Who's the Fairest of Them All? A Comparison of Methods for Classifying Tone and Attribution in Earnings-related Management Discourse

Martin Walker[‡]

Steven Young[†]

Vasiliki Athanasakou[#]

Mahmoud El-Haj[*]

Paul Rayson[*]

Thomas Schleicher[‡]

Version: February 2020

Preliminary: Please do not circulate without permission

[†] Corresponding author: Lancaster University Management School, Lancaster University, UK. Email: s.young@lancaster.ac.uk. [‡] Alliance Manchester Business School, University of Manchester, UK. [#] University of Halifax, Nova Scotia, Canada. [*] School of Computing and Communications, Lancaster University, UK. We are grateful to Andrew Moore for excellent support training the learning classifiers. We are also grateful for comments and suggestions from seminar participants at the 2019 Summer Programme in Accounting Research (WHU) and the 2nd ESRC Workshop on Textual Analysis in Accounting and Finance (Lancaster). Financial support for the research was provided by the Economic and Social Research Council (contracts ES/J012394/1, ES/K002155/1 and ES/R003904/1) and the Research Board of the Institute of Chartered Accountants in England and Wales. The dataset of preliminary earnings announcement performance sentences and manual annotations used to train learning classifiers described in the paper, together with python code and guidelines for implementing the classifiers, are available at https://github.com/apmoore1/pea_classification.

Who's the Fairest of Them All? A Comparison of Methods for Classifying Tone and Attribution in Earnings-related Management Discourse

Abstract

We compare the relative and absolute performance of various machine learning algorithms and wordlists at replicating manual coding results for tone and attribution by domain experts in management performance commentary. Our suite of learning classifiers comprises Naïve Bayes, random forest, support vector machines, and an artificial neural network called multilayer perceptron. We use wordlists proposed by Henry (2006, 2008) and Loughran and McDonald (2010) to classify tone. Wordlists for attribution are based on the causal reasoning list from Language Inquirer and Word Count (LIWC), together with two self-constructed lists. We use a self-constructed wordlist to distinguish between internal and external attributions. We train learning classifiers using a large sample of manually annotated performance sentences. Results for all classifiers are assessed using a separate manually annotated holdout sample. Conclusions regarding the best classification method vary according to the classification task. None of the approaches are capable of identifying the presence of an attribution reliably. Even for more reliable classification tasks such as tone and attribution type, absolute measurement errors often exceed 20%. We conclude that while automated textual analysis methods offer important opportunities in certain settings, manual content analysis remains an essential tool for researchers interested in studying the properties and consequences of financial discourse.

Who’s the Fairest of Them All? A Comparison of Methods for Classifying Tone and Attribution in Earnings-related Management Discourse

1. Introduction

Large-sample computerized analysis of text is now commonplace in mainstream capital markets research. Work in accounting is dominated by automated content analysis methods that count word frequencies using predetermined wordlists relating to particular linguistic constructs such as tone, uncertainty or future tense. Loughran and McDonald (2016) and Henry and Leone (2016) stress the transparency and parsimony advantages of simple word counts, and question the net benefits of applying more sophisticated learning algorithms in a financial context. This view contrasts with theory and evidence in computational linguistics, where machine learning methods are associated with substantially better results (Pang et al., 2002). We seek evidence on the relative and absolute performance of various procedures for scoring financial discourse. Specifically, we compare the ability of wordlists and learning algorithms to replicate manual classification by domain experts for tone and attribution in management commentary.

A large literature using manual scoring methods predicts and finds evidence of optimism and self-serving attribution bias in managerial narrative disclosures (Merkl-Davies and Brennan 2007). Recent work using large-sample automated methods replicates and extends these findings. Indeed, El-Haj et al. (2019) conclude that tone and attribution remain among the most widely studied aspects of financial discourse in the new wave of automated textual analysis research. Despite their continuing popularity, however, accounting and finance researchers have applied little effort to evaluating the reliability of automated methods for measuring tone and attribution relative to manual scoring.

A small group of studies evaluate methods for measuring the tone of management disclosures. Loughran and McDonald (2010) compare the performance of General Inquirer's positive and negative wordlists derived from general English language with domain-specific lists constructed from firms' 10-K fillings. Results highlight the benefits of using wordlists derived from the financial domain to score the tone of 10-K commentary. Li (2010a) trains a Naïve Bayes classifier for tone on over 30,000 manually-coded forward-looking sentences drawn from 10-K filings. Validation tests reveal a classification accuracy rate in the range of 60-66%. Henry and Leone (2016) extend Loughran and McDonald (2010) and Li (2010a) in two ways. First, they compare the performance of Loughran and McDonald's wordlists with more parsimonious wordlists constructed by Henry (2006, 2008). Results show the Henry (2006, 2008) lists proxy tone more reliably. Second, they compare the performance of wordlist-derived tone proxies with Li's (2010a) Naïve Bayes classifier and conclude that wordlists perform well in relative terms. Both Loughran and McDonald (2010) and Henry and Leone (2016) evaluate classification performance indirectly using the strength of the correlation between fundamental economic performance signals and their tone proxies. Meanwhile, El-Haj et al. (2016) report in-sample evidence on the performance of various classifiers for tone and attribution in UK earnings press releases.. To date, however, no study of which we are aware evaluates the ability of automated classifiers to replicate manual coding outcomes by domain experts on unseen data.

We assess the out-of-sample performance of automated scoring methods by directly comparing results for a broad range of classifiers against manual annotations. We use the dataset of performance-related sentences constructed by El-Haj et al. (2016) as the basis for training our learning classifiers and constricting feature wordlists. Briefly, our dataset comprises 8,805 three-sentence blocks drawn from a sample of earnings announcement press releases issued by non-financial firms listed on the London Stock Exchange, where the middle sentence in each block contains a performance keyword. Each sentence block is then classified manually by two domain experts working independently. Invalid sentences are discarded and the remaining performance triads are annotated for tone and the presence of attributions. Where an attribution is identified, coders determine whether management refer to internal factors (e.g., strategy) versus external factors (e.g., macroeconomic conditions) (Arets 2005, Kimbrough and Wang 2014). Disagreements between coders are reviewed and resolved by an independent judge.

We use the resulting dataset to train the following learning classifiers for tone, attribution and attribution type: Naïve Bayes, random forest, support vector machines (SVM), and an artificial neural network model known as multilayer perceptron. We evaluate performance across these algorithms and then compare the best performer against result using relevant wordlists. In the case of attribution and attribution type, we complement wordlists used on prior research with self-constructed lists derived from our training dataset. Following best practice in computational linguistics (Das 2014: 42), we then construct and annotate manually a second sample of performance sentences using an identical coding procedure. We use this second dataset to evaluate the out-of-sample performance of our alternative classifiers.

Our main findings are as follows. Maximum out-of-sample classification accuracy ranges from 84% for attribution type to 81% for tone, suggesting that automated methods are capable of replicating manual coding outcomes with a reasonable degree of accuracy for certain tasks. Relative comparisons reveal that the best learning classifier typically beats a simple word list approach. For tone and the presence of an attribution, the differential between the best learning classifier and the best performing wordlist is relatively small at 5-8%. For attribution type, the performance differential is much larger (≈25%). Among the suite of learning classifiers, Naïve

Bayes rarely performs best despite its prominence in accounting research. Henry (2008, 2006) wordlists for tone always outperform Loughran and McDonald (2010) by a substantial margin, while a self-constructed wordlist for attribution beats causal reasoning wordlists based on LIWC and use in prior research (Zhang et al. 2019, Dikolli 2017).

We make several contributions to the literature. We extend insights in Henry and Leone (2016) on the performance of computerized methods for scoring the tone of management commentary by providing the first evidence of which we are aware on the absolute out-of-sample ability of a suite of text classifiers to replicate manually coded outcomes by domain experts. Our best performing classifier for sentence-level tone (random forest) achieves accuracy rates of 81% in out-of-sample tests. While learning classifiers beat Henry's (2006, 2008) wordlists, the difference is economically small (<5%). In contrast, classification accuracy using Loughran and McDonald's (2010) wordlists does not exceed 60%.

Our findings for tone classification highlight the conflicting bright and dark sides of automated textual analysis. On the one hand, evidence that (some) automated classification methods are able to replicate domain-expert manual classification with reasonable accuracy suggests that carefully designed and executed computerized analyses of management sentiment are capable of providing useful insights that complement small-sample manual scoring approaches. On the other hand, error rates in the region of 20-25% highlight the risks associated with poorly executed textual analysis studies. Empirical researchers would not tolerate a scenario where one-in-four data items from Compustat or CRSP were incorrect and so why should different standards apply to measures derived from text?

We also contribute to extant research by extending evidence on classification performance beyond tone to consider the arguably more complex phenomenon of attribution.

Results reveal that neither our best performing learning classifier (multilayer perceptron) nor our various wordlist options are capably of identifying attributions reliably. Unlike tone, our findings suggest very limited scope for using automated textual analysis methods to detect the presence of attributions in performance commentary. Conditional on having identified the presence of an attribution manually, however, we find that learning classifiers (but not our self-constructed wordlist) are able to reliably distinguish between internal and external attributions.

We also extend Henry and Leone (2016) with evidence on the relative performance of learning versus simple wordlist classifiers. While results for tone are consistent with Henry and Leone's (2016) evidence of small relative gains to machine learning over a wordlist approach, our evidence for attribution type highlights the danger of generalizing this conclusion to other classification tasks. Specifically, we find that our best performing machine learning classifier (SVM) beats our self-constructed wordlist by 26% (58% accuracy for our wordlist versus 84% accuracy for SVM). The poor relative performance of our wordlist reflects the varied nature of the classified construct: attributions types take many forms, making identification of comprehensive wordlists for internal and external attributions very difficult. Learning classifiers offer significant performance improvements in such cases because they are better able to isolate latent features. Findings highlight how classifier choice is conditioned by the nature of the task.

Finally, we contribute to the accounting literature by providing the most comprehensive treatment of the text classification task to date. We show how classification performance varies for learning algorithms across different classification tasks, with no single machine learning algorithm is consistently best across all classification tasks (Goel et al. 2010). We also demonstrate the dangers of relying on average accuracy metrics to assess classification performance, and we provide evidence on the benefits of sample balancing when training learning classifiers on highly skewed datasets. Finally, we provide a range of including annotated datasets, python code, and step-by-step guidelines to help researchers replicate and extend our machine learning classifiers.

2. Background and research question

Research examining the properties, determinants and economic consequences of financial discourse has a long history in the accounting literature. Merkl-Davies and Brennan (2007) critique work in the area, which at the time of their review was dominated by manual content analysis methods applied to samples of hundreds (rather than thousands) of observations. Two of the most popular discourse features examined the literature are tone and causal reasoning. A significant body of research in accounting examines the tone of financial narratives. Results provide a mixed picture. While many studies conclude that tone is informative for future performance, a large body of work also provides evidence consistent with opportunistic tone management in various corporate communications including annual reports, earnings announcements and conference call presentations.

Work on causal reasoning and in particular self-serving attribution bias also has a long history in accounting research. Attribution bias occurs when management take credit for positive outcomes while blaming negative results on factors beyond their control. Clatworthy and Jones (2003) report evidence consistent with attribution bias for a sample of U.K. firms' Chairman's Statements, while Kimborough and Wang (2014) analyse apparently self-serving attributions in U.S. firms' quarterly earnings announcements and conclude that investors are not fooled by such behaviour. Consistent with the informativeness view, Baginski et al. (2004) find that the decision by management to issue attributions alongside their earnings forecasts does not reflect managerial self-serving opportunism.

The interval since Merkl-Davies and Brennan's (2007) review has witnessed an explosion of papers in accounting and finance applying automated scoring techniques to measure the properties of financial discourse (Loughran and McDonald 2016). Li (2010b) proposes the following benefits of automated textual analysis over manual content analysis: lower data collection costs because algorithms are able to score text more quickly than human annotators; higher statistical power as a result of being able to work with larger sample sizes; greater objectivity and replicability because algorithms do not involve the same level of subjective judgement; and more generalizeable insights due to larger and more representative samples. El-Haj et al. (2019) question the validity of these claims, arguing that in certain circumstances research designs applying a manual annotation strategy to a small sample can generate higher power tests that are no less objective or costly. Little direct evidence currently exists regarding the accuracy with which automated text processing methods applied in a large sample setting measure verifiable properties of financial discourse relating to content and linguistic style. While many studies apply automated text scoring methods to financial data, very few evaluate the precision of the resulting empirical proxies.

Li (2010a) trains a Naïve Bayes classifier for tone on 30,000 of forward-looking sentences drawn from firms' 10-K and 10-Q filings. Threefold cross-validation tests reveal that the learning algorithm classifies tone correctly for 60-66% of sentences depending on the number of classes predicted. While classification performance is substantially better than chance, absolute measurement error rates of 34-40% are nevertheless high. Li's (2010a) analysis is nevertheless notable because he provides a direct comparison between outcomes derived from an automated classifier and a manually annotated "gold standard" comparator group.

Henry and Leone (2016) seek further evidence on the performance of computerized approaches to classifying tone in earnings announcement discourse. They compare a suite of wordlists for positive and negative language, including popular dictionaries developed by Henry (2006, 2008) and Loughran and McDonald (2010). Henry and Leone (2016) use an indirect method to evaluate classification performance based on the correlation between tone scored using a particular wordlist and economic fundamentals known to co-vary with the polarity of management commentary. Higher correlations are interpreted as evidence that tone is measured more accurately. Results show that the Henry (2006, 2008) dictionaries outperform the Loughran and McDonald (2010) wordlists. Supplementary tests using Li's (2010a) Naïve Bayes classifier applied to 10-K discourse indicate that the learning algorithm does not provide a large performance gain over the Henry (2006, 2008) wordlists. While Henry and Leone's (2016) findings provide important insights on relative classification performance, their method does not shed light on absolute classification accuracy due to the absence of a gold standard benchmark.

The ability of human annotators to take account of context and meaning when interpreting the (often subtle) messages in corporate discourse is a potentially vital advantage associated with a manual content analysis strategy. Rapid growth in the application of automated text processing methods raises an inevitable and critical question: how well are automated classification methods able to replicate manual annotation by domain experts? The remainder of our paper seeks evidence on this question in relation to measuring tone and causal reasoning.

3. Research design, data and descriptive statistics

We seek evidence on the ability of automated text classification procedures to replicate manual coding for tone and attribution in management performance commentary. Our approach involves training a suite of classifiers on a manually annotated dataset of performance sentences and then applying the classifiers to a second sample of manually annotated sentences to evaluate out-of-sample classification performance. Figure 1 summarizes the key elements of our research design. The remainder of this section explains our sampling strategy and manual coding procedure, and provides summary statistics for final datasets. Details of our text classifiers are described in section 4.

3.1 Training and holdout samples

Our training sample is based on El-Haj et al.'s (2016) dataset of performance sentences sampled from annual earnings press releases for fiscal year 2011 issued by London Stock Exchange-listed non-financial firms with analyst coverage on IBES. El-Haj et al. (2016) use a stratified sampling approach to maximize variation in reported performance. Specifically, firms are ranked based on their change in earnings from continuing operations (scaled by lagged market capitalization). A sample of 150 firms is identified comprising the 50 highest ranked cases, the 50 lowest ranked cases, and 50 cases selected at random from firms in quartiles two and three. Earnings press releases are retrieved from Perfect Information. (Eight documents were unavailable on Perfect Information.) The narrative component of each press release (i.e., excluding financial statements, footnotes and residual regulatory content) is extracted from each file and the text is split into 32,449 sentences. Candidate earnings-related performance sentences are identified using a keyword list designed to minimize Type II errors.[1] The resulting 8,805 sentences containing at least one performance-related keyword are retained together with adjacent lead and lag sentences. The final sample for manual coding comprises 26,415 individual

[1] The keyword list consists of the following elements: "sales", "revenue", "revenues", "turnover", "trading", "cost", "costs", "expense", "expenses", "income", "earnings", "eps", "e.p.s", "profit", "profits", "profitability", "loss", "losses", "margin", "margins", "result", and "results".

sentences for 8,805 tri-sentence blocks centred on a potential performance sentence. Two factors motivate the decision to code tri-sentence blocks. First, test coding reveals that performance-related statements are often complex constructs involving multiple sentences. For example, management routinely provide a direct statement on performance in one sentence, with related explanations (i.e., attributions) presented in the preceding or subsequent sentence. Second, even in the absence of any complex multi-sentence attributions, adjacent sentences frequently provide important contextual information required to determine the polarity of a performance sentence.

We follow a similar approach to construct our holdout sample using earnings press releases for fiscal year 2012. The primary difference is that we select 150 firms with non-zero analyst following at random (rather than stratifying by scaled earnings changes) to maximize representativeness. We follow the same procedure applied in the training sample to extract text, split sentences, and identify candidate performance sentences. As only a subset of candidate performance sentences are valid and only a fraction of those contain an attribution, attributions are the limiting discourse feature when constructing our holdout sample. We proceed by setting the target number of attributions to 1,000 to ensure tests of classification accuracy are reliable, and continue sampling tri-sentence blocks randomly until this threshold is reached. The strategy requires us to score 6,242 tri-sentence blocks or 18,726 individual sentences.

3.2 Manual coding strategy

Figure 2 summarizes the manual coding strategy applied to the training and holdout samples. Tri-sentence blocks in both samples are coded manually for tone, the presence of an attribution, and attribution type. Tone measures the polarity of a valid performance sentence. We classify performance sentence tone as either positive, negative, neutral, or unclear (Li 2010a). Attribution occurs when management relate a performance outcome to at least one fundamental determinant such as operating efficiency, product development, adverse trading conditions, etc.[2] The presence of an attribution is treated as a binary outcome equal to one if management explicitly link performance with one or more fundamental determinants and zero otherwise. Finally and consistent with prior research (Aerts 2005, Kimbrough and Wang 2014), we categorize valid attributions according to whether the fundamental determinant(s) cited by management relate to internal or external factors. We classify internal factors as those over which management exercise direct control. Examples of internal factors include strategic reorientation, cost control, product design, marketing initiatives, labor relations, etc. Conversely, we classify external factors as those over which management are not expected to exercise direct control such as market competition, input prices, exchange rates, weather, etc.[3] We capture attribution type using separate indicator variables for internal factors and external factors because a sentence may contain both attribution types.

We develop a draft coding scheme to guide the classification process and refine the guidelines through several iterations where multiple coders classify 100 sentences, compare results, and modify coding rules accordingly. Annotators first determine whether the target sentence in a tri-sentence block is valid earnings-focused performance sentence. Conditional on identifying a valid performance sentence, annotators then determine tone, the presence of

[2] We distinguish between valid attributions and vague statements or tautologies. For example, a statement ascribing profit growth to lower costs is not treated as an attribution for the purposes of our study because the fundamental factor(s) causing costs to fall are not specified. In contrast, a statement linking profit growth to a specific cost reduction programme is classified as a valid attribution because the source of cost efficiencies is identified.

[3] Some factors such as supply chain are ambiguous and context-specific. For example, where management explicitly cite unforeseen problems in the supply chain resulting from extraordinary circumstances as the determinant we code the attribution as external on the basis that management is seeking to distance itself from the cause. On the other hand, where management highlight ongoing supply problems as the cause we treat the attribution as internal on the grounds that the firm has failed to resolve known problems.

attribution, and attribution type. Tone relates solely to the middle sentence in a block, whereas attribution and attribution type may involve any of the three sentences.[4]

The finalized manual coding scheme is implemented using double-blind classification by four domain experts from the author team, working in pairs. Tri-sentence blocks are divided equally among coder-pairs in extraction order. Coders view all sentences in a block together and in the correct sequence using a Microsoft Access form. Classification results are recorded through the same interface, with coders selecting numeric indicators from dropdown menus to limit the risk of typographical errors. A free text field is also available for coders to record explanatory notes as part of an audit trail. On completion of the double-blind coding task, results for coder-pairs are compared. Inconsistencies are identified and resolved with the aid of independent judge. Observations are coded as unclear for the small number of cases where the independent judge is unable to resolve the disagreement. The final manually annotated training and holdout samples available at https://github.com/apmoore1/pea_classification to support replication and further research.

3.3 Sample and descriptive statistics

Table 1 summarizes manual coding results for the training sample (first row) and holdout sample (second row). Of the initial 8,805 target sentences in the training sample, 1,604 target cases (18%) are invalid sentences (e.g., a list of performance metrics). The remaining 7,201 target sentences comprise 3,396 performance sentences and 3,805 target sentences judged not to be valid performance sentences because they either discuss non-earnings features such as cash, debt, inventory, production, etc. or they refer to results for the current fiscal year. The high rate

[4] We differentiate performance sentences including attributions and those where the attribution appears in an adjacent sentence when training our machine learnings classifiers for tone. It is an empirical issue whether classification accuracy varies as a function of sentence complexity.

of non-performance sentences reflects our performance sentence selection strategy which is purposely designed to minimize the risk of excluding valid performance sentences at the expense of a high Type I error rate. We therefore code the remaining 3,396 tri-sentence blocks (10,188 sentences) for tone, attribution and attribution type. Positive tone dominates as evidenced in prior research (Loughran and McDonald 2016): 2,393 target performance sentences (70%) are classified independently by two domain experts as positive, compared with 784 sentences (23%) that are classified as negative. The remaining 219 sentences are classified manually as either neutral (123) or mixed tone (96). Frequencies in these latter two categories are too sparse to classify reliably using learning algorithms and therefore we drop these cases from the analysis. Our final automated classification task therefore involves predicting a binary outcome (positive versus negative tone) using a sample of 3,177 sentences.

We code attributions associated with the 3,177 target performance sentences. Attributions may appear directly in the performance sentence or in one of two adjacent sentences in the tri-sentence block. Two coders working independently identify 1,594 sentences that contain at least one attribution, of which 1,161 are target performance sentences and 433 are sentences adjacent to a performance sentence. We define the sample of non-attribution sentences equal to the remaining 2,235 target performance sentences that do not contain an attribution plus the 3,372 target sentences that contain neither an attribution nor a valid performance sentence.

Finally, we code attribution type for the 1,594 sentences identified as containing at least one attribution. We identify 948 internal attributions and 800 external attributions. The aggregate number of attribution types 1,748 exceeds the number of attribution sentences because 180 sentences make reference to both internal and external causal factors. We exclude these 180 combined sentences from the training sample because they do not provide incremental information for the binary classification task. We also identify and exclude 26 attribution sentences where two coders plus a judge are unable to agree on the nature of the attribution. Our attribution type final training sample therefore comprises 768 (948 – 180) clean internal attributions and 620 (800 – 180) clean external attributions.

The second row in Table 1 provides comparable information for the holdout sample. The holdout sample consists of 1,774 valid target performance sentences, 1,200 (68%) of which are positive versus 574 (32%) that are classified as negatively toned. The total number of tri-sentence blocks containing at least one attribution is 966, comprising 720 target performance sentences and 246 adjacent nonperformance sentences. We use this sample of attribution sentences to identify 338 sentences that contain a clean internal attribution and 491 sentences that contain a clean external attribution.

4. Classifiers and classification performance

This section presents information on our sentence classification strategies and explains the metrics we use to evaluate classification performance for both the training and holdout samples. We use a suite of machine learning algorithms to classify tone, attribution and attribution type and then select the best performing model based on the training sample results to classify sentences in the holdout sample. We compare machine learning classification performance against classifications generated by applying separate wordlists for tone, attribution and attribution type. Table 2 summarizes our classification methods.

4.1 Machine learning algorithms

We train four popular machine learning algorithms on each of our three discourse features. Theory provides little guidance on which algorithm is likely to perform best and therefore NLP researchers typically start with a suite of classifiers and then select the best empirical performer as their baseline model (Goel et al. 2010). The four machine learning algorithms used in our analysis are Naïve Bayes, random forests, support vector machines (SVM), and a form of artificial neural network known as multilayer perceptron.[5]

Naïve Bayes is a probabilistic classifier that represents one of the simplest and most effective inductive machine learning algorithms. The Naïve Bayes approach uses the joint probabilities of words and categories to estimate the probabilities of categories when a document is given (McCallum and Nigam 1998). The NB classifier assigns the most likely class to a given example described by its feature vector. The underlying assumption of the Naïve Bayes approach is that the probability of each word occurring in a document is independent of the occurrence of other words in the document and the probability that a document is generated in some class depends only on the probabilities of the words given the context of the class. Even though it is a probabilistic classifier, its classification performance is competitive with the performance of other sophisticated learning methods (Mitchell 1997). Naïve Bayes is the method used by almost all studies in accounting and finance that use a machine learning classifier (Li 2010a, Huang et al. 2014, Buehimaier and Whited 2018).

Random forest is a supervised ensemble machine method that that fits a number of decision tree classifiers on various sub-samples of the dataset. Each decision tree generates a

[5] We also tested the Stanford SentimentAnnotator which implements Socher et al's (2013) sentiment model based on a new type of Recursive Neural Network that builds on top of grammatical structures with a fine grained sentiment treebank (https://nlp.stanford.edu/sentiment/; https://www.quora.com/How-does-the-sentiment-analysis-in-Stanford-NLP-work-Is-there-a-way-for-Stanford-NLP-to-take-the-overall-sentiment-of-multiple-sentences). The model is trained on movie reviews wherein a reviewer might discuss both positive and negative movie aspects in the same sentence (e.g. "the plot was slow but the acting was great"). The accuracy of predicting fine-grained sentiment labels for all phrases reaches 80.7%. The Stanford Sentiment tool is available in python. We classify performance sentence tone in our holdout sample using the Stanford tool. Accuracy rates never exceed 55% despite the sophisticated deep learning features of the model. Findings highlight the importance of training machine learnings models on relevant datasets annotated by domain experts (rather than applying complex models developed using language drawn from other, less relevant, domains).

prediction. Votes associated with different decision trees are then combined to determine the final class. The process of averaging across decision trees improves predictive accuracy relative to using a single decision tree while also controlling over-fitting.

SVM is a supervised machine learning technique that is based on statistical learning theory. The SVM algorithm learns by example to classify outcomes into predefined classes. SVMs are based on the Structured Risk Minimization (SRM) method for model selection that provides a trade-off between hypothesis space complexity and the quality of fitting the training data to guarantee the lowest true error on an unseen and randomly selected test example. SVMs determine a hyperplane in the feature space that best separates the data according to the predefined classes.

Our fourth classifier is a multilayer perceptron (MLP), which is a class of feedforward artificial neural network. MLP consists of at least three layers: an input layer (data), a hidden layer, and an output layer (classification). MLP utilizes a supervised learning technique called backpropagation for training and is able to distinguish data that is not linearly separable. The technique can be viewed as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation. This transformation projects the input data into a space where it becomes linearly separable. This intermediate layer represents the hidden layer.

Implementing each machine learning algorithm involves selecting a large number of model parameters. To maximise performance and minimize the number of imposed arbitrary choices, we train 40 different versions of each classifier to incorporate parameter variation and then select the version that maximises average classification performance.

4.2 Wordlists

We classify performance sentences for tone, attribution and attribution type using a suite of wordlists comprising dictionaries drawn in prior research plus several self-constructed lists designed to capture domain-specific characteristics. This section introduces the wordlists and explains how they are used to classify sentences. Appendix B provides further details of each wordlist used in our analysis, including the constituents of each list. Table 2 summarises the wordlists used to classify each feature.

Tone wordlists

We use positive and negative wordlists developed by Henry (2006), Henry (2008), and Loughran and McDonald (2011) to classify performance sentence tone. All lists are adjusted for British English spelling where appropriate. We do not use the positive and negative lists from General Inquirer because research demonstrates that the resulting tone measures perform poorly in a finance context (Loughran and McDonald 2011, Henry and Leone 2016). For each sentence *i* we count the number of positive and negative words associated with a tone measure and classify tone as positive (negative) where the positive word count exceeds the negative word count (negative word count exceeds the positive word count). Tone for sentence *i* is set to neutral where the difference between positive and negative word counts is equal to zero.

Attribution wordlists

Two wordlists have been employed in prior accounting research to measure attributions and causal reasoning. Zhang et al. (2019) measure the incidence of causal reasoning in earnings-related commentary from the MD&A section of firms' 10-K filings using a subset of causation words from the LIWC causation dictionary word stems. Dikolli et al. (2017) measure the incidence of CEO causal reasoning in shareholder letters. Dikolli et al. (2017) use a modified version of the LIWC causation wordlist. They construct their list by first identifying 505 causation words from the LIWC causation dictionary word stems. Each element in this initial wordlist is then reviewed for appropriateness in their corpus of shareholder letters and remove words are not associated with causal reasoning in a financial context. Evaluation tests reported by Dikolli et al. (2017, appendix) their modified dictionary correctly classifies 68% causal reasoning sentences compared with 60% accuracy using the original LIWC wordlist. We use both the Zhang et al. (2019) and Dikolli et al. (2017) lists as alternative sentence-level classifiers for the presence of an attribution. Sentence *i* is classified as containing an attribution using a given wordlist when the sentence contains at least one word from the corresponding list.

Prior research highlights the importance of using domain-specific wordlists. To the best of our knowledge, no wordlist designed to measure attributions in UK earnings press releases currently exists. We therefore construct two new attributions wordlist based on our manually-annotated training sample. The procedure for constructing our attribution wordlists is outlined in Appendix B. Sentence *i* is classified as containing an attribution using our self-constructed lists if it contains at least one word from the corresponding wordlist.

Attribution type wordlist

We are not aware of any wordlist in the extant literature that is designed to capture the presence of internal or external management attributions. We therefore develop two new lists for internal and external attributions, respectively, based on our manually-annotated training sample. The procedure to construct each wordlist is described in Appendix B and follows the same approach as that used to construct our attribution list. Sentence *i* is classified as containing at least one internal (external) attribution when the frequency count for the internal (external) list is greater than zero. Note that the classification procedure for attribution type does not generate mutually exclusive categories. A single sentence may contain multiple attributions and therefore may be classified as containing both an internal and an external attribution.

4.3 Classification performance

We evaluate classification performance using several metrics. Accuracy is defined as the ratio of correctly predicted outcomes (true positives plus true negatives) to total outcomes predicted, and arguable represents the most intuitive measure of performance:

$$Accuracy_p^k = \frac{N(tp) + N(tn)}{N(tp) + N(fp) + N(tn) + N(fn)}, \qquad (1)$$

where *Correctly classified* is equal to one where classifier *k* replicates the human coding outcome for binary feature *p* (*p* equals tone, attribution or attribution type) in sentence *i (i* = 1 to *N), N(tp)* is the number of true positives, *N(tn)* is the number of true negatives, *N(fp)* is the number of false positives, and *N(fn)* is the number of false negatives.[6] Accuracy values for wordlist classifiers applied to the training sample and all *k* classifiers applied to the holdout sample are computed directly from equation (1). Accuracy for machine learning classifiers applied to the training sample is equal to mean accuracy computed using the 10-fold cross-validation method. The procedure involves selecting 90% of sentences at random for training and using the remaining 10% for validation. The process is repeated 10 times (folds), with accuracy equal to the mean of the equation (1) values from each of the 10 iterations.

Using accuracy to evaluate classification performance in highly unbalanced samples is problematic because a model that predicts the high frequency category well (i.e., low false positive rate) will generate a high accuracy score even if it has little ability to predict the low

[6] Equation (1) is equivalent to the sum of true positives plus true negatives divided by the sum of true positives plus true negatives plus false positives plus false negatives.

frequency category (i.e., high false negative rate). The *F1 Score* metric addresses this weakness by accounting for false positives and false negatives:

$$F1\ Score_p^k = \frac{2 \times \left(Recall_p^k \times Precision_p^k\right)}{\left(Recall_p^k + Precision_p^k\right)}, \qquad (2)$$

where *Recall* is the ratio of correctly predicted positive outcomes to all cases in a class and *Precision* is the ratio of correctly predicted positive outcomes to total predicted positive cases:

$$Recall = \frac{N(tp)}{N(tp) + N(fn)},$$

$$Precision = \frac{N(tp)}{N(tp) + N(fp)},$$

Finally, we use *Macro F1* to evaluate overall performance for classifier *k* for binary feature *p*, where *Macro F1* is the arithmetic mean of the *F1 score* from equation (2) computed for both classes associated with binary feature *p*. *Macro F1* for wordlist classifiers applied to the training sample and all *k* classifiers applied to the holdout sample are computed directly from equation (2), whereas *Macro F1* for machine learning classifiers applied to the training sample is equal to mean of the Macro F1 values from each of the 10 cross-validation folds.

5. Main results

5.1 Tone

Classification performance for the training sample is summarized in Table 3. Results for the four machine learning algorithms are presented along with those for our three wordlist approaches. Macro F1 and accuracy metrics for machine learning classifiers represent average values based on the 10 cross-validation folds, whereas comparable statistics for the three wordlists reflect a single classification pass. Applying wordlist classifiers to the training sample yields out-of-sample tests because all tone wordlists are derived from exogenous sources. Concern about upward bias in performance statistics due to overfitting is therefore not a consideration for wordlist results. We also report the F1-score for each class (i.e., positive tone and negative tone) to shed light further light on the source of classification performance.

Several notable findings are evident in Table 3. First, with Macro F1 (accuracy) values typically around 75% (80%), results suggest it is possible to score sentence-level tone with a reasonable level of reliability using automated methods. Second, the F1-score for positive tone is higher than the comparable metric for negative tone across all classifiers. Most classifiers with the exception of the L&M wordlist classify over 80% of positive sentences correctly, compared with less than 66% of negative sentences. The median difference in F1-scores across all seven classifiers is 26% and ranges from a high of 36% for Naïve Bayes to a low of 8% for the L&M wordlist. These results provide consistent evidence that negative toned sentences are more difficult to classify automatically. This may be the result of inherently greater complexity associated with negative outcomes and descriptions thereof, or it may reflect a greater tendency for management to obfuscate bad news (Bloomfield 2008).

A further notable result in Table 3 is that the Henry (2006, 2008) classifiers perform well relative more sophisticated machine learning classifiers. Macro F1 values for both wordlist methods are only 2% lower than the best performing machine learning classifier. The evidence supports results and conclusions reported by Henry and Leone (2016). Also consistent with Henry and Leone (2016) is the poor relative performance of the L&M classifier, which underperforms both Henry classifiers and worst performing machine learning classifier by more than 20% (25%) based on Macro F1 (accuracy). Indeed, with only 53% of sentences correctly classified, the L&M classifier performs little better than chance.

Differences in classification performance across the four machine learning algorithms are small in absolute terms (approximately 2%), suggesting the choice of specific algorithm is of second order importance when scoring sentence-level tone in earnings announcements. Random forest displays the highest Macro F1 and accuracy performance among the four machine learning algorithms. Accordingly, we select this algorithm as our best performing machine learning classifier for subsequent out-of-sample tests.

Table 4 reports classification performance for the holdout sample. This analysis provides a more reliable test of classification performance for the learning approach. Results for the random forest classifier are qualitatively identical to those documented in Table 3 for the training sample. The classifier has a Macro F1 (accuracy) value of 76.4% (81.2%). Performance is only marginally better than the Henry (2006, 2008) classifiers. These findings support Henry and Leone’s (2016) conclusion that simple wordlists perform almost as well as more sophisticated classifiers despite being more straightforward to implement. In contrast and similar to results documented in Table 3, the L&M classifier is associated with accuracy levels below 60%. These findings have important implications for current research given widespread reliance on the L&M approach to measuring tone. Further, with error rates between 20-25%, even the best performing automated approaches generate material measurement error relative to manual coding.

Consistent with evidence reported for the training sample, all classifiers yield significantly less reliable results for negative sentences compared with positive sentences. All classifiers with the exception of L&M display F1-scores above 80% for positive sentences; and the incremental performance of the machine learning approach over the wordlist approach exceeds 5% for positive sentences. In contrast, F1-scores for negative sentences are below 70%, with the Henry classifiers outperforming the machine learning approach by approximately 3%.

Collectively, our evidence suggests that the reliability of automated methods for scoring sentence tone depends critically on the particular research question at hand. If the primary focus is on measuring positive statements by management then results for automated methods are likely to approximate those from manual coding reasonably well. In contrast, reliance on automated methods may prove problematic when the main focus is on negative language.

5.2 Attribution

Table 5 summarizes classification performance for management attributions using the training sample. Columns 2 and 3 report separate F1-scores for the attribution and no attribution classes, respectively, while Macro F1 and accuracy values for overall classification performance are presented in the final two columns. Results for the LIWC and DIK wordlist classifiers reflect out-of-sample tests because both wordlists are derived exogenously, whereas findings for the machine learning classifiers and our two self-constructed wordlist classifiers represent in-sample tests and may therefore be prone to upward bias due to overfitting.

The most striking feature of Table 5 is the very low F1-scores for the attribution class. Only the ATT_ALL and ATT_50 classifiers generate F1-scores above 50% for the presence of an attribution. None of the four machine learning classifiers yield an F1-score above 50%: the best performing algorithm is the neural network at just 49.4%. Note also the large variation in performance across machine learning algorithms, with Naïve Bayes displaying the worst performance at 23%. The DIK classifier performs close to the best machine learning model while the LIWC classifier displays particularly poor ability to identify the presence of an attribution (15%). In contrast, F1-scores for the no attribution class exceed 80% for all classifiers with the exception of ATT_ALL (78.2%) and DIK (67.6%). The asymmetry in classification performance highlights the danger of using a simple accuracy metric to evaluate reliability. Overall accuracy values reported in the final column suggest most classifiers are able to capture the presence of an attribution with a reasonable degree of accuracy (> 70%). Macro-F1 values present a less optimistic picture but even here overall performance often exceeds 65%. Only through the analysis of individual class F1-scores does the true picture emerge. All classifiers do a poor job of identifying attributions, with apparently high classification performance a consequence of the models correctly classifying the subset of no attribution cases that account for 78% of sentences in the training sample. In such a scenario, a naïve model that classifies all sentences as containing no attribution will display an overall accuracy rate of 78% despite having no genuine ability to detect the presence of an attribution. Results reported in Table 5 suggest that all classifiers struggle to outperform such a naïve model.

Table 6 presents results for out-of-sample classification performance. The story is consistent with findings document in Table 5 for the training sample. F1-scores reflecting classifiers' ability to detect the presence of a valid attribution range from a high of 52.8% for ATT_ALL to a low of 19.6% for LIWC. Conversely, F1-scores reflecting classifiers' ability to correctly detect sentences that do not contain an attribution range from a low of 70% for DIK to a high of 88.8% for the best-performing machine learning algorithm. Both Macro-F1 and accuracy values significantly overstate the reliability with which our classifiers are able to replicate manual coding. Indeed, evidence indicates that reliance on any of our automated classifiers is likely to yield unreliable large sample evidence on managers' attribution behaviour.

5.3 Attribution type

Tables 7 and 8 provide evidence on classification performance for attribution type based on the training and holdout samples, respectively. We develop classifiers to distinguish between attributions relating to internal and external factors conditional on a sentence being manually classified as containing at least one attribution. Table 7 reports overall classification performance in the training sample, along with separate F1-scores for internal and external classes.

Macro F1 and simple accuracy metrics reported in the final two columns of Table 7 provide a consistent picture. All four machine learning models are associated with accuracy levels around 84%. Results provide prima facie evidence that reliable classification of attribution type is possible using automated methods. Performance is broadly similar across the four models, with SVM yielding the highest Macro F1 value (84.8%) and our neural network algorithm producing the highest accuracy value (84.9%). Analysis of individual F1-scores associated with the internal and external classes provide weak evidence that models are better able to classify internal attributions, although the performance gap is typically less than 5%. All machine learning algorithms outperform our self-constructed wordlist classifier by a substantial amount. Absolute performance for our classifier based on internal and external wordlists is nevertheless respectable at 70%.

Comparable results using the holdout sample are presented in Table 8. Overall performance of the best machine learning classifier (multilayer perceptron) remains impressive at over 82%. Similar to the evidence presented in Table 7, the algorithm displays slightly superior performance when classifying internal attributions (86%) versus external attributions (78%). The performance of our self-constructed wordlist classifiers is significantly worse in the holdout sample compared with results reported in Table 7 for the training sample, suggesting that findings for the latter may reflect overfitting. With Macro F1 and accuracy values of 56.4% and 57.8%, respectively, results casts doubt on the ability of our wordlists to replicate manual coding outcomes for attribution type, with any value being limited to external attributions. Collectively, findings reported in Table 8 suggest that reliable identification of attribution type is possible using a machine learning approach but not a wordlist approach.

6. Supplementary analysis

6.1 Sentence complexity

Results for tone and attribution reported in section 5 are based on samples that include a subset of target performance sentences where at least one attribution is also present. Performance sentences that also include an attribution are arguably more linguistically complex than sentences that contain a single discourse feature. On the one hand, complexity may reduce classification performance if the presence of multiple discourse features reduces the signal-to-noise ratio for each distinct feature. If this is the case then classification performance may be superior when classifiers are trained on “clean” sentences containing a single feature. On the other hand, co-occurring features may aid the classification task for an individual feature. For example, the presence of an attribution in a performance sentence may provide additional information on polarity if attributions are associated with specific tonal features. Whether and how sentence complexity affects classification performance is an empirical issue on which we seek evidence.

Figure 3 presents findings from tests examining the impact of sentence complexity on classification performance for tone and the presence of an attribution. Panel A compares tone classification results (Macro F1) in the holdout sample for machine learning classifiers trained on all performance sentences versus classifiers trained on the subset of clean performance sentences where no attribution is present. A consistent pattern of results is evident across all four machine learning algorithms: Macro F1 values are materially higher for models trained on the aggregate performance sentence sample. For completeness Panel A also includes line plots of Macro F1 values for the training sample, where a similar a pattern is again evident. Specifically, in-sample classification performance for all four algorithms is superior when the training sample includes all performance sentences regardless of whether or not they also contain an attribution. The evidence is consistent with the view that co-occurring attributions may contain additional information that aids the task of classifying sentence polarity. The pattern is also consistent with a positive association between classification performance the size of the training sample.

Panel B of Figure 3 compares attribution classification performance (Macro F1) in the holdout sample for classifiers trained on all attribution sentences versus those trained on the subset of clean attribution performance sentences that do not contain a performance statement. Findings and conclusions contrast with those presented in Panel A insofar as out-of-sample classification performance for all four machine learning classifiers is materially higher for the subsample of attribution-only sentences. Results suggest that in the case of attribution, the benefits of lower sentence complexity outweigh any costs associated with a reduction in the size of the training sample. Macro F1 values for the training sample also demonstrate superior classification performance using the subset of attribution-only sentences.

Next we assess whether attribution type affects the probability of identifying the presence of an attribution. Specifically, we test whether the likelihood of detecting causal reasoning varies conditional on whether the attribution refers to internal versus external factors. We train machine learning algorithms for the presence of an attribution separately on subsamples of internal- and external-only attribution cases and then test whether the ability to detect attributions in the holdout sample varies with the training sample. Results are summarized in Figure 4. Panel A reports Macro F1 values for models trained on internal and external attributions. Classifiers trained with the subsample of external attributions are associated with higher out-of-sample classification accuracy for all four machine learning algorithms, with the effect being especially pronounced for SVM. The same pattern is also evident in the training samples, although to a lesser degree. Panel B of Figure 4 provides further insight on the source of the performance improvement. Individual F1 scores for the attribution and no attribution classes suggest that superior performance is the result of improvements in the ability to detect attributions and not the accuracy with which non-attribution cases are identified.

Results reported in Figures 3 and 4 collectively indicate how the impact of sentence complexity on classification performance varies with the nature of the classification problem. In the case of classifying sentence tone, the presence of a concurrent discourse feature such as causal reasoning serves to improve classification performance. In contrast, lower sentence complexity is associated with superior classification performance in the case of causal reasoning. Our findings highlight the conditional nature of classification strategies and the corresponding difficulty of developing universal guidelines for multiple discourse features.

6.2 Sample balancing

It is well established in the machine learning literature that large differences in class size can affect classification performance. All else equal, highly unbalanced samples can result in a classifier anchoring on the high frequency class at the expense of reliable feature detection in the low frequency class. Table 1 reveals substantial sample imbalance for two of our discourse features. Tone is heavily skewed towards positive sentences, with only 25% of the training sample classified as negative. Similarly, only 22% of the 7,201 sentences in the attribution training sample contain an attribution compared. Imbalance in our training samples for tone and attribution may generate classification outcomes that are biased towards the majority classes of positive tone and no attribution, respectively.

Figure 5 documents the impact of sample imbalance on the classification performance of our machine learning algorithms. Panel A reports results for tone while Panel B reports evidence for attribution. In each case we compare out-of-sample classification performance for algorithms trained on the corresponding full (i.e., unbalanced) sample with results using two alternative balancing methods. Undersampling sets the size of the high frequency class equal to the maximum number of cases in the low frequency class by randomly sampling from the high frequency class. Oversampling generates a larger number of observations for the minority class using the Synthetic Minority Oversampling Technique (SMOTE) algorithm. Both Panels report separate F1 scores majority and minority classes, together with Macro F1 values measuring overall classification performance.

Results highlight the potential importance of sample balancing. As expected, classification performance for the majority class in each Panel is invariant to sampling method: results for the positive tone (no attribution) class in Panel A (B) are qualitatively identical regardless of whether classifiers are trained on balanced or unbalanced samples. In contrast, classification accuracy for the minority class improves substantially in both Panels when algorithms are training on balanced samples. In both cases, undersampling tends to generate better results than oversampling, although effects vary with discourse feature and algorithm. For negative tone in Panel A, the benefits of balancing are most pronounced for SVM, while the particular balancing strategy appears largely irrelevant. Undersampling is associated with more pronounced effects for the remaining three algorithms, and in particular Naïve Bayes. For attribution in Panel B, undersampling yields superior results for all four classifiers, and in particular for Naïve Bayes and random forest.

The graph on the far right in each Panel captures the impact of sample balancing on overall classification performance. Balancing is associated with superior overall classification performance (driven by the minority class) for all four machine learning algorithms in both Panels. Undersampling yields the largest performance gain and the impact appears particularly important for Naïve Bayes. Collectively, evidence presented in Figure 5 highlights the potential importance of balancing when training classifiers on highly imbalanced samples.

7. Summary and conclusions

Our analysis extends prior work in several important ways. To the best of our knowledge, ours is the first study to provide direct evidence on the accuracy of alternative sentence-level approaches for measuring tone and attribution. Extant research, in contrast, relies on correlations with predicted determinants to evaluate measures of tone (Loughran and McDonald 2011, Henry and Leone 2016). We show that classification accuracy rarely exceeds 80% for tone, suggesting that even the best performing classifiers are associated with substantial measurement error relative to manual coding. Collectively, our findings highlight opportunities and limitations associated with automated textual analysis methods. Critically, we conclude that manual content analysis remains an essential tool for researchers interested in studying the properties and consequences of financial discourse.

We also extend Henry and Leone (2016) by demonstrating that machine learning classifiers for tone can outperform word-frequency measures in some settings. Our evidence is consistent with established results in the NLP literature documenting the accuracy gains over simple word counts of more sophisticated discourse methods that account for word meaning and context. As such, our findings speak directly to the question pose by Loughran and McDonald

(2016: 1199) about the incremental value of applying deeper semantic parsing tools in accounting and finance research.

Finally, we also show how fundamental implementation choices can affect the performance of machine learning classifiers. First, we find that alternative classifiers often outperform Naïve Bayes despite the latter's dominant position in extant accounting and finance research. Second, we demonstrate the importance of imposing sample balance when training classifiers to measure features such as tone where real-world norms are biased heavily in favour of a particular outcome category (e.g., positive tone). In such cases, failure to use a balanced sample leads to serious overfitting and poor out-of-sample classification performance.

Appendix: Wordlist construction

Feature: **Tone**

1. L&M: Loughran & McDonald positive (POS) and negative (NEG) wordlists. Classification rule: Sentence polarity determined by relative frequency counts of POS vs NEG.
2. HEN08: Henry (2008) positive (POS) and negative (NEG) wordlists. Classification rule: Sentence polarity determined by relative frequency counts of POS vs NEG.
3. HEN06: Henry (2006) positive (POS) and negative (NEG) wordlists. Classification rule: Sentence polarity determined by relative frequency counts of POS vs NEG.

Feature: **Attribution**

4. LIWC (as implemented by Zhang et al. 2019): This is a subset of the complete LIWC causal list but no details are provided by Zhang et al. (2019) how the subset is generated. Classification rule: Sentence classified as causal if keyword count > 0.
5. DIK (as implemented by Dikolli et al. 2017): start by identifying 505 unique causation words from the LIWC causation dictionary word stems, from which elements are removed where they appear to represent missclassifaction in a business setting. This subset is then augmented with words that LIWC omits but which likely denote causation in a business setting. Classification rule: Sentence classified as causal if keyword count > 0.
6. ATT_ALL: List created by authors based on manual analysis of classified sentences in the training sample. Classification rule: Sentence classified as causal if keyword count > 0.
7. ATT_50: 50 most frequently occurring words from MW_ALL Classification rule: Sentence classified as causal if keyword count > 0.

Feature: **Attribution type**

8. ATT_TYPE (Internal): Used to classify internal attributions. List created by authors based on analysis of classified sentences in the training sample. Classification rule: Sentence classified as internal (external) attribution if keyword count > 0 (highest relative frequency count where counts for internal and external are both > 0).
9. ATT_TYPE (External): Used to classify external attributions. List created by authors based on analysis of classified sentences in the training sample. Classification rule: Sentence classified as internal (external) attribution if keyword count > 0 (highest relative frequency count where counts for internal and external are both > 0).

References

Aerts, W. (2005). Picking up the pieces: impression management in the retrospective attributional framing of accounting outcomes. *Accounting, Organizations and Society* 30(6): 493-517

Baginski, S. P., Hassell, J. M., Kimbrough, M. D. (2004). Why do managers explain their earnings forecasts? *Journal of Accounting Research*, 42(1): 1-29

Buehimaier, M., Whited, T. (2018). Are financial constraints priced? Evidence from textual analysis. *Review of Financial Studies* 31(7): 2693–2728

Clatworthy, M. Jones, M.J. (2003). Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and Business Research* 33(3): 171-185

Das, S. (2014). Text and context: Language analytics in finance. *Foundations and Trends in Finance* 8 (2014): 145–261

Dikolli, S., Keusch, T., Mayew, W., Steffen, T. (2017). Using shareholder letters to measure CEO integrity. http://ssrn.com/abstract=2131476

El-Haj, M., Rayson, P., Young, S., Moore A., Walker M., Schleicher T., Athanasakou V. (2016). Learning tone and attribution for financial text mining. 10th edition of the Language Resources and Evaluation Conference (LREC'16). May 2016. Portoroz, Slovenia

El-Haj,, M., Rayson, P., Walker, M., Simaki, V., Young, S. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance and Accounting* 46(3-4) 2019: 265-306

Goel, S., Gangolly, J., Faerman, S., Uzuner, O. (2010). Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7(1): 24-46

Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3(1): 1-19

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication* 45(4): 363-40

Henry, E., Leone, A.J. (2016) Measuring qualitative information in capital markets research: comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91(1): 153-178

Huang, A., Zang, A. Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review* 89(6): 2151–2180

Kimbrough, M.D., Wang, I.Y. (2014). Are seemingly self-serving attributions in earnings press releases plausible? Empirical evidence. *The Accounting Review* 89(2): 635–667

Li, F. (2010a). The information content of forward-looking statements in corporate filings: A Naive Bayesian machine learning approach. *Journal of Accounting Research* 48(5): 1049-1102

Li, F. (2010b). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29: 143-165

Loughran, T., McDonald, B. (2011), When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66(1): 35–65

Loughran, T., McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54(4): 1187-1230

Merkl-Davies, D.M., Brennan, N.M. (2007). Discretionary disclosure strategies in corporate narratives: incremental information or impression management? *Journal of Accounting Literature* 26: 116–194

McCallum, A., Nigam, K. (1998). A comparison of event models for naive bayes text classification. Proceedings in Workshop on Learning for Text Categorization. AAAI Technical Report WS-98-05: 41-48

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002): 79–86. doi 10.3115/1118693.1118704

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)

Zhang, S., Aerts, W., Pan, H. (2019). Causal language intensity in performance commentary and financial analyst behaviour. *Journal of Business Finance and Accounting* 46(1-2): 3-31

Table 1: Training and holdout samples by language feature.

	Tone			Attribution			Attribution type		
		Class			Class			Class	
Sample	N	Positive	Negative	N	Yes	No	N	Internal	External
Training sample	3,177	2,393	784	7,201	1,594	5,607	2,217	768	620
Holdout sample	1,774	1,200	574	4,382	966	3,416	829	338	491

Performance sentences in the training sample are drawn from annual earnings announcements made by 150 firms in 2011 ranked by their change in earnings from continuing operations (scaled by lagged market capitalization). The sample comprises the 50 highest ranked firms, the 50 lowest ranked firms, and 50 cases selected at random from firms in quartiles two and three. Candidate earnings-related performance sentences are identified using a keyword list resulting in 8,805 target performance sentences together with adjacent lead and lag sentences for manual coding (26,415 individual sentences). We eliminate 1,604 invalid target sentences. The remaining 7,201 target sentences comprise 3,396 performance sentences and 3,805 target sentences judged not to be valid performance sentences because they either discuss non-earnings features such as cash, debt, inventory, production, etc. or they refer to results for the current fiscal year, Sentences coded as neutral or mixed tone (N = 219) are removed from the final training sample for Tone. The presence of an attribution is treated as a binary outcome equal to one if management explicitly link performance with one or more fundamental determinants and zero otherwise. Attribution type distinguishes between internal and externals factors. Internal factors as those over which management exercise direct control. External factors as those over which management are not expected to exercise direct control. Both attribution types may be present in a single sentence. Such cases (N = 180) are excluded from the Attribution type training sample because they do not provide incremental information for the binary classification task. Twenty-six cases where two coders plus a judge are unable to agree on the nature of the attribution are also removed. The holdout sample is constructed using earnings announcements released in 2012, following the same coding strategy.

Table 2. Classification methods of discourse features.

	Sentence-level discourse feature:		
Classifier	Tone	Attribution	Attribution type
Machine learning algorithm:			
Naïve Bayes	✓	✓	✓
Random forests	✓	✓	✓
Support vector machines	✓	✓	✓
Recursive neural network	✓	✓	✓
Wordlist:			
HEN06	✓		
HEN08	✓		
L&M	✓		
LIWC		✓	
DIK		✓	
ATT_ALL		✓	
ATT_50		✓	
ATT_TYPE			✓

Tone measures the polarity of target performance sentences. *Tone* as either positive, negative, neutral, or unclear. *Attribution* occurs when management relate a performance outcome to at least one fundamental determinant such as operating efficiency, product development, adverse trading conditions. *Attribution* is a binary outcome equal to one if management explicitly link the performance outcome with one or more fundamental determinants and zero otherwise. *Attribution type* categories attributions according to whether the fundamental determinant(s) cited by management relate to internal or external factors. Internal factors are those over which management has direct control, such as strategic reorientation, cost control, product design, marketing initiatives, labor relations, etc. External factors are those over which management are not expected to exercise direct control such as market competition, input prices, exchange rates, weather, etc. *Attribution type* comprises separate binary outcomes equal to one for the presence of at least on internal (external) attribution and zero otherwise. Four machine learning algorithms are used to classify *Tone*, *Attribution* and *Attribution type*. Details of each algorithm are provided in an appendix. Three wordlists are used to classify *Tone*. HEN06 and HEN08 are the wordlists from Henry (2007) and Henry (2008), and L&M comprises the positive and negative wordlists developed by Loughran and McDonald (2011). Four wordlists are used to classify *Attribution*. LIWC is a version of the causal wordlist from Linguistic Inquirer and Word Count and applied by Zhang et al. (2019). DIK is causation wordlist developed by Dikolli et al. (2017). ATT_ALL is a self-constructed domain-specific attribution wordlist, further details of which are provided in the Appendix. ATT_50 comprises the 50 most frequently occurring words from ATT_ALL in the training sample. A single wordlist is used to classify *Attribution type*. ATT_TYPE is a self-constructed domain-specific wordlist, further details of which are described in the Appendix.

Table 3: Classification results for performance sentence tone using training sample

	F1-Scores by tone class:		Classification averages (%):	
Classifiers	Positive	Negative	Macro-F1	Accuracy
Machine learning:				
Naïve Bayes	89.09	53.56	74.06	82.47
Random forest	90.41	63.97	76.97	84.23
Support vector machines	89.40	57.01	76.40	81.87
MLP	88.86	60.29	74.58	81.84
Wordlists:				
HEN06	84.40	65.16	74.78	78.47
HEN08	84.06	64.72	74.39	78.06
L&M	56.73	49.08	52.90	53.26

Values for machine learnings models reflect average values computed using results from each fold in the 10-fold cross-validation. Macro-F1 scores are not the average of Positive and Negative class F1-Scores due to averaging across the 10 folds. Accuracy and Macro F1 values for wordlists are the actual fraction of sentences where the wordlist prediction equals the manual annotation in a single classification pass. HEN06 and HEN08 refer to the dictionaries proposed by Henry (2006) and Henry (2008), respectively. L&M refers to the dictionaries of positive and negative words proposed by Loughran and McDonald (2010).

Table 4. Classification results for performance sentence tone for the holdout sample.

	Classification performance metrics by tone class:							
	Positive			Negative			Classification averages:	
Classification method	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Macro-F1	Accuracy
Random forest	94.92	80.72	87.25	52.61	83.20	64.46	76.36	81.23
HEN06	77.00	87.09	81.73	76.13	61.29	67.91	74.82	76.72
HEN08	76.83	86.74	81.48	75.44	60.90	67.39	74.44	76.38
L&M	43.83	90.85	59.13	90.77	43.60	58.90	59.02	59.02

Random forest is selected for comparison purposes as the best performing learning algorithm for the training sample based on highest average Accuracy score.

Table 5. Classification results for the presence of attribution using training sample**.**

	F1-Scores by attribution class:		Classification averages (%):	
Classifiers	Yes	No	Macro-F1	Accuracy
Machine learning:				
Naïve Bayes	23.00	86.97	62.14	77.12
Random forest	35.39	88.44	71.85	81.95
Support vector machines	44.36	88.13	71.52	78.77
MLP	49.43	87.87	70.61	80.83
Wordlists:				
LIWC	15.76	83.44	49.60	72.32
DIK	44.99	67.64	56.32	59.28
ATT_ALL	55.10	78.20	66.65	70.67
ATT_50	51.90	80.80	66.35	72.57

Values for machine learnings models reflect average values computed using results from each fold in the 10-fold cross-validation. Macro-F1 scores are not the average of Positive and Negative class F1-Scores due to averaging across the 10 folds. Accuracy and Macro F1 values for wordlists are the actual fraction of sentences where the wordlist prediction equals the manual annotation in a single classification pass. LIWC and DIK are the causal reasoning wordlists derived using Language Inquirer and Word Count by Zhang et al. (2019) and Dikolli et al. (2017), respectively. ATT_ALL is a self-constructed domain-specific attribution wordlist, further details of which are provided in the Appendix. ATT_50 comprises the 50 most frequently occurring words from ATT_ALL in the training sample. .

Table 6. Classification results for the presence of attribution for the holdout sample.

	Classification performance metrics by attribution class:							
	Attribution present			Attribution not present			Classification averages:	
Classification method	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Macro-F1	Accuracy
Random forest	26.50	67.72	38.10	96.43	82.27	88.79	63.44	81.01
LIWC	14.60	29.81	19.60	90.28	78.89	84.20	51.90	73.60
DIK	78.16	33.72	47.11	56.56	90.15	69.51	58.31	61.32
ATT_ALL	77.12	40.14	52.80	67.48	91.25	77.58	65.19	69.60
ATT_50	69.05	42.14	52.33	73.18	89.32	80.45	66.39	72.27

Random forest is selected for comparison purposes as the best performing learning algorithm for the training sample based on highest average Accuracy score.

Table 7. Classification results for attribution type (conditional on the presence of an attribution) using training sample.

	F1-Scores by attribution class:		Classification averages (%):	
Classifiers	Internal	External	Macro-F1	Accuracy
Machine learning:				
Naïve Bayes	86.01	82.56	84.45	84.29
Random forest	84.89	79.71	82.95	82.85
Support vector machines	86.75	81.93	84.78	84.73
MLP	85.96	81.91	84.29	84.94
Wordlists:				
ATT_TYPE	66.05	74.38	70.21	70.82

Values for machine learnings models reflect average values computed using results from each fold in the 10-fold cross-validation. Macro-F1 scores are not the average of Positive and Negative class F1-Scores due to averaging across the 10 folds. Accuracy and Macro F1 values for wordlists are the actual fraction of sentences where the wordlist prediction equals the manual annotation in a single classification pass. ATT_TYPE is a self-constructed domain-specific wordlist, further details of which are described in the Appendix.

Table 8. Classification results for attribution type (conditional on the presence of an attribution) for the holdout sample.

	Classification performance metrics by attribution type class:							
	Internal			External			Classification averages:	
Classification method	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Macro-F1	Accuracy
Multilayer perceptron	87.58	84.15	85.83	76.04	80.82	78.35	82.09	82.87
ATT_TYPE	33.60	87.30	48.53	92.90	49.06	64.21	56.37	57.78
P-values for pairwise difference			0.01			0.01	0.01	0.01

Multilayer perceptron is selected for comparison purposes as the best performing learning algorithm for the training sample based on highest average Accuracy score.

Figure 1: Summary of research design

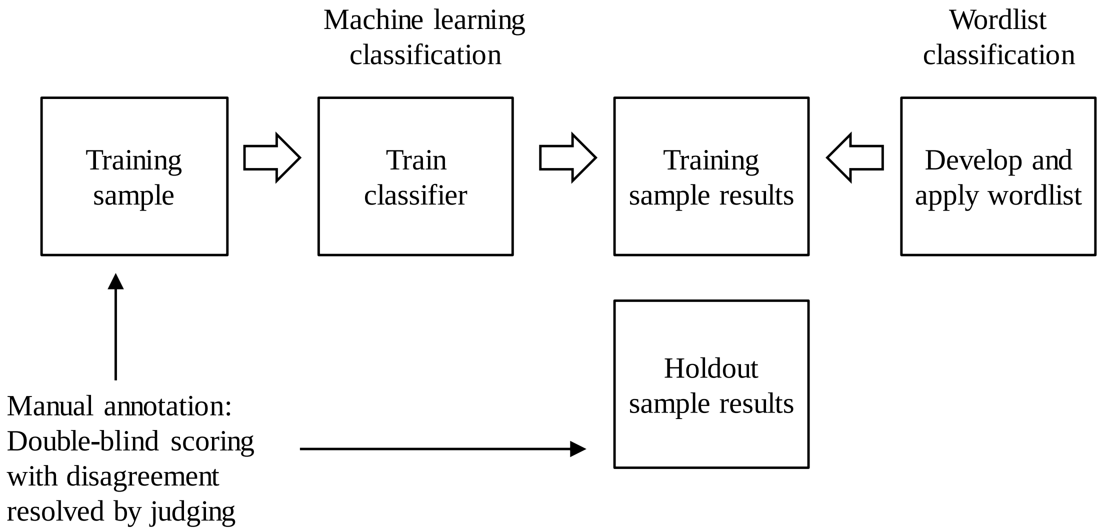
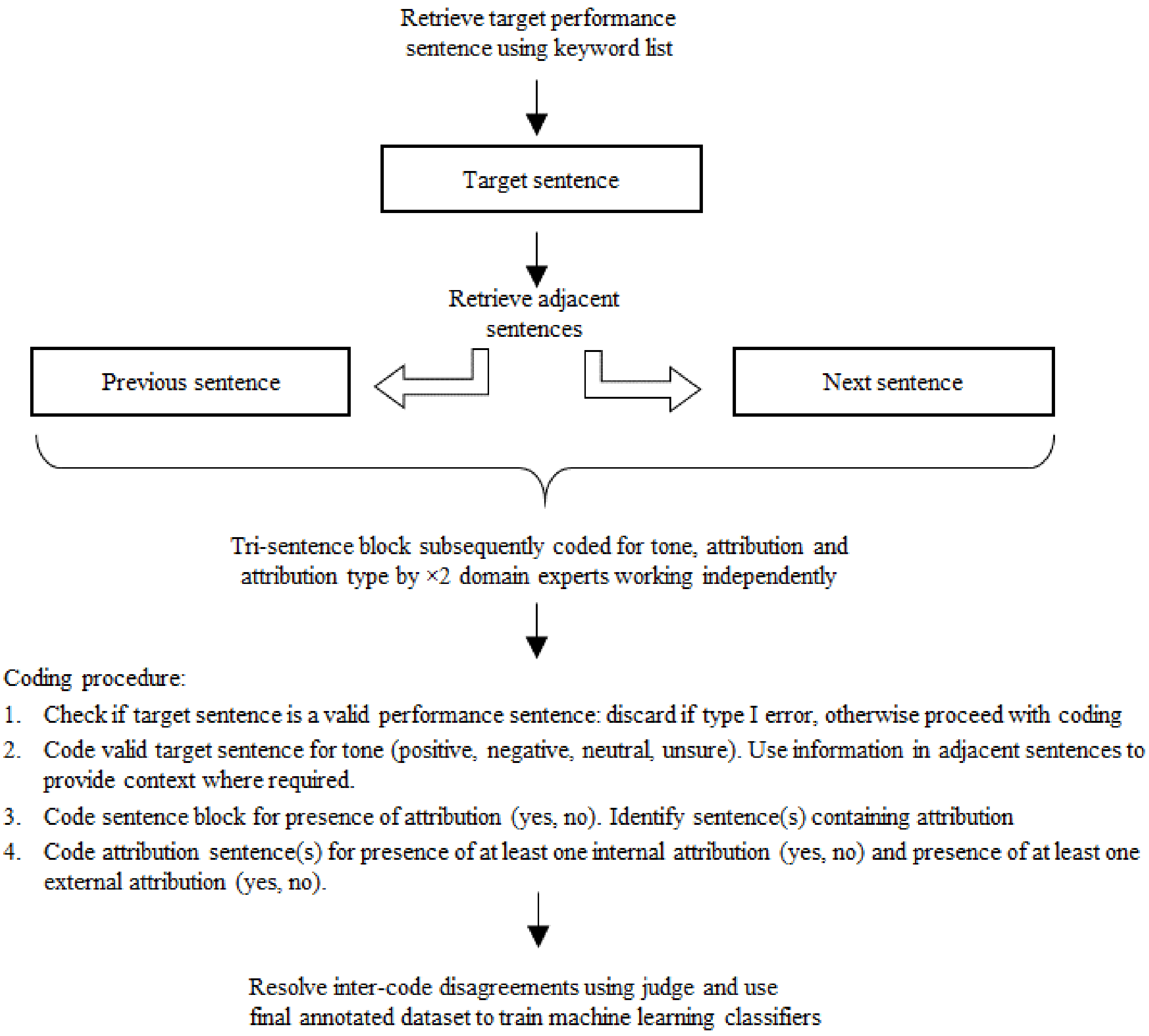

Machine learning classification
Wordlist classification
Training sample
Train classifier
Training sample results
Develop and apply wordlist
Holdout sample results
Manual annotation: Double-blind scoring with disagreement resolved by judging

Figure 2. Details of manual coding procedure

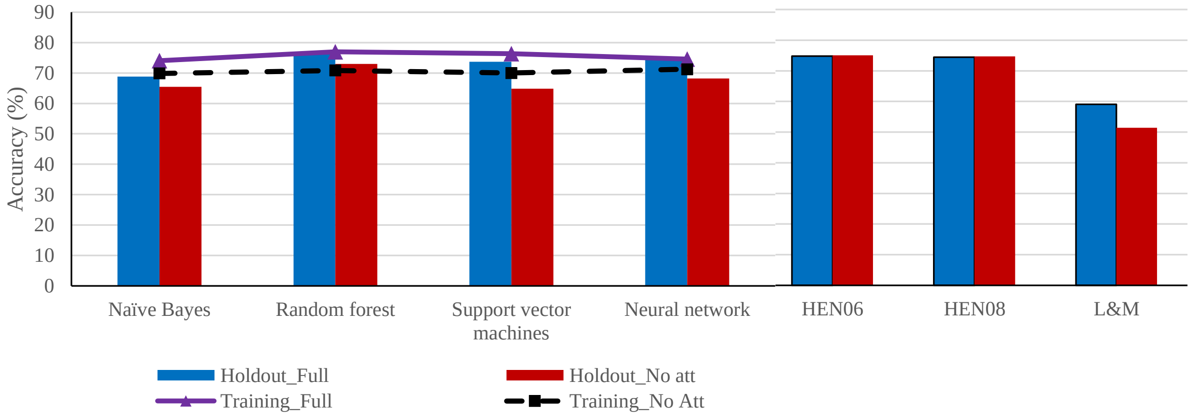

Retrieve target performance sentence using keyword list
Target sentence
Retrieve adjacent sentences
Previous sentence
Next sentence
Tri-sentence block subsequently coded for tone, attribution and attribution type by ×2 domain experts working independently
Resolve inter-code disagreements using judge and use final annotated dataset to train machine learning classifiers

Coding procedure:

1. Check if target sentence is a valid performance sentence: discard if type I error, otherwise proceed with coding
2. Code valid target sentence for tone (positive, negative, neutral, unsure). Use information in adjacent sentences to provide context where required.
3. Code sentence block for presence of attribution (yes, no). Identify sentence(s) containing attribution
4. Code attribution sentence(s) for presence of at least one internal attribution (yes, no) and presence of at least one external attribution (yes, no).

Figure 3: Impact of sentence complexity on classification performance using the holdout sample

Panel A: Tone conditional on the presence of attribution



Panel B: Attribution conditional on the presence of a performance statement

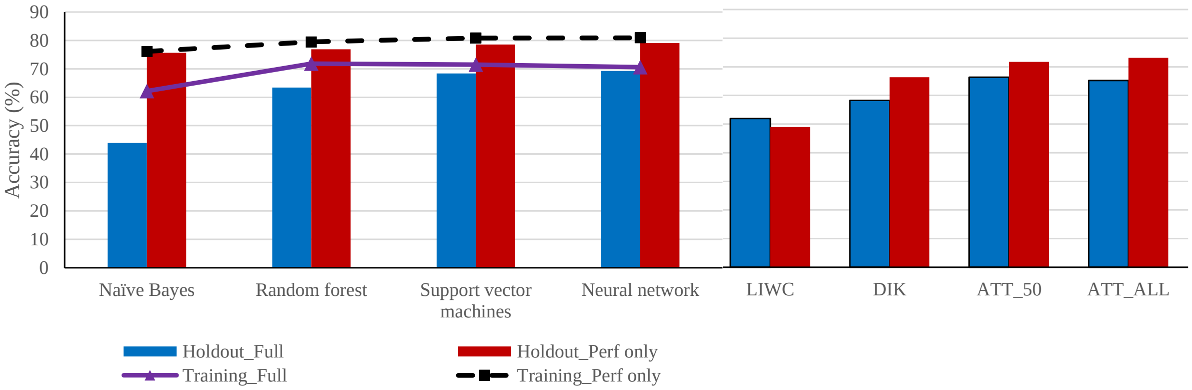
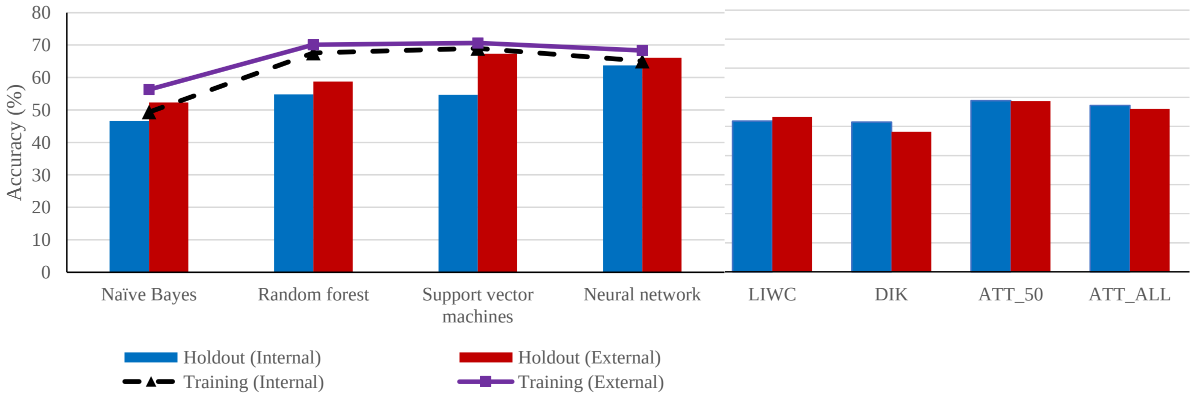


Figure 4. Impact of attribution type on classification performance for attribution detection using holdout sample.

Panel A: Macro F1 scores



Panel B: Class F1 scores

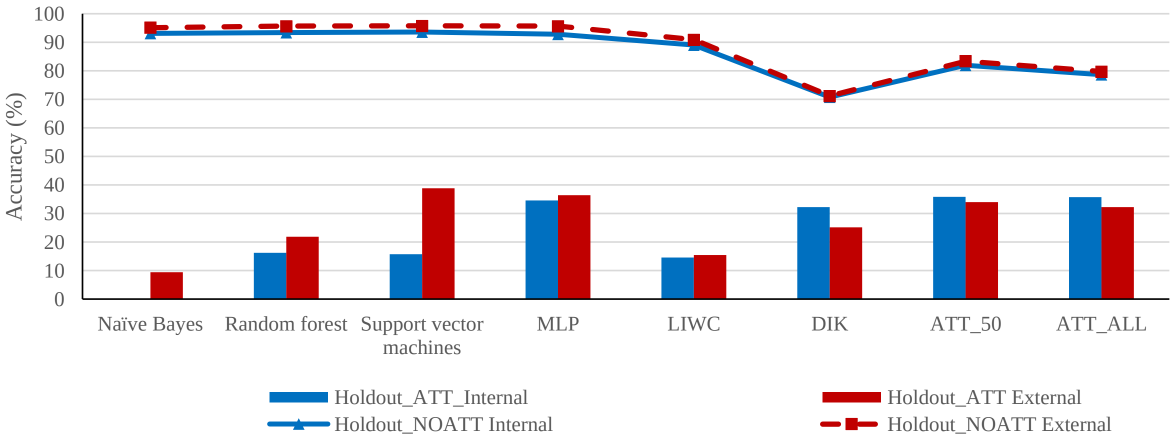
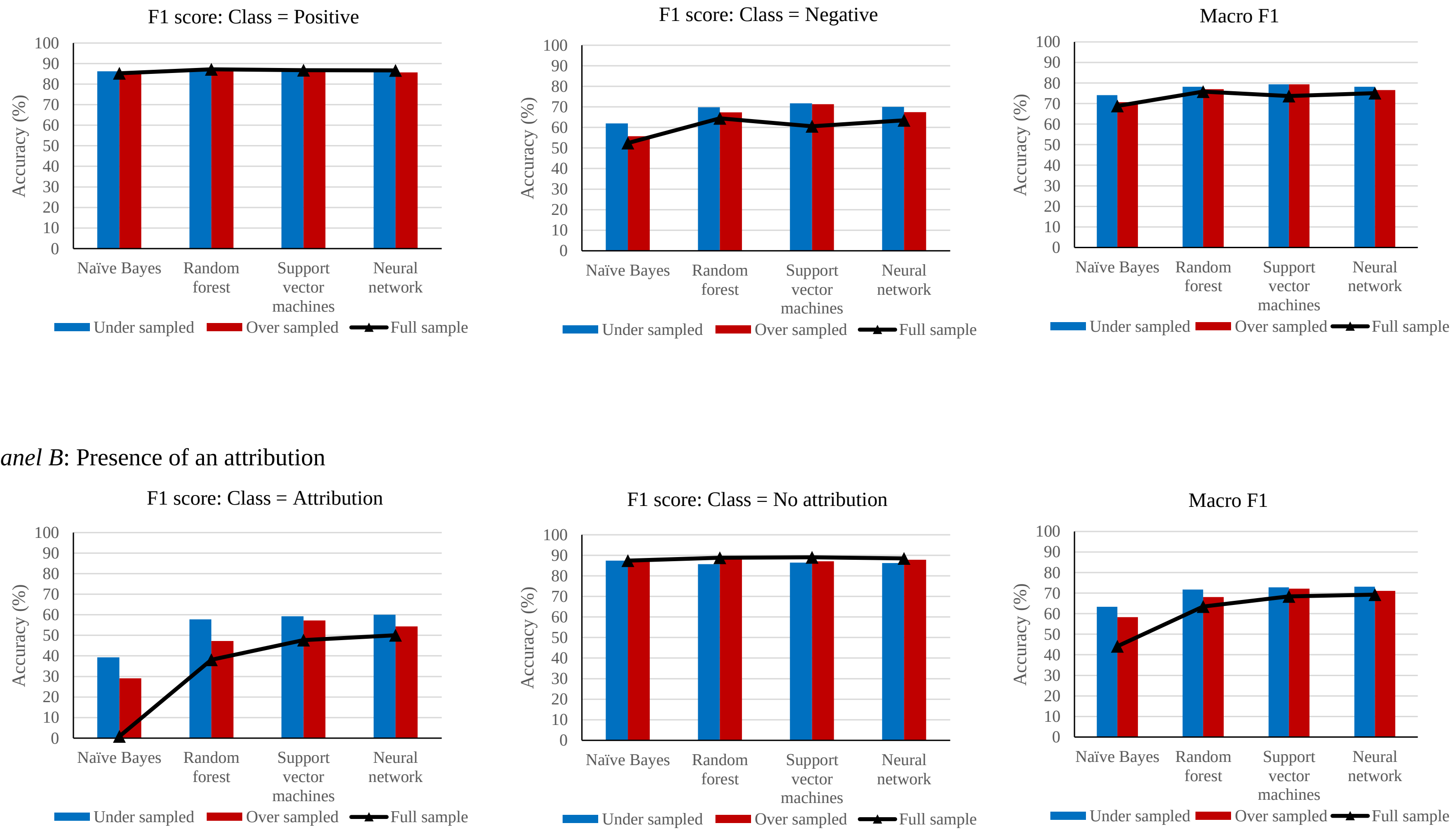


Figure 5: Impact of sample balancing on classification performance of machine learning algorithms using the holdout sample.



Panel A: Sentence tone

Panel B: Presence of an attribution