# Is it Offensive or Abusive? An Empirical Study of Hateful Language Detection of Arabic Social Media Texts

**Salim Al Mandhari, Mo El-Haj** and **Paul Rayson**
UCREL, School of Computing and Communications
Lancaster University, UK
{s.m.almandhari, m.el-haj, p.rayson}@lancaster.ac.uk

## Abstract

Among many potential subjects studied in Sentiment Analysis, widespread offensive and abusive language on social media has triggered interest in reducing its risks on users; children in particular. This paper centres on distinguishing between offensive and abusive language detection within Arabic social media texts through the employment of various machine and deep learning techniques. The techniques include Naïve Bayes (NB), Support Vector Machine (SVM), fastText, keras, and RoBERTa XML multilingual embeddings, which have demonstrated superior performance compared to other statistical machine learning methods and different kinds of embeddings like fastText. The methods were implemented on two separate corpora from YouTube comments totalling 47K comments. The results demonstrated that all models, except NB, reached an accuracy of 82%. It was also shown that word *tri-grams* enhance classification performance, though other tuning techniques were applied such as *TF-IDF* and *grid-search*. The linguistic findings, aimed at distinguishing between offensive and abusive language, were consistent with machine learning (ML) performance, which effectively classified the two distinct classes of sentiment: offensive and abusive.

## 1 Introduction

Social media streams such as X (previously known as Twitter) and YouTube apply individual policies to control the content posted by Internet users (Kolla et al., 2024). Despite having in place automatic methods, there is no guarantee that all the unsuitable content is detected. That is because it is challenging to completely filter out all slang, misspelling, and dialectal terms. The filters, based on sentiment analysis techniques, detect the targeted content typically depend on keyword lists, rule-based approaches, and machine learning algorithms to classify sentiments.

The significance of sentiment analysis escalates with the growth of unsuitable content disseminated across social media platforms on a daily basis. It is a necessary pre-requisite for categorising personal opinions into positive, negative, or neutral. Accordingly, this paper concentrates on establishing consistent definitions of unsuitable language prevalent on social media discourse and developing a robust sentiment analysis classifier to detect offensive and abusive language in Arabic.

Classifying positive and negative text poses plenty of challenges. However, one of the most difficult challenges is dealing with dialectal terms in Arabic. Millions of Arab users on social media use a combination of Modern Standard Arabic (MSA) and dialects to post their opinions. The complexity of Arabic dialects is underscored by their common linguistic characteristics intertwined with polysemous words that share identical structures but harbour multiple meanings, which can influence the classification process in machine learning. For instance, the verb يُبَقِّي (*yubaqi*), typically meaning "to remain" or "to keep in one's possession", acquires a distinct interpretation in Omani dialects, particularly in Buhla and Al-Hamra[1].

The influence of polysemous words among dialects is not a major problem in a non-offensive context. However, it is a significant issue when it occurs in sensitive contexts such as gender, religion and race (Khan et al., 2024). The following sentence: وتجيك الممحونه تقول هذا فن = "wetjík al-mamúnah taqúl hadhá fann"- which translates to "A lustful comes to you to say that this is a kind of art" is found in the corpus used for training a classifier in this study. The adjective ممحون (*mamhún*) signifies "afflicted" in Modern Standard Arabic (MSA) and in many Arabic dialects. However, in this context, it takes on the meaning of "libidinous",

[1]For more details about Buhla, Al-Hamra, and other Omani dialects, see: https://en.wikipedia.org/wiki/Omani_Arabic

influenced by specific Gulf dialects.

Accordingly, disseminating offensive, obscenities, profanities and insulting content on social media by using informal or dialectal language may possibly constitute a challenge to classifiers trained on Sentiment Analysis to detect harmful content. Inexperienced social media users such as children and teenagers could be affected by viewing undetected abusive and offensive sentiments. Making consistent definitions for offensive, abusive and clean content by specifying their exact linguistic and cultural features is a significant technical challenge in terms of classifying the three classes by means of the existing tools for Arabic text.

As Roache (2019) commented, there is no clear explanation regarding why offense is an inappropriate way to behave. It is enough to say that it is not part of the culture and moral rules. Coughlan (2016), gathered interesting results in a case study showing that three-quarters of social media users aged between 11 to 12 years old had faked their ages to browse adult content, while two-thirds of children did not report offensive content on social media. The results of this behaviour go in par with the findings of Millwood-Hargrave (2000), who states that the use of strong language by children may possibly reflect on their ability to be ethical and responsible parents in the future.

### 1.1 Contributions

The main objective of this study is to build an offensive and abusive language detection classifier which is robust to the challenges of the mixed texts containing MSA and dialectal Arabic which is commonly used on social medial platforms. To deal with the limitations of the previous studies, we have built an efficient detection approach for offensive and abusive language. The main contributions of this study are as follows:

- Our study confirms the capability of machine learning models, embeddings, and libraries to differentiate offensive and abusive sentiments including MSA and Arabic dialects through a thorough exploration of the linguistic delineations between these terms.

- We trained several machine learning models, including Naive Bayes (NB), Support Vector Machine (SVM), fastText, Keras, and multilingual RoBERTa XML embeddings, using various features. Across these models, we achieved an accuracy of 82% in the majority of cases.

- Ascertaining the advantages and disadvantages of cleaning and pre-processing data in relation to enhancing the classification performance.

- Examination of an open access multi-class dataset including labelled offensive, abusive, and clean classes. It contains 32K comments in MSA and dialects collected from Aljazeera channel on YouTube encompassing a range of subjects including politics, society, and economics.

The rest of the paper is organised as follows, in section 2 we describe the `Related Work`. In section 3, we describe the `Data`. Section 4 explains the study `Methodology`, following by section 5 to release the `Results` and section 6 for the `Conclusion`.

## 2 Related work

### 2.1 Related terms in previous studies

As described in this section, the literature of hate language detection within Sentiment Analysis field reveals that some of the most popular terms used to represent this kind of language on social media are offensive, abusive, cyberbullying, and swearing. It appears that there is no such agreement in the literature to define these terms. The next subsections show evidence of this disagreement.

#### 2.1.1 Diversity in meaning

Significant research in Sentiment Analysis proves that offensive language represents diversity in meaning to a language that includes remarks of obscene, inflammatory and profane targeting peoples race, religion, nationality, and gender (Alsafari et al., 2020), (Mubarak et al., 2017), (Mubarak et al., 2020), (Jay and Janschewitz, 2008), (Abozinadah et al., 2015), (Abozinadah, 2017), (Waseem et al., 2017). According to that, offensive language seems to be a synonym for hate and aggressive speech. They share common features that use strong language in discussions without paying attention to peoples emotions. Mubarak et al. (2017), studied Arabic abusive speech, and excluded the sort of language that promotes hateful words and named it offensive. Practically, it showed that combining SeedWords (SW) with Log Odds Ratio (LOR) by using word unigram outperformed

the others and obtained an F-score of 0.60, which is not a promising performance. The term 'abusive language appearing in (Mubarak et al., 2017) studies, is a language that uses vulgar and obscene words on social media. In fact, this differentiation between offensive and abusive language is based on three classes of offensive speech identified by Jay and Janschewitz (2008); specifically, vulgar, pornographic and odious.

Abozinadah et al. (2015); Abozinadah (2017) and (2017), also studied abusive language found on Arabic social media and defined the term "abusive language" as what produces obscenity, profanity, or insulting words, which seems to be more related to sexual content. The study focused on detecting abusive Twitter accounts that distribute adult content in Arabic tweets.

The trained algorithms for the model were SVM, NB and decision trees (J48). The results demonstrated that the NB classifier with 10 tweets and 100 features obtained the best performance, with an average accuracy of 90%.

Waseem et al. (2017)'s study, uses the term 'abusive' to refer to both aggressive and sexual content such as cyberbullying, trolling, racial categories, and sexual orientations.

#### 2.1.2 Unity in meaning

A significant number of studies have adapted (Jay and Janschewitz, 2008), the concept of offensive language, again including vulgar, pornographic, and hateful speeches. Based on that, there is no considerable need to have a special term for abusive speech regarding sexual content.

Alakrot et al. (2018), used the term offensive language as the main term which may include various forms of inflammatory language, profanities, obscenities and insults in aggressive and sexual contexts. An SVM classifier was trained on this dataset in multiple stages by using word-level and N-gram features. The study determined that the pre-processing steps and features returned a better accuracy of 90.05% than others reported in the literature related to the classification of Arabic text. Unlike in this study, where pre-processing did not yield a significant alteration in the results.

Mouheb et al. (2018), used the term offensive language to refer to harassing messages that include rude, insulting and life-threatening texts. This study proposed a cyberbullying detector for Arabic comments on YouTube and Twitter based on a dataset of bullying and aggressive keywords. It weighted the bullying comments according to their strength into three categories; specifically, mild, medium and strong in order to help determine the best action to take against bullying comments. The study reported that the proposed detector could accurately detect most of the bullying comments without applying statistical tests for evaluation.

Mathur et al. (2018), used the term offensive speech to cover hate speech and abusive speech which includes sexual content. Even though the study differentiates between offensive and abusive language as aggressive and sexual respectively, it denies an offensive term as an umbrella for both terminologies. The model uses a transfer learning technique on pre-trained CNN architecture to classify two datasets: English and Hinglish (without transfer learning (w/o TFL) and with transfer learning (TFL). It is concluded that the model significantly improved the results with TFL demonstrating an accuracy of 83.90%, which surpassed the results of the English dataset.

#### 2.1.3 Interchange in meaning

It is also worth mentioning that other terms interchange in meaning with offense and abuse. The cyberbullying term has a presence in the literature also. The principal difference between it and other terms like offensive and abusive is that offensive and abusive contents generally describe texts comprising bad language, whereas cyberbullying is a more general description of texts comprising bad language, images and videos.

NCPC (2019), explores cyberbullying as the use of different technologies, for example, cell phones, video games and the Internet to post a threat, an embarrassing video or image, or a rumor about someone.

Miller and Hufstedler (2009) and Beale and Hall (2007), clarify that electronic bullying, online bullying, and/or cyberbullying are new strategies of bullying including forms of bullying considered as harassment using technology, such as mobile phone texting and cameras, email, social media websites (MySpace, Facebook, etc.), chat rooms, picture messages (involving sexting), blogs and/or IM (instant messages).

Haidar et al. (2017), resumes from the same previous definition of cyberbullying and designed a machine learning system to detect and stop ongoing cyberbullying attacks for Arabic and

English languages. Seeing that no other work had been completed on Arabic cyberbullying prior to this paper, it is the first study that proposed a system to solve Arabic cyberbullying problems. The study utilised NB and SVM classifiers for binary classification by using a WEKA toolkit. The results showed that SVM outperformed NB in overall classification including classified and misclassified instances. The highest F-score was 0.927.

Based on what is mentioned here, anti-social behavioural language is studied by using various concepts. Certain studies use offensive language to describe hateful and aggressive speech only, and use abusive language to describe sexual speech. Others follow Jay and Janschewitz (2008), concept of offensive language that include both hateful and sexual speech. The literature describes other terms used to describe the targeted language that are cyberbullying and swearing. It is reasonably hard to adapt certain concepts without returning to the dictionaries to establish the linguistic potential and original meanings of the mentioned terminologies. Therefore, the following part will discuss the terms found in the dictionaries.

### 2.2 Linguistic perspective

Starting with the terms; offense and abusive, Collins (2019), links “offense" to any public wrong or crime, attack, and assault. It also identifies it as a behaviour that causes people to be upset or embarrassed such as: The book might be published without creating offense. The adjective “offensive" therefore is something that upsets or embarrasses people because it is rude or insulting. Such as; “some friends of his found the play horribly offensive". The dictionary mentions that using the word indicates how angry the person is about something. It can be inferred from this that offensive language is more related to a language that seems to be hateful and aggressive. A good point to demonstrate here is that “offensive language" is not related only to sexual content as other terms like “abusive language" for example. This conclusion is in accordance with Mubarak et al. (2017)’s understanding of offensive language.

Moving to understanding the terms; abuse and abusive, Cobuild and of Birmingham (2003), provides two meanings for abuse; specifically special and general. It mentions that abuse can be directed at the sexual treatment of someone and it is cruel and violent treatment. It can be said from there, victims of sexual and physical abuse. Sex, therefore, is related to the concept of abuse.
Whereas, general abuse offers a general meaning for extremely rude and insulting things which a person may say when he or she is angry. For example, I was left shouting abuse as the car sped off.
The adjective of abuse is regularly used to describe certain content that is extremely rude and insulting by expressing abusive language. Abusive language appears to have a higher degree of assault than offensive language. Hence, there is little surprise that it is linked more to sexual content and any sort of behaviour that is deemed to be unacceptable in society.

The last terminology to identify in this section is obscene. Cobuild and of Birmingham (2003), demonstrates that obscene is close to abuse in meaning. Both share relevant semantics that relate to sex or violence occurring in shocking and unpleasant offensive way. For example, He continued to use obscene language and also to make threats.

Consequently, offensive and abusive language is similar to each other in terms of being adjectives for texts consisting of bad language. However, offensive is more likely to include inflammatory language, profanities, obscenities and insults, whereas abusive language is more likely to include obscene and sexual insults.

### 2.3 Offensive language in Arabic

Many research studies on the detection of offensive and abusive language have been conducted on English datasets but only a small number on Arabic due to its morphological complexity and limitation regarding software support for Arabic (Abozinadah et al., 2015). Several cases in Arabic build its complexity while dealing with software as digital content. The following are common challenging cases: free word order, gendered pronouns, dual subject, and lemmatisation (Salem et al., 2008), (Aoun et al., 2009), (Muaad et al., 2023). In Sentiment Analysis and inappropriate language prevalent on social media platforms, there is a misinterpretation of numerous Modern Standard Arabic (MSA) and dialectical terms in Arabic. Consequently, classifiers encounter difficulties in accurately categorizing content as offensive, abu-

sive, bullying, or clean.

This research decides to take all these challenges and attempt to work on a corpus of YouTube Arabic comments and implement a different ML algorithm to detect offensive and abusive language in the corpus.

## 3 Data

By reviewing the literature about offensive language, two separate datasets from previous studies form the corpus of this paper; Alakrot et al. (2018) and Mubarak et al. (2017). Table 1 shows details of both datasets including each class size[2]. Alakrot's dataset contains 15,050 annotated comments by three annotators.

| Dataset | Alakrot | Mubarak |
|---|---|---|
| Size | 15,050 | 32,000 |
| Source | YouTube | YouTube |
| Off. class | 39% | 79% |
| Non-off./clean class | 71% | 19% |
| Obscene size | NA | 2% |

Table 1: Details of Alakrot and Mubarak datasets

It was collected from various YouTube channels in an effective way, where the videos uploaded on those channels display celebrities in controversial footage with the aim of provoking viewers to use strong language in response. This led to a rich corpus of offensive words being collected. The annotation which is binary has only two classes: offensive and non-offensive. The inter-annotation agreement is reasonably good (71%). A strong point in this study is that it did not collect the data based on predefined profane words as the previous studies have done, for the reason that it lessens the ability of the predictivity of the tools proposed.
Despite the fact that it being highlighted as the largest dataset in tackling Arabic offensive language, it appears that the dataset utilised by Mubarak et al. (2017), is larger than Alakrots. Mubaraks includes 1100 tweets and 32K comments collected from the Aljazeera channel on YouTube covering various topics, such as politics, society, the economy and science. The annotation classes are obscene, offensive or clean. The inter-annotation agreement is relatively high, 87%.

[2]The datasets are publicly available at: https://github.com/EtcoNLP/Offensive-detection.git

Moving the argument along, remarkably, there are certain offensive ideas that are found in many discussions on Arabic social media regardless of what the users are commenting about. For example, it is common to notice offensive remarks on ideas relating to the Sunni-Shii conflict[3], complaints about terrorism and comparing people to Jews when their behaviour is very poor.

### 3.1 Pre-processing

Text pre-processing is an essential step to start with the data in the text mining field. Regardless of the field of research, it may include different techniques to split or clean the text, such as tokenisation, segmentation, normalisation, filtering and part of speech tagging (Mathiak and Eckstein, 2004). In this paper, segmentation, normalisation and filtering were applied to manage some linguistic remarks that may negatively affect the accuracy of classification in the experimental section.

#### 3.1.1 Segmentation

Segmentation generally is splitting white-space delimited units in the text. The function of the segmenter is to perform stemming that is splitting each linked element from the stem of the word.
For morphological segmentation, this paper chose the Arabic-SOS tool: Segmentation, Stemming and Orthography Standardization for Classical and pre-Modern Standard Arabic (Mohammed, 2019) to conduct the segmentation. The Arabic SOS builder reported 98.47% of accuracy in comparison with other tools employed for Arabic segmentation, such as Mohamed, (2018) (96.8%), MADAMIRA (94.7%) and SAPA (86.47%).

On closer examination of this segmenter, it performed the job correctly in many cases to segment the stem from the article, feminine sign, and preposition such as (نفط + ال/ al + nefa/ oil), (مصفا + ة/ mesfá + h/ refinery) ,(نفط + ل + ل/ le + l + nefta/ for oil).

#### 3.1.2 Normalisation

What normalisation operations do is to unify common misspellings in writing to allow the classi-

[3]Both Sunni and Shia are the largest Islamic schools. Their conflict has deep historical roots and is fueled by political tensions between the two parties. To read more, see: https://shorturl.at/Z3b9I.

fier to recognise similar words that have only misspellings in a few of them. Misspellings confuse ML classification in the step of recognising the words. One word might be considered as different words because it has multiple spellings, which affect the accuracy of classification. Recently, there have been some attempts to normalise dialects by using pre-trained Transformer based models such as BERT (Alnajjar and Hämäläinen, 2024) and (Hämäläinen et al., 2022). In this paper, normalisation has been implemented using MSA orthography based. We, therefore, replaced characters such as, أ , آ , إ with ا, replacing ة with ه, replacing ي with ى, replacing اردوجان with اردوغان, and replacing انكليزيه with انجليزيه.

#### 3.1.3 Filtering

Filtering is removing diacritics, punctuation, commas, symbols and stop words that are prepositions, conjunctions and articles. The main function of filtering is to minimise the size of features in the dataset, otherwise there will be impediments in the classification process (Saad and Ashour, 2010). In this paper, a regular expressions method was used to filter the corpus of Latin strings, diacritics, symbols, stop words list provided by NLTK for Arabic (Bird et al., 2009).

## 4 Methodology

### 4.1 Methods and features

All the experiments conducted in this paper dealt with the two datasets separately. Mubarak et al. (2017)'s dataset will be called later (A) and Alakrot et al. (2018)'s dataset will be called (B). Our research investigates the efficacy of binary sentiment classification, distinguishing between offensive and non-offensive sentiments. Additionally, we address the challenge of multi-class sentiment classification, encompassing offensive, abusive, and clean languages. This paper used some machine learning models: Naïve Bayes (NB) by using Multinomial NB variant, Support Vector Machine (SVM) by using different variants: linear kernel, SGD Classifier, SVC and Radial basis function (rbf), and Fast Text word embedding[4]. We also run Keras ANN, ANN with embedding layer, embedding layer with max pooling, and ConcNets with max pooling. A deep learning model has also run which is Roberata XML multilingual embedding (Conneau et al., 2019). Roberata XML is a multilingual model trained on 100 different languages (including Arabic). It has proved to achieve significant performance gains for a wide range of classification tasks in languages other than English. The model was trained on four epochs and fine-tuned with the following hyperparameters: lr=2e-5, $\epsilon = 1e-8$. For the experiments on NB and SVM, two tuning techniques were selected: TF-IDF and Grid Search. TF-IDF (2019) denotes term frequency-inverse document frequency. It is commonly used for information retrieval and text mining. It evaluates how important a word is in a document or a corpus based on a statistical calculation. Another technique used in the NB and SVM experiments is Grid Search (Lutins, 2019) that scan the data to figure which parameters are the most appropriate for the model being employed.

Turning to the fastText method, the experiments included features of word n-gram *(-word Ngrams)* from 1 to 7 words to acquire the closest existing words for offensive and abusive language. Features also contain different experiments for epoch parameter *(-epoch)* from 5 to 5000, which controls the looping times of training over the data. While the default epoch is 5, the performance of the poor quality data might improve by increasing the looping times. A further parameter applied in fastText classification is learning rate *(-lr)*, which ranges from 0.1 to 1.0. It helps to fasten coverage to a solution by way of the model (FastText, 2019). To control the size of the vectors, we used *(-dim 300)*. Furthermore, independent binary classifiers *(-loss one-vs-all)* were used for each class in the dataset to handle multiple classes.

### 4.2 Evaluation

To examine the effectiveness of the three algorithms used in this paper to classify offensive, abusive and clean texts, a confusion matrix was utilised to demonstrate the accuracy of classification. It returns numbers concerning actual and predicted classifications carried out by the proposed classifiers (Patil et al., 2013). Table 2 provides an example of a confusion matrix for SVM implemented on Alakrots dataset in this paper.

The Table reveals that the total number of predicted instances is 4492; 728 instances are predicted as YES (offensive) and 1518 instances are predicted as NO (non-offensive). In reality, 1286 instances are YES (offensive) and 960 are NO (non-

[4]The models are publicly available at: https://github.com/EtcoNLP/Offensive-detection.git

| n= 4492 | Predicted: NO | Predicted: YES | |
|---|---|---|---|
| Actual: NO | TN = 1177 | FP = 109 | 1286 |
| Actual: YES | FN = 341 | TP = 619 | 960 |
| Total | 1518 | 728 | |

Table 2: An example of a confusion matrix for a binary classification

offensive). Three measures are utilised in this paper: precision, recall and F-score, in addition to the general measure, accuracy.

## 5 Results

To examine how pre-processing and stemming affect the classification performance positively or negatively, four versions of each dataset were examined in the implementation of NB and SVM. The versions included the following: pure dataset (pipeline), stemmed dataset, pre-processed dataset, in addition to the stemmed and pre-processed dataset. The best results were obtained by linear kernel and rbf kernel (B) by assigning various features for instance cache size (200), gamma (scale), and max iterations (-1). Their F-score is 79%. The results show that the improvement in performance of fastText is slightly higher than NB and SVM. Despite the fact those results are the best in each feature, the best F-score was obtained by using the word tri-gram feature.

### 5.1 Error analysis

We inspected the stems that were segmented by the Arabic-SOS segmenter incorrectly. The tool failed to recognise several cases in the segmentation such as the following:

1. While it segmented the appended *ya* of the present verb in some cases such as (حرق + ي/ ya + req / burn) and (وفق + ي/ yu + waffeq/ reconcile), it did not recognise it when another prefix occasionally comes before it such as (بيهرب/ beyahrub/ to escape) and (سيلعن/ sayalan/ will curse).

2. The tool struggled with missed spaces in between some words for example (لماوردفيحوار/ lemá warada fí ewár/ as stated in the dialogue).

3. Many words appeared to be segmented incorrectly. There is no obvious reason why they were segmented in this particular way. This occurred with MSA words and dialectal words.

| Rank | Freq | Keyness | Effect | Keyword |
|---|---|---|---|---|
| 1 | 1040 | + 109.75 | 0.0046 | |
| 2 | 337 | + 81.77 | 0.0015 | عملاء |
| 3 | 473 | + 77.37 | 0.0021 | |
| 4 | 374 | + 67.39 | 0.0016 | الصهاينه |
| 5 | 3819 | + 56.53 | 0.0166 | يا |
| 6 | 594 | + 56.37 | 0.0026 | هؤلاء |
| 7 | 303 | + 54.95 | 0.0013 | الصهيونى |
| 8 | 259 | + 54 | 0.0011 | لعنا |
| 9 | 156 | + 50.98 | 0.0007 | عميل |

Figure 1: A sample of the offensive keyword list

| Rank | Freq | Keyness | Effect | Keyword |
|---|---|---|---|---|
| 1 | 78 | + 211.72 | 0.0188 | ابن |
| 2 | 44 | + 210.5 | 0.0107 | كس |
| 3 | 192 | + 208.64 | 0.0434 | يا |
| 4 | 42 | + 170.03 | 0.0102 | |
| 5 | 35 | + 167.41 | 0.0085 | القحبه |
| 6 | 31 | + 139.56 | 0.0076 | امك |
| 7 | 27 | + 129.12 | 0.0066 | ولاد |
| 8 | 62 | + 127.6 | 0.0149 | قناه |
| 9 | 22 | + 105.2 | 0.0054 | العاهره |

Figure 2: A sample of the abusive keyword list

### 5.2 Keyword lists

We have generated two keyword lists for the most frequent words appearing in offensive and abusive sentences in the corpus. This could be beneficial for other applications in Arabic. Therefore, the AntConc system was utilised to analyse the corpus and generate keyword lists. Consequently, two keyword lists were generated:

1. A keyword list of offensive language based on Mubaraks definition of offense.

2. A keyword list of abusive language based on Mubaraks definition of abuse.

For ease of illustration, Figures 1 and 2 show examples of the two keyword lists. For the full lists, see: https://t.ly/PhbiC.

### 5.3 Discussion

The tests for offensive and abusive language classification demonstrate that all models obtained the same accuracy except NB, as shown in Table 3, where A is Mubarak's dataset and B is Alakrot's dataset.

In terms of how data quality affects classification performance, the implementation tests on different data show that raw data is reasonable enough to estimate the quality of classification. Even though pre-processing demonstrated improvement in the classification performance, the improvement in the best condition between the pre-processing data and

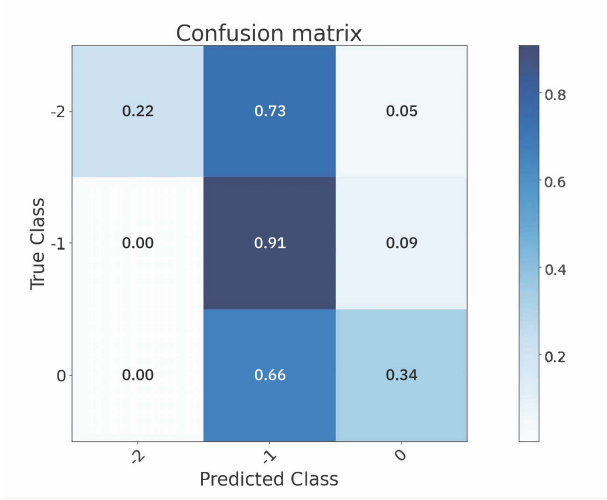


Figure 3: Confusion matrix for the multi-class classification

raw data is only by 4%. It is also worth mentioning that pre-processing is a time consuming expensive step.

During training the multi-class dataset on Keras sequential, we noticed how the model is fast learning the offensive language represented in 2 as (-1) and slow in learning the abusive language (-2). It is quite clear that the imbalance between the two classes affected the training process.

frequent a lot in offensive and abusive contexts. Words such as (يا= *ya*/ o)h), (ابن= *bn*/ son of) and (ولاد= *welad*/ sons of) are popular parts in offensive and abusive expressions in Arabic. Examples of that are: (ابن المتعة= *Ibn elmutá*/ the critiqued relationship known as mutaa), (يا حقير= *yá haqír*/ oh tacky), (ولاد الكلب= *weládel kalb*/ sons of dogs). Moreover, this classification of offensive and abusive words emphasises the definitions raised above concerning offense and abuse, where offense is a general assault, but abuse is a sexual assault. It is evident that the lists were able to distinguish between them.

## 6 Conclusion

This paper concentrated on classifying and distinguishing offensive and abusive language on social media, YouTube in particular. NB, SVM and fastText, keras, and Roberta XML multilingual embeddings algorithms were implemented on two separated datasets, binary and multi-class, comprising 47k comments in total, and demonstrated high performance in relation to classification. The fastText algorithm surpassed the others by achieving 82% accuracy. The tests on fastText confirmed that using the word tri-gram feature improves the accuracy of this classification topic.

| | Accuracy | | Recall | | Precision | | F-score | |
|---|---|---|---|---|---|---|---|---|
| Algorithm | A | B | A | B | A | B | A | B |
| NB | 0.81 | 0.78 | 0.81 | 0.78 | 0.82 | 0.79 | 0.74 | 0.77 |
| SVM | **0.82** | 0.80 | 0.82 | 0.80 | 0.81 | 0.80 | 0.76 | 0.79 |
| fastText | **0.82** | 0.76 | 0.82 | 0.76 | 0.82 | 0.76 | 0.82 | 0.76 |
| Logistic Regression | 0.81 | 0.80 | 0.95 | 0.80 | 0.84 | 0.80 | 0.81 | 0.80 |
| Roberta XML multilingual embeddings | **0.82** | | 0.82 | | 0.82 | | 0.82 | |
| Keras: Sequential | 0.82 | 0.76 | 0.80 | 0.76 | 0.80 | 0.76 | 0.82 | 0.76 |
| ANN with embedding layer | 0.81 | 0.77 | 0.81 | 0.77 | 0.81 | 0.77 | 0.81 | 0.77 |
| + max pooling | 0.81 | 0.77 | 0.81 | 0.77 | 0.81 | 0.77 | 0.81 | 0.77 |
| ConvNets + max pooling | **0.82** | 0.79 | 0.82 | 0.79 | 0.82 | 0.79 | 0.82 | 0.79 |

Table 3: A comparison in results among the three algorithms

The importance of keyword lists is not only in relation to gathering domain words, it is also vital to recognise the features of the language in the domain. Certain words in the lists are neither offensive nor abusive, such as (قناة = *qanh*/ channel) and some political leaders that are greyed out text in the sampled Figures 1 and 2 to prevent potential sensitivities. However, it reveals that these words

It is also important to note that ability of classifying offensive and abusive languages shown in the results tried to prove the definitions of offensive and abusive language agreed in the literature review, that language which contains hateful and aggressive remarks is offensive, whereas language that includes vulgar, pornographic and sexual remarks is abusive. However, lack of balance in the amount of offensive and abusive comments led to

lower accuracy.

In the future, we will work on collecting a multi-class dataset that is large enough and balanced to run more deep learning models to enhance classification.

## Limitations

While this study endeavors to advance Arabic text classification of offensive and abusive sentiments, several limitations have been acknowledged:

1. Limited data size: The two separate datasets utilized for training in this study are employed independently. If there was enough time, more experiments on the binary dataset could implemented to separate the offensive class into offensive and abusive, and then combine this dataset with the other one to have a singular, larger dataset with a more substantial number of instances for each class.

2. Limited error analysis: While the study includes comprehensive error analysis for the segmentation step outcomes, a lesser degree of analysis is devoted to the results of the machine learning (ML) experiments.

3. Limited neural network experiments: This work implements various classical ML and neural network models. However, Implementing more deep learning models and LLM might come up with better results.

## Ethics statement

Throughout data collection, experimentation, and analysis of this study, the ACL Ethics were upheld. We have taken careful attention to implement ethical guidelines regarding copyrights and intellectual property. We are committed to responsible research practices that contribute positively to the field of Natural Language Processing while prioritizing ethical standards.

## References


Ehab Abozinadah. 2017. *Detecting Abusive Arabic Language Twitter Accounts Using a Multidimensional Analysis Model*. Ph.D. thesis.

Ehab A Abozinadah, Alex V Mbaziira, and J Jones. 2015. Detection of abusive accounts with arabic tweets. *Int. J. Knowl. Eng.-IACSIT*, 1(2):113–119.

Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Towards accurate detection of offensive language in online communication in arabic. *Procedia computer science*, 142:315–320.

Khalid Alnajjar and Mika Hämäläinen. 2024. Normalization of arabic dialects into modern standard arabic using bert and gpt-2. *Journal of Data Mining & Digital Humanities*.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096.

Joseph E Aoun, Elabbas Benmamoun, and Lina Choueiri. 2009. *The syntax of Arabic*. Cambridge University Press.

Andrew V Beale and Kimberly R Hall. 2007. Cyberbullying: What school administrators (and parents) can do. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 81(1):8–12.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analysing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Collins Cobuild and University of Birmingham. 2003. *Collins Cobuild advanced learner's English dictionary*.

W Collins. 2019. *Collins English Dictionary*. Harper Collins: Glasgow.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

S Coughlan. 2016. Safer internet day: Young ignore social media age limit. *BBC News*.

FastText. 2019. What is fasttext? *[Online]. Available: https://fasttext.cc. [Accessed 04 September 2019]*.

Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. 2017. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6):275–284.

Mika Hämäläinen, Khalid Alnajjar, and Tuuli Tuisk. 2022. Help from the neighbors: Estonian dialect normalization using a finnish dialect generator. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 61–66.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Atif Khan, Abrar Ahmed, Salman Jan, Muhammad Bilal, and Megat F Zuhairi. 2024. Abusive language detection in urdu text: Leveraging deep learning and attention mechanism. *IEEE Access*.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation?

Evan Lutins. 2019. Grid searching in machine learning: Quick explanation and python implementation. *[Online]. Available: https://medium.com/@elutins/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596. [Accessed 06 September 2019].*

Brigitte Mathiak and Silke Eckstein. 2004. Five steps to text mining in biomedical literature. In *Proceedings of the second European workshop on data mining and text mining in bioinformatics*, volume 24.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

Jerold D Miller and Shirley M Hufstedler. 2009. Cyberbullying knows no borders. *Australian Teacher Education Association*.

Andrea Millwood-Hargrave. 2000. *Delete expletives?* Advertising Standards Authority London.

Sayyed Z. Mohammed, E. 2019. Arabic-sos: Segmentation, stemming, and orthography standardization for classical and pre-modern standard arabic. *[Forthcoming]*.

Djedjiga Mouheb, Rutana Ismail, Shaheen Al Qaraghuli, Zaher Al Aghbari, and Ibrahim Kamel. 2018. Detection of offensive messages in arabic social media communications. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 24–29. IEEE.

Abdullah Y Muaad, Shaina Raza, Usman Naseem, and Hanumanthappa J Jayappa Davanagere. 2023. Arabic text detection: a survey of recent progress challenges and opportunities. *Applied Intelligence*, 53(24):29845–29862.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

NCPC. 2019. Cyberbullying tip sheets. *[Online]. Available: shorturl.at/atADJ. [Accessed 07 August 2019].*

Tina R Patil, SS Sherekar, et al. 2013. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2):256–261.

R Roache. 2019. Offensive language. [Podcast].

Motaz K Saad and Wesam M Ashour. 2010. Arabic text classification using decision trees. *Arabic text classification using decision trees*, 2.

Yasser Salem, Arnold Hensman, and Brian Nolan. 2008. Implementing arabic-to-english machine translation using the role and reference grammar linguistic model.

TF-IDF. 2019. What does tf-idf mean? *[Online]. Available: http://www.tfidf.com. [Accessed 06 September 2019].*

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.