion. LeXmo and RoBERTa correctly classify the sentence as anticipation, focusing on the key phrase "fully expecting", which signals a future-oriented emotion. They seem to have accurately interpreted the forward-looking emotion, recognising that the reviewer is anticipating an efficient, well-functioning system.

However, EmoNet and ETT both classify the emotion as joy, which seems unsuitable in this context. While the sentence does convey a positive expectation, joy is not the most accurate label, as the reviewer is not expressing current happiness but, rather, looking forward to something in the future. Pysentimiento annotates the sentence with surprise, which is also incorrect. The sentence does not indicate any unexpected or shocking events; rather, it clearly expresses a calm, confident expectation of what the reviewer believes will happen.

Sample Review 5: "I do not feel the money I paid was worth the experience even though I paid the discounted rate". [Gold standard annotation: Sadness].

The sentence clearly conveys the reviewer's disappointment and regret over an experience that failed to meet their expectations, despite paying a reduced price. The gold standard annotation for this sentence is sadness. It reflects a deeper emotional let-down, where the reviewer feels that the experience fell short of what they had hoped for, regardless of the cost.

As shown in Table 4.23, the classification models produce a wide range of

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anticipation	Joy	Anger	Joy	Sadness

Table 4.23: The predicted emotion class of the fifth sentence produced by the emotion detection classifiers

classifications. LeXmo labels the sentence with anticipation, which is a significant misclassification. There is no forward-looking emotion or expectation in the sentence. Both EmoNet and ETT classify the sentence as joy, which also unsuitable. These tools may have focused on the positive connotation of a "discounted rate" and misunderstood the broader context of regret expressed by the reviewer.

Meanwhile, Pysentimiento classifies the emotion as anger, which is more understandable but still not entirely accurate. While the reviewer is frustrated at the lack of value received, the emotion expressed is more passive, leaning towards disappointment rather than active anger. The tool correctly identified the reviewer's dissatisfaction but misinterpreted the emotional nuance, confusing it with anger when sadness better reflects the overall tone.

However, only RoBERTa is fully aligned with the gold standard, classifying the emotion as sadness. This model effectively captures the reviewer's feeling of regret and unfulfilled expectations. Rather than focusing on isolated aspects of the sentence, such as the discount or the money paid, RoBERTa recognises the overall emotional tone of dissatisfaction, acknowledging that the reviewer felt let down by the experience.

Sample Review 6: "I hope my next visit will have the feel good factor not the feel bad factor!". [Gold standard annotation: Anticipation].

This sentence expresses the reviewer's forward-looking desire for a more positive experience in the future, contrasting it with a previous visit, that was clearly disappointing. The gold standard annotation for this sentence is anticipation, which reflects the reviewer's hope and expectation for a better outcome on the next visit.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Surprise	Joy	Joy	Joy	Anticipation

Table 4.24: The predicted emotion class of the sixth sentence produced by the emotion detection classifiers

As Table 4.24 shows, RoBERTa correctly classify the sentence as anticipation, in line with the gold standard. The model effectively captures the future-oriented emotion, recognising the reviewer's expression of hope and expectation for a better experience. The models probably picked up on key phrases such as "I hope" and "next visit", which strongly indicate that the reviewer is looking forward to an improved situation in the future.

However, LeXmo classifies the sentence as surprise, which is a misclassification in this context. The sentence does not contain any elements of shock or unexpectedness. EmoNet, Pysentimiento, and ETT all annotate the sentence as joy, which is not the most accurate classification. The reviewer expresses hope for a better experience, which is different from experiencing joy in the present. These models may have focused on the optimistic nature of the term "feel-good factor" and interpreted it as expressing current happiness, but overlook the fact that the reviewer is talking about future expectations rather than a present emotional state.

Sample Review 7: "Never in my life have I ever been made to feel so bad, and paying fortune for tickets to feel like that". [Gold standard annotation: Anger].

This sentence expresses strong dissatisfaction and frustration, particularly regarding the high cost of tickets and the emotional impact of this experience. The gold standard annotation for this sentence is anger, which describes the reviewer's outrage and sense of having been wronged after spending a large sum of money on a negative experience.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Joy	Joy	Sadness	Sadness	Anger

Table 4.25: The predicted emotion class of the seventh sentence produced by the emotion detection classifiers

The emotion detection tools, as shown in Table 4.25, produced a variety of responses, with only RoBERTa correctly classifying the sentence as anger. RoBERTa accurately captures the intense negative emotion in the reviewer's words, particularly the phrases "feel so bad" and "paying a fortune", which indicate both emotional and financial frustration. The model's correct classification demonstrates its ability to identify the strong sense of injustice described in the sentence.

EmoNet and LeXmo misclassified the sentence as joy. This is a significant misunderstanding, as the sentence clearly conveys dissatisfaction and anger. On the

other hand, Pysentimiento and ETT classify the emotion as sadness, which appears reasonable, but fail to capture the full emotional intensity of anger. While the sentence does convey an underlying feeling of having been let down, the reviewer's tone is more accusatory and furious than purely sad. Sadness captures the emotional disappointment, but it overlooks the reviewer's outrage at having paid a large sum for such a poor experience. These tools seem to have picked up on the "feel so bad" portion of the sentence but failed to recognise the frustration that amplifies the emotional feeling.

Sample Review 8: "Prices were ridiculously high on some stalls but it was very busy and clearly there were a lot of enthusiasts prepared to pay top money". [Gold standard annotation: Anger].

This sentence conveys the reviewer's frustration at the high prices while simultaneously observing the enthusiasm of others who are willing to pay these. The gold standard annotation for this sentence is anger, which primarily derives from the reviewer's dissatisfaction with the inflated prices.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anticipation	Joy	Joy	Joy	Joy

Table 4.26: The predicted emotion class of the eighth sentence produced by the emotion detection classifiers

As Table 4.26 shows, the classification tools produce two different classifications. LeXmo labels the sentence with anticipation, which is a significant misclassification. There is no forward-looking emotion or expectation in the sentence. EmoNet, Pysentimiento, ETT, and RoBERTa all classify the sentence as joy, which is incorrect. Interestingly, none of the classification tools capture anger, and the majority, instead, labelled the sentence as joy. This misclassification underscores how difficult it is for the tools to interpret more subtle expressions of frustration when paired with observations that are not entirely negative. Although the reviewer does acknowledge the enthusiasm of others, the tone of the sentence conveys annoyance about the high prices. The use of the term "ridiculously high" clearly indicates dissatisfaction, yet these models may have been influenced by the phrases "very busy" and "a lot of enthusiasts", which might seem positive, taken in isolation. Nevertheless, the emotion felt by the reviewer is not joy, but frustration over the costliness of the stalls, despite the positive energy in the environment.

Sample Review 9: "We found a very helpful guide with a bag saying "Ask Me". We did and she was most helpful so our experiences did improve". [Gold standard annotation: Joy].

This sentence reflects a positive, improving experience. The gold standard annotation for this sentence is joy, which depicts the relief and satisfaction that the reviewer felt after their experience became more enjoyable.

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Joy	Joy	Joy	Joy	Joy

Table 4.27: The predicted emotion class of the ninth sentence produced by the emotion detection classifiers

In this case, all of the emotion detection tools correctly classify the emotion as joy, as Table 4.27 shows. This demonstrates a strong consensus and reliability among the tools in identifying the positive emotion expressed in the sentence. Key phrases such as "helpful", "most helpful", and "did improve" clearly describe the reviewer's happiness and appreciation regarding the guide's assistance, and all of the models effectively capture this emotion. This sentence is a straightforward example, where the positive tone is unambiguous, allowing the tools to perform accurately. The clear expressions of satisfaction and improvement in the reviewer's experience make joy the most obvious and suitable label.

Sample Review 10: "Went through the usual security measures, there was a huge security presence so felt quite secure but not once was my ticket checked!". [Gold standard annotation: Joy].

This sentence indicates a mixture of feelings, where the reviewer expresses satisfaction with the security measures but also notes the particularity of not having their ticket checked. The gold standard annotation for this sentence is joy, which reflects the reviewer's overall sense of safety and security, despite the observation about the unchecked ticket.

As Table 4.28 shows, ETT and RoBERTa correctly classify the emotion as joy, in line with the gold standard. These tools seem to have captured the reviewer's primary sentiment of feeling secure, which is a positive emotional response. The key phrase "felt quite secure" suggests a sense of comfort and relief, and these models correctly interpret the overall tone as one of satisfaction.

EmoNet and Pysentimiento, however, classify the sentence as surprise, which is a reasonable interpretation, although it misses the broader emotional tone of joy. The

The Classifier	**LeXmo**	**EmoNet**	**Pysentimiento**	**ETT**	**RoBERTa**
Predicted Class	Anger	Surprise	Surprise	Joy	Joy

Table 4.28: The predicted emotion class of the tenth sentence produced by the emotion detection classifiers

fact that the reviewer's ticket was not checked is unusual and could evoke surprise. These tools focused on that specific detail and interpreted it as the dominant emotion. While surprise is a probable reaction to part of the sentence, it overlooks the positive sentiment about the security presence that defines the overall emotional tone.

LeXmo labels the sentence as anger, which is a clear misclassification. There is no indication of either frustration or dissatisfaction in the sentence, and the reviewer does not express any negative feelings about the security measures or lack of ticket-checking.

In conclusion, the analysis of these sentences provides several examples of both the strengths and limitations of various emotion detection tools, especially in the context of tourism-related text. Across the sentences, considerable variability in the tool's performance was observed, with some tools being more adept at capturing the nuances of certain emotions, while others frequently misclassified the dominant emotion. RoBERTa generally performed well in terms of alignment with the gold standard, demonstrating its ability to understand complex emotional cues and context. For instance, it consistently captured negative emotions like anger and sadness, while other models often misinterpreted them as joy or surprise.

Misclassifications were particularly evident in sentences where multiple emotional layers existed or where positive and negative elements coexisted, such as in the cases of joy versus surprise and anger versus sadness. For example, tools like LeXmo and EmoNet frequently misinterpreted sentences by focusing too heavily on isolated words or tonal shifts, rather than considering the broader context of the emotion. This indicates that, while these tools can perform in straightforward sentences with clear emotional cues, they prove less accurate when faced with more nuanced expressions, where emotions may be less explicitly stated.

The results also highlight a common challenge related to distinguishing between certain emotions, such as anger and sadness, or joy and anticipation. As seen in several examples, Pysentimiento tended to capture a portion of the emotional tone but missed the primary emotion category.

Overall, while many tools display promise with regard to the ability to classify emo-

tions accurately, RoBERTa consistently demonstrated strong performance, capturing emotional nuances more effectively than the other models. This analysis underscores the value of using fine-tuned LLMs, specifically RoBERTa, to achieve more accurate emotion recognition, especially in subjective and diverse domains like tourism.

4.5 Chapter Summary

This chapter provides an in-depth exploration of the field of emotion detection within the context of tourism reviews, encompassing a detailed assessment of the mainstream tools, and comparative performance evaluations of these tools. It begins by outlining the existing emotion detection tools, highlighting the variety in their operational methodologies and analytical approaches. It introduces key tools, ranging from ordinary emotion analysis engines to complex deep learning models, each serving as an essential component for this chapter's experiments, that were designed to identify and classify emotional content within tourism-related textual data. Furthermore, a detailed examination of the mainstream emotion detection tools was undertaken, highlighting their significance in identifying and categorising emotions within the tourism reviews. The selection criteria for these tools have been carefully defined, considering factors such as their public availability, and their degree of alignment with the selected emotion scheme. Then, a comprehensive testing methodology is applied to evaluate the performance of each tool across several emotion categories, using the F-score as a key measure of effectiveness.

This part of the thesis addressed the RQ2 based on a thorough evaluation of these tools, revealing significant gaps in their ability to capture the complex emotional nuances that occur in tourism reviews. The results demonstrated that, while these tools showed moderate performance in identifying basic emotions such as "joy" and "anger", they were less accurate in the case of more complex emotions, such as "surprise". This limitation derives from their reliance on generalised models that fail to account for the domain-specific linguistic patterns and layered emotional expressions that are inherent in tourism reviews. These findings emphasise the inadequacy of the general-purpose emotion detection tools for detecting emotions in tourism text, emphasising the need for advanced, context-specific models that are capable of addressing the unique challenges associated with emotion analysis in tourism-related text.

Therefore, this chapter further examines the fine-tuning of LLMs to improve emotion classification within tourism reviews. To address the imbalance in the emotion categories within the TORCEv2 dataset, various data augmentation methods were employed. Three language models including BERT, DistilBERT and RoBERTa, were fine-tuned with different learning rates to examine the effect of fine-tuning on emotion detection accuracy. The impact of language model fine-tuning using data

augmentation techniques was examined on the performance of emotion classification. The results indicate that the fine-tuning of LLMs, supported by appropriate data augmentation methods, substantially enhances the performance of these models with regard to accurately classifying emotions, demonstrating the benefits of adapting LLMs for emotion detection within the tourism domain.

Furthermore, RQ3 was addressed at this part of the thesis by undertaking an in-depth analysis of the adaptability and performance of LLMs that have been fine-tuned specifically for the emotion classification of tourism reviews. By fine-tuning these models on a tourism-related dataset, the experiment demonstrated considerable improvements in classification accuracy. The fine-tuning process incorporated optimisations such as adjusting the learning rates and applying data augmentation techniques, which enhanced the model's capability to identify the emotion categories. The results showed that fine-tuned LLMs outperformed the existing emotion detection tools, achieving higher F-scores across all emotion classes.

In conclusion, this chapter demonstrates the feasibility of emotion detection within the tourism context, presenting an extensive overview of the mainstream tools and advantages that arise from fine-tuning LLMs with various learning rates and data augmentation techniques to optimise their classification performance.

Chapter 5

Impact of Emotion Information on Fake Review Detection

5.1 Introduction

Emotion information is crucial for detecting fake reviews because emotions are deeply tied to the authenticity of human experiences. Genuine reviews often reflect the reviewer's real emotional responses, based on their actual experience with a product or service. In contrast, fake reviews, especially those written with the intent to manipulate opinions, frequently lack this emotional depth. These reviews may exhibit unnatural or inconsistent expressions of emotions, making them less credible. By analysing the expressions of emotions in a review, detection models can assess whether the review is either truthful or deceptive.

The following two example reviews illustrate how expressions of emotions may influence the perception of these reviews' credibility. The example truthful review is as follows:

> *"We always stay at the Sheraton Chicago Hotel & Towers when we take our kids to the city. You can't beat the location which is in walking distance of Navy Pier and Michigan Ave. The hotel is also just a couple of blocks from the fabulous Fox & Obel, a high end grocery store with many prepared items available to go and a great cafe that serves a stellar breakfast. The hotel staff are professional and courteous and the rooms are more than adequate with very comfy beds. My kids enjoy the pool with it's windows that overlook the Chicago river."*

The example fake review is as follows:

> *"Very beautiful hotel. The historic features of this hotel make it absolutely amazing. This hotel offers many great guest accommodations of which I*

> *enjoyed the spacious, tidy rooms. The staff here are very friendly, helpful, and professional. The concierge went out of her way to make sure we knew about the area and not just what was in the hotel. I especially enjoyed the service at The Lockwood Restaurant here. Our server was excellent and the food was delicious! I must say this hotel exceeded my expectations. All in all, my stay was awesome!"*

When analysing the expressions of emotions in these two reviews, the first one comes across as more authentic due to its balanced, natural expression of emotions. Studies by Choudhry et al. (2022) and Zhang et al. (2023a) highlight that genuine content often features softened emotional expressions, avoiding exaggerated sentiments. In this case, the reviewer conveys their satisfaction in a calm, measured way, without any overstatement or forced enthusiasm, which aligns with the characteristics of authenticity that have been identified in prior studies. Phrases like "professional and courteous staff" and "very comfy beds" communicate positive feelings, but the tone remains grounded. The reviewer shares genuine excitement about specific aspects, such as the pool that the children enjoy, but this excitement feels appropriate and tied to a real experience. The emotional phrases flow naturally, in line with the details provided, and there is no sense of exaggeration or excessive praise. This balance suggests that the emotions expressed are aligned with a real, nuanced experience at the hotel.

In contrast, fake content often exhibits increased emotional intensity, as suggested by Choudhry et al. (2022) and Zhang et al. (2023a). The expressions of emotions in the second review appear exaggerated and overly enthusiastic, which can be indicative of deception. Phrases like "absolutely amazing", "exceeded my expectations", and "my stay was awesome" suggest an attempt to amplify positivity in a way that feels forced and unnatural. The expressions of emotions are uniformly positive, without any variation, making the review seem overly glowing. This kind of language often signals inauthenticity because real experiences, even highly positive ones, typically evoke more nuanced emotional reactions. The lack of subtlety in the emotional phrases creates an impression that the review was written with the intent to promote the hotel rather than reflect a genuine, personal experience.

Moreover, the second review uses emotionally-charged words like "very beautiful" and "amazing" repeatedly, creating a sense that the reviewer is trying too hard to make an emotional impact. This kind of hyperbolic emotional language is often observed in deceptive reviews because it aims to convince the reader of the hotel's superiority, rather than simply to share a personal, heartfelt experience. The reviewer's tone lacks the emotional realism that would typically be present in a more genuine review.

Overall, the key difference between these two reviews lies in the subtle or exaggerated emotional tone. The first review's balanced emotional tone is in line

with the authenticity associated with a positive yet realistic experience. In contrast, the second review's excessive emotional language raises suspicion, as it feels more like an attempt to overstate the hotel's quality, which is often characteristic of deceptive reviews.

Detecting fake reviews is crucial in the tourism industry due to the significant influence of user-generated content on consumers' decisions. Fake reviews, also known as deceptive opinion spam, can mislead potential customers and damage the credibility of genuine businesses (Ott et al., 2011). Generally, detecting fake reviews is paramount for maintaining the integrity of the review platforms, protecting consumers, and ensuring fair competition among businesses. This chapter investigates the feasibility of integrating emotion information into fake review classifiers based on LLMs to enhance detection accuracy. The objective is to demonstrate how emotional cues can improve the performance of machine learning models in identifying deceptive reviews in the tourism domain.

5.2 Dataset of Genuine and Fake Reviews

The GeFaRe (Genuine and Fake Reviews) dataset was implemented for this aspect of the fake review detection experiments, which is drawn from multiple sources to ensure a balanced representation of both fake and genuine reviews. This dataset consisted of two parts: the genuine reviews were collected for this project via a process that will be discussed in Section 5.2.1; and the fake reviews were sourced from two studies (Ott et al., 2011, 2013) and will be discussed in Section 5.2.2. The process by which data are collected is critical for ensuring the reliability and validity of the findings, as the quality of the data directly impacts the effectiveness of the detection methods developed. The GeFaRe dataset comprises 1,600 reviews, equally divided into the fake and genuine categories, with each category further split into positive and negative reviews.

By combining genuine reviews from a reputable source with carefully curated fake reviews, I aimed to create a comprehensive, representative dataset. This dataset serves as a valuable resource for developing and testing machine learning models for fake review detection.

5.2.1 Genuine Review Dataset

The truthful reviews were collected from TripAdvisor, a trusted, widely-used review platform. The collection process ran from the 1^{st} of February to the 25^{th} of May 2022. To ensure the authenticity of these reviews, specific criteria were applied. Reviews were considered truthful if they had received at least five ***likes*** by different users through the like tag provided by the platform. The threshold of five likes was selected

as a robust indicator of authenticity for several reasons. Firstly, achieving five likes implies that multiple independent users found the review valuable and/or credible, providing a form of community validation that is less likely to be influenced by a single fake opinion. Secondly, this threshold strikes a balance between being too lenient, where a lower number of likes might not offer sufficient corroboration, and too strict, which might exclude genuine reviews that have not yet received significant engagement. Additionally, I observed that reviews that receive at least five likes tend to be genuine. By implementing this criterion, the study ensures that the dataset comprised reviews with a reasonable level of user endorsement, thereby enhancing the overall authenticity of the collected data. The reviews were split into positive and negative categories, with either 4-5 or 1-2 star ratings by the review author, respectively, each containing 400 reviews. This balanced approach ensured that the dataset is representative of genuine user experiences across different sentiment spectra.

The selection criteria ensured that each genuine review contained a minimum of 150 characters, similar to the deceptive reviews (will be discussed in Section 5.2.2), with an average word count of 123 for positive reviews, and 179 for negative reviews. The higher word count for negative reviews is consistent with a previous research (Ott et al., 2013) finding that users tend to provide more detailed explanations when dissatisfied.

Example Truthful Positive Review:

> *"My husband and I were in the Fairmont Chicago recently for a conference. We stayed in a spacious suite. All amenities appeared recently updated and in excellent shape. The bed was very comfortable. Views were great. I love their products in the bathroom and used them in the pristine tub two out of three nights. The room was so quiet, it was very relaxing. We ate in the restaurant downstairs (Aria) and although pricey, it was excellent. The staff were consistently attentive and responsive. It is evident that everyone is very well-trained. I would love to stay here again."*

Example Truthful Negative Review:

> *"I was in Chicago for a convention and stayed at the Sheraton towers. I had a negative experience because I got about one dozen very painful bites by bed bugs on my torso. It made attending the conference uncomfortable. The staff changed the sheets but could not move me to another room, as the conference had booked the whole hotel. After the sheets were changed, I got no more bites, thankfully. Despite this, I managed to enjoy the stay. I liked the location next to the river, and the fact that it was easy to walk to Magnificent Mile."*

Table 5.1 presents detailed statistics about the genuine reviews that were collected for the study. It includes information on the source of the reviews, the number of

hotels covered, the number of reviews per hotel, and the average word count. This table highlights the efforts made to collect a diverse, representative sample of genuine reviews.

Review Type	**Source**	**Hotels Covered**	**Reviews per Hotel**	**Minimum Word Count**	**Maximum Word Count**	**Average Word Count**
Positive Truthful	TripAdvisor	20	20	30	423	123
Negative Truthful	TripAdvisor	20	20	35	749	179

Table 5.1: Statistics for the truthful reviews

5.2.2 Fake Review Dataset

Collecting fake reviews posed a unique challenge, as this research needed to ensure that the fabricated content closely resembled authentic reviews while maintaining diversity and variability. It was decided that utilising an existing fake review dataset instead of creating a new one offered several methodological and practical advantages. The dataset had already been analysed and validated, ensuring its reliability and consistency. A key factor in its selection was its alignment with the objectives of this research, as its source, content, and stylistic characteristics closely resemble those of the genuine reviews collected. Additionally, constructing a new dataset would require substantial resources, including data validation and considerable funding to hire reviewers to generate fake reviews. By leveraging an established dataset, this research maintains its methodology while building upon the existing work in the field of fake review detection.

The fake reviews were obtained from two pivotal studies, conducted by Ott et al. (2011) and Ott et al. (2013). These studies are widely recognised in the field of deceptive review detection and provided a robust foundation for the dataset used in this research. Ott et al. (2011) collected positive fake reviews using Amazon Mechanical Turk (MTurk), a crowdsourcing platform that was discussed earlier in 3.3.2 Workers on MTurk were instructed to write deceptive positive reviews for selected hotels. Each review needed to be at least 150 characters long to ensure sufficient content for analysis. This subset of the dataset includes 400 positive reviews, with an average word count of 116 words per review. Similarly, Ott et al. (2013) collected negative fake reviews using the same method, ensuring consistency in the data collection process. The negative reviews were also required to be at least 150 characters long. This subset includes 400 reviews, averaging 178 words per review.

Example Fake Positive Review:

> *"Very beautiful hotel. The historic features of this hotel make it absolutely amazing. This hotel offers many great guest accommodations of which I enjoyed the spacious, tidy rooms. The staff here are very friendly, helpful, and professional. The concierge went out of her way to make sure we knew about the area and not just what was in the hotel. I especially enjoyed the service at The Lockwood Restaurant here. Our server was excellent and the food was delicious! I must say this hotel exceeded my expectations. All in all, my stay was awesome!"*

Example Fake Negative Review:

> *"I'd expect a 'luxury' hotel to pay more attention to details. The room really hadn't been thoroughly cleaned– I found hairballs under the bed, the mirrors were streaked, etc. The desk staff took forever to even acknowledge me when I was checking in; three of them, two on the phone, and nobody even made eye contact for 5 minutes. Then, they gave the vaguest answers to some pretty basic questions about the neighborhood; you'd think they'd know where to find a Starbucks. The internet was ridiculously slow the whole two days I was there."*

Table 5.2 provides a comprehensive overview of the fake reviews that were obtained for the study. Similar to Table 5.1, it includes information on the sources, number of hotels covered, number of reviews per hotel, and the average word count.

Review Type	**Source**	**Hotels Covered**	**Reviews per Hotel**	**Minimum Word Count**	**Maximum Word Count**	**Average Word Count**
Positive Fake	Ott et al. (2011)	20	20	25	425	116
Negative Fake	Ott et al. (2013)	20	20	32	784	178

Table 5.2: Statistics for the fake reviews

5.2.3 Combined Dataset of Genuine and Fake Reviews

The fake positive and fake negative reviews, each consisting of 400 entries, were obtained from external sources, as discussed earlier. To align with this structure and maintain equal representation across different review types, the genuine dataset

contained 400 truthful positive and 400 truthful negative reviews. This equal distribution of 400 reviews per category ensures that the GeFaRe dataset is not skewed towards any particular type of sentiment or authenticity, allowing for a more accurate assessment of the detection models.

In constructing the GeFaRe dataset, deceptive reviews were sourced from established studies, while genuine reviews were collected specifically for this research. The decision not to utilise the genuine reviews from prior studies was driven by critical methodological considerations regarding temporal validity. The genuine reviews from previous datasets were collected over a decade ago (2011–2013), and given the substantial evolution of online platforms, user behaviour patterns, linguistic conventions, and reviewing practices during this period, incorporating such outdated data would fundamentally compromise the dataset's representativeness of contemporary authentic review characteristics. The digital landscape has undergone profound transformations, including changes in platform interfaces, review guidelines, user demographics, and communication styles, making decade-old genuine reviews inadequate proxies for current authentic reviewing behaviour.

Conversely, creating a new deceptive dataset presents challenging methodological and ethical challenges that render existing validated datasets the optimal choice. Generating reliable deceptive reviews requires extensive validation protocols, and substantial funding resources to generate such data. The established deceptive reviews represent gold-standard benchmarks that have undergone rigorous validation and peer review across multiple studies. The collection of fresh genuine reviews was methodologically advantageous, ensuring the dataset captures contemporary linguistic patterns and modern user behaviour while maintaining critical consistency with the deceptive review dataset, as both subsets are sourced from the same tourism domain. This hybrid approach maximises both the reliability of deceptive classifications through established validation and the ecological validity of genuine reviews through contemporary collection, creating a more robust and practically relevant dataset that accurately reflects the fake review detection challenge.

The GeFaRe dataset, as summarised in Table 5.3, has been carefully designed to provide a comprehensive, balanced representation of different review types and sentiments, ensuring a solid foundation for the development and evaluation of the detection methods. A key consideration in the construction of this dataset was maintaining a balance across the various categories of reviews, which is critical for ensuring fair, unbiased model performance.

In addition to balancing the review categories, care was taken to ensure consistency in other aspects of the GeFaRe dataset. Specifically, 20 hotels were included, each of which was represented by exactly 20 reviews. This standardisation was implemented to match the structure of the imported fake reviews and further strengthen the dataset's uniformity. By maintaining these controlled variables, the dataset offers

Review Type	Hotels Covered	Reviews per Hotel	Reviews in the Dataset
Fake Positive	20	20	400
Fake Negative	20	20	400
Truthful Positive	20	20	400
Truthful Negative	20	20	400
The GeFaRe Dataset			1600

Table 5.3: Overall GeFaRe dataset statistics

a well-rounded, robust testing ground for the detection models, allowing reliable comparisons of model performance across different review types and ensuring that no particular category disproportionately influences the results.

Additionally, Table 5.4 provides a detailed breakdown of the word count for each review type, indicating the minimum, maximum, and average word counts. This breakdown is crucial for understanding the variability and depth of the reviews, which can impact the detection algorithms.

Review Type	Minimum Word Count	Maximum Word Count	Average Word Count
Fake Positive	25	425	116
Fake Negative	32	784	178
Truthful Positive	30	423	123
Truthful Negative	35	749	179
The GeFaRe Dataset	25	784	149

Table 5.4: Word counts for each review type

The statistics listed in Table 5.4 show that the fake positive reviews have a minimum word count of 25 and a maximum of 425, with an average of 116 words. The fake negative reviews are slightly longer, with a minimum of 32 words, a maximum of 784 words, and an average of 178 words. For the truthful reviews, the positive reviews have a minimum word count of 30, a maximum of 423, and an average of 123 words. The truthful negative reviews, similar to the fake negative reviews, tend to be longer, with a minimum of 35 words, a maximum of 749 words, and an average of 179 words.

5.2.4 Importance of Balanced Data

A balanced dataset is crucial for developing reliable, unbiased machine learning models. By ensuring an equal number of positive and negative reviews within both the fake and genuine categories, the risk of skewed results, that could arise from an imbalanced dataset, is mitigated. On the other hand, it also eliminates the necessity to use any augmentation method. This balance allows the detection algorithms to learn equally from all types of reviews, improving their ability to generalise across different scenarios. When a dataset is imbalanced, with a significant disparity between the number of instances per class, the machine learning models tend to become biased towards the majority class, leading to poor performance regarding the minority class.

Moreover, a balanced dataset helps to address the issue of class imbalance, which can occur when one class is significantly underrepresented compared to others. This imbalance can cause the model to overlook or misclassify instances from the minority class, resulting in suboptimal performance and potentially harmful consequences, especially in the case of critical applications, such as fraud detection or medical diagnosis (Fernández et al., 2018). By ensuring an equal representation of both classes, a balanced dataset encourages the model to learn the distinctive patterns and features of each class, leading to more accurate, fairer predictions.

Furthermore, a balanced dataset is essential for evaluating the performance of the machine learning models accurately. Performance metrics, such as accuracy, precision, and recall, can be misleading when applied to imbalanced datasets. A model that has been trained on an imbalanced dataset may appear to have a high degree of accuracy simply by predicting the majority class for all instances, while performing poorly on the minority class. With a balanced dataset, these metrics provide a more reliable, interpretable measure of the model's true performance across all classes.

5.3 Emotion Information

Incorporating emotion information into the analysis of reviews provides richer, more diverse content, which can significantly enhance the accuracy of the fake review detection models. Emotion analysis involves identifying and categorising the emotions expressed in text, allowing a deeper insight into the reviewer's sentiment and intent. Incorporating emotion information into the analysis of reviews can provide valuable insights that exceed the traditional positive or negative sentiment classification. For example, a genuine negative review may express "sadness", "anticipation" or "surprise", while a fake negative review might exhibit more intense emotions like "anger". Similarly, a genuine positive review may convey "joy" or "surprise", while a fake positive review might display exaggerated enthusiastic emotions. This section details the process of extracting emotion information from the reviews.

5.3.1 Emotion Classification

Each review in the GeFaRe dataset was segmented into sentences using the SpaCy English language model. Segmentation is the process of breaking down text into smaller units, such as sentences or words, to facilitate more detailed analysis. For this study, sentence-level segmentation was chosen to capture the emotional context within different parts of a review, as emotions can vary significantly throughout a single piece of text.

Once segmented, each sentence was classified into one of five basic emotions: "joy", "anger", "sadness", "surprise", or "anticipation". The emotion classification model utilised in this step is the one that demonstrated the highest F-score, as developed and thoroughly evaluated in the previous chapter. This model was selected based on its superior performance, ensuring the most accurate and reliable emotion classification for this phase of the study.

Consider the following review:

> *"I am so glad I decided to stay at the Intercontinental Chicago for my first trip to the city. The staff is very attentive, I felt like I was the only person they had to take care of! The location is great, right on the Magnificent Mile and so close to major attractions. My suite was well appointed and very clean. I even looked behind some furniture for dust and couldn't find any. The hotel is very luxurious. The sheets and towels were very soft and the bed very comfortable. The next time I travel to Chicago, I will definitely stay at the Intercontinental again."*

This review was segmented into sentences and classified into its suitable emotion class, as illustrated in Table 5.5.

Table 5.6 presents the distribution of emotions across the sentences in the example review. This detailed breakdown illustrates how different parts of the review convey distinct emotional expressions, which can be crucial for detecting the authenticity of a review.

5.3.2 Integration of Emotion Information into the Detection Models

The emotion information extracted from each review was integrated into the fake review detection models to enhance their performance. Several configurations were tested to determine the most effective way of incorporating this information into the classifiers:

1. **Base Models**: These models did not incorporate any information related to emotions. Specifically, they were designed and implemented without the

The sentence	Emotion Class
"I am so glad I decided to stay at the Intercontinental Chicago for my first trip to the city."	Joy
"The staff is very attentive, I felt like I was the only person they had to take care of!"	Joy
'The location is great, right on the Magnificent Mile and so close to major attractions."	Joy
"My suite was well appointed and very clean."	Joy
"I even looked behind some furniture for dust and couldn't find any."	Surprise
"The hotel is very luxurious."	Joy
"The sheets and towels were very soft and the bed very comfortable."	Joy
"The next time I travel to Chicago, I will definitely stay at the Intercontinental again."	Joy

Table 5.5: An example review containing multiple sentences with emotions

Emotion Class	Joy	Anger	Sadness	Surprise	Anticipation
Number of Sentences	6	0	0	1	0

Table 5.6: The emotion distribution of the sentences within a review

integration of emotional data or features, ensuring that their functionality and performance remained independent of any emotional context or variables.

2. **Dominant Emotion Class**: Only the dominant emotion class of each review was included in the experiment. This approach simplifies the emotional information provided to the classifier by concentrating exclusively on the most prominent emotion expressed in each review. By focusing on the primary emotional category, we can reduce the complexity and enhance the clarity, making it easier to include more straightforward emotional information without interference from secondary emotions.

3. **All Emotion Classes**: All five emotion classes were included, providing a comprehensive, nuanced emotional profile for each review. This thorough inclusion ensures that the full spectrum of emotional information is captured, offering a detailed, multifaceted understanding of the emotional dynamics within each review. The analysis is deepened by accounting for all emotion classes, reflecting a complete emotional landscape.

4. **Dominant Emotion Class with General Text Features**: This configuration integrated the dominant emotion class with general textual features, including counts for the sentences, words, characters, uppercase words, and exclamation marks. As discussed in Section 2.5.1.2, these structural features were extracted following established methodologies in text analysis and deception detection research. Sentence and word counts provide fundamental measures of text complexity and verbosity, which have been identified as discriminative features in fake review detection. Character count serves as a granular measure of text length that captures detailed writing patterns beyond simple word counts (Le et al., 2022; Qiao and Rui, 2023; Abri et al., 2020). Uppercase word frequency has been recognised as an indicator of emotional intensity and emphasis in text communication, particularly relevant for detecting sentiment in deceptive reviews (Hajek and Sahut, 2022). Exclamation mark frequency represents a punctuation-based feature that has proven effective in sentiment analysis and authenticity detection, as it often correlates with manufactured enthusiasm or emotional manipulation in fake content (Ott et al., 2011). By combining these elements with emotion information, the model leverages both the emotional and structural aspects of the text, enhancing its ability to capture nuanced patterns and provide a more robust analysis that incorporates both linguistic complexity and stylistic indicators established in the deception detection literature.

5. **All Emotion Classes with All Class Textual Features**: This configuration combined all five emotion classes with the textual features for each emotion class, offering a detailed analysis of how different emotions are expressed in the text. By incorporating these elements, the model provides a comprehensive understanding of the emotional content, capturing the unique characteristics and patterns associated with each emotion that is expressed within the text. This approach allows a more granular, insightful analysis of the text.

6. **All Emotion Classes with General Text Features**: This approach combined all of the emotion classes with general textual features to provide a holistic view of the review's emotional and textual characteristics. By integrating a full range of emotions with structural elements, such as counts for the sentences, words, characters, uppercase words, and exclamation marks, the model delivers a comprehensive analysis that captures both the emotional depth and the textual information of each review. This method ensures a well-rounded understanding of the content, encompassing the full spectrum of expressions of emotions alongside the fundamental textual properties.

5.4 Experiments

Multiple experiments were conducted to investigate the impact of incorporating emotion information on the accuracy of the fake review detection models. The key difference between these experiments lay in the type of emotion information used and how it was integrated into the classifier. By varying the emotion information, the experiments aimed to determine the most effective way to enhance the fake review detection performance through the addition of emotional data.

Three LLMs were included in the experiments, namely BERT, RoBERTa, and DistilBERT. To ensure the robustness and reliability of the results, each experiment was trained and tested using a standard train-validation-test split methodology. This approach involved dividing the GeFaRe dataset into three distinct subsets: 60% for training, 20% for validation, and 20% for testing. The training set was used to optimise model parameters, the validation set monitored performance during training to prevent overfitting and guide hyperparameter selection, while the test set provided an unbiased evaluation of final model performance. For each configuration, the models underwent 20 epochs of training to ensure thorough learning.

In addition to this training process, the models were fine-tuned using multiple learning rates [1e-5, 3e-5, 5e-5] to determine the optimal rate for gradient descent, which directly impacts the models' ability to learn from the data. Furthermore, two different dropout rates [0.2 and 0.3] were tested to address any overfitting by randomly dropping units from the neural network during training. This comprehensive training regimen, encompassing various configurations and hyperparameters, aimed to identify the most effective setup for enhancing the accuracy of the fake review detection models by incorporating detailed emotional information.

This systematic experimental design underscores the commitment to optimising model performance and understanding the nuanced impact of emotional information on fake review detection. The train-validation-test methodology ensures that the reported results reflect genuine model capabilities on unseen data, providing reliable estimates of performance that would be expected in real-world deployment scenarios. In the following, each experiment will be discussed in detail, together with its results.

5.4.1 Structure of the Classification Models

5.4.1.1 Models Without Emotion Information

Two structures of classification methods were tested for the base models without emotion information. The first structure included one dense layer, and the second included multiple dense layers. These structures were designed to evaluate the baseline performance of the models without any emotion-related features.

1. Single Dense Layer Architecture

The model architecture for fake review detection without emotion information follows the same fundamental structure as described in Section 4.3.4 for emotion detection, leveraging the capabilities of pre-trained language models like BERT, RoBERTa, or DistilBERT. However, this architecture differs in two critical aspects that adapt it specifically for binary classification of fake review detection.

Input Text, Tokenisation and Masking: Although both models utilise identical tokenisation, masking, embedding, CLS vector extraction, fully connected layers, and ReLU activation components, the fake review detection model processes the same features but optimises them specifically for deceptive detection rather than emotion classification. Ultimately, the fundamental difference lies in the classification objective and the learned feature representations that the model develops during training for the specific task of distinguishing truthful from deceptive reviews.

Classification Label: While the emotion detection model outputs predictions across five emotion categories including "joy", "anger", "sadness", "anticipation", and "surprise", this fake review detection model performs binary classification, determining whether a review is fake or genuine. The final classification layer is therefore configured for two classes rather than five, with the fully connected layer dimensions adjusted accordingly.

```
class classifier(nn.Module):
  def __init__(self,model_name,num_classes,dropout_val):
    super(classifier, self).__init__()
    self.fModel = sModel.from_pretrained(model_name)
    self.dropout = nn.Dropout(dropout_val)
    self.classifier = nn.Linear(768 , num_classes)

  def forward(self, input_ids, attention_mask):
    outputs = self.fModel(input_ids=input_ids,
                          attention_mask=attention_mask)
    pooler = outputs[0][:, 0]
    pooler = nn.ReLU()(pooler)
    pooler = self.dropout(pooler)
    logits = self.classifier(pooler)
    return logits
```

Listing 5.1: Classification function of one dense layer architecture without emotion information

Where the model's name can take any of the following values: "bert-base-uncased", "roberta-base", and "distilbert-base-uncased", based on the chosen LLM.

In summary, the model with one dense layer leverages the sophisticated language understanding of BERT, RoBERTa, or DistilBERT, combined with a straightforward dense layer and activation function, to classify reviews effectively. Each component of the model plays a crucial role in transforming raw text into a meaningful classification, demonstrating the power and efficiency of LLMs in addressing the task of fake review detection.

Moreover, Listing 5.1 presents the classifier class from the Python code of the experiment. This class is a core component of the model, responsible for processing input data then categorising it, based on the pre-defined classification parameters. Specifically, this classification class takes the original hidden layer from the LLM, which has 768 dimensions, and transforms it into a 2-dimensional vector, representing the authenticity of the given text.

2. Multiple Dense Layer Architecture

The model architecture with multiple dense layers builds upon the simpler, one dense layer model by introducing additional, fully-connected layer. The model with multiple dense layers extends the capabilities of the simpler architecture by adding depth to the fully connected layers. This added depth allows the model to capture more complex patterns and relationships within the data, potentially leading to improved predictive performance by the model.

```
class classifier(nn.Module):
  def __init__(self,model_name,num_classes,dropout_val):
    super(classifier, self).__init__()
    self.fModel = sModel.from_pretrained(model_name)
    self.dropout = nn.Dropout(dropout_val)
    self.preclassifier = nn.Linear(768 , 192)
    self.classifier = nn.Linear(192 , num_classes)

  def forward(self, input_ids, attention_mask):
    outputs = self.fModel(input_ids=input_ids,
                          attention_mask=attention_mask)
    pooler = outputs[0][:, 0]
    pooler = self.preclassifier(pooler)
    pooler = nn.ReLU()(pooler)
    pooler = self.dropout(pooler)
    logits = self.classifier(pooler)
    return logits
```

Listing 5.2: Classification function of multiple dense layers architecture without emotion information

Where the model's name can take any of the following values: "bert-base-uncased", "roberta-base", and "distilbert-base-uncased", based on the chosen LLM.

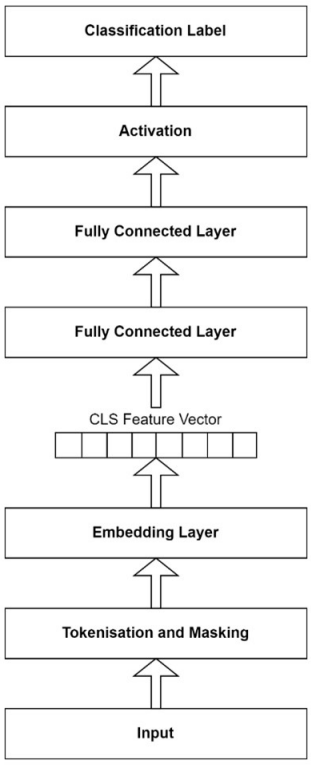


Figure 5.1: Flowchart of the model with multiple dense layers without emotion information

Unlike the simpler model, that uses a single dense layer, this architecture, as shown in Figure 5.1, incorporates multiple, fully-connected layers. The output from the CLS feature vector is first fed into the initial dense layer, which performs a linear transformation on the input data. This transformation reduces the high dimensional contextual representation into a more manageable form. However, the subsequent dense layer further process these transformed data, allowing the model to learn and capture more complex patterns and relationships within the input text. Each layer builds upon the features extracted by the previous layer, progressively refining the representation of the input data.

Additionally, Listing 5.2 demonstrates the classifier class from the Python code used in the experiment. This class is a crucial part of the model, and is responsible for processing the input data and categorising them according to the predefined classification parameters. Explicitly, this classification class takes the original hidden layer from the LLM, which has 768 dimensions, and transforms it into a 2-dimensional vector, representing the authenticity of the given text. The process consists of two steps: first, it reduces the 768-dimensional input to an intermediate 192-dimensional

vector, then further transforms this into a final 2-dimensional vector.

5.4.1.2 Models with Emotion Information

The models incorporating emotion information represent a more advanced approach to text classification tasks. These architectures aim to leverage additional emotional information that are present in the input text to enhance the model's predictive capabilities. Two distinct architectures were explored, one with a single dense layer and another with multiple dense layers, for processing the combined text and emotion representations. These models incorporated different configurations of emotion information to assess their impact on fake review detection accuracy.

1. Single Dense Layer Architecture

The single dense layer architecture, as depicted in Figure 5.2, introduces a dedicated input for emotion information alongside the main text input. This emotion information can take various forms, such as emotion categories or other textual features that are extracted from the text, as discussed earlier in the emotion information section.

In this architecture, separate embedding layers are employed to convert the text and emotion information into dense vector representations. These separate embeddings are then combined to form a joint feature vector, which serves as the input for a single, fully-connected layer. This dense layer aims to learn a meaningful representation that fuses the emotional and textual information. By including emotion information directly into the model's architecture, this approach allows the model to capture and utilise the emotional context that is present in the input text, potentially improving the classification performance.

The following provides an overview of the additional steps of the model that incorporates emotion information, in contrast to the models that do not.

Emotion Information: In addition to the raw text input, this model incorporates emotion-related features, derived from the input text. These features include emotion categories, which classify the emotions present in the text, including "joy", "anger", "surprise", "anticipation" and "sadness", or textual features, including the counts for sentences, words, characters, uppercase words, and exclamation marks. Extracting these features can provide valuable contextual information for the model, potentially enhancing its ability to understand and classify the input data more accurately.

Combining Features: After processing the text embeddings, the resulting contextual representations are combined with the emotional features that were

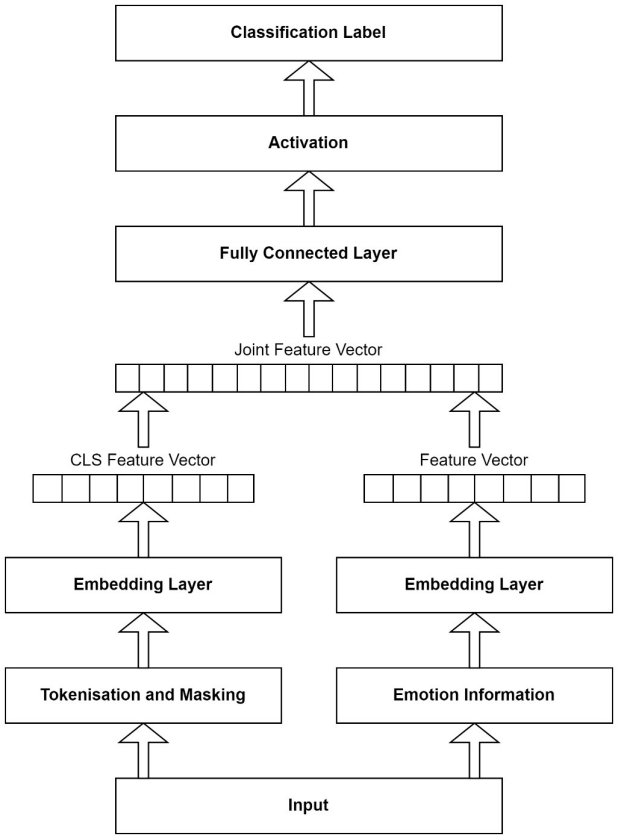


Figure 5.2: Flowchart of the model with one dense layer with emotion information

extracted earlier. This combination of contextual, textual, and emotional information aims to equip the model with a more comprehensive understanding of the input data, enabling it to capture both the semantic content and the emotional information present in the text.

By incorporating emotion-related features alongside textual data, this architecture endeavours to attain a more holistic understanding of the input, potentially enhancing the accuracy and efficacy of the fake review classification task, given the significant role of emotional information.

Listing 5.3 introduces an example of the classifier class from the Python code used in the *dominant emotion class* experiment, which is one of the experiments that integrate emotion information with the classifier. This class plays a vital role in the model, processing input data and categorising it, based on predefined classification parameters. Moreover, this classification class takes the original hidden layer from the LLM, which has 768 dimensions, and merges it with emotion information that is

transformed first into a vector of a predefined size (five dimensions, in this example), and then into a 2-dimensional vector, representing the authenticity of the given text.

```
class classifier(nn.Module):
  def __init__(self,model_name,num_classes,dropout_val):
    super(classifier, self).__init__()
    self.fModel = sModel.from_pretrained(model_name)
    self.emotion_feature = nn.Embedding(5,10)
    self.dropout = nn.Dropout(dropout_val)
    self.classifier = nn.Linear(768 + 10 , num_classes)

  def forward(self,input_ids, attention_mask,
              emotion_feature_ids):
    outputs = self.fModel(input_ids=input_ids,
                          attention_mask=attention_mask)
    pooler = outputs[0][:, 0]
    emotion_emb=self.emotion_feature(emotion_feature_ids)
    pooler = torch.cat((pooler, emotion_emb), dim=-1)
    pooler = nn.ReLU()(pooler)
    pooler = self.dropout(pooler)
    logits = self.classifier(pooler)
    return logits
```

Listing 5.3: Classification function of one dense layer architecture with emotion information

Where the model's name can take any of the following values: "bert-base-uncased", "roberta-base", and "distilbert-base-uncased", based on the chosen LLM. Also, the emotion feature variable can be redefined to different settings, based on the method of incorporating emotion information with the classifier.

2. Multiple Dense Layer Architecture

The multiple dense layer architecture follows a similar principle but introduces an additional level of complexity. Instead of a single fully connected layer, this architecture employs two dense layers to process the combined text and emotion representations, as shown in Figure 5.3. The first dense layer operates on the text embeddings, while the second takes the concatenation of the first layer's output and the emotion embeddings as its input.

By integrating emotion-related features with textual data and leveraging multiple dense layers, this architecture aspires to achieve a more nuanced understanding of the input text, potentially enhancing the accuracy and performance of the fake review classification task. The combination of textual and emotional information, along with the increased complexity introduced by employing multiple dense layers, may provide deeper insights and a more refined contextual understanding. This multi-faceted approach ensures that the model captures a broader spectrum of information,

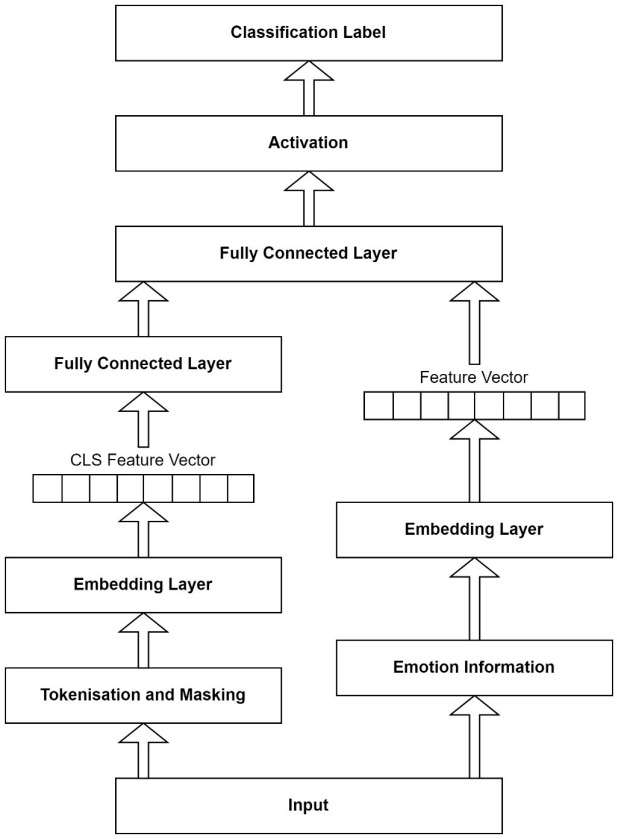


Figure 5.3: Flowchart of the model with multiple dense layers with emotion information

encompassing both the semantic content and the emotional information present in the text. Given the significant role that emotional information plays in human communication, this integration is particularly advantageous for the classification task. By enriching the model's capacity to discern subtle emotional information, it will be able to differentiate between genuine and deceptive reviews more effectively, thus improving the robustness and reliability of the classification outcomes.

Listing 5.3 presents an example of the classifier class from the Python code used in the *dominant emotion class* experiment, which is one of the experiments that integrate emotion information with the classifier. This class is a key component of the model, responsible for processing input data and categorising it according to the established classification criteria. Additionally, this classification class takes the original hidden layer from the LLM, which has 768 dimensions, transforms it into a 192-dimensional vector, then merges it with emotion information that transformed

into a vector of a predefined size (five dimensions, in this example). The merged vector is then transformed into a 2-dimensional vector, representing the authenticity of the given text.

```
class classifier(nn.Module):
  def __init__(self,model_name,num_classes,dropout_val):
    super(classifier, self).__init__()
    self.fModel = sModel.from_pretrained(model_name)
    self.emotion_feature = nn.Embedding(5,10)
    self.dropout = nn.Dropout(dropout_val)
    self.preclassifier = nn.Linear(768 , 192)
    self.classifier = nn.Linear(192 + 10 , num_classes)

  def forward(self, input_ids, attention_mask,
              emotion_feature_ids):
    outputs = self.fModel(input_ids=input_ids,
                          attention_mask=attention_mask)
    pooler = outputs[0][:, 0]
    pooler = self.preclassifier(pooler)
    emotion_emb=self.emotion_feature(emotion_feature_ids)
    pooler = torch.cat((pooler, emotion_emb), dim=-1)
    pooler = nn.ReLU()(pooler)
    pooler = self.dropout(pooler)
    logits = self.classifier(pooler)
    return logits
```

Listing 5.4: Classification function of multiple dense layers architecture with emotion information

Where the model's name can take any of the following values: "bert-base-uncased", "roberta-base", and "distilbert-base-uncased", based on the chosen LLM. Also, the emotion feature variable can be redefined to different settings, based on the method of incorporating emotion information with the classifier.

5.4.2 Experimental Results and Evaluation

5.4.2.1 Base Models

The base models classified the GeFaRe dataset without the incorporation of emotion-based features. The comprehensive performance metrics obtained using this method are presented in Tables 5.7-5.14, encompassing accuracy, precision, recall, and F-score values across different architectural configurations. These findings establish the baseline performance levels for subsequent evaluation of emotion-enhanced architectures using both single and multiple dense layer implementations.

The results outlined in Tables 5.7-5.10 reveal distinct performance patterns across the three language models in single dense layer configurations. RoBERTa achieves

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.903	0.906
	3e-5	0.922	0.906	0.916
	5e-5	0.913	0.922	0.919
0.3	1e-5	0.916	0.900	0.922
	3e-5	0.919	0.894	0.906
	5e-5	0.922	0.928	0.913

Table 5.7: Accuracy scores of the base model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.917	0.907	0.908
	3e-5	0.922	0.906	0.916
	5e-5	0.914	0.924	0.919
0.3	1e-5	0.918	0.902	0.922
	3e-5	0.919	0.894	0.907
	5e-5	0.922	0.929	0.913

Table 5.8: Precision scores of the base model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.903	0.906
	3e-5	0.922	0.906	0.916
	5e-5	0.913	0.922	0.919
0.3	1e-5	0.916	0.900	0.922
	3e-5	0.919	0.894	0.906
	5e-5	0.922	0.928	0.913

Table 5.9: Recall scores of the base model with one dense layer

superior scores across multiple evaluation metrics, with accuracy values reaching 0.928 at 5e-5 learning rate combined with 0.3 dropout. The model's precision performance peaks at 0.929 under the same hyperparameter configuration, while recall metrics

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.912	0.903	0.906
	3e-5	0.922	0.906	0.916
	5e-5	0.912	0.922	0.919
0.3	1e-5	0.915	0.900	0.922
	3e-5	0.919	0.894	0.906
	5e-5	0.922	0.928	0.913

Table 5.10: F-score values of the base model with one dense layer

attain their highest value of 0.928 at 5e-5 learning rate with 0.3 dropout. F-score values corroborate these findings, with optimal performance of 0.928 achieved under same conditions. BERT produces comparably robust results across all metrics, though with moderately lower peak values. Accuracy scores range from 0.913 to 0.922, with the maximum achieved at both 3e-5 and 5e-5 learning rates when combined with 0.2 and 0.3 dropout respectively. Precision values span 0.914 to 0.922, demonstrating particular strength at 3e-5 and 5e-5 learning rates. Recall scores exhibit similar patterns, ranging from 0.913 to 0.922, while F-score values remain within the 0.912-0.922 range across various hyperparameter combinations. DistilBERT generates the most modest results among the three models, yet produces acceptable performance levels across all metrics. Accuracy measurements range from 0.906 to 0.922, with optimal performance occurring at 1e-5 learning rate combined with 0.3 dropout. Precision scores span 0.907 to 0.922, while recall values range from 0.906 to 0.922. F-score values fall within 0.906 to 0.922, indicating balanced performance between precision and recall across different configurations.

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.881	0.913
	3e-5	0.913	0.900	0.913
	5e-5	0.897	0.928	0.916
0.3	1e-5	0.928	0.903	0.906
	3e-5	0.913	0.925	0.938
	5e-5	0.913	0.916	0.925

Table 5.11: Accuracy scores of the base model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.926	0.884	0.914
	3e-5	0.916	0.900	0.913
	5e-5	0.898	0.928	0.916
0.3	1e-5	0.930	0.905	0.908
	3e-5	0.913	0.925	0.938
	5e-5	0.913	0.916	0.925

Table 5.12: Precision scores of the base model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.881	0.913
	3e-5	0.913	0.900	0.913
	5e-5	0.897	0.928	0.916
0.3	1e-5	0.928	0.903	0.906
	3e-5	0.913	0.925	0.938
	5e-5	0.913	0.916	0.925

Table 5.13: Recall scores of the base model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.881	0.912
	3e-5	0.912	0.900	0.912
	5e-5	0.897	0.928	0.916
0.3	1e-5	0.928	0.903	0.906
	3e-5	0.912	0.925	0.937
	5e-5	0.912	0.916	0.925

Table 5.14: F-score values of the base model with multiple dense layers

The multiple dense layer implementation, as documented in Tables 5.7-5.14, introduces notable variations in performance characteristics compared to single

layer architectures. RoBERTa exhibits pronounced performance fluctuations, with accuracy scores varying from 0.881 to 0.928. The model achieves its peak accuracy of 0.928 at 5e-5 learning rate with 0.2 dropout. However, precision values demonstrate similar variability, ranging from 0.884 to 0.928, with the highest precision of 0.928 occurring at 5e-5 learning rate with 0.2 dropout. Recall scores span from 0.881 to 0.928, while F-score values range from 0.881 to 0.928. BERT demonstrates more stable performance characteristics in multiple dense layer configurations. Accuracy scores range from 0.897 to 0.928, with precision values spanning 0.898 to 0.930. The model achieves its highest precision of 0.930 at 1e-5 learning rate with 0.3 dropout. Recall measurements vary from 0.897 to 0.928, while F-score values remain within the 0.897 to 0.928 range, indicating balanced performance across different evaluation metrics. DistilBERT produces the most stable results in multiple dense layer architectures. Accuracy scores range from 0.906 to 0.938, with precision values spanning 0.908 to 0.938. The model's recall performance varies from 0.906 to 0.938, while F-score measurements range from 0.906 to 0.937, demonstrating consistent behaviour across various hyperparameter configurations.

Dropout rate effects reveal model-specific response patterns. In single dense layer configurations, RoBERTa benefits from higher dropout values of 0.3, particularly evident in the 5e-5 learning rate condition where performance improvements are observed across multiple metrics. BERT shows limited responses to dropout variations, with certain configurations favouring 0.2 while others perform optimally with 0.3. DistilBERT demonstrates preference for lower dropout rates of 0.2 in multiple configurations, though exceptions exist.

Learning rate influence proves more substantial and consistent across models. For single dense layer architectures, 5e-5 learning rate frequently produces optimal results for RoBERTa, while 3e-5 provides stable performance for BERT. DistilBERT shows favourable performance at 5e-5 learning rate in numerous metric evaluations. In multiple dense layer configurations, learning rate selection becomes more critical, with 1e-5 enabling RoBERTa's lowest performance, while BERT and DistilBERT show improved stability at moderate learning rates.

The comparison between single and multiple dense layer architectures reveals mixed interchanges between performance potential and stability. One dense layer implementations demonstrate higher consistency across all models, with performance variations typically remaining within narrow ranges for each model. RoBERTa's single-layer performance fluctuates between 0.894 and 0.928 across all metrics, representing manageable variability. Multiple dense layer architectures introduce considerable performance variance, particularly for RoBERTa, where the performance differential between optimal and suboptimal configurations exceeds 4 points. This architectural enhancement enables higher metric scores while simultaneously increasing the risk of poor performance under suboptimal hyperparameter selection. BERT and

DistilBERT exhibit greater resilience to architectural modifications, preserving more stable performance ranges regardless of layer depth.

5.4.2.2 Dominant Emotion Class

This experimental approach incorporates the dominant emotion class for each review as a feature within the classification framework. The underlying assumption assumes that the dominant emotion class represents the primary emotional expression of each review. The comprehensive performance metrics obtained through this methodology are presented in Tables 5.15-5.22, encompassing accuracy, precision, recall, and F-score measurements across single and multiple dense layer architectural configurations.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.916	0.938
	3e-5	0.950	0.919	0.938
	5e-5	0.931	0.919	0.931
0.3	1e-5	0.944	0.906	0.938
	3e-5	0.913	0.934	0.931
	5e-5	0.922	0.941	0.928

Table 5.15: Accuracy scores of the dominant emotion class model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.935	0.916	0.938
	3e-5	0.951	0.924	0.938
	5e-5	0.931	0.919	0.931
0.3	1e-5	0.945	0.907	0.938
	3e-5	0.914	0.934	0.932
	5e-5	0.922	0.941	0.929

Table 5.16: Precision scores of the dominant emotion class model with one dense layer

The results illustrated in Tables 5.15-5.18 reveal distinct performance patterns when emotion-based features are incorporated into one dense layer architectures. BERT achieves superior performance across most evaluation metrics, with accuracy

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.916	0.938
	3e-5	0.950	0.919	0.938
	5e-5	0.931	0.919	0.931
0.3	1e-5	0.944	0.906	0.938
	3e-5	0.913	0.934	0.931
	5e-5	0.922	0.941	0.928

Table 5.17: Recall scores of the dominant emotion class model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.916	0.937
	3e-5	0.950	0.918	0.937
	5e-5	0.931	0.919	0.931
0.3	1e-5	0.944	0.906	0.938
	3e-5	0.912	0.934	0.931
	5e-5	0.922	0.941	0.928

Table 5.18: F-score values of the dominant emotion class model with one dense layer

scores reaching a peak of 0.950 at 3e-5 learning rate combined with 0.2 dropout. This represents a notable enhancement compared to the base model performance. Precision measurements further support BERT's improved capability, achieving 0.951 at 3e-5 learning rate with 0.2 dropout, while recall values reach 0.950 under same conditions. F-score metrics corroborate these findings, with BERT attaining 0.950 at the same hyperparameter configuration. RoBERTa produces mixed results in single dense layer implementations, with accuracy scores ranging from 0.906 to 0.941. The model achieves its optimal accuracy of 0.941 at 5e-5 learning rate with 0.3 dropout. Precision values span from 0.907 to 0.941, reaching their peak at the same configuration. Recall measurements demonstrate similar patterns, with values ranging from 0.906 to 0.941, while F-score results maintain comparable ranges across different hyperparameter combinations. DistilBERT exhibits stable performance characteristics across single dense layer configurations. Accuracy measurements remain within the 0.928 to 0.938 range, demonstrating consistent behaviour across various hyperparameter settings. Precision scores span from 0.929 to 0.938, while recall values range from 0.928 to 0.938.

F-score measurements fall within similar ranges, indicating balanced performance between precision and recall metrics.

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.903	0.944
	3e-5	0.934	0.931	0.925
	5e-5	0.931	0.938	0.938
0.3	1e-5	0.919	0.903	0.938
	3e-5	0.919	0.925	0.950
	5e-5	0.919	0.925	0.938

Table 5.19: Accuracy scores of the dominant emotion class model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.909	0.945
	3e-5	0.935	0.933	0.926
	5e-5	0.933	0.939	0.939
0.3	1e-5	0.920	0.905	0.938
	3e-5	0.920	0.927	0.955
	5e-5	0.919	0.927	0.939

Table 5.20: Precision scores of the dominant emotion class model with multiple dense layers

The multiple dense layer results, as presented in Tables 5.19-5.22, introduce notable shifts in performance hierarchies compared to single layer implementations. DistilBERT emerges as the best performer in this architectural configuration, achieving exceptional accuracy scores of up to 0.950 at 3e-5 learning rate with 0.3 dropout. The model's precision performance reaches 0.955 under the same conditions. Recall measurements peak at 0.950, while F-score values attain 0.950, confirming DistilBERT's superior performance in multiple dense layer architectures. BERT produces solid results in multiple dense layer configurations, with accuracy scores ranging from 0.919 to 0.934. The model achieves its peak accuracy of 0.934 at 3e-5 learning rate with 0.2 dropout. Precision values span from 0.919 to 0.935, reaching

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.903	0.944
	3e-5	0.934	0.931	0.925
	5e-5	0.931	0.938	0.938
0.3	1e-5	0.919	0.903	0.938
	3e-5	0.919	0.925	0.950
	5e-5	0.919	0.925	0.938

Table 5.21: Recall scores of the dominant emotion class model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.922	0.903	0.944
	3e-5	0.934	0.931	0.925
	5e-5	0.931	0.931	0.925
0.3	1e-5	0.919	0.903	0.937
	3e-5	0.919	0.925	0.950
	5e-5	0.919	0.925	0.937

Table 5.22: F-score values of the dominant emotion class model with multiple dense layers

optimal performance at 3e-5 learning rate with 0.2 dropout. Recall scores demonstrate similar patterns, ranging from 0.919 to 0.934, while F-score measurements maintain comparable performance ranges. RoBERTa exhibits relatively modest performance in multiple dense layer architectures compared to the other models. Accuracy measurements range from 0.903 to 0.938, with peak performance achieved at 5e-5 learning rate combined with dropout of 0.2. Precision scores span from 0.905 to 0.939, while recall values range from 0.903 to 0.938. F-score results maintain similar ranges, indicating consistent performance across different evaluation metrics.

Dropout rate effects demonstrate model-specific responses when emotion features are incorporated. BERT shows clear preference for lower dropout rates in single dense layer configurations, achieving optimal performance across all metrics. This suggests that the additional emotion information reduces the need for aggressive regularisation. DistilBERT exhibits preference for higher dropout rates in multiple dense layer

architectures, indicating that the model benefits from increased regularisation when processing both textual and emotional features simultaneously.

Learning rate selection reveals consistent patterns across models. The 3e-5 learning rate frequently produces optimal results for both BERT and DistilBERT across different architectural configurations. This mid-range learning rate appears to provide an effective balance for processing the combined textual and emotional feature. RoBERTa shows preference for the 5e-5 learning rate in several configurations, suggesting different optimisation requirements when emotional features are incorporated.

The comparison between single and multiple dense layer architectures reveals interesting outcomes when emotion features are included. Single dense layer architectures enable BERT to achieve its peak performance, suggesting that the emotion information provides sufficient additional context without requiring increased architectural complexity. Multiple dense layer configurations favour DistilBERT, which appears to benefit from the enhanced processing capacity when handling combined feature representations. The performance differential between architectural configurations varies across models. BERT shows minimal difference between single and multiple dense layer implementations, with single layer configurations often producing superior results. DistilBERT demonstrates clear benefits from multiple dense layers when emotion features are incorporated, achieving its highest performance levels in these configurations. RoBERTa exhibits relatively stable performance across both architectural variants, though with generally lower absolute performance levels.

The incorporation of dominant emotion class features yields measurable improvements across multiple model configurations when compared to baseline performance. BERT demonstrates substantial enhancements, particularly evident in single dense layer architectures where accuracy improvements of approximately 2-3 percentage points are observed. For instance, BERT achieves 0.950 accuracy at 3e-5 learning rate with 0.2 dropout in the emotion-enhanced model, compared to 0.922 in the corresponding base configuration. DistilBERT shows remarkable improvement in multiple dense layer configurations, achieving peak performance levels that surpass both BERT and RoBERTa. The model's accuracy enhancement from base performance is particularly pronounced, with improvements of up to 1.2 percentage points observed in optimal configurations. This suggests that DistilBERT benefits substantially from the additional emotional context provided by the dominant emotion class feature. RoBERTa presents a more complex pattern of performance changes. While the model shows improvements in certain configurations, the enhancement is less pronounced compared to BERT and DistilBERT. In some instances, particularly in multiple dense layer architectures, RoBERTa's performance remains relatively stable or shows modest improvements compared to baseline results.

The integration of dominant emotion class features produces meaningful performance enhancements across the majority of experimental configurations. Statistical analysis reveals improvements in 30 out of 36 tested settings compared to base models, representing an 83% success rate for emotion-enhanced performance. This substantial improvement rate validates the hypothesis that dominant emotion information provides valuable context for fake review classification tasks. The experimental findings support the integration of emotion-based features as a viable enhancement strategy for fake review detection systems, with clear benefits observed across multiple model architectures and hyperparameter configurations.

5.4.2.3 All Emotion Classes

This experimental methodology incorporates comprehensive emotional profiling through the inclusion of five vectors representing the frequency distribution of sentences across each emotion class within individual reviews. This approach provides a granular representation of the emotional composition of each review, capturing both the diversity and intensity of emotional expressions. The comprehensive performance metrics obtained through this approach are documented in Tables 5.23-5.30, encompassing accuracy, precision, recall, and F-score values across single and multiple dense layer architectural implementations.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.906	0.900	0.931
	3e-5	0.919	0.925	0.938
	5e-5	0.913	0.941	0.938
0.3	1e-5	0.925	0.919	0.934
	3e-5	0.925	0.941	0.931
	5e-5	0.931	0.931	0.928

Table 5.23: Accuracy scores of the all emotion classes model with one dense layer

The results presented in Tables 5.23-5.26 reveal distinct performance characteristics when comprehensive emotional profiling is integrated into single dense layer architectures. RoBERTa achieves superior performance across most evaluation metrics, with accuracy scores reaching 0.941 at both 3e-5 and 5e-5 learning rates when combined with 0.3 and 0.2 dropout respectively. This represents a substantial enhancement in classification capability compared to baseline configurations. The model's precision performance peaks at 0.943 at 3e-5 learning rate with 0.3 dropout,

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.908	0.904	0.932
	3e-5	0.919	0.929	0.939
	5e-5	0.916	0.941	0.940
0.3	1e-5	0.925	0.919	0.934
	3e-5	0.926	0.943	0.932
	5e-5	0.935	0.931	0.936

Table 5.24: Precision scores of the all emotion classes model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.906	0.900	0.931
	3e-5	0.919	0.925	0.938
	5e-5	0.913	0.941	0.938
0.3	1e-5	0.925	0.919	0.934
	3e-5	0.925	0.941	0.931
	5e-5	0.931	0.931	0.928

Table 5.25: Recall scores of the all emotion classes model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.906	0.900	0.931
	3e-5	0.919	0.925	0.937
	5e-5	0.912	0.941	0.937
0.3	1e-5	0.925	0.919	0.934
	3e-5	0.925	0.941	0.931
	5e-5	0.931	0.931	0.928

Table 5.26: F-score values of the all emotion classes model with one dense layer

while recall measurements achieve their highest value of 0.941 at both 3e-5 and 5e-5 learning rates with 0.3 and 0.2 dropout respectively. F-score metrics corroborate these findings, with RoBERTa attaining 0.941 under same optimal configurations. BERT

produces moderate results across single dense layer implementations, with accuracy scores ranging from 0.906 to 0.931. The model achieves its peak accuracy of 0.931 at 5e-5 learning rate combined with 0.3 dropout. Precision values span from 0.908 to 0.935, reaching their maximum at 5e-5 learning rate with 0.3 dropout. Recall measurements demonstrate similar patterns, with values ranging from 0.906 to 0.931, while F-score results maintain comparable ranges across different hyperparameter combinations, peaking at 0.931. DistilBERT exhibits robust and stable performance characteristics throughout single dense layer configurations. Accuracy measurements range from 0.928 to 0.938, demonstrating consistent high-level performance across various hyperparameter settings. Precision scores span from 0.932 to 0.940, while recall values range from 0.928 to 0.938. F-score measurements fall within the 0.928 to 0.937 range, indicating well-balanced performance between precision and recall metrics across all configurations.

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.900	0.950
	3e-5	0.931	0.925	0.938
	5e-5	0.931	0.913	0.953
0.3	1e-5	0.925	0.894	0.938
	3e-5	0.919	0.919	0.947
	5e-5	0.919	0.925	0.950

Table 5.27: Accuracy scores of the all emotion classes model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.901	0.950
	3e-5	0.931	0.926	0.940
	5e-5	0.933	0.915	0.956
0.3	1e-5	0.929	0.899	0.938
	3e-5	0.921	0.919	0.949
	5e-5	0.922	0.926	0.950

Table 5.28: Precision scores of the all emotion classes model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.900	0.950
	3e-5	0.931	0.925	0.938
	5e-5	0.931	0.913	0.953
0.3	1e-5	0.925	0.894	0.938
	3e-5	0.919	0.919	0.947
	5e-5	0.919	0.925	0.950

Table 5.29: Recall scores of the all emotion classes model with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.900	0.950
	3e-5	0.931	0.925	0.937
	5e-5	0.931	0.913	0.953
0.3	1e-5	0.925	0.893	0.937
	3e-5	0.919	0.919	0.947
	5e-5	0.919	0.925	0.950

Table 5.30: F-score values of the all emotion classes model with multiple dense layers

The multiple dense layer implementation results, as documented in Tables 5.27-5.30, reveal interesting performance patterns. DistilBERT emerges as the best performer in this architectural configuration, achieving exceptional accuracy scores of up to 0.953 at 5e-5 learning rate with 0.2 dropout. The model's precision performance reaches 0.956 at the same configuration, representing the highest precision recorded across all experimental conditions in this methodology. Recall measurements peak at 0.953, while F-score values attain 0.953, establishing DistilBERT as the optimal choice for multiple dense layer architectures when comprehensive emotional profiling is employed. BERT demonstrates solid performance in multiple dense layer configurations, with accuracy scores ranging from 0.919 to 0.931. The model achieves its peak accuracy of 0.931 at both 3e-5 and 5e-5 learning rates with 0.2 dropout. Precision values span from 0.921 to 0.933, reaching optimal performance at 5e-5 learning rate with 0.2 dropout. Recall scores exhibit similar patterns, ranging from 0.919 to 0.931, while F-score measurements maintain comparable performance ranges, consistently achieving strong results across different hyperparameter configurations.

RoBERTa produces relatively modest results in multiple dense layer architectures when compared to its single layer performance and other models in this configuration. Accuracy measurements range from 0.894 to 0.925, with peak performance achieved at 3e-5 learning rate with 0.2 dropout and 5e-5 learning rate with 0.3 dropout. Precision scores span from 0.899 to 0.926, while recall values range from 0.894 to 0.925. F-score results maintain similar ranges, indicating that RoBERTa may not benefit as substantially from architectural complexity when processing comprehensive emotional feature representations.

Dropout rate effects reveal model-specific optimisation patterns when comprehensive emotional features are incorporated. RoBERTa shows clear preference for higher dropout rates in single dense layer configurations, achieving optimal performance across multiple metrics. This suggests that the additional emotional information requires moderate regularisation to prevent overfitting while maintaining generalisation capability. DistilBERT exhibits preference for lower dropout rates in multiple dense layer architectures, particularly when achieving peak performance levels. This indicates that the model benefits from reduced regularisation when processing both comprehensive emotional features and increased architectural complexity, suggesting efficient feature integration without excessive regularisation requirements.

Learning rate selection demonstrates consistent patterns across models and architectural configurations. The 5e-5 learning rate frequently produces optimal results for both RoBERTa and DistilBERT across different settings, suggesting that this higher learning rate facilitates effective optimisation when processing comprehensive emotional feature representations. BERT shows more balanced performance across different learning rates, indicating robust optimisation characteristics regardless of learning rate selection.

The comparison between single and multiple dense layer architectures reveals significant model-specific outcomes when comprehensive emotional features are incorporated. Single dense layer architectures enable RoBERTa to achieve peak performance, indicating that the comprehensive emotional information provides sufficient additional context without requiring increased architectural complexity. The model's superior performance in simpler architectures suggests efficient processing of the five-vector emotional representation. Multiple dense layer configurations strongly favour DistilBERT, which achieves its highest performance levels in these settings. This architectural preference suggests that DistilBERT benefits from the enhanced processing capacity when handling comprehensive emotional feature representations, effectively utilising the additional computational complexity to improve classification performance.

The integration of comprehensive emotional profiling yields measurable improvements across multiple model configurations when compared to baseline performance levels. The enhancement patterns vary noticeably across models, with each architec-

ture responding differently to the additional emotional context provided by the five-vector emotional representation. RoBERTa demonstrates notable improvements in single dense layer configurations, particularly achieving its highest recorded accuracy of 0.941 under optimal conditions. This represents a substantial enhancement from baseline performance, suggesting that RoBERTa effectively leverages the comprehensive emotional information when architectural complexity remains manageable. However, the model shows more modest improvements in multiple dense layer configurations, indicating potential challenges in processing both increased architectural complexity and comprehensive emotional features simultaneously. BERT exhibits consistent improvements across both architectural configurations, though the enhancements are more obvious in multiple dense layer implementations. The model benefits from the additional emotional context, achieving performance gains that suggest effective integration of textual and emotional feature representations. The improvements are particularly evident in precision metrics, where BERT achieves notable gains compared to baseline configurations. DistilBERT presents the most remarkable improvement pattern, particularly excelling in multiple dense layer architectures where it achieves peak performance across all metrics. This suggests that DistilBERT's architectural characteristics are particularly well-suited for processing comprehensive emotional profiling information, especially when combined with increased classifier complexity. The model's ability to maintain high performance while processing extensive emotional feature representations demonstrates effective feature integration capabilities.

The incorporation of comprehensive emotional profiling through five-vector representation produces substantial performance enhancements across the majority of experimental configurations. Statistical analysis reveals improvements in 29 out of 36 tested settings compared to base models, representing an 81% success rate for comprehensive emotion-enhanced performance. This substantial improvement rate validates the hypothesis that detailed emotional profiling provides valuable contextual information for fake review classification tasks.

5.4.2.4 Dominant Emotion Class with General Text Features

This experimental methodology represents a hybrid approach that combines the dominant emotion class information with general textual feature. The approach integrates emotional context through dominant emotion classification while simultaneously incorporating textual characteristics including the counts for the sentences, words, characters, uppercase words, and exclamation marks in the review. The comprehensive performance metrics obtained through this integrated methodology are documented in Tables 5.31-5.38, encompassing accuracy, precision, recall, and F-score measurements across single and multiple dense layer architectural configurations.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.903	0.941
	3e-5	0.913	0.931	0.938
	5e-5	0.919	0.938	0.950
0.3	1e-5	0.913	0.919	0.925
	3e-5	0.934	0.919	0.950
	5e-5	0.934	0.938	0.950

Table 5.31: Accuracy scores of the dominant emotion class with the general text features model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.929	0.903	0.942
	3e-5	0.914	0.932	0.938
	5e-5	0.919	0.938	0.951
0.3	1e-5	0.917	0.919	0.925
	3e-5	0.935	0.922	0.950
	5e-5	0.942	0.938	0.950

Table 5.32: Precision scores of the dominant emotion class with the general text features model with one dense layer

The results presented in Tables 5.31-5.34 reveal distinctive performance patterns when both emotional and textual features are incorporated into single dense layer architectures. DistilBERT appears as the best performer across most evaluation metrics, with accuracy scores reaching 0.950 at multiple configurations including 5e-5 learning rate with 0.2 dropout, and 3e-5 and 5e-5 learning rates with 0.3 dropout. This represents exceptional performance consistency across different hyperparameter settings. The model's precision performance achieves a peak of 0.951 at 5e-5 learning rate with 0.2 dropout, while recall measurements reach their highest value of 0.950 at multiple optimal configurations. F-score metrics corroborate these findings, with DistilBERT attaining 0.950 under several hyperparameter combinations, indicating robust performance across all evaluation configurations. RoBERTa produces variable results in single dense layer implementations, with accuracy scores ranging from 0.903 to 0.938. The model achieves its peak accuracy

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.903	0.941
	3e-5	0.913	0.931	0.938
	5e-5	0.919	0.938	0.950
0.3	1e-5	0.913	0.919	0.925
	3e-5	0.934	0.919	0.950
	5e-5	0.934	0.938	0.950

Table 5.33: Recall scores of the dominant emotion class with the general text features model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.925	0.903	0.941
	3e-5	0.912	0.931	0.937
	5e-5	0.919	0.937	0.950
0.3	1e-5	0.912	0.919	0.925
	3e-5	0.934	0.919	0.950
	5e-5	0.934	0.937	0.950

Table 5.34: F-score values of the dominant emotion class with the general text features model with one dense layer

of 0.938 at both 5e-5 learning rate with 0.2 dropout and 5e-5 learning rate with 0.3 dropout. Precision values span from 0.903 to 0.938, reaching their maximum at 5e-5 learning rate with both dropout configurations. Recall measurements demonstrate similar patterns, with values ranging from 0.903 to 0.938, while F-score results maintain comparable performance ranges, indicating consistent behaviour across different metrics. BERT demonstrates solid performance characteristics throughout single dense layer configurations. Accuracy measurements range from 0.913 to 0.934, with peak performance achieved at both 3e-5 and 5e-5 learning rates combined with 0.3 dropout. Precision scores span from 0.914 to 0.942, reaching their maximum at 5e-5 learning rate with 0.3 dropout. Recall values range from 0.913 to 0.934, while F-score measurements fall within the 0.912 to 0.934 range, demonstrating balanced performance between precision and recall metrics.

The multiple dense layer implementation results, as documented in Tables 5.35-

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.938	0.900	0.944
	3e-5	0.944	0.925	0.938
	5e-5	0.925	0.919	0.944
0.3	1e-5	0.925	0.894	0.947
	3e-5	0.931	0.925	0.944
	5e-5	0.938	0.928	0.938

Table 5.35: Accuracy scores of the dominant emotion class with the general text features with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.938	0.901	0.944
	3e-5	0.945	0.925	0.940
	5e-5	0.927	0.921	0.944
0.3	1e-5	0.926	0.895	0.949
	3e-5	0.935	0.926	0.944
	5e-5	0.939	0.929	0.939

Table 5.36: Precision scores of the dominant emotion class with the general text features with multiple dense layers

5.38, reveal interesting performance patterns. DistilBERT maintains its superior results, achieving accuracy scores of up to 0.947 at 1e-5 learning rate with 0.3 dropout. The model's precision performance reaches 0.949 at the same configuration, while recall measurements peak at 0.947. F-score values attain 0.947, establishing DistilBERT as the optimal choice for multiple dense layer architectures when hybrid feature integration is employed. BERT produces robust results in multiple dense layer configurations, with accuracy scores ranging from 0.925 to 0.944. The model achieves its peak accuracy of 0.944 at 3e-5 learning rate with 0.2 dropout. Precision values span from 0.926 to 0.945, reaching optimal performance at 3e-5 learning rate with 0.2 dropout. Recall scores exhibit similar patterns, ranging from 0.925 to 0.944, while F-score measurements maintain comparable performance ranges, consistently achieving strong results across different hyperparameter configurations. RoBERTa exhibits more modest performance in multiple dense layer architectures compared to its single

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.938	0.900	0.944
	3e-5	0.944	0.925	0.938
	5e-5	0.925	0.919	0.944
0.3	1e-5	0.925	0.894	0.947
	3e-5	0.931	0.925	0.944
	5e-5	0.938	0.928	0.938

Table 5.37: Recall scores of the dominant emotion class with the general text features with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.938	0.900	0.944
	3e-5	0.944	0.925	0.937
	5e-5	0.925	0.919	0.944
0.3	1e-5	0.925	0.894	0.947
	3e-5	0.931	0.925	0.944
	5e-5	0.937	0.928	0.937

Table 5.38: F-score values of the dominant emotion class with the general text features with multiple dense layers

layer implementations. Accuracy measurements range from 0.894 to 0.928, with peak performance achieved at 5e-5 learning rate with 0.3 dropout. Precision scores span from 0.895 to 0.929, while recall values range from 0.894 to 0.928. F-score results maintain similar ranges, indicating that RoBERTa may face challenges in processing both architectural complexity and hybrid feature representations simultaneously.

The experimental results reveal notable hyperparameter sensitivity, particularly for RoBERTa under specific configurations. A remarkable performance anomaly is observed for RoBERTa at 1e-5 learning rate with 0.3 dropout, where accuracy drops substantially to levels remarkably below expected performance ranges. This anomalous behaviour highlights the critical importance of comprehensive hyperparameter optimisation when implementing hybrid feature integration approaches. The considerable performance variations observed under minor hyperparameter adjustments demonstrate the complex interaction effects between learning rate,

dropout regularisation, and feature complexity in language model architectures.

Dropout rate effects demonstrate model-specific response patterns throughout the experimental conditions. DistilBERT shows robust performance across both dropout configurations, indicating resilience to regularisation variations when processing hybrid feature representations. RoBERTa's response to dropout variations proves more complex, with optimal performance achieved at different dropout levels depending on learning rate selection.

The comparison between single and multiple dense layer architectures reveals distinct optimisation patterns when hybrid features are incorporated. Single dense layer architectures enable DistilBERT to achieve highest performance levels, indicating that the combined emotional and textual features provide sufficient representational complexity without requiring increased architectural depth. The model's superior performance in simpler architectures suggests efficient processing of hybrid feature representations. Multiple dense layer configurations provide benefits for BERT, which achieves its highest performance levels in these settings. This architectural preference indicates that BERT benefits from enhanced processing capacity when handling combined feature representations, effectively utilising additional computational complexity to improve classification accuracy.

The integration of dominant emotion class with general textual features yields substantial improvements across multiple model configurations when compared to baseline performance levels. Statistical analysis reveals performance enhancements in 29 out of 36 tested configurations, representing an 81% success rate for hybrid feature integration. This validates the effectiveness of combining emotional and textual feature representations.

RoBERTa demonstrates notable improvements in specific configurations, particularly achieving accuracy enhancements from baseline levels of 0.900 to 0.925 at 3e-5 learning rate with 0.2 dropout in multiple dense layer settings. This represents a meaningful performance gain that suggests effective integration of hybrid feature representations under appropriate hyperparameter conditions. However, the model's sensitivity to hyperparameter selection remains a consideration. BERT exhibits consistent improvements across both architectural configurations. The model benefits substantially from the hybrid feature integration, achieving performance enhancements that suggest effective processing of combined emotional and textual representations. The improvements are particularly pronounced in multiple dense layer implementations. DistilBERT presents remarkable improvement patterns, achieving peak performance levels that establish it as the optimal model for hybrid feature processing. The model's accuracy enhancements from baseline levels demonstrate exceptional capability in integrating and processing complex feature representations while maintaining stable performance across different hyperparameter configurations.

The experimental findings provide strong support for hybrid feature integration as an effective enhancement strategy for fake review detection systems. The methodology's success in combining emotional and textual feature representations contributes meaningfully to improved classification performance deception detection applications.

5.4.2.5 All Emotion Classes with Class Textual Features

This methodology extends the all emotion classes architecture by incorporating class-specific textual features for each emotion category. The approach combines all emotion classes information with the five textual features mentioned earlier which computed separately for each emotion class. The comprehensive performance metrics obtained through this methodology are documented in Tables 5.39-5.46, encompassing accuracy, precision, recall, and F-score measurements across single and multiple dense layer architectural configurations.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.922	0.944
	3e-5	0.922	0.925	0.956
	5e-5	0.938	0.913	0.938
0.3	1e-5	0.919	0.919	0.950
	3e-5	0.919	0.944	0.956
	5e-5	0.919	0.938	0.938

Table 5.39: Accuracy scores of the all emotion classes with the class textual features model with one dense layer

The results presented in Tables 5.39-5.42 reveal distinct performance patterns when class-specific textual features are incorporated into single dense layer architectures. DistilBERT achieves superior performance across most evaluation metrics, with accuracy scores reaching 0.956 at both 3e-5 learning rate with 0.2 dropout and same learning rate with 0.3 dropout. The model's precision performance peaks at 0.957 at 3e-5 learning rate with both 0.2 and 0.3 dropout, while recall measurements reach their highest value of 0.956 at same optimal configurations. F-score metrics corroborate these findings, with DistilBERT attaining 0.956 under several hyperparameter combinations, indicating robust performance across all evaluation metrics. RoBERTa produces variable results in single dense layer implementations, with accuracy scores ranging from 0.913 to 0.944. The model achieves its peak accuracy of 0.944 at

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.937	0.922	0.945
	3e-5	0.922	0.925	0.957
	5e-5	0.938	0.914	0.938
0.3	1e-5	0.922	0.919	0.950
	3e-5	0.919	0.944	0.957
	5e-5	0.919	0.938	0.938

Table 5.40: Precision scores of the all emotion classes with the class textual features model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.922	0.944
	3e-5	0.922	0.925	0.956
	5e-5	0.938	0.913	0.938
0.3	1e-5	0.919	0.919	0.950
	3e-5	0.919	0.944	0.956
	5e-5	0.919	0.938	0.938

Table 5.41: Recall scores of the all emotion classes with the class textual features model with one dense layer

3e-5 learning rate with 0.3 dropout. Precision values span from 0.914 to 0.944, reaching their maximum at 3e-5 learning rate with 0.3 dropout. Recall measurements demonstrate similar patterns, with values ranging from 0.913 to 0.944, while F-score results maintain comparable performance ranges, indicating balanced behaviour across different metrics. BERT demonstrates stable performance characteristics throughout single dense layer configurations. Accuracy measurements range from 0.919 to 0.938, with peak performance achieved at 5e-5 learning rate combined with 0.2 dropout. Precision scores span from 0.919 to 0.938, reaching their maximum at 5e-5 learning rate with 0.2 dropout. Recall values range from 0.919 to 0.938, while F-score measurements fall within the 0.919 to 0.937 range, demonstrating consistent performance between precision and recall metrics.

The multiple dense layer implementation results, as documented in Tables 5.43-5.46, reveal shifts in model performance hierarchies. DistilBERT maintains its

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.931	0.922	0.944
	3e-5	0.922	0.925	0.956
	5e-5	0.937	0.912	0.937
0.3	1e-5	0.919	0.919	0.950
	3e-5	0.919	0.944	0.956
	5e-5	0.919	0.937	0.937

Table 5.42: F-score values of the all emotion classes with the class textual features model with one dense layer

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.909	0.913	0.944
	3e-5	0.931	0.913	0.938
	5e-5	0.913	0.944	0.925
0.3	1e-5	0.913	0.906	0.944
	3e-5	0.919	0.906	0.944
	5e-5	0.928	0.931	0.944

Table 5.43: Accuracy scores of the all emotion classes with the class textual features with multiple dense layers

high values, achieving accuracy scores of up to 0.944 at multiple learning rate and dropout configurations. The model's precision performance reaches 0.944, while recall measurements peak at 0.944. F-score values attain 0.944, establishing DistilBERT as a reliable choice for multiple dense layer architectures when class-specific textual features are employed. BERT produces mixed results in multiple dense layer configurations compared to single layer implementations, with accuracy scores ranging from 0.909 to 0.931. The model achieves its peak accuracy of 0.931 at 3e-5 learning rate with 0.2 dropout. Precision values span from 0.910 to 0.933, reaching optimal performance at 3e-5 learning rate with 0.2 dropout. Recall scores exhibit similar patterns, ranging from 0.909 to 0.931, while F-score measurements maintain comparable performance ranges. RoBERTa exhibits more variable performance in multiple dense layer architectures compared to its single layer implementations. Accuracy measurements range from 0.906 to 0.944, with peak performance achieved

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.910	0.914	0.944
	3e-5	0.933	0.915	0.943
	5e-5	0.913	0.945	0.927
0.3	1e-5	0.913	0.907	0.944
	3e-5	0.919	0.908	0.944
	5e-5	0.931	0.932	0.944

Table 5.44: Precision scores of the all emotion classes with the class textual features with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.909	0.913	0.944
	3e-5	0.931	0.913	0.938
	5e-5	0.913	0.944	0.925
0.3	1e-5	0.913	0.906	0.944
	3e-5	0.919	0.906	0.944
	5e-5	0.928	0.931	0.944

Table 5.45: Recall scores of the all emotion classes with the class textual features with multiple dense layers

at 5e-5 learning rate with 0.2 dropout. Precision scores span from 0.907 to 0.945, while recall values range from 0.906 to 0.944. F-score results maintain similar ranges, indicating that RoBERTa's performance with class-specific features requires careful hyperparameter optimisation in complex architectures.

The experimental results reveal notable hyperparameter sensitivity patterns across different model architectures. Learning rate effects prove substantial and consistent across models. For single dense layer architectures, 3e-5 learning rate frequently produce optimal results across all models, while 5e-5 generally yields lower performance levels. The mid learning rate appears necessary for effectively processing the increased feature complexity introduced by class-specific textual information.

Dropout rate effects demonstrate model-specific response patterns. DistilBERT shows robust performance across both dropout configurations in single layer architectures, with slight preference for 0.3 dropout in several configurations. BERT

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.909	0.912	0.944
	3e-5	0.931	0.912	0.937
	5e-5	0.912	0.944	0.925
0.3	1e-5	0.912	0.906	0.944
	3e-5	0.919	0.906	0.944
	5e-5	0.928	0.931	0.944

Table 5.46: F-score values of the all emotion classes with the class textual features with multiple dense layers

exhibits preference for 0.2 dropout in most scenarios, particularly in single dense layer implementations. RoBERTa shows mixed responses to dropout variations, with optimal performance achieved at different dropout levels depending on learning rate selection and architectural depth.

The comparison between single and multiple dense layer architectures reveals distinct patterns when class-specific textual features are integrated. Single dense layer architectures enable DistilBERT to achieve peak performance levels, indicating that the class-specific features provide sufficient representational complexity without requiring increased architectural depth. The model's superior performance in simpler architectures suggests efficient processing of the enhanced feature. However, the performance differential between architectural configurations varies across models. RoBERTa, for example, maintains variable performance across both architectural variants, with optimal results dependent on specific hyperparameter selections and the interaction between architecture depth and regularisation.

The integration of class-specific textual features with all emotion classes information yields measurable improvements across multiple model configurations when compared to baseline models. The enhancements are most pronounced at learning rates 3e-5. DistilBERT demonstrates notable improvements, particularly achieving accuracy enhancements from baseline levels around 0.919 to 0.956 in optimal single dense layer settings. This represents a meaningful performance gain of approximately 3.7 percentage points that suggests effective integration of class-specific textual features with all emotion classes information. The improvements are consistent across precision, recall, and F-score metrics, indicating genuine enhancement in classification capability. BERT exhibits improvements across both architectural configurations, with accuracy gains ranging from baseline levels around 0.922 to 0.938 in optimal conditions. The model benefits from these features integration, achieving performance

enhancements that suggest effective processing of combined emotional and textual representations.

The integration of class-specific textual features with all emotion classes information yields substantial improvements across multiple model configurations when compared to baseline performance levels. Statistical analysis reveals performance enhancements in 29 out of 36 tested configurations, representing an 81% success rate for this advanced feature integration approach. The experimental findings provide support for combining detailed emotional profiling with class-specific textual features as a valuable enhancement strategy for fake review detection systems. The methodology's success in combining granular emotional analysis with detailed linguistic characteristics contributes to improved classification performance, suggesting this approach as a viable advancement in emotion-based deception detection applications.

5.4.2.6 All Emotion Classes with General Text Features

This methodology extends the all emotion classes architecture by incorporating general textual features computed across the entire review. The approach combines the five-vector emotion representation with five general textual features mentioned previously that calculated at the review level. The comprehensive performance metrics obtained through this methodology are displayed in Tables 5.47-5.54, encompassing accuracy, precision, recall, and F-score measurements across single and multiple dense layer architectural configurations.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.925	0.931
	3e-5	0.931	0.931	0.928
	5e-5	0.919	0.938	0.931
0.3	1e-5	0.913	0.913	0.928
	3e-5	0.928	0.916	0.913
	5e-5	0.925	0.919	0.938

Table 5.47: Accuracy scores of the all emotion classes with the general text features model with one dense layer

The results presented in Tables 5.47-5.50 reveal distinct performance patterns when general textual features are incorporated alongside all emotion classes information in single dense layer architectures. RoBERTa achieves variable performance across different hyperparameter combinations, with accuracy scores ranging from

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.926	0.933
	3e-5	0.934	0.931	0.928
	5e-5	0.922	0.939	0.932
0.3	1e-5	0.913	0.913	0.928
	3e-5	0.928	0.916	0.913
	5e-5	0.927	0.919	0.938

Table 5.48: Precision scores of the all emotion classes with the general text features model with one dense layer

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.925	0.931
	3e-5	0.931	0.931	0.928
	5e-5	0.919	0.938	0.931
0.3	1e-5	0.913	0.913	0.928
	3e-5	0.928	0.916	0.913
	5e-5	0.925	0.919	0.938

Table 5.49: Recall scores of the all emotion classes with the general text features model with one dense layer

0.913 to 0.938. The model reaches its peak accuracy of 0.938 at 5e-5 learning rate with 0.2 dropout. Precision performance peaks at 0.939 at 5e-5 learning rate with 0.2 dropout, while recall measurements achieve their highest value of 0.938 at the same configuration. F-score metrics corroborate these findings, with RoBERTa attaining 0.937 under the same conditions. BERT demonstrates stable performance characteristics throughout single dense layer configurations. Accuracy measurements range from 0.913 to 0.931, with peak performance achieved at 3e-5 learning rate combined with 0.2 dropout rate. Precision scores span from 0.913 to 0.934, reaching their maximum at 3e-5 learning rate with 0.2 dropout. Recall values range from 0.913 to 0.931, while F-score measurements fall within the 0.912 to 0.931 range, demonstrating balanced performance between precision and recall metrics. DistilBERT produces the most stable results across single dense layer implementations, with accuracy scores ranging from 0.913 to 0.938.

Dropout	**Learning Rate**	**One Dense Layer**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.912	0.925	0.931
	3e-5	0.931	0.931	0.928
	5e-5	0.919	0.937	0.931
0.3	1e-5	0.912	0.912	0.928
	3e-5	0.928	0.916	0.912
	5e-5	0.925	0.919	0.938

Table 5.50: F-score values of the all emotion classes with the general text features model with one dense layer

The model achieves its peak accuracy of 0.938 at 5e-5 learning rate with 0.3 dropout. Precision values span from 0.913 to 0.938, while recall measurements demonstrate similar patterns, ranging from 0.913 to 0.938. F-score results maintain comparable performance ranges, indicating consistent behaviour across different evaluation metrics.

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.906	0.950
	3e-5	0.925	0.938	0.944
	5e-5	0.928	0.931	0.944
0.3	1e-5	0.913	0.913	0.925
	3e-5	0.938	0.928	0.938
	5e-5	0.925	0.953	0.931

Table 5.51: Accuracy scores of the all emotion classes with the general text features with multiple dense layers

The multiple dense layer implementation results, as documented in Tables 5.51-5.54, reveal some variations in model performance hierarchies compared to single layer configurations. RoBERTa exhibits notable performance variability in multiple dense layer architectures, with accuracy scores ranging from 0.906 to 0.953. The model achieves its peak accuracy of 0.953 at 5e-5 learning rate with 0.3 dropout, representing the highest performance recorded across all configurations for this model. Precision performance reaches 0.953 at the same configuration, while recall measurements

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.915	0.910	0.950
	3e-5	0.929	0.938	0.944
	5e-5	0.933	0.932	0.944
0.3	1e-5	0.924	0.919	0.926
	3e-5	0.938	0.934	0.940
	5e-5	0.925	0.953	0.936

Table 5.52: Precision scores of the all emotion classes with the general text features with multiple dense layers

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.913	0.906	0.950
	3e-5	0.925	0.938	0.944
	5e-5	0.928	0.931	0.944
0.3	1e-5	0.913	0.913	0.925
	3e-5	0.938	0.928	0.938
	5e-5	0.925	0.953	0.931

Table 5.53: Recall scores of the all emotion classes with the general text features with multiple dense layers

peak at 0.953. F-score values attain 0.953, establishing this as RoBERTa's optimal configuration for multiple dense layer architectures. BERT produces improved and more stable results in multiple dense layer configurations compared to single layer implementations, with accuracy scores ranging from 0.913 to 0.938. The model achieves its peak accuracy of 0.938 at 3e-5 learning rate with 0.3 dropout. Precision values span from 0.915 to 0.938, reaching optimal performance at 3e-5 learning rate with 0.3 dropout. Recall scores exhibit similar patterns, ranging from 0.913 to 0.938, while F-score measurements maintain comparable performance ranges. DistilBERT maintains strong performance in multiple dense layer architectures, achieving accuracy scores ranging from 0.925 to 0.950. The model reaches its peak accuracy of 0.950 at 1e-5 learning rate with 0.2 dropout. Precision performance peaks at 0.950 at the same configuration, while recall measurements achieve their highest value of 0.950. F-score metrics align similarly with these results, indicating balanced performance across all

Dropout	**Learning Rate**	**Multiple Dense Layers**		
		BERT	**RoBERTa**	**DistilBERT**
0.2	1e-5	0.912	0.906	0.950
	3e-5	0.925	0.937	0.944
	5e-5	0.928	0.931	0.944
0.3	1e-5	0.912	0.912	0.925
	3e-5	0.938	0.928	0.937
	5e-5	0.925	0.953	0.931

Table 5.54: F-score values of the all emotion classes with the general text features with multiple dense layers

evaluation metrics.

The experimental results reveal complex hyperparameter sensitivity patterns that vary considerably across models and architectural configurations. Learning rate effects prove particularly important for optimal performance. RoBERTa demonstrates strong performance at 5e-5 learning rate in both single and multiple layer configurations, though the optimal dropout rate varies between architectures. The model shows sensitivity to the interaction between learning rate and architectural complexity, particularly evident in the variable performance across different configurations.

Dropout rate effects demonstrate model-specific response patterns that interact with architectural depth. RoBERTa shows preference for 0.2 dropout in single layer configurations but achieves peak performance with 0.3 dropout in multiple dense layer settings. BERT exhibits more stable performance across different dropout rates, with slight preferences varying by learning rate selection. DistilBERT demonstrates robust performance across both dropout configurations, indicating good regularisation characteristics when processing combined emotion and general textual features.

The comparison between single and multiple dense layer architectures reveals distinct patterns when general textual features are combined with emotion information. Single dense layer architectures provide stable performance across all models, with RoBERTa and DistilBERT achieving comparable peak performance levels around 0.938. The simpler architecture appears sufficient for processing the combined feature representation effectively. Multiple dense layer configurations enable exceptional peak performance, particularly for RoBERTa, which achieves 0.953 accuracy under optimal conditions. This represents the highest performance recorded across all experimental configurations, suggesting that the additional processing capacity helps leverage complex interactions between emotion and textual features. BERT shows consistent improvement in multiple layer settings, while DistilBERT achieves strong

performance of 0.950.

The integration of general textual features with all emotion classes information yields measurable improvements across multiple model configurations when compared to baseline performance levels without emotion information. The enhancements are most pronounced in configurations where hyperparameters are optimally tuned, demonstrating the value of combining emotional and textual feature representations. RoBERTa demonstrates notable improvements, particularly achieving accuracy enhancements in multiple layer configurations where peak performance of 0.953 represents substantial improvement over baseline levels. This indicates that RoBERTa effectively utilises the combined feature representation when architectural complexity and hyperparameters are appropriately configured. BERT exhibits consistent improvements across both architectural configurations, with accuracy gains particularly evident in multiple layer implementations. The model benefits from the integration of general textual features with emotion information, achieving performance enhancements that suggest effective processing of the combined features. DistilBERT shows solid improvements, achieving peak performance of 0.950 in multiple layer configurations. The model demonstrates good stability across different hyperparameter combinations while benefiting from the enhanced feature representation. The consistent performance suggests efficient processing of combined emotional and textual information.

The integration of general textual features with all emotion classes information produces substantial improvements across multiple model configurations when compared to baseline performance levels. Statistical analysis reveals performance enhancements in 30 out of 36 tested configurations, representing an 83% success rate for this feature integration approach. The experimental findings provide strong support for the integration of general textual features with all emotion information as an effective enhancement strategy for fake review detection systems.

5.4.3 Results Overview

The summary Figures 5.4-5.7 present a comprehensive comparative analysis of the three fine-tuned language models across various methodologies for integrating emotion information into fake review classification models. The models are evaluated based on their performance across six distinct architectural approaches: (1) base models without emotion information, (2) dominant emotion class integration, (3) all emotion classes information, (4) dominant emotion class with general text features, (5) all emotion classes with all class textual features, and (6) all emotion classes with general text features. The y-axis measures the performance of the models based on their accuracy score, while the x-axis categorises the models and their respective learning rates, which take these values: 1e-5, 3e-5 and 5e-5.

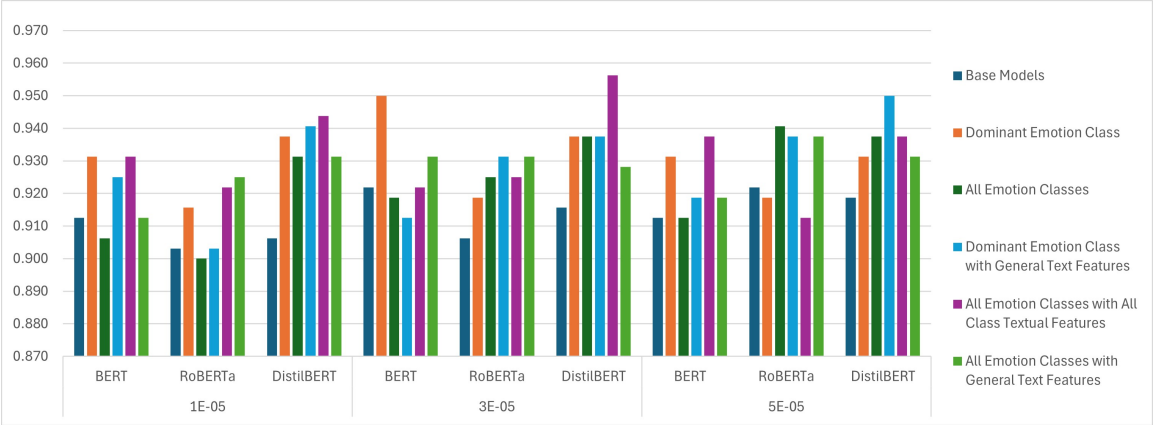


Figure 5.4: Accuracy scores of all models with one dense layer and a dropout of 0.2

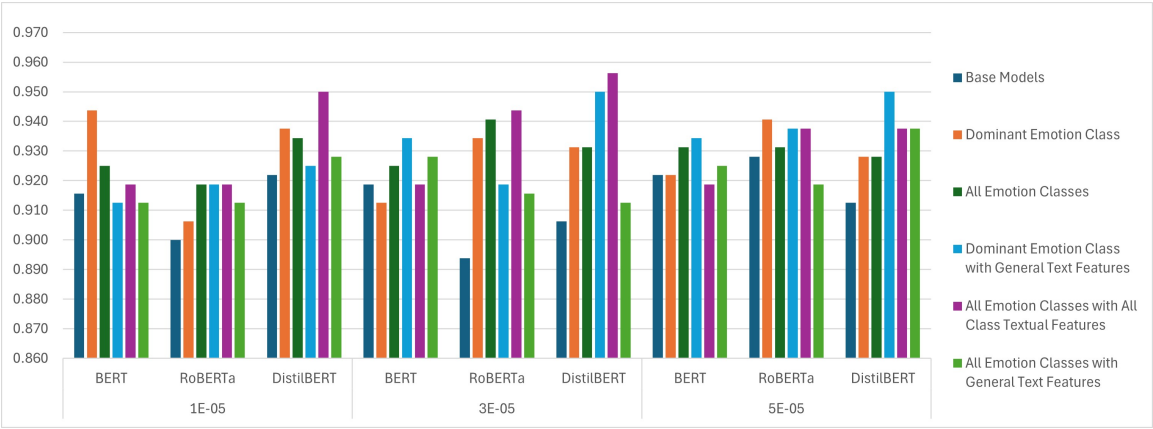


Figure 5.5: Accuracy scores of all models with one dense layer and a dropout of 0.3

Figure 5.4 demonstrates the accuracy results obtained from the experiments with the one dense layer architecture and a dropout value of 0.2, while Figure 5.6 shows the results for multiple dense layers with the same dropout rate. The experiments with a dropout value of 0.3 are presented in Figure 5.5 for the one dense layer architecture and Figure 5.7 for the multiple dense layer architecture.

The comprehensive analysis reveals several key patterns regarding the effectiveness of different emotion integration methodologies. The dominant emotion class approach provides consistent baseline improvements across all models and configurations, demonstrating the fundamental value of incorporating primary emotional context. The all emotion classes methodology offers enhanced performance through compre-

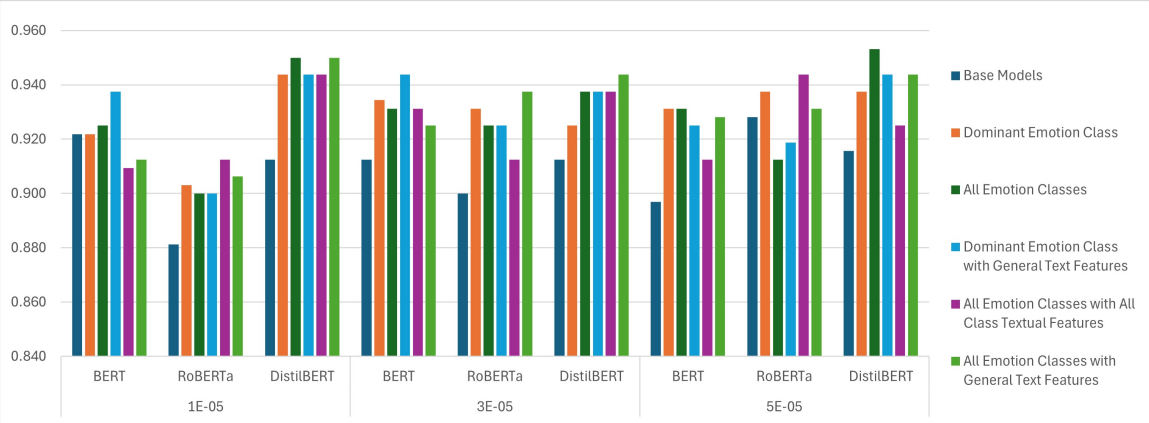


Figure 5.6: Accuracy scores of all models with multiple dense layers and a dropout of 0.2

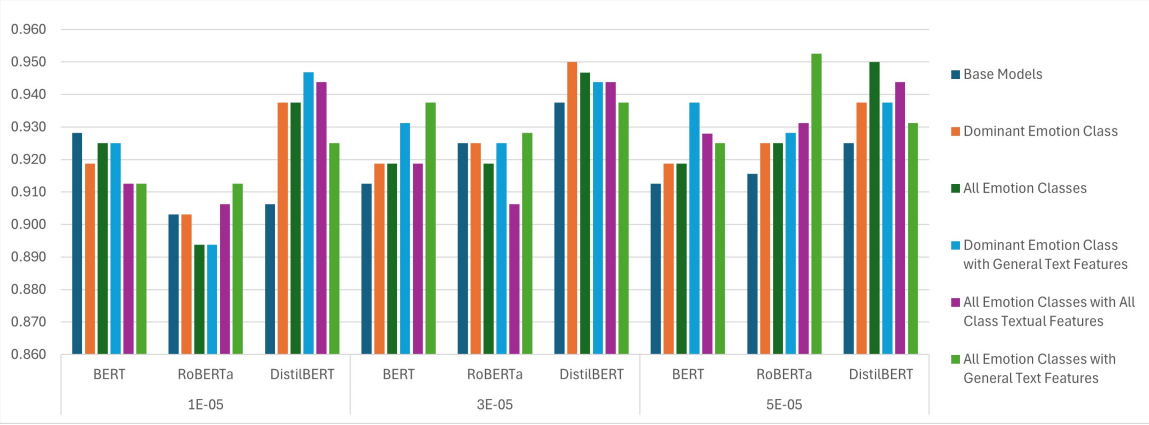


Figure 5.7: Accuracy scores of all models with multiple dense layers and a dropout of 0.3

hensive emotional profiling, with benefits varying by model architecture and hyperparameter configuration. The integration of textual features with emotion information produces the most substantial performance improvements. The dominant emotion class with general text features approach demonstrates particular effectiveness for BERT across multiple configurations, while the all emotion classes with class-specific textual features methodology shows exceptional performance for DistilBERT. The all emotion classes with general text features approach enables RoBERTa to achieve its peak performance.

The integration of emotion information significantly enhances fake review de-

tection model performance across all configurations. The degree of improvement varies by model, methodology, and configuration, with emotion-enhanced variants consistently outperforming base models. The experimental findings demonstrate that enriching language models with diverse emotional and linguistic features represents a valuable approach in fake review detection models, with specific combinations achieving substantial accuracy improvements of up to 4-5 percentage points over baseline models.

5.5 Comparison of Models' Performance

The experiments conducted in this study aimed to evaluate the effectiveness of incorporating emotion information into fake review detection models. The results obtained from employing the various configurations and model structures provide insightful conclusions about the impact of emotion information on fake review detection accuracy. The accuracy scores obtained from the experiments show that the models that incorporated emotion information consistently outperformed the base models. This section will explore the specific results obtained from the experiments, outlining what each configuration achieved.

5.5.1 One Dense Layer Architecture

Table 5.55 and Table 5.56 summarise the highest accuracy scores achieved with one dense layer across different models and configurations with a dropout value of 0.2 and 0.3, respectively. They highlight that incorporating emotion information significantly enhances model performance, particularly for the DistilBERT model, which achieved the highest accuracy level of 0.956 using all emotion classes information with class textual features combined with a learning rate of 3e-5 and a dropout rate of both 0.2 and 0.3.

In detail, for a dropout rate of 0.2 shown in Table 5.55, at a learning rate of 1e-5, BERT achieved its highest accuracy of 0.931 using the dominant emotion class and all emotion classes with class textual features methods. Under the same conditions, RoBERTa reached 0.925 accuracy when enhanced with all emotion classes and general text features, while DistilBERT peaked at 0.944 by employing all emotion classes with class textual features. When the learning rate increased to 3e-5, BERT's top score rose to 0.950 using the dominant emotion class method. RoBERTa recorded 0.931 accuracy via both dominant emotion class with general text features and all emotion classes with general text features approaches. In that same setting, DistilBERT achieved its maximum 0.956 accuracy when leveraging all emotion classes with class textual features. At a learning rate of 5e-5, BERT's accuracy reached 0.938 with all emotion classes with class textual features, RoBERTa attained 0.941 using all emotion classes,

Learning Rate	**LLM**	**Highest Accuracy Score**	**Method**
1e-5	BERT	0.931	Dominant Emotion Class - All Emotion Classes with Class Textual Features
	RoBERTa	0.925	All Emotion Classes with General Text Features
	DistilBERT	0.944	All Emotion Classes with Class Textual Features
3e-5	BERT	0.950	Dominant Emotion Class
	RoBERTa	0.931	Dominant Emotion Class with General Text Features - All Emotion Classes with General Text Features
	DistilBERT	0.956	All Emotion Classes with Class Textual Features
5e-5	BERT	0.938	All Emotion Classes with Class Textual Features
	RoBERTa	0.941	All Emotion Classes
	DistilBERT	0.950	Dominant Emotion Class with General Text Features

Table 5.55: Highest accuracy scores for one dense layer with a dropout of 0.2

and DistilBERT scored 0.950 under the dominant emotion class with general text features method.

On the other hand, for a dropout rate of 0.3 shown in Table 5.56, at a learning rate of 1e-5, BERT attains 0.944 accuracy with the dominant emotion class method, RoBERTa reaches 0.919 via both of all emotion classes, dominant emotion class with general text features, and all emotion classes with class textual features, and DistilBERT scores 0.950 when using all emotion classes with class textual features. When the learning rate increases to 3e-5, BERT achieves 0.934 accuracy with the dominant emotion class with general text features approach, RoBERTa scores 0.944 using all emotion classes with class textual features, and DistilBERT records 0.956 under the same method. At 5e-5, BERT again records 0.934 accuracy via dominant emotion class with general text features, RoBERTa attains 0.941 using the dominant emotion class method, and DistilBERT reaches 0.950 by employing dominant emotion class with general text features.

Overall, DistilBERT consistently demonstrated superior performance across all of the learning rates and methods. However, the specific methods used indicate that capturing a comprehensive emotional information with textual features of the reviews was crucial for achieving high accuracy.

Learning Rate	**LLM**	**Highest Accuracy Score**	**Method**
1e-5	BERT	0.944	Dominant Emotion Class
	RoBERTa	0.919	All Emotion Classes - Dominant Emotion Class with General Text Features - All Emotion Classes with Class Textual Features
	DistilBERT	0.950	All Emotion Classes with Class Textual Features
3e-5	BERT	0.934	Dominant Emotion Class with General Text Features
	RoBERTa	0.944	All Emotion Classes with Class Textual Features
	DistilBERT	0.956	All Emotion Classes with Class Textual Features
5e-5	BERT	0.934	Dominant Emotion Class with General Text Features
	RoBERTa	0.941	Dominant Emotion Class
	DistilBERT	0.950	Dominant Emotion Class with General Text Features

Table 5.56: Highest accuracy scores for one dense layer with a dropout of 0.3

5.5.2 Multiple Dense Layers Architecture

Table 5.57 and Table 5.58 show the highest accuracy scores for the models with multiple dense layers. Similar to the one dense layer results, incorporating emotion information noticeably improved the models' performance. The DistilBERT and RoBERTa models achieved the highest accuracy of 0.953 in two distinct settings. For DistilBERT it was when using the all emotion classes information combined with a learning rate of 5e-5 and a dropout rate of 0.2. While for RoBERTa it was when using the all emotion classes with general text features combined with a learning rate of 5e-5 and a dropout rate of 0.3.

In detail, for a dropout rate of 0.2 shown in Table 5.57, at a learning rate of 1e-5, BERT achieved an accuracy of 0.938 using the dominant emotion class with general text features method. RoBERTa reached 0.913 via all emotion classes with class textual features, while DistilBERT attained 0.950 using both all emotion classes and all emotion classes with general text features methods. When the learning rate increased to 3e-5, BERT's accuracy rose to 0.944 utilising the dominant emotion class with general text features approach. RoBERTa improved to 0.938 accuracy with all emotion classes with general text features, and DistilBERT maintained strong performance at 0.944 using the same method. At a learning rate of 5e-5, BERT achieved 0.931 accuracy employing the dominant emotion class and the all emotion

Learning Rate	**LLM**	**Highest Accuracy Score**	**Method**
1e-5	BERT	0.938	Dominant Emotion Class with General Text Features
	RoBERTa	0.913	All Emotion Classes with Class Textual Features
	DistilBERT	0.950	All Emotion Classes - All Emotion Classes with General Text Features
3e-5	BERT	0.944	Dominant Emotion Class with General Text Features
	RoBERTa	0.938	All Emotion Classes with General Text Features
	DistilBERT	0.944	All Emotion Classes with General Text Features
5e-5	BERT	0.931	Dominant Emotion Class - All Emotion Classes
	RoBERTa	0.944	All Emotion Classes with Class Textual Features
	DistilBERT	0.953	All Emotion Classes

Table 5.57: Highest accuracy scores for multiple dense layers with a dropout of 0.2

classes method. RoBERTa produced 0.944 accuracy via all emotion classes with class textual features, while DistilBERT recorded the highest score of 0.953 using the all emotion classes method.

Learning Rate	**LLM**	**Highest Accuracy Score**	**Method**
1e-5	BERT	0.928	Base Model
	RoBERTa	0.913	All Emotion Classes with General Text Features
	DistilBERT	0.947	Dominant Emotion Class with General Text Features
3e-5	BERT	0.938	All Emotion Classes with General Text Features
	RoBERTa	0.928	All Emotion Classes with General Text Features
	DistilBERT	0.950	Dominant Emotion Class
5e-5	BERT	0.938	Dominant Emotion Class with General Text Features
	RoBERTa	0.953	All Emotion Classes with General Text Features
	DistilBERT	0.950	All Emotion Classes

Table 5.58: Highest accuracy scores for multiple dense layers with a dropout of 0.3

In contrast, for a dropout rate of 0.3 shown in Table 5.58, at a learning rate of 1e-5, BERT achieved an accuracy of 0.928 using the base model approach. RoBERTa scored 0.913 when classifying fake reviews with all emotion classes with general text features, while DistilBERT reached 0.947 using the dominant emotion class with general text features technique. At a learning rate of 3e-5, BERT's accuracy increased to 0.938 using All emotion classes with general text features. RoBERTa performed at 0.928 accuracy with the same all emotion classes with general text features method, while DistilBERT reached 0.950 utilising the dominant emotion class approach. Furthermore, at a learning rate of 5e-5, BERT maintained 0.938 accuracy using the dominant emotion class with general text features method. RoBERTa demonstrated the highest accuracy, scoring 0.953 with all emotion classes with general text features, while DistilBERT scored 0.950 using the all emotion classes method.

Overall, DistilBERT achieved superior accuracy across all of the learning rates and methods. The results indicate that all of the emotion classes combined with general textual features is particularly effective in enhancing model performance.

5.5.3 Highest Accuracy Score for Each Method

The highest accuracy scores across different configurations indicate that models incorporating emotion information generally outperform the base model. As shown in Table 5.59, DistilBERT achieved the highest overall accuracy score of 0.956 when using all emotion classes with class textual features, while RoBERTa reached 0.953 with all emotion classes with general text features. DistilBERT consistently demonstrated superior performance, achieving 0.953 with all emotion classes method, 0.950 with both dominant emotion class, tied with BERT, and dominant emotion class with general text features methods.

Notably, the peak accuracy of 0.956 achieved by the DistilBERT model with all emotion classes with class textual features represents a significant advancement over existing approaches in fake review detection. This performance surpasses all previously reported results in the literature review section, particularly in Section 2.5.4, establishing a new benchmark for fake review classification models. The achievement of 0.956 accuracy demonstrates that the integration of comprehensive emotional profiling with class-specific textual features can push the boundaries of detection performance beyond what other methods have accomplished, validating the theoretical importance of emotional context in distinguishing authentic from deceptive review content.

These findings highlight that DistilBERT's architecture is particularly well-suited to leveraging emotion information, while RoBERTa shows exceptional performance when combining comprehensive emotional profiling with general textual features. The results underscore the importance of emotion information in enhancing fake review

Method	LLM	Accuracy Score
Base Model	DistilBERT	0.938
Dominant Emotion Class	BERT - DistilBERT	0.950
All Emotion Classes	DistilBERT	0.953
Dominant Emotion Class with General Text Features	DistilBERT	0.950
All Emotion Classes with Class Textual Features	DistilBERT	0.956
All Emotion Classes with General Text Features	RoBERTa	0.953

Table 5.59: Summary of the highest accuracy scores

detection, with the most sophisticated feature combinations yielding the highest accuracy scores.

5.5.4 Analysis of Selected Model Configurations

The confusion matrices, as illustrated in Figure 5.8, provide detailed insights into the classification performance of sample of each model configuration, revealing distinct patterns in their ability to distinguish between deceptive and truthful reviews. The DistilBERT base model with multiple dense layers, 3e-5 learning rate, 0.3 dropout establishes the foundational performance reference with 151 correctly classified deceptive reviews and 149 correctly classified truthful reviews. However, the model exhibits notable classification errors with 9 deceptive reviews misclassified as truthful and 11 truthful reviews incorrectly labelled as deceptive. This error distribution suggests that without emotion information, the model struggles equally with both false negatives and false positives, indicating limited discriminative power for subtle linguistic patterns that distinguish truthful from fake reviews.

The integration of emotion information demonstrates clear improvements in classification accuracy across all enhanced configurations. The dominant emotion class model with BERT, one dense layer, 3e-5 learning rate, 0.2 dropout shows enhanced deceptive review detection with 156 correct classifications while reducing false negatives to 4. However, this improvement comes with a slight increase in false positives to 12, suggesting that the dominant emotion features help identify deceptive patterns. The all emotion classes model with DistilBERT, multiple dense layers, 5e-5 learning rate, 0.2 dropout achieves exceptional performance in deceptive review identification, correctly classifying 159 deceptive reviews with only 1 false negative. This near-perfect recall for deceptive content comes at the cost of increased false positives 16 reviews, indicating that comprehensive emotion profiling creates highly sensitive detection but may over-classify some truthful reviews as deceptive.

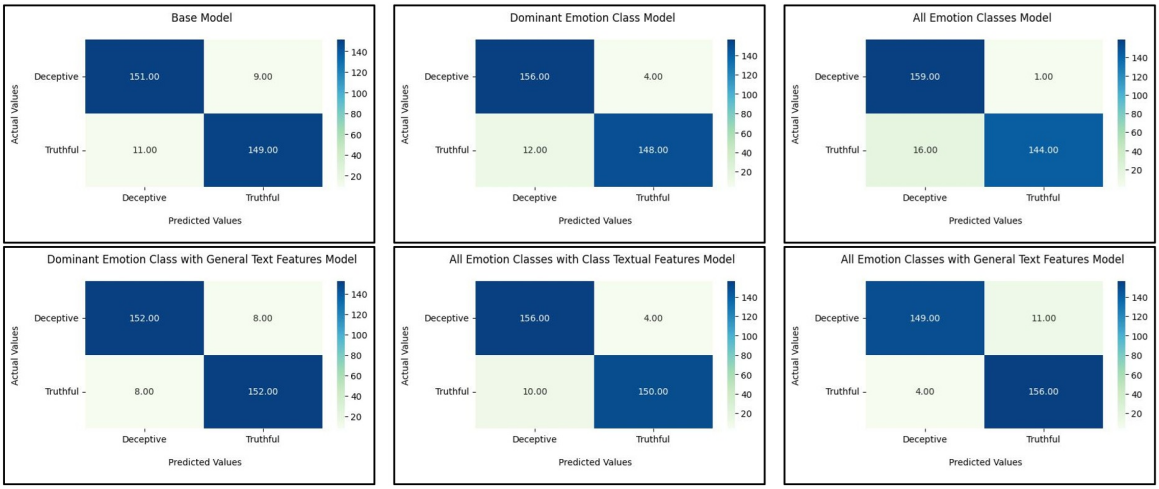


Figure 5.8: The confusion matrices produced by sample of model architectures

The addition of textual features to emotion information produces nuanced changes in classification patterns. The dominant emotion class with general text features model with DistilBERT, one dense layer, 5e-5 learning rate, 0.3 dropout maintains strong deceptive detection with 152 correct classifications and reduces false negatives to 8, while achieving excellent truthful review recognition with 152 correct classifications and only 8 false positives. This configuration demonstrates optimal balance between sensitivity and specificity. The all emotion classes with class textual features model with DistilBERT, one dense layer, 3e-5 learning rate, 0.3 dropout achieves notable deceptive detection accuracy with 156 correct classifications and minimal false negatives 4 reviews, while maintaining strong truthful classification performance 150 correct classifications and only 10 false positives. This suggests that class-specific textual features provide refined discrimination capabilities that enhance both precision and recall.

The all emotion classes with general text features model with RoBERTa, multiple dense layers, 5e-5 learning rate, 0.3 dropout demonstrates good balanced high-performance classification. With 149 correctly classified deceptive reviews and 156 correctly classified truthful reviews, combined with only 11 false negatives and 4 false positives. The low false positive rate indicates strong precision in truthful review identification, while the moderate false negative rate suggests good but not perfect recall for deceptive content. Finally, across all configurations, the consistent pattern emerges that emotion-enhanced models significantly improve deceptive review detection compared to the base model.

5.5.5 Statistical Significance of Emotion Integration Classification

McNemar's test was conducted to determine whether the observed difference in performance was statistically significant. Specifically, the comparison involved the model that provides the highest accuracy score without incorporating emotion information and the model that provides the highest accuracy score with emotion information incorporated which was the all emotion classes with class textual features model. This test is specifically designed for paired classification results and evaluates whether two models differ in their proportions of correct and incorrect predictions on the same test set. By focusing on discordant classification outcomes, where one model is correct and the other is incorrect, McNemar's test provides a measure of whether one classifier significantly outperforms the other. The McNemar test statistic is defined as:

$$\chi^2 = \frac{\big(|b - c| - 1\big)^2}{b + c}$$

where b denotes the number of reviews misclassified by the baseline but correctly classified by the emotion-enhanced model, and c denotes the number of reviews correctly classified by the baseline but misclassified by the enhanced model. After computing b and c from the paired predictions, the resulting p-value was 0.013, which is below the conventional 0.05 significance threshold. This p-value indicates that the probability of observing such a performance difference by chance is very low, providing strong evidence that the emotion-enhanced model's improvement is not due to random variation.

Overall, the McNemar's test confirms that integrating emotion information into the fake review classifier delivers a statistically significant gain in accuracy. The observed increase from 0.938 for the baseline model to 0.956 for the emotion-enhanced model demonstrates that emotional context contributes meaningful discriminative power for fake review detection.

5.6 Discussion of the Findings

The analysis of the results obtained from the experiments reveals several key insights into the efficacy of incorporating emotion information into fake review detection models. These insights are based on the accuracy scores obtained from different configurations and will be discussed in detail in this section.

5.6.1 Dominant Emotion Class

The dominant emotion class method frequently outperformed the base models. The findings reveal that focusing on the dominant emotion expressed in a review offers a powerful indicator for distinguishing between genuine and deceptive content. Emphasising the dominant emotional class provides a clear, concise indicator of a review's overall emotion and authenticity. This approach demonstrates that even a simplified representation of emotional content can yield significant enhancements, underscoring the value of capturing the core emotional expression in order to enhance the fake review detection models.

5.6.2 All Emotion Classes

Including all emotion classes led to considerable performance improvements. This method revealed that capturing the full emotional range in a review is effective for distinguishing between genuine and deceptive reviews. The obtained results indicate that a comprehensive analysis of the emotional classes is crucial for accurately detecting fake reviews. The method's effectiveness is further supported by its consistent high performance across various configurations. Ultimately, these findings underscore the importance of a detailed, granular analysis of the emotional content for improving fake review detection.

5.6.3 Combination with General Text Features

Combining emotional information with general textual features emerged as a powerful strategy for improving the detection of fake reviews, too. By integrating insights from both the emotional features and the underlying writing style, this approach provided a richer understanding of review authenticity. The findings suggest that the interplay between how emotions are expressed, and the overall textual characteristics can reveal subtle indicators of deception, ultimately offering a more robust means of distinguishing between genuine and fake reviews.

5.6.4 Combination with Class-specific Textual Features

Incorporating class-specific textual features alongside the emotion categories yielded to the most significant performance improvements. This approach allowed a deeper exploration of how emotions are expressed within the text, thereby enhancing the understanding of the interplay between expressions of emotions and writing style. The findings indicate that analysing the detailed textual features associated with each emotion can improve the detection of deceptive reviews by providing a more refined perspective on authenticity. Ultimately, this method demonstrates that a

granular analysis of emotion-specific textual features offers valuable insights into the relationship between the emotions expressed and a review's authenticity.

The results of this study demonstrate that incorporating emotion information into fake review detection models significantly enhances their performance. The highest accuracy scores were achieved by models that included comprehensive emotional information. DistilBERT, in particular, displayed exceptional performance when leveraging emotion information, suggesting its suitability for this task. These findings provide a strong foundation for future research and development in the field of fake review detection, emphasising the importance of emotional context in accurately identifying fake reviews. The detailed analysis of the different methods and their impact on model performance highlights the critical role of emotion information in enhancing the detection accuracy, paving the way for more effective and reliable fake review detection systems.

5.7 Chapter Summary

This chapter presented a comprehensive analysis of the impact of incorporating emotion information into the fake review detection models. The work began with the experimental setup, where a dataset of 1,600 reviews named GeFaRe, equally divided between fake and genuine categories, was collected, obtained and annotated. Three LLMs were employed to assess the efficacy of various configurations incorporating emotion information. The experiments were designed to evaluate how different types of emotion information and their integration into classifiers affect the detection accuracy. Two primary model structures were tested: one with a single dense layer and another with multiple dense layers. These structures provided a baseline with which to compare the performance improvements achieved by including emotion information.

Several methods of incorporating emotion information were explored. The base models, which did not include any emotion information at all, served as the baseline models for measuring the impact of adding emotional context. The *Dominant Emotion Class* method included the most prominent emotion expressed in each review, providing a simplified yet effective emotional information. By including all five emotion classes, the *All Emotion Classes* method captured the full range of emotions expressed in the reviews. The *Dominant Emotion Class with General Text Features* configuration combined the dominant emotion class with general textual features, such as the sentence and word counts, to provide additional context. The *All Emotion Classes with the Class Specific Textual Features* method added detailed textual features for each emotion class, providing a holistic view of both emotional and textual characteristics. The *All Emotion Classes with General Text Features* approach combined all of the emotion classes with general text features, allowing a different analysis of emotional expressions.

The results demonstrated that incorporating emotion information significantly enhances the accuracy of the fake review detection models. Notably, the DistilBERT model achieved the highest accuracy score of 0.956 using the all emotion classes with class textual features method with a learning rate of 3e-5, a dropout rate of 0.3, and a one dense layer architecture. This indicates that capturing comprehensive emotional information is crucial for accurately detecting fake reviews.

The detailed analysis produced several key insights. Leveraging the dominant emotion class improved the model performance noticeably, with models achieving high accuracy scores by focusing on the primary emotional category of the reviews. Including all emotion classes provided better performance improvements, highlighting the importance of capturing the full emotional context. Enriching emotional information with general textual features further enhanced the model accuracy, demonstrating the value of combining multiple sources of information. Adding detailed textual features for each emotion class yielded the most substantial improvements, facilitating a different understanding of how emotions are expressed in the text.

One of the most notable findings was the impact of incorporating emotion information on model performance. The results showed that even a simple inclusion of the dominant emotion class led to remarkable improvements in detection accuracy across all of the three language models. This highlights the importance of capturing the emotional information within a review, as it can provide valuable insights into its authenticity.

The all emotion classes method, which includes information about the presence of the five different emotion classes, consistently outperformed the base models. This suggests that capturing comprehensive emotional information is crucial for accurately detecting fake reviews. By considering the full range of emotions expressed, the models were better equipped to distinguish between genuine and deceptive reviews, which often exhibit distinct emotional patterns.

Furthermore, the inclusion of detailed textual features for each emotion class also yielded significant improvements in model performance. This method allows for a detailed analysis of how different emotions are expressed in the text, capturing subtle linguistic patterns and features that are associated with each emotion class. By leveraging this granular information, the models can better differentiate between genuine and deceptive reviews, based on the specific ways in which emotions are conveyed within them.

It is noteworthy that the DistilBERT model consistently outperformed BERT and RoBERTa models across the majority of the configurations, suggesting its suitability for tasks requiring emotional analyses. The DistilBERT model's architecture and pre-training approach may have enabled it to capture and leverage the emotional information present in the reviews more effectively. These results provide a strong foundation for future research to continue to explore and refine the methods for

integrating emotion information into detection models.

Through the work presented in this chapter, RQ4 was addressed by analysing the potential contribution of emotional features towards enhancing the application of the fake review detection models to tourism reviews. The experiments demonstrated how integrating emotional features, derived from the Plutchik Wheel of Emotions, improved the accuracy and reliability of the detection systems compared to the baseline models that did not consider emotional information. The findings emphasise the role of emotion information in improving the accuracy of the fake review detection models, particularly when applied to the tourism domain.

In conclusion, this chapter demonstrates the performance gains that can be achieved by integrating emotion information into fake review detection models. The comprehensive analysis of the different methods and their impact on model accuracy highlights the important role that the emotional context plays in enhancing detection accuracy.

Chapter 6

Conclusions and Future Research Directions

6.1 Thesis Summary

This thesis explores the integration of emotion-based features into LLM classifiers for detecting fake reviews, addressing a critical challenge associated with maintaining the integrity and reliability of online review platforms. By bridging theoretical insights from psychology with computational advancements in NLP, this research advances our understanding of how emotional information can enhance the accuracy of fake review detection classifiers. The thesis is structured across six chapters, with each contributing to a cohesive investigation of emotion-enhanced review classification.

The initial phase of this research involved developing a novel dataset comprising tourism-related content annotated with Plutchik's emotion scheme and evaluate the reliability of employing crowdsourcing methods for emotion classification, this study employed a comprehensive, rigorous methodology. Initially, 2,500 sentences were systematically extracted from TripAdvisor reviews, which platform was selected due to its rich content of user reviews related to the tourism domain. These sentences were then annotated via Amazon MTurk, where annotators were provided with detailed instructions and illustrative examples for each of the eight primary emotions: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", and "trust", with the option to assign up to two labels per sentence or indicate the absence of any apparent emotion.

To ensure the quality and reliability of these annotations, robust quality control measures were implemented, including the use of gold standard questions and the required history of accepted HITs of the workers. The evaluation, particularly the almost perfect agreement observed via the PEA High metric, confirmed the reliability of the crowdsourcing process. Additionally, a majority voting strategy was utilised

to resolve conflicting annotations, concluding in the creation of two dataset versions: TORCEv1, which included all eight emotion categories for detailed analysis, and TORCEv2, which focused on only five emotions: "anger", "anticipation", "joy", "sadness" and "surprise", to promote computational efficiency. Overall, this part of the research demonstrates how crowdsourcing, when coupled with careful quality control and majority voting techniques, can produce a high-quality, domain-specific dataset that meets the complex needs of emotion detection in tourism reviews.

The second major component focused on testing and comparing the efficacy of mainstream emotion detection tools on tourism-related data through systematic, comprehensive evaluation. A wide range of emotion detection tools, from basic lexicons to advanced deep learning models, was thoroughly assessed against the TORCEv2 dataset in Chapter 4. The evaluation process involved carefully selecting tools that met criteria such as public availability and alignment with Plutchik's primary emotions, ensuring each tool's performance relevance to the required emotion categories.

A systematic testing methodology, which employed an F-score metric, revealed that, while many tools were able to capture basic emotional categories, such as "joy" and "anger", they often struggled with the more complex or subtle emotional expressions that are inherent in tourism-related content. These findings highlighted the considerable limitations in the current mainstream methodologies and underscored the need to develop tools that are specifically designed to address the complexity of domain-specific expressions of emotions. This comprehensive analysis thus provides a strong foundation for understanding the gaps in the existing emotion detection systems and confirms the need for future improvements in this area.

The third component involved designing and implementing emotion classification models leveraging LLM architectures through fine-tuning BERT, DistilBERT, and RoBERTa on the TORCEv2 dataset, which was discussed in Chapter 4. The process began with a thorough evaluation of these models' baseline performances, followed by systematic experiments refining capabilities through hyperparameter adjustment, particularly learning rates, and incorporating several data augmentation techniques. By fine-tuning these models on domain-specific corpora enriched with emotional annotations, the classifiers enhanced the models' ability to capture both simple and complex emotional categories.

The experimental results demonstrated that fine-tuned LLMs, especially RoBERTa, outperformed the mainstream emotion detection tools by achieving higher F-scores across all emotion classes, thereby validating the transformative potential of the LLM architectures in this specialised context. This approach addressed the challenges posed by the distinct emotional landscape of tourism reviews, and provided empirical evidence that advanced, domain-adapted LLMs can substantially improve the accuracy and reliability of the emotion classification systems.

The final component methodically investigated and compared multiple strategies for incorporating emotion-based features into fake review detection models, subsequently developing new detection models using LLM architectures. For this part of the work, the GeFaRe dataset was used, which contains 1,600 tourism reviews that are evenly distributed between fake and genuine reviews, served as a robust training and testing ground. Systematic experimental configurations explored how various approaches might enhance classifier ability to identify deception. Two distinct model architectures were evaluated, one using a single dense layer and the second employing multiple dense layers, which provided a comparative framework against the baseline models, that lacked any emotional features.

The systematic experimentation revealed that emotion-enriched models consistently outperformed their counterparts, with the DistilBERT model configured with all emotion classes and class textual features demonstrating the highest accuracy of 0.956. This process underscored the transformative potential of integrating emotional information to improve fake review detection in the tourism domain, thereby fulfilling the goal of developing powerful, domain-specific fake review detection models.

In addition, this thesis makes the following contributions to the NLP field:

1. Develop a novel dataset comprising tourism-related content annotated with emotion information.

2. Design and implement an emotion classification model leveraging LLMs' architecture.

3. Develop a fake review detection model using LLM architectures and analyse the influence of emotion-based features on the performance of the classification models.

All of the datasets and source codes for this research are available at: *https://github.com/Mansour765/Experiments*.

6.2 Research Questions Revisited

This project sought to answer four research questions, as outlined in section 1.2. In the following, the research questions that guided this study are revisited to evaluate the outcomes of the investigation and the findings that help to answer these questions.

RQ1: How can crowdsourcing annotation help to produce a reliable tourism emotion dataset?

This research question was answered in Chapter 3. Crowdsourcing annotation enhances the reliability of a tourism emotion dataset by leveraging the diverse

perspectives of multiple annotators and incorporating robust quality control measures. By distributing the annotation task across multiple, varied workers, each review is evaluated by several individuals from different backgrounds, which helps to counterbalance individual biases and subjectivities. By providing clear instructions and designing intuitive annotation interfaces, the task ensured that consistent, meaningful outputs were produced. This process is further strengthened by employing methods like majority voting to resolve differing opinions, ensuring that the final labels reflect a consensus rather than isolated interpretations. Additionally, using validation techniques such as IAA metrics ensured that the annotations were consistent and reliable. Achieving a high IAA score indicating that the workers could reliably classify complex emotional content in tourism-related reviews. Finally, my research results show that crowdsourcing is a powerful tool for generating high-quality, domain-specific datasets that meet the specific requirements of tasks like emotion detection.

RQ2: How are the existing emotion detection tools effective for tourism reviews?

The evaluation in Chapter 4 demonstrates that the existing emotion detection tools, while moderately effective in identifying basic emotions such as "joy" and "anger", struggle to capture the complex, nuanced emotional expressions that are typical of tourism reviews. This outcome was achieved by systematically measuring performance using F-scores, revealing that tools based on generalised models are frequently incapable of handling the context-specific language found in tourism data. My analysis shows that these tools tend to overlook complex emotions, particularly the more complex ones like "surprise", due to their reliance on simplified classification frameworks. Consequently, my findings highlight that, while the current methodologies are useful to an extent, their effectiveness is limited by a lack of domain-specific adaptation. This underscores the critical need to develop context-aware, tailored emotion detection models that can better address the unique linguistic and emotional characteristics inherent in tourism reviews.

RQ3: How can fine-tuned LLMs improve emotion detection?

This question was addressed in Chapter 4. Fine-tuned LLMs enhance emotion detection by adapting to the unique, context-dependent language found in tourism reviews. By carefully refining the hyperparameters, and implementing data augmentation techniques for scarce emotion categories, the LLM classifiers are better equipped to capture the complex interplay between the emotions expressed in tourism-related text. This fine-tuning process enables the model to overcome the limitations of generic emotion detection tools, allowing it to differentiate accurately between complex emotional categories. The empirical results, which include an overall F-score of 0.80 and high scores for specific emotions such as 0.93 for "joy" and 0.85 for "anger", clearly

demonstrate that these modified adjustments significantly improve the classification accuracy. This study confirms that fine-tuning LLMs is a highly effective strategy for addressing the specific challenges associated with emotion detection in specialised domains like tourism.

RQ4: How can emotion information help to detect fake tourism reviews?

By integrating emotion information into the detection models, the classifiers can more effectively differentiate between genuine and fake tourism reviews, as demonstrated by the findings presented in Chapter 5. The thesis systematically explored various strategies for embedding emotional features, thereby creating models with a richer, multidimensional understanding of emotion. This approach allows the models to recognise the patterns that are indicative of genuine or fake reviews. My experiments revealed that the emotion-aware classifiers consistently outperformed the baseline models that relied solely on linguistic features, highlighting that the inclusion of emotion data addresses a critical gap in fake review detection. This outcome underscores that enriching LLMs with emotional information enables the more robust and more reliable identification of fake reviews in user-generated content, leading to more effective detection systems in the tourism domain.

6.3 Limitations of the Research

This thesis has some limitations, that should be mentioned. A significant challenge was the limited availability of tourism-related datasets that are annotated with both emotional and fake review labels. The scarcity of such specialised corpora necessitated extensive efforts to compile and annotate data using crowdsourcing methods. However, this approach, while effective, brings its own set of complexities, such as ensuring consistency across annotators and mitigating potential biases arising from subjective judgments. This limitation highlights the need for further data collection and annotation for specialised domains like tourism.

Furthermore, constructing a dedicated fake review dataset was a challenging task due to the difficulty of generating convincing fake reviews that accurately mimic authentic user experiences. Creating such data entails simulating the nuanced linguistic, contextual, and emotional characteristics that real reviews naturally possess, a task that is inherently complex and likely to introduce bias. This process is further complicated by the need to ensure that the synthetic reviews do not deviate too far from the realistic patterns observed in genuine user reviews, thereby maintaining their relevance and utility for training robust detection models.

Another limitation of this research was the constraint imposed by the available computational resources. The experiments involving LLMs were restricted by the

computer hardware capacity, which in turn limited the scale and depth of the model training and evaluation processes. Due to insufficient computing power and memory, the study had to operate with smaller batch sizes, reduced training epochs, and less complex model architectures. This constraint hindered the opportunity for extensive hyperparameter tuning and the exploration of larger models, that might have provided deeper insights into emotion detection and fake review classification.

In addition, while models such as BERT and its variants were considered state-of-the-art at the commencement of this work, the rapid evolution of LLMs means that more advanced architectures have since emerged. These models might offer improved performance in terms of capturing nuanced linguistic and contextual information, which is crucial for accurately detecting emotions and identifying fake reviews. The newer architectures are pre-trained on larger corpora and feature more sophisticated finetuning mechanisms, potentially leading to significant gains in terms of their detection accuracy and reliability.

6.4 Future Research Directions

Building on the findings of this project, future research might explore several avenues to enhance and extend the methodologies and applications developed in this research. One avenue is the expansion and enhancement of the dataset. The limited availability of tourism-related datasets annotated with both emotional and fake review labels proved a significant constraint. Future work might explore novel data collection strategies, including the use of semi-automated annotation methods and synthetic data generation techniques, to build larger, more representative corpora.

Furthermore, the challenges associated with crowdsourcing annotation for emotion detection warrant additional attention. Although the crowdsourcing approach used in this study proved effective, it also introduced potential bias due to the subjective nature of emotion annotation. Future research should investigate methods for further standardising the annotation protocols by developing improved training modules for annotators or incorporating advanced quality control mechanisms that dynamically adjust in order to compensate for annotator variability. The integration of machine-assisted annotation tools can also help to streamline the process and reduce the reliance on human judgment, thereby increasing both the efficiency and consistency of the annotations.

Another potential future research direction involves applying emotion enhanced models to diverse datasets beyond the tourism domain. While this study focused on tourism-related reviews, testing the models on datasets associated with other sectors,

such as healthcare [23] or entertainment [24], can reveal their cross-domain applicability and robustness. Another promising area for future exploration is the assessment of cross-domain generalisability. By training models on datasets from one domain and testing them on another, it will be possible to evaluate the transferability of emotion-based features and determine the extent to which these features capture universal versus domain-specific patterns. This transferability may enhance the development of generalised fake review detection frameworks that are capable of adapting to various contexts without extensive retraining.

Another direction is the exploration of the advanced LLMs that have emerged since this research began. While this study utilised models including BERT and its variants, the rapid evolution of LLM architectures means that the newer models might offer superior performance with regard to capturing complex linguistic and contextual nuances. Future research should integrate these state-of-the-art models, which are pre-trained on more extensive and diverse datasets, to evaluate their potential for improving emotion detection accuracy and fake review classification.

In addition to textual data, integrating multimodal data sources, such as images, videos, and audio, offers a compelling direction for advancing the fake review detection systems. For instance, combining textual emotional features with visual signs, such as sentiments conveyed in accompanying emojis, together with images or facial expressions in video reviews, may provide a more holistic view of authenticity. Similarly, audio data, including the tone and rhythm, can uncover subtle emotional inconsistencies that are indicative of deception. The fusion of these modalities with textual analysis can lead to new methods of detection systems that are more comprehensive and accurate.

Furthermore, future studies might explore the ethical considerations and practical challenges associated with deploying emotion-enhanced detection systems. This includes addressing potential biases in the datasets, ensuring the interpretability of the models, and examining the implications of their usage by users, businesses, and platforms. Developing transparent, fair frameworks for implementation will be essential to maximise the impact of these technologies.

Finally, to explore the potential of advanced LLM's for emotion-enhanced fake review detection without extensive fine-tuning, a preliminary pilot study was conducted using the Qwen LLM in an instructed zero-shot setting on the GeFaRe dataset's test set. This investigation aimed to assess whether state-of-the-art generative models could leverage their inherent language understanding capabilities to distinguish between authentic and deceptive reviews when provided with dominant emotion class information through prompts. The experimental setup involved providing the Qwen model with clear instructions to classify reviews as fake or

[23] Such as Google Maps and https://www.nhs.uk/

[24] Such as https://www.tripadvisor.com/ and https://www.ign.com/

truthful, where each review was accompanied by its corresponding dominant emotion class information, without any prior training. Despite incorporating emotional information that proved highly effective in the fine-tuned models presented in this thesis, the results revealed considerable limitations in this zero-shot approach, with the model achieving an accuracy of only 0.506 and an F-score of 0.347, indicating performance barely above random chance. These findings suggest that while LLMs have sophisticated language understanding abilities, the nuanced detection of deceptive content in tourism reviews requires more than general linguistic knowledge, highlighting the importance of specialised training approaches and the systematic integration of comprehensive emotional features as demonstrated in this research.

In future work, further investigation will focus on developing strategies that benefit of the general interpretation power of LLMs and the domain-specific requirements of fake review detection. This will include systematically exploring fine-tuning pipelines that integrate emotional, semantic, and contextual representations derived from the GeFaRe dataset to enhance model sensitivity to subtle indicators of deception. Additionally, prompt design and instruction-tuning approaches will be examined to determine how targeted guidance can enable LLMs to adopt emotional information more effectively without requiring large-scale retraining.

Appendix A

A.1 Samples of the Crowdsourcing Annotations

Sample Review: "This man had made the effort to get himself out and about despite presumably these scenarios being so difficult for him yet he was being treated with such disrespect and unkindness."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Surprise	Sadness	Surprise	Anger	Surprise
Emotion 2	Disgust	Surprise	-	Sadness	Sadness

Sample Review: "Before entering the theatre, I prepared myself for a feel good show full of bright costumes, lights and excellent dancing."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Trust	Joy	Anticipation	Anticipation
Emotion 2	-	Anticipation	-	Joy	-

Sample Review: "Given the amount of money I paid to enter the property, I was expecting to be treated a significant deal better than I was."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anticipation	Anticipation	Anticipation	Anger	Disgust
Emotion 2	-	Sadness	-	Sadness	Anger

Sample Review: "They are, as a collective, so sour, dour, miserable, unyieldingly inflexibly reliably unhelpful as to make them a special class of people."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Sadness	Disgust	Disgust	Anger	Anger
Emotion 2	Disgust	-	-	Disgust	Sadness

Sample Review: "I felt a rush of panic then overwhelming disappointment at what a dull shadow this fabulous museum has become today."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Disgust	Surprise	Sadness	Surprise	Sadness
Emotion 2	-	Sadness	Disgust	-	-

Sample Review: "The whole zoo is decaying, the animals look unhappy and neglected, cages, surroundings and everything is in decay and so depressing."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Disgust	Joy	Surprise	Anger	Surprise
Emotion 2	Fear	Surprise	Fear	Disgust	-

Sample Review: "Contrary to other reviews we both felt the food was excellent, really tasty and good honest food!"

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Trust	Trust	Surprise	Joy
Emotion 2	Surprise	Joy	-	-	Surprise

Sample Review: "I feel far more threatened when I see groups of drunks in the park and feel that more should be done to ban them!"

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Fear	Anger	Fear	Disgust	Disgust
Emotion 2	-	Fear	-	-	Fear

Sample Review: "Even though you were not there during the war, you feel the pain and deprivation those people felt."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Disgust	Trust	Sadness	Anticipation	Trust
Emotion 2	Sadness	Surprise	Disgust	Disgust	Surprise

Sample Review: "We were treated with such dignity & respect & never did anyone feel a burden or uncomfortable."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Joy	Trust	Trust	Surprise	Trust
Emotion 2	Surprise	-	-	-	Joy

Sample Review: "but I feel in this case describing my experience is a necessary evil in the hope it leads to an improvement."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anticipation	Anticipation	Anticipation	Anticipation	Anticipation
Emotion 2	Sadness	-	Sadness	Sadness	-

Sample Review: "I have always felt on a high after coming out of an escape game, with this one I felt so sad that I had wasted my money."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Anger	Anger	Disgust	Anticipation	Sadness
Emotion 2	Disgust	Disgust	Anger	Surprise	Anger

Sample Review: "I almost choked on a bit as well and that would have a very bad death, choking on overpriced non-tasty hot dog."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Fear	Sadness	Disgust	Fear	Anger
Emotion 2	Anger	Fear	Fear	-	-

Sample Review: "Jay makes you feel comfortable and in case you feel nervous you quickly get the feeling you can do anything."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Trust	Trust	Joy	Joy	Trust
Emotion 2	-	Joy	-	-	-

Sample Review: "What ought to have been a tale of unbridled happy families was somehow tinged with an air of unspoken sadness."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Sadness	Disgust	Sadness	Sadness	Sadness
Emotion 2	-	Surprise	Surprise	Fear	Anticipation

Sample Review: "Aside from the eatery, the rest of the shops are lovely, nice and quaint exactly what you'd expect to find in the countryside."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Surprise	Joy	Anticipation	Trust	Trust
Emotion 2	Trust	-	Joy	Joy	-

Sample Review: "We felt for the money we paid we had a good deal and enjoyed our day."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Surprise	Joy	Joy	Joy	Trust
Emotion 2	Joy	Surprise	-	-	Joy

Sample Review: "A word of advice- pay a bit more for a proper spa as this isn't worth the time or money."

Annotator	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**	**Annotator5**
Emotion 1	Disgust	Anger	Anger	Disgust	Surprise
Emotion 2	Surprise	-	Disgust	Anger	-

Appendix B

B.1 Samples of the TORCEv1 and the TORCEv2 Datasets

sentence	TORCEv1		TORCEv2 Emotion Class
	Emotion 1	Emotion 2	
This man had made the effort to get himself out and about despite presumably these scenarios being so difficult for him yet he was being treated with such disrespect and unkindness.	Surprise	-	Surprise
Before entering the theatre, I prepared myself for a feel good show full of bright costumes, lights and excellent dancing.	Anticipation	Joy	Joy
Given the amount of money I paid to enter the property, I was expecting to be treated a significant deal better than I was.	Anticipation	-	Anticipation
They are, as a collective, so sour, dour, miserable, unyieldingly inflexibly reliably unhelpful as to make them a special class of people.	Disgust	-	Anger
I felt a rush of panic then overwhelming disappointment at what a dull shadow this fabulous museum has become today.	Sadness	-	Sadness

sentence	**TORCEv1**		**TORCEv2 Emotion Class**
	Emotion 1	**Emotion 2**	
The whole zoo is decaying, the animals look unhappy and neglected, cages, surroundings and everything is in decay and so depressing.	Surprise	-	Surprise
Contrary to other reviews we both felt the food was excellent, really tasty and good honest food!	Joy	Surprise	Joy
I feel far more threatened when I see groups of drunks in the park and feel that more should be done to ban them!	Fear	-	Anger
Even though you were not there during the war, you feel the pain and deprivation those people felt.	Disgust	-	Anger
We were treated with such dignity & respect & never did anyone feel a burden or uncomfortable.	Trust	-	Joy
but I feel in this case describing my experience is a necessary evil in the hope it leads to an improvement.	Anticipation	-	Anticipation
I have always felt on a high after coming out of an escape game, with this one I felt so sad that I had wasted my money.	Anger	-	Anger
I almost choked on a bit as well and that would have a very bad death, choking on overpriced non-tasty hot dog.	Fear	-	Anger
Jay makes you feel comfortable and in case you feel nervous you quickly get the feeling you can do anything.	Trust	-	Joy
What ought to have been a tale of unbridled happy families was somehow tinged with an air of unspoken sadness.	Sadness	-	Sadness
Aside from the eatery, the rest of the shops are lovely, nice and quaint exactly what you'd expect to find in the countryside.	Joy	Trust	Joy

sentence	**TORCEv1**		**TORCEv2 Emotion Class**
	Emotion 1	**Emotion 2**	
We felt for the money we paid we had a good deal and enjoyed our day.	Joy	-	Joy
A word of advice- pay a bit more for a proper spa as this isn't worth the time or money.	Disgust	Anger	Anger

Appendix C

C.1 Samples of the GeFaRe Dataset

Review Text	Dominant Emotion	Number of Sentences with Emotion Class				
		Joy	Anger	Sadness	Surprise	Anticipation
I booked this via Priceline, and was not sure I'd like it, as some Sheratons have fallen off the cliff. This hotel absolutely surpassed my expectations. Despite booking with Priceline, I was treated like a king. They let me check in early, gave me a beautiful room, and even went so far as to look up my Starwood number so I could get some credit for the stay. The staff was unfailingly pleasant, the property immaculate, and the room very comfortable. I highly recommend this hotel.	Joy	4	1	1	0	1
In my experience the Ambassador hotel didn't seem to be a 3 1/2 star establishment. The building and lobby were nice but the hotel room was extremely dated. The carpeting was very old and worn. The floor - carpeting and tile both seemed dirty which made me very uncomfortable. I would imagine that many years ago this was a very nice hotel. Currently, it's indespirate need of renovation.	Anger	2	3	0	0	1
I recently stayed at the Hard Rock Hotel in Chicago, Il. From the start, the experience was bad. The room was filthy, there were no towels, and the front desk did nothing to rectify the situation. I will never stay there again. I could not have been more dissatisfied.	Sadness	1	1	3	0	0
I would like to add a comment, maybe this will help others when choosing for a wedding destination. My daughter was married in May of 2010 and we decided to have the special event at the Sofitel Water Tower; we were amazed. She always wanted a ballroom type of wedding, and Sofitel made this happen for her. They took care of all the invitations, cake decor, and so much more. The special day for my daughter was worth everything. If you are going to be in the Chicago area, check them out; you will not regret it!	Surprise	2	0	0	3	1
We stayed at this hotel for our last vacation to the windy city. Although the hotel looks nice, the service is terrible. The staff was rude and unprofessional and not helpful at all. The room service closes at 8pm so don't plan on getting hungry after that time. Also, even though we stayed for a whole week we had to ask for new towels and extra toiletries. I will never go back to that place	Anger	2	3	1	0	0

Review Text	Dominant Emotion	Number of Sentences with Emotion Class				
		Joy	Anger	Sadness	Surprise	Anticipation
Stayed at this hotel with 3 friends or 4 nights. The hotel was clean and tidy, throughout. The Concierge Christopher was excellent and helped as with all our needs, gave us discount vouchers etc. Hotel was in excellent position. 3 blocks from John Hancok Building and more importantly The Cheesecake Factory , Bloomingdales 3 blocks away and the Water Tower Shopping Centre 2 blocks away. We went to the huge I Max cinema which is about ten minutes away. Cant wait to go back to Chicago and The Affinia	Joy	5	0	0	1	1
While making the obligatory trip to the Chicago area to visit family, we decided to make it fun for the teenage members of our group and reserved a couple of rooms at the Hard Rock Hotel Chicago. What a great decision! The rooms are very upscale at a reasonable price with great amenities. And, according the younger members of the family, 'very cool'! We can't wait to come back.	Anticipation	2	0	0	1	2
Named my price on priceline. $50.00 Bucks. Hotel room was great. Clean, Clean and new. Fresh crisp sheets, comfortable bed, flat screen TV, clean carpet, nice bath, etc. Short distance to food and sightseeing. I highly recommend this property. Be prepared to pay over $40.00 per night to park. Hey, Its a hyatt. They also charge $5.00 for a bottle of water that is normally a buck. For a total of $100.00 nightly stay in Chicago this place is it just don't drink the bottled water. Happy Travels	Joy	7	3	0	1	1
i recently stayed at the amalfi hotel in Chicago and was very disappointed right from the beginning. the girl at the front desk was busy talking to a one of her friends and when i finally did get her to wait on me she was rude and seemed bothered to have to wait on me. the room wasnt at all what i was expecting. it was dingy, dirty, and just seems like an old hotel. just not what i was looking for in a nice hotel. seemed like the quality you'd expect from a cheap motel.	Sadness	1	1	4	0	0

Review Text	Dominant Emotion	Number of Sentences with Emotion Class				
		Joy	Anger	Sadness	Surprise	Anticipation
My son & I joined my husband on a work trip. We planned to swim while he was working. We overheard a lady asking when the pool would open as as we were checking in. Though my son was eager to swim as soon as we checked in the pool was still closed at 630 pm. We gave up & went out for the evening. We had a very noisy sleepless night. We weren't facing MI ave, but still heard sirens and I'm not sure what the other noise was, but it sounded like a malfunctioning hand dryer in our ceiling that would sound off about 20x an hour throughout the entire night. I called the front desk the next morning and was told the pool was open, so my (very excited) son got his bathing suit on and we got to the fitness center only to be told by housekeeping that the pool was closed for construction. Upon calling the front desk again the lady claimed she was unaware of the maintenance crew's actions & on my insistence she said she'd look into it. A gentleman called back an hour later & said the pool should open at 5pm...it was 9am. It is now 6pm & we still haven't even seen the pool. We check out tomorrow. We passed up time with grandparents who are in for the holiday to 'swim in Chicago.' Needless to say we are VERY disappointed.	Surprise	1	3	1	6	4
This hotel was not at all what I expected it would be. The website and it's pictures portray the hotel in a much better condition than it is actually in. The amenities were not what I was looking for, and the price of the hotel was not a good value for what was offered there. My AAA membership did not get me a very good discount like it does at most hotels. The staff was not very friendly or helpful when I had questions about the area, and they were not very prompt when responding to requests. I understand that there were a lot of guests staying there while I was there, but they need to have more staff on hand if they cannot provide their hotel guests with quality service. I don't think I will be staying at this hotel again anytime soon.	Sadness	1	2	3	1	0

Review Text	Dominant Emotion	Number of Sentences with Emotion Class				
		Joy	Anger	Sadness	Surprise	Anticipation
We absolutely loved the Knickerbocker. Now, if you expect your hotels to be cookie cutter comfortable, forget this place. If Holiday Inns are like the rows of housing developments in the suburbs, the Knickerbocker is like those beautiful old houses in the center of the city. The location is great, the hotel bar is cozy and classy, and the rooms feel like guest bedrooms in Victorian mansions. For the price, you simply can't do any better than this in downtown Chicago.	Joy	3	1	0	1	0
The hotel is very impressive upon entering and the staff was very friendly, however we felt our room was very dated and worn looking. Our air conditioning didn't seem to be working well, but we turned it down and thought that it would eventually... cool off. Our first night was interrupted by a phone call at one A.M. That was a fax, it happened 4 times. By this time we realized that the air was not working again, so after fussing with the controls it did kick on. The next morning I mentioned to the concierge the issues with the phone, but didn't think to mention the issues with the air because it appeared to be working. After being out seeing the sites of Chicago for several hours and being very hot we were looking forward to resting in a cool room before going to a show . Our room was 78 degrees, I called and maintenience did come to check. After determining that the unit needed a motor it was another hour so we did not get to rest in a cool room before going out. I feel an offer should have been made to us by the hotel. Also was very surprised that WiFi was not a free service !! We will not be staying at the Hilton on our next trip to Chicago.	Surprise	1	5	0	6	1
The InterContinental Chicago hotel terrible. First of all, the continental breakfast was a joke. The food was unhealthy and unsatisfactory. After our first night, we expected the maid to come clean up our room. We came back from site-seeing to find that our room was in the same condition we left it in. I will never go back to this hotel again.	Anger	0	4	1	1	0

Review Text	Dominant Emotion	Number of Sentences with Emotion Class				
		Joy	Anger	Sadness	Surprise	Anticipation
Whenever I decided to stay at Fairmont Chicago Millennium Park, the only thing that I had on my mind was the idea of taking that time to relax and enjoy some time with my husband. However, the hotel choice wasn't the best. Right in the front desk when we arrived there we had problems with our reservation. The room that we had reserved was occupied and so we had to switch to some other room, which was much smaller than the original one. The hotel did give us a discount, but if I'd paid for the large room was because I wanted the it. A hotel with this name shouldn't put its guests in these kind of situation. I wont be reserving room there again.	Surprise	1	2	1	3	0
Staying here currently. Had aspired to sleep. The paper thin walls allow even the slightest peep from your neighbor to feel like you're right in bed with them. Couldn't sleep anyway because of the cheap, giant fluffy pillows. Called housekeeping, asked for any help...was simply and accomodatingly told, after my plea for help, could they, would they have any ability to supply me with a reasonable pillow? Complete answer before hanging up: 'We do not.' Hyatt, may your well executed service lead to your demise.	Anger	2	7	1	1	0
I stayed at the Hard Rock hotel in Chicago last year and can say i was not satisfied with my experience. I feel that it didn't live up to what was advertised and the accommodations were sub-par at best. I had the chance to catch a show and the music was too loud and the sound guy's work resembled that of an amateur. My personal rating is 2/5 stars and i would not recommend this hotel to a friend.	Sadness	0	1	2	1	0

Bibliography

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Faranak Abri, Luis Felipe Gutiérrez, Akbar Siami Namin, Keith S. Jones, and David R. W. Sears. 2020. Linguistic features for detecting fake reviews. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 352–359.

Sajad Saraygord Afshari, Fatemeh Enayatollahi, Xiangyang Xu, and Xihui Liang. 2022. Machine learning-based methods in structural reliability analysis: A review. *Reliability Engineering System Safety*, 219:108223.

Tariq Ahmad, Allan Ramsay, and Hanady Ahmed. 2019. Detecting emotions in english and arabic tweets. *Information*, 10(3).

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *SECURITY AND PRIVACY*, 1(1):e9.

Hani Al-Omari, Malak A. Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232.

Ali Reza Alaei, Susanne Becken, and Bela Stantic. 2019. Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research*, 58(2):175–191.

Sara Ali, Bushra Naz, Sanam Narejo, Sahil Ali, and Jitander Kumar Pabani. 2024. Transformers unveiled: A comprehensive evaluation of emotion detection in text transcription. In *2024 Global Conference on Wireless and Optical Technologies (GCWOT)*, pages 1–7.

Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label emotion analysis in conversation via multimodal knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6090–6100, New York, NY, USA. Association for Computing Machinery.

K. Anoop, Manjary P. Gangan, Deepak P, and V. L. Lajish. 2019. *Leveraging Heterogeneous Data for Fake News Detection*, pages 229–264. Springer International Publishing, Cham.

K Anuja, P C Reghu Raj, and Remesh Babu K R. 2022. State-of-the-art methods for fine-grained emotion detection from malayalam text using deep learning: A survey. In *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pages 1–5.

Ugo Arbieu, Kathrin Helsper, Maral Dadvar, Thomas Mueller, and Aidin Niamir. 2021. Natural language processing as a tool to evaluate emotions in conservation conflicts. *Biological Conservation*, 256:109030.

Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. 2011. Detecting emotions in social affective situations using the emotinet knowledge base. In *Advances in Neural Networks – ISNN 2011*, pages 611–620, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rodrigo Barbado, Oscar Araque, and Carlos A. Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing Management*, 56(4):1234–1244.

Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. 2019. Semantic-emotion neural network for emotion recognition from text. *IEEE Access*, 7:111866–111878.

Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. 2022. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Heang-Ping Chan, Lubomir M. Hadjiiski, and Ravi K. Samala. 2020. Computer-aided diagnosis in the era of deep learning. *Medical Physics*, 47(5):e218–e227.

Arjun Choudhry, Inder Khatri, Arkajyoti Chakraborty, Dinesh Vishwakarma, and Mukesh Prasad. 2022. Emotion-guided cross-domain fake news detection using adversarial domain adaptation. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 75–79, New Delhi, India. Association for Computational Linguistics.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Alan S Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. 2021. Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841):251—257.

Akash Das, Kartik Nair, and Yukti Bandi. 2022. Emotion detection using natural language processing and convnets. In *Data Science and Security*, pages 127–135, "Singapore. Springer Nature Singapore.

Felicitas Datz, Guoruey Wong, and Henriette Löffler-Stastka. 2019. Interpretation and working through contemptuous facial micro-expressions benefits the patient-therapist relationship. *International Journal of Environmental Research and Public Health*, 16.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2018. LT3 at SemEval-2018 task 1: A classifier chain to detect emotions in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 123–127, New Orleans, Louisiana. Association for Computational Linguistics.

Dirk Deichmann, Thomas Gillier, and Marco Tonellato. 2021. Getting on board with new ideas: An analysis of idea commitments on a crowdsourcing platform. *Research Policy*, 50(9):104320.

Flor Miriam Plaza del Arco, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and Ruslan Mitkov. 2020. Improved emotion recognition in spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*, 110:1000–1008.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patel Dhruv and Subham Naskar. 2020. Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): A review. In *Machine Learning and Information Processing*, pages 367–381, Singapore. Springer Singapore.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 135–143, New York, NY, USA. Association for Computing Machinery.

Ramadhani Ally Duma, Zhendong Niu, Ally S. Nyamawe, Jude Tchaye-Kondi, Nuru Jingili, Abdulganiyu Abdu Yusuf, and Augustino Faustino Deve. 2024. Fake review detection techniques, issues, and future research directions: a literature review. *Knowledge and Information Systems*, 66(9):5071–5112.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Mohammed Ennaouri, Mohamed Raoui, Ahmed Zellou, and Moulay Hafid El Yazidi. 2024. Fake review detection using extravagant words: A comparative study of k-means, k-mode, and hierarchical classification. In *2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, pages 1–7.

K. Ezzameli and H. Mahersia. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847.

Roger Ferdinan, Krestella Margareta, Stevan Christyan, and Said Achmad. 2024. Utilizing inter-annotator agreement for effective fake-review detection in e-commerce. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 953–958.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from imbalanced data sets*, volume 10. Springer.

Alessandro Gambetti and Qiwei Han. 2023. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media.

globenewswire. 2022. Brand rated: "nine out of ten customers read reviews before buying a product". *GlobeNewswire News Room*.

Dhruvi D Gosai, Himangini J Gohil, and Hardik S Jayswal. 2018. A review on a emotion detection and recognization from text using natural language processing. *International journal of applied engineering Research*, 13(9):6745–6750.

Wilfredo Graterol, Jose Diaz-Amado, Yudith Cardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino. 2021. Emotion detection for social robots based on nlp transformers and an emotion ontology. *Sensors*, 21(4).

Jia Guo. 2022. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, 31(1):113–126.

Priyanka Gupta, Shriya Gandhi, and Bharathi Raja Chakravarthi. 2021. Leveraging transfer learning techniques- bert, roberta, albert and distilbert for fake review detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 75–82.

Priyanka Gupta, Shriya Gandhi, and Bharathi Raja Chakravarthi. 2022. Leveraging transfer learning techniques- bert, roberta, albert and distilbert for fake review detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 75–82, New York, NY, USA. Association for Computing Machinery.

Petr Hajek, Lubica Hikkerova, and Jean-Michel Sahut. 2023. Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research*, 167:114143.

Petr Hajek and Jean-Michel Sahut. 2022. Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection. *Technological Forecasting and Social Change*, 177:121532.

Jameson L. Hayes, Brian C. Britt, William Evans, Stephen W. Rush, Nathan A. Towery, and Alyssa C. Adamson. 2021. Can social media listening platforms' artificial intelligence be trusted? examining the accuracy of crimson hexagon's (now brandwatch consumer research's) ai-driven analyses. *Journal of Advertising*, 50(1):81–91.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):38–50.

Katie Hoemann, Catie Nielson, Ashley Yuen, JW Gurera, Karen S Quigley, and Lisa Feldman Barrett. 2021. Expertise in emotion: A scoping review and unifying framework for individual differences in the mental representation of emotional experience. *Psychological bulletin*, 147(11):1159—1183.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yue Huang and Lichao Sun. 2024. Fakegpt: Fake news generation, explanation and detection of large language models.

Zhao Huiqi, Abdullah Khan, Xu Qiang, Shah Nazir, Yasir Ali, and Farhad Ali. 2021. Mcdm approach for assigning task to the workers by selected features based on multiple criteria in crowdsourcing. *Scientific Programming*, 2021:4600764.

Petr Hájek, Aliaksandr Barushka, and Michal Munk. 2020. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 32.

Meherunnesa Ibnath, Khan Hasib, Md Parvez, and M. Ph. D. 2025. Enhancing multi-emotion detection in text: A comparative study of feature extraction techniques and machine learning models. pages 1–6.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews.

Abdullah Ilyas, Khurram Shahzad, and Muhammad Kamran Malik. 2023. Emotion detection in code-mixed roman urdu - english text. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 219–230, New York, NY, USA. Association for Computing Machinery.

Rashid Kamal, Munam Ali Shah, Carsten Maple, Mohsin Masood, Abdul Wahid, and Amjad Mehmood. 2019. Emotion classification and crowd source sensing; a lexicon based approach. *IEEE Access*, 7:27124–27134.

Nithin M Kannal, N Asmathunnisa, and Jagadish S Kallimani. 2024. Impacts of fake reviews on dietary supplements and healthcare products in social media. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–4.

Gurpreet Kaur and Kamal Malik. 2021. A comprehensive overview of sentiment analysis and fake review detection. In *Mobile Radio Communications and 5G Networks*, pages 293–304, Singapore. Springer Singapore.

Shagufta Khalif and Kishor Mane. 2024. Exploring machine learning and deep learning techniques for fake review detection: A comprehensive literature review. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 2:1669–1677.

Md. Imran H. Khan, Shyam S. Sablani, M. U. H. Joardder, and M. A. Karim. 2022. Application of machine learning-based approach in food drying: opportunities and challenges. *Drying Technology*, 40(6):1051–1067.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. WWW '19, New York, NY, USA. Association for Computing Machinery.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Thi-Kim-Hien Le, Yi-Zhen Li, and Sheng-Tun Li. 2022. Do reviewers' words and behaviors help detect fake online reviews and spammers? evidence from a hierarchical model. *IEEE Access*, 10:42181–42197.

Sophia Yat Mei Lee. 2019. Emotion and cause. Springer Singapore.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. 2012. Cdas: A crowdsourcing data analytics system. *Proc. VLDB Endow.*, 5(10):1040–1051.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhifang Lu. 2023. Research on python crawling algorithm in model data visualization. In *2023 World Conference on Communication Computing (WCONF)*, pages 1–6.

Bundit Manaskasemsak, Jirateep Tantisuwankul, and Arnon Rungsawang. 2023. Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. *Neural Computing and Applications*, 35(2):1169–1182.

Juan María Martínez Otero. 2021. Fake reviews on online platforms: perspectives from the us, uk and eu legislations. *SN Social Sciences*, 1(7):181.

Sushil Kumar Maurya, Dinesh Singh, and Ashish Kumar Maurya. 2023. Deceptive opinion spam detection approaches: a literature survey. *Applied Intelligence*, 53(2):2189–2234.

Arvind Mewada, Sushil Kumar Maurya, Mohd. Aquib Ansari, Om Prakash Sharma, Suman Avdhesh Yadav, and Shahnawaz Ahmad. 2025. Deceptive opinion detection using stacking-based deep ensemble learning. In *2025 3rd International Conference on Disruptive Technologies (ICDT)*, pages 1614–1617.

Abrar Qadir Mir, Furqan Yaqub Khan, and Mohammad Ahsan Chishti. 2023. Online fake review detection using supervised machine learning and bert model.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017b. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Rami Mohawesh, Haythem Bany Salameh, Yaser Jararweh, Mohannad Alkhalaileh, and Sumbal Maqsood. 2024. Fake review detection using transformer-based enhanced lstm and roberta. *International Journal of Cognitive Computing in Engineering*, 5:250–258.

Animesh Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013a. What yelp fake review filter might be doing? *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 7:409–418.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013b. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 632–640, New York, NY, USA. Association for Computing Machinery.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 191–200, New York, NY, USA. Association for Computing Machinery.

Subhabrata Mukherjee, Sourav Dutta, and Gerhard Weikum. 2016. Credible review detection with limited information using consistency features. In *Machine Learning and Knowledge Discovery in Databases*, pages 195–213, Cham. Springer International Publishing.

Irina Nalis and Julia Neidhardt. 2023. Not facial expression, nor fingerprint – acknowledging complexity and context in emotion research for human-centered personalization and adaptation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 325–330, New York, NY, USA. Association for Computing Machinery.

Kichan Nam, Jeff Baker, Norita Ahmad, and Jahyun Goo. 2020. Dissatisfaction, disconfirmation, and distrust: an empirical examination of value co-destruction through negative electronic word-of-mouth (ewom). *Information Systems Frontiers*, 22(1):113–130.

Quang Nguyen, Hai-Bang Ly, Ho Lanh, Nadhir Al-Ansari, Hip Lê, Tran Van Quan, Indra Prakash, and Binh Pham. 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Brussels, Belgium. Association for Computational Linguistics.

Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on paid crowdsourcing platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Jenny Pak. 2020. Integrating psychology, religion, and culture: The promise of qualitative inquiry. *Brill Research Perspectives in Religion and Psychology*, 2:1–86.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Himangshu Paul and Alexander Nikolaev. 2021. Fake review detection on online e-commerce platforms: A systematic literature review. *Data Min. Knowl. Discov.*, 35(5):1830–1881.

Sancheng Peng, Rong Zeng, Hongzhan Liu, Guanghao Chen, Ruihuan Wu, Aimin Yang, and Shui Yu. 2021. Emotion classification of text based on bert and broad learning system. In *Web and Big Data*, pages 382–396, Cham. Springer International Publishing.

Mathivanan Periasamy, Rohith Mahadevan, Bagiya Lakshmi S, Raja CSP Raman, Hasan Kumar S, and Jasper Jessiman. 2024. Finding fake reviews in e-commerce platforms by using hybrid algorithms.

Bhakti Pithava, Abhay Magar, and Santosh Bharti. 2024. Unveiling sentiment dynamics: Emotion detection in social media. In *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, pages 1–6.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anima Pradhan, Manas Ranjan Senapati, and Pradip Kumar Sahu. 2023. Comparative analysis of lexicon-based emotion recognition of text. In *Machine Learning, Image Processing, Network Security and Data Sciences*, pages 671–677, Singapore. Springer Nature Singapore.

Raffaele Pugliese, Stefano Regondi, and Riccardo Marini. 2021. Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, 4:19–29.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*.

Dandan Qiao and Huaxia Rui. 2023. Text performance on the vine stage? the effect of incentive on product review text quality. *Information Systems Research*, 34(2):676–697.

Tapasy Rabeya, Sanjida Ferdous, Himel Suhita Ali, and Narayan Ranjan Chakraborty. 2017. A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–7.

Umar Rashid, Muhammad Waseem Iqbal, Muhammad Akmal Skiandar, Muhammad Qasim Raiz, Muhammad Raza Naqvi, and Syed Khuram Shahzad. 2020. Emotion detection of contextual text using deep learning. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–5.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 985–994, New York, NY, USA. Association for Computing Machinery.

Paul Rayson, Dawn Archer, Scott Piao, and Tony Mcenery. 2004. The ucrel semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, Lisbon, Portugal, 2004*, pages 7–12.

Ravula Tarun Reddy and Suja Palaniswamy. 2024. Hidden emotion detection using speech, text, facial expressions and neural networks. In *2024 5th International Conference on Communication, Computing Industry 6.0 (C2I6)*, pages 1–6.

Adrián Reviriego and Ralitza Raynova. 2024. Naïveve bayes and logistic regression for sentiment analysis and emotion detection from text. In *2024 XXXIII International Scientific Conference Electronics (ET)*, pages 1–6.

Johnovon Richards, Saumya Dabhi, Faryaneh Poursardar, and Sampath Jayarathna. 2023. Poster: Leveraging data analysis and machine learning to authenticate yelp reviews through user metadata patterns. In *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, MobiHoc '23, page 580–582, New York, NY, USA. Association for Computing Machinery.

Néstor Rodríguez, David López, Alberto Fernández, Salvador García, and Francisco Herrera. 2021. Soul: Scala oversampling and undersampling library for imbalance classification. *SoftwareX*, 15:100767.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Ahmed Abdullah Alqarni Nizar Alsharif Theyazn H. H. Aldhyani Fawaz Waselallah Alsaade Osamah I. Khalaf Saleh Nagi Alsubari, Sachin N. Deshmukh. 2022. Data analytics for the identification of fake reviews using supervised learning. *Computers, Materials & Continua*, 70(2):3189–3204.

Mohammad Salehan and Dan J. Kim. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, 81:30–40.

Saba Salehi-Esfahani and Ahmet Bulent Ozturk. 2018. Negative reviews: Formation, spread, and halt of opportunistic behavior. *International Journal of Hospitality Management*, 74:138–146.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon gyo Jung, and Bernard J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Wataru Sato, Sylwia Hyniewska, Kazusa Minemoto, and Sakiko Yoshikawa. 2019. Facial expressions of basic emotions in japanese laypeople. *Frontiers in Psychology*, 10.

Supriya Saxena. 2025. Transformers (bert) based framework for web recommendations using sentiment-enriched web data. *Journal of Information Systems Engineering and Management*, 10:445–455.

Brita Schemmann, Andrea M. Herrmann, Maryse M.H. Chappin, and Gaston J. Heimeriks. 2016. Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development. *Research Policy*, 45(6):1145–1154.

Klaus R. Scherer. 2022. Theory convergence in emotion science is timely and realistic. *Cognition and Emotion*, 36(2):154–170.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Bjorn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st edition. Wiley Publishing.

Alfonso Semeraro, Salvatore Vilella, and Giancarlo Ruffo. 2021. Pyplutchik: Visualising and comparing emotion-annotated corpora. *PLOS ONE*, 16(9):1–24.

G. M. Shahariar, Swapnil Biswas, Faiza Omar, Faisal Muhammad Shah, and Samiha Binte Hassan. 2019. Spam review detection using deep learning. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, page 0027–0033. IEEE.

Md Shajalal, Md Atabuzzaman, Alexander Boden, Gunnar Stevens, and Delong Du. 2024. What matters in explanations: Towards explainable fake review detection focusing on transformers.

Siddhanth Chandrahas Shetty. 2019. Learning to detect fake online reviews using readability tests and text analytics. Master's thesis, Dublin, National College of Ireland, August. Submitted.

Ali Simaei, Rudy Hirschheim, and Helmut Schneider. 2023. Idea crowdsourcing platforms for new product development: A study of idea quality and the number of submitted ideas. *Decision Support Systems*, 175:114041.

Digvijay Singh, Minakshi Memoria, and Rajiv Kumar. 2023. Deep learning based model for fake review detection. In *2023 International Conference on Advancement in Computation Computer Technologies (InCACCT)*, pages 92–95.

Mai Magdy M. Sleim. 2022. Sentiments and cognition interdependence: An exploratory study of sentiment analysis and image schema. In *2022 20th International Conference on Language Engineering (ESOLEC)*, volume 20, pages 1–5.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Katie Spoon, Hsinyu Tsai, An Chen, Malte J. Rasch, Stefano Ambrogio, Charles Mackin, Andrea Fasoli, Alexander M. Friz, Pritish Narayanan, Milos Stanisavljevic, and Geoffrey W. Burr. 2021. Toward software-equivalent accuracy on transformer-based deep neural networks with analog memory devices. *Frontiers in Computational Neuroscience*, 15.

Sanja Stajner. 2021. Exploring reliability of gold labels for emotion detection in Twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1350–1359, Held Online. INCOMA Ltd.

Teodor Stoev, Kristina Yordanova, and Emma L. Tonkin. 2023. Experiencing annotation: Emotion, motivation and bias in annotation tasks. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 534–539.

Himanshu Suyal and Avtar Singh. 2024. Multilabel classification using crowdsourcing under budget constraints. *Knowledge and Information Systems*, 66(2):841–877.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. In *24th Annual Conference of the International Speech Communication Association*.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Nicole Amada Toby Lea and Henrik Jungaberle. 2020. Psychedelic microdosing: A subreddit analysis. *Journal of Psychoactive Drugs*, 52(2):101–112. PMID: 31648596.

Viet Trinh, Vikrant More, Samira Zare, and Sheideh Homayon. 2020. Quarantine deceiving yelp's users by detecting unreliable rating reviews.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Seppe vanden Broucke and Bart Baesens. 2018. *From Web Scraping to Web Crawling*, pages 155–172. Apress, Berkeley, CA.

Shivani Vora and Dr Mehta. 2024. Hcnnxgboost: A hybrid cnn-xgboost approach for effective emotion detection in textual data. *International Journal of Innovative Technology and Exploring Engineering*, 13:12–17.

Jessica G. J. Vuijk, Jeroen Klein Brinke, and Nikita Sharma. 2023. Utilising emotion monitoring for developing music interventions for people with dementia: A state-of-the-art review. *Sensors*, 23(13).

Erin Yirun Wang, Lawrence Hoc Nang Fong, and Rob Law. 2022a. Detecting fake hospitality reviews through the interplay of emotional cues, cognitive cues and review valence. *International Journal of Contemporary Hospitality Management*, 34(1):184–200.

Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. 2011. Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining*, pages 1242–1247.

Pengqi WANG, Yue LIN, and Junyi CHAI. 2023. Unmasking deception: A comparative study of tree-based and transformer-based models for fake review detection on yelp. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1848–1853.

Xiangwen Wang, Xianghong Lin, and Xiaochao Dang. 2020. Supervised learning in spiking neural networks: A review of algorithms and evaluations. *Neural Networks*, 125:258–280.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022b. A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83-84:19–52.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 823–831, New York, NY, USA. Association for Computing Machinery.

Zongben Xu, Niansheng Tang, Chen Xu, and Xueqi Cheng. 2021. Data science: connotation, methods, technologies, and development. *Data Science and Management*, 1(1):32–37.

Xinting Yang, Song Zhang, Jintao Liu, Qinfeng Gao, Shuanglin Dong, and Chao Zhou. 2021. Deep learning for smart fish farming: applications, opportunities and challenges. *Reviews in Aquaculture*, 13(1):66–90.

Ying Yang, Ying yi Hong, and Jeffrey Sanchez-Burks. 2019. Emotional aperture across east and west: How culture shapes the perception of collective affect. *Journal of Cross-Cultural Psychology*, 50(6):751–762.

Samira Zad, Maryam Heidari, James H Jr Jones, and Ozlem Uzuner. 2021. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261.

Allison Zengilowski, Brendan A. Schuetze, Brady L. Nash, and Diane L. Schallert. 2021. A critical review of the refutation text literature: Methodological confounds, theoretical problems, and possible solutions. *Educational Psychologist*, 56(3):175–195.

Fuzhi Zhang, Shuai Yuan, Peng Zhang, Jinbo Chao, and Hongtao Yu. 2022. Detecting review spammer groups based on generative adversarial networks. *Information Sciences*, 606:819–836.

Shunxiang Zhang, Aoqiang Zhu, Zhu Guangli, Wei Zhongliang, and Li KuanChing. 2023a. Building fake review detection model based on sentiment intensity and pu learning. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–14.

Wen Zhang, Qiang Wang, Jian Li, Zhenzhong Ma, Gokul Bhandari, and Rui Peng. 2023b. What makes deceptive online reviews? a linguistic analysis perspective. *Humanities and Social Sciences Communications*, 10(1):769.

Zheng Zhang, Jun Wan, Mingyang Zhou, Zhihui Lai, Claudio J. Tessone, Guoliang Chen, and Hao Liao. 2023c. Temporal burstiness and collaborative camouflage aware fraud detection. *Information Processing Management*, 60(2):103170.

Ying Zhen, Abdullah Khan, Shah Nazir, Zhao Huiqi, Abdullah Alharbi, and Sulaiman Khan. 2021. Crowdsourcing usage, task assignment methods, and crowdsourcing platforms: A systematic literature review. *Journal of Software: Evolution and Process*, 33(8):e2368.

Saide Zhu, Zhipeng Cai, Huafu Hu, Yingshu Li, and Wei Li. 2020. zkcrowd: A hybrid blockchain-based crowdsourcing platform. *IEEE Transactions on Industrial Informatics*, 16(6):4196–4205.

Saide Zhu, Huafu Hu, Yingshu Li, and Wei Li. 2019. Hybrid blockchain design for privacy preserving crowdsourcing platform. In *2019 IEEE International Conference on Blockchain (Blockchain)*, pages 26–33.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

Şule Öztürk Birim, Ipek Kazancoglu, Sachin Kumar Mangla, Aysun Kahraman, Satish Kumar, and Yigit Kazancoglu. 2022. Detecting fake reviews through topic modelling. *Journal of Business Research*, 149:884–900.