View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Lancaster E-Prints

Learning Tone and Attribution for Financial Text Mining

Mahmoud El-Haj[1] **, Paul Rayson**[1] **, Steven Young**[1]**, Andrew Moore**[1]**,**
Martin Walker[2]**, Thomas Schleicher**[2]**, Vasiliki Athanasakou**[3]
[1]Lancaster University, UK
[2]University of Manchester, UK
[3]London School of Economics, UK

Abstract

Attribution bias refers to the tendency of people to attribute successes to their own abilities but failures to external factors. In a business context an internal factor might be the restructuring of the firm and an external factor might be an unfavourable change in exchange or interest rates. In accounting research, the presence of an attribution bias has been demonstrated for the narrative sections of the annual financial reports. Previous studies have applied manual content analysis to this problem but in this paper we present novel work to automate the analysis of attribution bias through using machine learning algorithms. Previous studies have only applied manual content analysis on a small scale to reveal such a bias in the narrative section of annual financial reports. In our work a group of experts in accounting and finance labelled and annotated a list of 32,449 sentences from a random sample of UK Preliminary Earning Announcements (PEAs) to allow us to examine whether sentences in PEAs contain internal or external attribution and which kinds of attributions are linked to positive or negative performance. We wished to examine whether human annotators could agree on coding this difficult task and whether Machine Learning (ML) could be applied reliably to replicate the coding process on a much larger scale. Our best machine learning algorithm correctly classified performance sentences with 70% accuracy and detected tone and attribution in financial PEAs with accuracy of 79%.

Keywords: PEA, financial, narratives, machine learning, string vectors, NLP, semantic, part of speech, tagger

1. Introduction

In social science, attribution bias refers to human beings' tendency to attribute successes to their own abilities but failures to external factors. In the financial domain, internal factors for an organisation could include cost reduction programmes and employee training, while external factors include movements in exchange or interest rates. In accounting research, the presence of an attribution bias has been demonstrated for the narrative sections of annual financial reports (AFRs). For instance, Clatworthy and Jones (Clatworthy and Jones, 2003) confirm the existence of an attribution bias in the UK Chairman's Statement sections of AFRs when they find that statements about last year's positive financial performance are typically explained through superior management skills whereas statements about negative financial performance are typically attributed to unfavourable factors outside the firm's control. Similar findings are reported in Aerts (Aerts, 1994), Aerts (Aerts, 2005) and Hooghiemstra (Hooghiemstra, 2000). All previous studies have applied manual content analysis to the examination of attribution bias in narratives. The experiment presented here breaks new ground by quantifying the difficulty of the manual annotation task in a new type of financial narrative and by attempting to automate the analysis of attribution bias in order to scale up the analysis for a much larger number of companies. In the UK, firms typically release the Preliminary Earning Announcement (PEA) several weeks before the annual report. Investor relations experts report that this is more important than the AFR in terms of influencing market perceptions, hence our focus on PEAs for this experiment.

Teaching the computer to automatically detect attributions obviously requires a set of manually coded sentences. Manual coding in any domain is not an exact science and different coders might not agree on the coding of a particular sentence. In the financial narrative domain in particular, different coders might respond to ambiguities in sentences in different ways. While case law rules can minimise disagreement among coders, with only small scale studies taking place in previous work, we hypothesise that it may be impossible to cater for all possible scenarios and keywords, and hence some degree of disagreement will always persist. Our research here will focus on the evaluation of inter-rater reliability to see how hard a problem this is and then the results of our ML experiments to evaluate how well such algorithms can be trained on this problem.

The remainder of the paper is organised as follows. In section 2. we review previous applications of NLP and ML to financial narratives. Section 3. describes the dataset that we have collated for our experiments. We explain our process of manual coding in section 4. and the ML models in section 5. Results and discussion appear in section 6. and we conclude in section 7.

2. Related Work

There has been growing interest in recent years in the application of Natural Language Processing (NLP) and text analysis techniques in the financial domain. One of the largest areas of recent research has been the application of sentiment analysis to stock–related tweets with the intention of predicting the stock market performance (Devitt and Ahmad, 2007; Schumaker, 2010; Im et al., 2013; Ferreira et al., 2014; Neuenschwander et al., 2014). Some studies have taken this a step further by trying to explain the impact on investors' behaviour from negative reports in the financial media of corporate actions (Moniz and de Jong, 2014). The work by Shuyu (2016) describes the use of computer techniques to measure causal reasoning in financial earnings-related outcomes of a large sample of 10-K (annual reports) filings of US firms. Their work showed positive and sig-

nificant association between firms' causal reasoning intensity and other analyst earning and forecasts. In their work they focused on using non-language dependent approaches by applying simple text analysis techniques and frequency count using PERL.

Whitelaw and Patrick (2004) and Goel and Gangolly (2012) have investigated systemic and other predictive features for the task of identifying financial scam documents. NLP researchers have also attempted to develop empirical techniques for ranking risk (Kogan et al., 2009; Tsai and Wang, 2012) particularly in the American context. In the Accounting and Finance literature, there is an increasing body of work using basic word-list and ML approaches to examine the information content of forward-looking statements (e.g., Li (Li, 2010)) and other work has looked at the relationship of optimistic versus mild statements in the context of positive and negative financial performance (Chen et al., 2013). However, there is no large-scale empirical work on detecting and measuring tone and attribution in financial narratives. In terms of novelty in the ML experiments over and above the state of the art in Accounting and Finance literature, we include part-of-speech and semantic features alongside traditional bag-of-word models.

3. Dataset

We collected 500 Preliminary Earning Announcements (PEAs) released between 2010 and 2012 by firms listed on the London Stock Exchange.[1] The analysis of this paper focuses on the PEAs for the middle year for which we had 140 PEAs. The PEAs were manually downloaded from Perfect Information.[2] PEAs typically contain a commentary section followed by a section containing the summary financial statements. We extracted the commentary section since that is more likely to contain attribution by excluding any highlight,[3] bullet points or formal financial statements, including financial statement footnotes since those sections typically lack any attribution sentences.

First, we converted the original PEA Word documents to HTML file format, and then we split the PEAs into sentences using the Stanford NLP sentence splitter.[4] Expert annotators marked each PEA with tags indicating the start and end of the commentary section since that is more likely to contain attribution.

The total number of sentences extracted was 98,958 and the annotators then reduced the number of sentences to 32,449 by ignoring sentences with zero frequency of performance keywords.[5] We apply the machine learning algorithms on the reduced data only. We calculated recall accuracy of the performance sentences by asking a fifth annotator to go through a random sample of sentences that have been ignored automatically, to classify them as performance/non-performance using the same definition followed by the original annotators[6]. Recall results (94%) showed high accuracy and a very low false-negatives, which suggests the absence of any of the hand-crafted performance keywords by domain experts to be a good indicator of non-performance sentences. This process estimated false-negatives, and in section 4. we show how we estimated false-positives.

4. Manual Coding and Inter-rater Reliability

The process starts with the annotators confirming whether or not a sentence is performance related – agreeing or disagreeing with the keywords list (estimating false-positives). The annotators will only code performance related sentences and ignore those that are not. Figure 1 shows a sample of the form used by the annotators to code the sentences.

Figure 1: Annotation Form

Codes used by the annotators (Figure 2) were as follows:

Performance – whether or not a sentence is performance related.

Sent_Tone – positivity or negativity of the annotated sentences. The human annotators assigned Negative (NEG), Positive (POS) or Neutral (NEU).

Att – whether a sentence contains attribution and where it does whether this attribution is internal (INT) or external (EXT). An example of internal attribution would be "our continuing focus on tight cost control.". External attribution is something driven by external circumstances e.g. "challenging consumer environment".

Attr_Tone – whether the impact of the internal or external attribution on performance is positive (POS), negative (NEG), or neutral (NEU).

4.1. Performance Keywords

The expert annotators formed a list of keywords to identify performance in sentences as follows: 'sales', 'revenue*',

[1] We are making the dataset publicly available: http://ucrel.lancs.ac.uk/cfie/lrec2016/

[2] http://www.perfectinformation.com/

[3] The first section in financial reports is typically a highlight section where the firm 'highlights' the progress, both financial and non-financial, over the last 12 months.

[4] http://nlp.stanford.edu/software/tokenizer.shtml

[5] The expert annotators formed a list of keywords to identify performance in sentences as follows: 'sales', 'revenue*', 'turnover', 'trading', 'cost*', 'expense*', 'income', 'earnings', 'E.P.S.', 'profit*', 'loss*', 'margin*', 'result*'. The * acting as a wildcard for zero or more characters at the end of the word. We manually checked a small number of sentences while amending this list in order to maximise recall.

[6] Performance: "The economic outcome of the operating activities during the financial year along with an indication of the quality of the operation outcome"

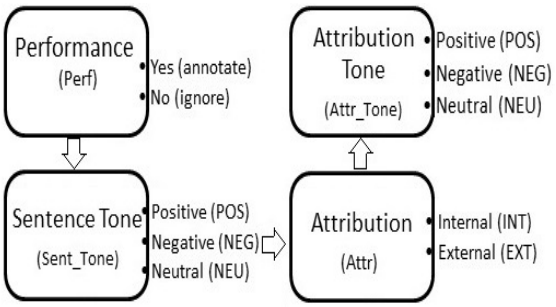

Figure 2: Annotation process

'turnover', 'trading', 'cost*', 'expense*', 'income', 'earnings', 'E.P.S.', 'profit*', 'loss*', 'margin*', 'result*'.[7]
The annotators created this keyword list above based on the following observations:

- To focus on Profit/s and its two main components i.e. Sales minus Costs.
- Income and Earnings are common alternative words for Profit.
- When a firm reports its profits for the year this often is referred to as the "results" announcement and therefore included result and results.
- Profitability expresses profits as a ratio. It is very commonly used.
- Margin/Margins is typically used to express profits as a proportion of sales.
- Many firms refer to profits per share using the term earnings per share or EPS or E.P.S.
- Loss/Losses allows for the possibility of negative profits.
- Revenue/Revenues/Trading are common alternative wordings for Sales.
- Expenses are often referred to as an alternative to Costs.

4.2. Annotation Process

Sentences from the 140 PEAs were divided equally between the four annotators. The split was randomised by document and not by sentences in order to ensure that the annotation could be carried out in context with the previous and the following sentences.
The process starts with the annotators confirming whether a sentence is performance related or not – agreeing or disagreeing with the keywords list. For example, the sentence '2010 adjusted earnings per share is a 53 week number' includes a performance-related keyword, namely 'earnings per share' but was discarded as it does not comment on last year's earnings performance. Instead it simply defines the length of last year's accounting period. In contrast, the statement 'the summer weather disappointed and impacted soft drinks sales in each of our markets' indicates that last year's sales performance was disappointing and, hence, was coded, under *Sent_Tone* (tone of the sentence), as a negative performance statement, NEG. The sentence was also coded as explaining the negative sales performance as being impacted by bad weather leading to a coding of *EXT* for *Attr* (i.e. external attribution) and *NEG* for *Attr_TONE* (i.e. negative tone of the attribution), because the impact of the bad weather on soft drink sales is clearly negative.
Explicitly coding the negative impact of weather on soft drink sales is necessary as it is also quite common that the tone in the performance statement is not consistent with the impact of the external (or internal) factor as demonstrated, for example, in 'our profit margins increased despite higher raw material prices'. This sentence is coded *POS*, *EXT*, and *NEG*, under *Sent_Tone*, *Attr* and *Attr_Tone*, respectively. Specifically, *POS* refers to the increase in the profit margin while NEG refers to the increase in raw material prices that negatively affects profit margins (as margins are defined as the difference between sales and direct costs, divided by sales). So, in many ways the increase in profit margin could be seen as even more positive in the presence of a raw material price increase.
Of course, one might argue that the raw material price increase cannot explain the increase in profit margin and hence should not be considered in an attribution study. While, theoretically, this is an interesting argument we have decided to ignore it during the coding process in this experiment.
In a small number of cases one performance-related statement was associated with multiple attribution statements. For example, 'Caltech [...] has continued to perform well with growing demand in existing markets and the successful entry into new markets being reflected in a significant increase in sales and profit' suggests that the positive sales and profit performance can be explained by both additional demand in the external market and the firm's internal decision to enter new markets, and hence it is coded twice (as two different sentences) one under *Sent_Tone*, *Attr*, *Attr_Tone*, as *POS*, *EXT*, *POS* and again as *POS*, *INT*, *POS*, respectively. As always, attributions need to be included in the same sentence as the performance-related statement, or in one of the two previous or two following sentences around the performance statement, giving a maximum of five sentences per performance-related statement that are considered for the coding of attributions.
Performance statements are defined as statements about performance in the last twelve months about sales, costs, and profits. In contrast, we ignored statements about future performance as in 'Regarding current trading he added: "The Group has continued to perform in-line with the Board's expectations".' and instead marked such a statement as *FL*, for 'forward-looking', under *Sent_Tone*. Also, statements about financial performance, non-recurring operating performance, and balance sheet and cash flow statement line item were not coded and instead an *OFC*, for 'other financial commentary', was entered under *Sent_Tone*. In both cases attributions were not coded, even if present.
Manual coding is not an exact science and different coders

[7]The * acting as a wildcard for zero or more characters at the end of the word.

might not agree on the coding of a particular sentence. Imagine, for example, that a firm simply states that 'the weather impacted soft drinks sales in each of our markets'. Unlike the earlier example statement, there is no longer an explicit reference to the 'disappointing' weather and this lack of explicit reference makes the statement ambiguous. Different coders might respond to this ambiguity in different ways. For example, one coder might well decide that the direction of the sales change is no longer sufficiently clear, as is the impact of the weather on the sales change, and, as a result, code this statement as neutral, that is as *NEU*, *EXT*, *NEU* under *Sent_Tone*, *Attr*, *Attr_Tone*. In contrast, another coder might well decide to make the ambiguity inherent in the statement explicit by marking this sentence as *UNSURE*, *EXT*, *UNSURE*. Finally, a third coder might draw on her experience from previous manual content analysis and decide that the word 'impact' is used, more often than not, in negative tone statements, and thus decide that this statement must imply a negative tone, thus leading to a coding of *NEG*, *EXT*, and *NEG*. While coding rules can minimise disagreement among coder, it is impossible to cater for all possible scenarios and keywords, and hence some degree of disagreement will always persist. To minimise this, we developed annotation guidelines and revised them in light of an initial round of coding.[8] Out of 32,449 the annotators found 3,500 performance sentences and 1,480 non-performance sentences (those *FL* or *OFC*).[9] The remaining sentences were found to be unrelated/incomplete sentences and hence were not coded by the annotators. These numbers serve to illustrate the drawbacks of the word list approach to identifying performance sentences.

4.3. Inter-rater Reliability

To measure the degree of agreement among the raters we asked the four annotators 'A' to 'D' to blind recode 1000 sentences for each other. We measured the percentage of agreement between coders on four levels a) performance (Perf), b) sentence tone (Sent_Tone), c) attribution (Attr), d) attribution tone (Attr_Tone). Table 1 with Cohen's Kappa scores shows substantial (or better) agreement between the coders but in some pairs illustrates how hard such coding is even for expert judges. The results in the table show a one-way comparison (i.e. A vs B, B vs A), and this simply reflects the order in which the annotation was carried out.

Coder	Perf	Sent_Tone	Attr	Attr_Tone
A vs B	.73	.99	.90	.90
B vs A	.71	.98	.93	.93
C vs D	.64	.94	.71	.71
D vs C	.86	.99	.80	.80

Table 1: Inter-rater Reliability – Cohen's Kappa

[8]These are available at http://ucrel.lancs.ac.uk/cfie/lrec2016/ along with our manually coded dataset.

[9]*FL* 'forward-looking' and *OFC* 'other financial commentary' are considered non-performance statements. In both cases attributions were not coded, even if present.

5. Machine Learning Models

In our study we used the Weka software (Hall et al., 2009) and applied four machine learning models: SMO, Logistic Regression, Random Forest and Naïve Bayes. Naïve Bayes tends to be the most used technique in the business literature (Li, 2010; Zhang, 2004; Manning et al., 2008). We tested accuracy using a 10-fold cross validation, splitting the entire sample into 90% training and 10% testing for each fold.

5.1. Machine Learning Features

We combined linguistic features that have been assigned manually by the human annotators with automatically extracted features for training purposes. Manual features include all the codes assigned by the expert annotators as in section 4. Automatically assigned features include three sets of features, including the novel use of conceptual annotation:

Keywords Frequency – for each sentence we count the individual frequency for each word in the performance keyword list.

Part of speech tags (POST) – we used CLAWS (Garside and Smith, 1997) part of speech tagger to tag the sentences. We only consider a number of tags that the annotators were interested in and those include: present and past verbs, singular and plural pronouns in addition to counting the shift between the tags (e.g. number of times the writers switch from using singular to plural pronouns or switching from past to present verbs).

Semantic Tagging – we used the USAS semantic tagger for English (Rayson et al., 2004). We only count frequency for semantic tags A2.1 (Affect:- Modify, change) and A2.2 (Affect:- Cause/Connected) as these were thought to potentially provide good indicators for performance sentences. This is conceptual level annotation relying on an ontology of the domain (i.e. accounting and finance).

6. Results and Discussion

Overall, the machine learning algorithm correctly identified performance sentences with 70% accuracy. This was calculated by training the classifier using sentences manually coded as 'performance' or 'non-performance' (FL or OFC).

Table 2 illustrates the results we achieved running four machine learning algorithms to detect sentence tone, attribution and attribution's tone. Taking into account the inter-rater reliability scores, it is to be expected that detecting the sentence's tone has proven to be more accurate than detecting attribution or attribution's tone. The table also shows that using features such as part of speech tags did not have a significant impact when compared to the use of performance keywords but was enough to achieve accuracy similar to using expert-generated keyword list.

Table 3 shows that using attribution's tone has a significant effect on the detection process, which is in line with the experts in accounting and finance expectations where they believe that the tone of the attribution could help in detecting the attribution type whether it is external or internal with the later tending to be positive or neutral in most of the cases. Similarly, table 4 shows that attribution's type

ML	Feature	Key + POST	Key	POST	Sem
SMO	Sent_Tone	75.9	76.0	71.0	71.0
	Attr	57.1	57.1	56.7	56.8
	Attr_Tone	60.2	60.4	59.8	60.0
LR	Sent_Tone	71.2	70.9	71.0	70.9
	Attr	57.4	57.7	57.2	56.4
	Attr_Tone	59.6	60.3	59.1	60.3
RF	Sent_Tone	67.1	69.8	70.0	71.0
	Attr	57.2	54.3	56.7	55.4
	Attr_Tone	55.6	57.7	57.7	59.4
NB	Sent_Tone	67.9	68.2	71.8	70.5
	Attr	51.0	54.9	48.2	56.2
	Attr_Tone	50.3	55.8	44.9	60.3

Table 2: Effect of keywords, POS and Sem Taggers on attribution and sentence's/attribution's tone
LR: Logistic Regression, RF: Random Forest, NB: Naïve Bayes, Key: Keywords, Sem: Semantic Tagger

ML	Attr_Tone + Key + POST	Attr_Tone + Key	Attr_Tone + POST	Attr_Tone + Sem
SMO	76.3	76.2	76.3	76.3
LR	79.1	79.1	79.7	76.3
RF	76.9	76.1	78.3	75.5
NB	69.9	72.8	79.2	75.9

Table 3: Effect of attribution's tone on attribution

could help in detecting attribution's tone. In addition, the results show that part of speech tagging helped in slightly enhancing the detection process, this shows that singular and plural pronouns in addition to past and present verbs and the shift between them could help in telling when there exist a positive or negative internal or external attribution. This aligns with Pang et al. (Pang et al., 2002) where using POS tends to improve accuracy for a NB classifier. The use of singular and plural pronouns could be considered as an indicator of attribution's positivity where a company usually tends to speak positively when discussing attribution by its own management.

ML	Att + Key + POST	Att + Key	Att + POST	Att + Sem
SMO	75.7	75.7	75.8	75.8
LR	76.7	77.0	77.7	75.7
RF	74.0	73.8	76.6	75.3
NB	69.1	70.4	74.6	75.8

Table 4: Effect of attribution, keywords, POS and Sem Taggers on attribution's tone

Testing the effect of keywords, POST and Sent_Tone on attribution's tone suggests the tone of the sentence to be a good indicator of the impacting result of the attribution, which sounds reasonable to consider the impact of the attribution to be positive when the tone of the complete sentence is positive as well. We also performed conceptual annotation where we counted the frequency of semantic (Sem) tags *A2.1* (Affect:- 'Modify', 'Change') and *A2.2* (Affect:- 'Cause'/'Connected') those believed by the accounting and finance expert to be good indicators of attribution. This is proven true considering the results we obtained using conceptual annotation of only *A2.1* and *A2.2* semantic tags were enough to achieve an accuracy level consistent with that obtained by the keyword list as shown in Tables 2, 3, and 4.

ML	Sent_Tone	Attr	Attr_Tone
SMO	76.5	58.6	66.6
LR	75.7	58.5	64.5
RF	76.3	55.2	66.9
NB	72.1	57.8	64.9

Table 5: String to Word Vector Results

Most Frequent Class	Accuracy
Sent_Tone	71.0
Attr	56.8
Attr_Tone	59.9

Table 6: Most Frequent Class Results

We also ran another experiment using Weka's *StringToWordVector* unsupervised filter to convert all PEA sentences into vectors of words contained in the sentences. We kept the top 100 words according to TF.IDF weights. We also used the attribution selection filter as a middle man by applying the *InfoGainAttributeEval* which evaluates the worth of an attribute by measuring the information gain with respect to the class. Running the classifier with the applied filters we reached the accuracies in Table 5.
As a baseline we created a simple classifier that always selects the most frequent class. Table 6 shows the simple classifier results. Comparing the results of our machine learning experiment using manual and automatic features to the baseline shows that using the manually created keyword list does not usually increase the chance of detecting tone and attribution in PEAs. While on the other hand the use of top ranked keyword lists from the PEAs text along with machine learning attribution selection has helped by boosting the results above the baseline in most cases.

7. Conclusion

This is the first large scale study of its kind combining both human and machine evaluations of this classification task. We applied four Machine Learning models: SMO, Logistic Regression, Random Forest and Naïve Bayes and evaluated the results using 10 fold cross validation for each model. The linguistic features we used to train the systems were generated manually by the expert annotators and automatically using NLP tools such as using Part of Speech and Semantic Taggers, with the latter novel application of conceptual annotation showing promising results. The best machine learning algorithm correctly classified performance sentences with 70% accuracy and detected tone and attribution in financial PEAs with accuracy of 79%. The expert

annotators inter-rater reliability showed Kappa scores up to 0.86 and 0.94–0.99 on defining performance sentences and detecting sentence's tone respectively but in some cases much lower than that (0.71) when it comes to detecting attribution. As expected, this is shown to be a hard problem even for experts in accounting and finance. Further work will need to be carried out to improve inter-rater reliability and to continue to develop case-law for manual annotation of the attribution categories. This in turn will help us develop better training and test corpora for future experiments.

W. Aerts. 1994. On the use of accounting logic as an explanatory category in narrative accounting disclosures. *Accounting, Organizations and Society*, pages 337–353.

W. Aerts. 2005. Picking up the pieces: Impression management in the retrospective attributional framing of accounting outcomes. *Accounting, Organizations and Society*, pages 493–517.

Chien-Liang Chen, Chao-Lin Liu, Yuan-Chen Chang, and Hsiangping Tsai. 2013. Opinion mining for relating multiword subjective expressions and annual earnings in us financial statements. *Journal of Information Science and Engineering*, 29(3).

M. Clatworthy and M. J. Jones. 2003. Financial reporting of good news and bad news: Evidence form accounting narratives. *Accounting and Business Research*, pages 171–185.

A. Devitt and K. Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991.

J. Z. Ferreira, J. Rodrigues, M. Cristo, and D. F. de Oliveira. 2014. Multi-entity polarity analysis in financial documents. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 115–122, New York, NY, USA. ACM.

R. Garside and N. Smith. 1997. A hybrid grammatical tagger: CLAWS4. *Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121.

S. Goel and J. Gangolly. 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Int. J. Intell. Syst. Account. Financ. Manage.*, 19(2):75–89, April.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

R. Hooghiemstra. 2000. Corporate communication and impression management – new perspectives why companies engage in social reporting. *Journal of Business Ethics*, pages 55–68.

T. L. Im, P. W. San, C. K. On, R. Alfred, and P. Anthony. 2013. Analysing market sentiment in financial news using lexical approach. In *Open Systems (ICOS), 2013 IEEE Conference on*, pages 145–149, Dec.

S. Kogan, D. Levin, Bryan R. Routledge, J. S. Sagi, and N. A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 272–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

A. Moniz and F. de Jong. 2014. Classifying the influence of negative affect expressed by the financial media on investor behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 275–278, New York, NY, USA. ACM.

B. Neuenschwander, A. C.M. Pereira, W. Meira, and D. Barbosa. 2014. Sentiment analysis for streams of web data: A case study of brazilian financial markets. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 167–170, New York, NY, USA. ACM.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

P. Rayson, D. Archer, S. Piao, and A. McEnery. 2004. The ucrel semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, pages 7–12.

R. P. Schumaker. 2010. An analysis of verbs in financial news articles and their impact on stock price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 3–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huifeng P. Shuyu, Z. and A. Walter. 2016. Causal language intensity in performance commentary and financial analyst behavior. *Business Finance and Accounting*, 42(10).

M. Tsai and C. Wang. 2012. Visualization on financial terms via risk ranking from financial reports. In *Proceedings of COLING 2012: Demonstration Papers*, pages 447–452, Mumbai, India, December. The COLING 2012 Organizing Committee.

C. Whitelaw and J. Patrick. 2004. Selecting systemic features for text classification. In *Australasian Language Technology Workshop*, pages 93–100.

H. Zhang. 2004. The Optimality of Naive Bayes. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference*. AAAI Press.