

# Sentiment Analysis of Financial News Articles using Performance Indicators

Srikumar Krishnamoorthy

Received: 24 Sep 2016 / Revised: 30 May 2017 / Accepted: 29 Oct 2017

**Abstract** Mining financial text documents and understanding the sentiments of individual investors, institutions and markets is an important and challenging problem in the literature. Current approaches to mine sentiments from financial texts largely rely on domain specific dictionaries. However, dictionary based methods often fail to accurately predict the polarity of financial texts. This paper aims to improve the state-of-the-art and introduces a novel sentiment analysis approach that employs the concept of financial and non-financial performance indicators. It presents an association rule mining based hierarchical sentiment classifier model to predict the polarity of financial texts as positive, neutral or negative. The performance of the proposed model is evaluated on a benchmark financial dataset. The model is also compared against other state-of-the-art dictionary and machine learning based approaches and the results are found to be quite promising. The novel use of performance indicators for financial sentiment analysis offers interesting and useful insights.

**Keywords** Sentiment analysis · Financial news · Performance indicators · Text mining · Machine learning · Classification

## 1 Introduction

Sentiment analysis and opinion mining [31,37] has received significant attention in the literature due to its wide applicability in business, management and social science disciplines. It has been effectively applied in domains such as movies [31, 37], product reviews [5], travel reviews [7,37] and finance [2, 11, 23, 24, 34, 35].

---

This is an accepted version of the article published in Knowledge and Information Systems Journal. The final authenticated version is available online at: <https://doi.org/10.1007/s10115-017-1134-1>

Indian Institute of Management Ahmedabad, India Tel.: +91-79-66324834  
E-mail: srikumark@iima.ac.in

Financial sentiment analysis is considered to be an important and challenging problem in the literature [23, 24]. Current approaches to financial sentiment analysis utilize generic dictionary [34, 35], domain specific dictionary [10, 17, 19, 22, 25] or statistical/machine learning methods [2, 11, 15, 25, 30, 38] to determine polarity in financial texts. Some of the common dictionaries used in the financial sentiment analysis literature include Harvard GI (HGI) [33], MPQA [39], Sentiwordnet [8], SenticNet [6], SentiStrength2 [36], LM [22], and Financial Polarity Lexicon (FPL) [25]. LM & FPL are finance-specific dictionaries used in the recent literature. Other dictionaries (such as HGI [33], MPQA [39], SenticNet [6]) are generic in nature and are likely to mis-classify common financial words. For example, a detailed study of HGI [33] dictionary by [22] shows that more than 75% of the negative words used in HGI are non-negative in a financial context. Furthermore, recent research studies demonstrate superior results while using domain specific dictionary over general dictionary [11, 19, 23, 25].

In this paper, we aim to improve the state-of-the-art in domain specific dictionary based financial sentiment analysis. We motivate the need for the current study with the help of a set of illustrative examples. Let us consider the following financial text sentences extracted from financial phrase-bank dataset [25].

1. Halonen’s office acknowledged receiving the letter but declined comment.
2. Financial details were not disclosed.
3. The serial bond is part of the plan to refinance the short-term credit facility.
4. Aspo’s strong company brands ESL Shipping, Leipurin, Telko and Kauko-markkinat aim to be the market leaders in their sectors.
5. DnB Nord of Norway is the most likely Nordic buyer for Citadele, while Nordea would be a good strategic fit, according to published documents.
6. We are also pleased to welcome the new employees.

In the LM dictionary [22], declined, disclosed and refinance are negative words; similarly, strong, good and pleased are positive words. Therefore, each of the above sentences will be classified as positive or negative sentences by methods that use LM or other related domain dictionaries. However, it is quite evident that all of the above sentences are neutral statements from an investor/analyst perspective.

A few research studies have attempted to go beyond just polarity based dictionaries and utilized financial entities (custom words [25] or noun phrases [17] or named entities [32]) to improve the quality of polarity detection. But, the use of financial entities often generate a lot of false positives and false negatives. Let us consider the following sentences that contain at least one financial entity.

1. The company’s board of directors has proposed a *dividend* of EUR 0.12 per share.
2. Amanda *Capital* has *investments* in 22 private *equity funds* and in over 200 unquoted companies mainly in Europe.
3. A corresponding increase of 85,432.50 euros in Ahlstrom’s *share capital* has been entered in the Trade Register today.

All of the above sentences are neutral statements, though they contain financial terms/entities (*italized* words in text). Sentence 3 also indicates increase of a financial entity (share capital). Hence, the sentence is likely to be considered as positive by systems that use a combination of LM dictionary and financial entities (or noun phrases/named entities) [17, 25].

The following sentences contain financial or non-financial terms and directionality related words (e.g. raise, fell, lower, increase and so on).

1. *Turnover* rose to EUR 21mn from EUR 17mn.
2. The *EPS* outlook was increased by 5.6
3. Unit *costs* for flight operations fell by 6.4 percent.
4. The company intends to raise *production capacity* in 2006.
5. Rapala said it estimates it will make savings of 1-2 million euros a year by centralising its French *operations* at one site.
6. *Demand* for fireplace products was lower than expected, especially in Germany.

Sentences 4, 5 and 6 do not use financial entities. The systems that use only financial entities are likely to treat them as neutral sentences. An investor/analyst, however, is likely to consider the above sentences as polarized ones (positive or negative). A closer examination of these sentences reveals that they: (1) use non-financial indicators like demand, production capacity, and operations, and (2) refer to improvement or decline of the non-financial indicators.

From the foregoing discussions, it is evident that the use of domain specific dictionary (on a stand-alone basis or in conjunction with financial entities) is inadequate to accurately predict the polarity in financial texts.

It is quite intuitive to understand that an investor or a financial analyst is likely to analyze the performance of a company in terms of both financial and non-financial performance indicators. An investor/analyst often looks for performance indicators or measures while assessing polarity in financial texts and making investment decisions. Therefore, we posit that the use of financial and non-financial performance indicators can improve the quality of financial sentiment analysis.

Our main objective in this research work is to examine the use of performance indicators (lagging/leading financial and non-financial indicators) to predict polarity in financial texts. We conjecture that the use of performance indicators [13] is likely to be more meaningful and closely aligned with the way humans interpret financial texts while making investment decisions. We present a new approach to perform financial sentiment analysis based on performance indicators and conduct rigorous experiments to assess its usefulness. Such an approach, to the best of our knowledge, has not been explored in the literature.

This paper primarily makes two key contributions to the literature. First, the paper introduces the use of performance indicators to predict polarity in financial texts. Second, the paper presents a hierarchical sentiment classifier model based on the concept of association rule mining [1]. The primary benefit of using association rule mining for polarity prediction is that the generated rules can be used to easily explain predictions, unlike more complex discriminative models (like SVM, Neural networks) that are black-boxes. The performance of the proposed sentiment classifier model is assessed on a benchmark financial dataset. The model is also compared against other state-of-the-art methods to demonstrate its usefulness.

The rest of the paper is organized as follows: In the next section, the related work in the literature is described. Subsequently, in section 3, the paper outlines the proposed method in detail. The experimental evaluation of the method and a discussion of the key findings are presented in section 4. The paper concludes with a summary and directions for future research work.

## 2 Related Literature

Financial sentiment analysis approaches in the literature can be broadly categorized as (a) generic dictionary based methods, (b) domain specific dictionary based methods, and (c) statistical or machine learning based methods. Generic dictionaries such as Harvard GI [33] was used in some of the early works in financial sentiment analysis [34,35]. The use of generic dictionaries, however, lead to misclassification of common words in financial text and impacts the performance of sentiment prediction [22] or stock price movements [17,19]. Recent works in the literature [10,16,17,19,25] predominantly use domain specific dictionary such as LM dictionary [22], and FPL [25]. A detailed review of different dictionaries used in the literature can be found in [23]. Recent surveys on sentiment analysis in finance can be found in [14,24].

Statistical or machine learning based methods [2,11,15,25,30,38] use bag-of-words or n-grams as features and apply generative or discriminative classifier models for predicting sentiments. Internet message postings were used by [2] to classify financial text as buy, hold or sell. The classifier results were then aggregated for a pre-defined time period and bullishness & agreement indices were computed. The authors also conducted a study on relationship between the computed indices and financial measures such as stock returns and market volatility. The experimental results show that stock messages help predict market volatility.

A naive bayes classifier model is used in [11] to classify sentences as positive, negative or neutral. The predicted sentence level opinions are aggregated to derive report level opinions. The authors show that their method outperforms both generic and domain specific dictionary methods. A naive bayes classifier model based on bag-of-words features is trained in [15] to predict the tone (as positive, neutral, negative or uncertain) of forward looking statements in corporate filings. The author presents evidence that dictionary based methods (both generic and domain specific) are unsuitable for analyzing tone of corporate filings.

Naive bayes and support vector machine methods were used to predict sentiment of bloggers towards companies and their stocks [30]. The sentiment prediction was done using selected topic terms. Their work analyzes sentiment at the topic level and not at the sentence level as in [15,25]. Other recent works in the literature that perform sentiment analysis include [16,17,19,28,32]. These works utilize sentiment dictionaries and extract sentiments from text and predict stock price movements or market returns. The proposed paper is distinct from such works as the focus of the current paper is primarily on predicting the polarity of news articles as positive, neutral or negative. A comparative analysis of the financial sentiment analysis works in the literature is presented in Table 1.

The polarity sequence model proposed in [25] is an extension of [29]. The authors propose an LPS model and utilize both generic dictionary, MPQA [39] and domain specific LM dictionary [22]. The authors also enrich the finance lexicon by including (a) financial concept, (b) directional verbs such as increase, decrease, and (c) polarity for the interaction between financial concept and directional verb e.g. cost-increase is pre-labelled in the dictionary as negative; profit-increase is pre-labelled as positive. The authors also contribute to the literature by making an annotated financial sentiment corpus publicly available.

The LPS model proposed in [25] primarily works in three phases. In the first phase, entities are extracted along with their semantic orientations from the given

**Table 1** Comparison of financial sentiment analysis literature

Literature	Nature of text	Dictionary	Features	Method	Objective
Tetlock [34]	News	HGI	HGI words	Regression	Stock returns
Tetlock [35]	News	HGI	HGI words	Regression	Earnings, Returns
Loughran [22]	10-K	LM	LM words	Regression	Returns
Antweiler [2]	Internet messages	-	BoW	NB Regression	Volatility, Returns
Huang [11]	Analyst reports	HGI,LM LIWC	BoW	NB Regression	Polarity, Earnings
Li [15]	10-K 10-Q	-	BoW	NB	Polarity, Info. Content
O'Hare [30]	Blogs	-	Topic terms, Words near topics	NB,SVM	Polarity prediction
Marjan [38]	News	Dutch	Entities Polarity expression	Unsupervised	Polarity detection
Li [16]	News, Social media	-	Nouns, Sentiment words	Tensor regression	Stock trends
Li [19]	News	HGI,LM	Dictionary words	SVM	Stock returns
Li [17]	News, Internet messages	HGI, LM	Entities, Polarity words	SVR	Stock trends
Mo [28]	News	SWN	SWN words	Regression	Stock returns
RPS [32]	News	-	Nouns	SVR	Stock trends
Malo [25]	News	FE,MPQA HGI,LM DI	Entities, Phrase structure	SVM	Polarity prediction
Our work	News	PI, LM DI	PI tags	ARM	Polarity prediction

BoW - Bag of words; FE - Financial entity; DI - Directionality; SWN - Sentiwordnet  
PI - Performance indicator words; LIWC - Linguistic Inquiry and Word Count

text sentences. Several heuristics are applied to extract different entities (financial and general) and semantic orientations from financial text sentences. Subsequently, in the second phase, the model applies a phrase structure projection to project the extracted entities into a sequence in  $l^2$  space. Finally, in the third phase, a multi-label classifier is applied on the generated  $l^2$  sequence. For multi-label

classification, support vector machines (SVM) with one-against-one strategy was used.

The proposed work is distinct from existing works in the literature on the following aspects: (1) It presents a financial text polarity prediction method based on financial and non-financial performance indicators. This is different from the works of [25] that primarily use financial concepts and are likely to generate a lot of false positives and false negatives. Our experimental results corroborates this claim and show the usefulness of the proposed method. (2) Polarity for the interaction between financial concept and directional verb is not pre-defined as in [25]. (3) Presents an association rule [1] based hierarchical classifier to predict sentiment. To the best of our knowledge, this is the first work that uses association rule mining based method for financial sentiment analysis. A few recent works in non-financial domain have explored the use of a non-hierarchical associative classifier for sentiment analysis of webreviews [26] and product reviews [40].

We demonstrate the usefulness of the proposed approach over other state-of-the-art methods through rigorous experimental evaluation.

### 3 Hierarchical Sentiment Classifier (HSC)

This section describes the proposed method for financial sentiment analysis. The method broadly consists of the following four key aspects.

1. **Domain specific lexicon** The proposed method uses a standard domain specific dictionary, LM dictionary [22]. In addition, the lexicon defines words related to performance indicators and directionality.
2. **Text tagging using lexicon** The given financial text is parsed and tagged by looking up words in the lexicon.
3. **Polarity classifier model** The tagged financial text is used to build a hierarchical classifier model. The classifier model utilizes the concept of association rule mining.
4. **Predict sentiment** The polarity classifier model (association rules) is used to make polarity predictions for new financial text sentences.

Each of the above aspects is described in detail in the following pages.

#### 3.1 Domain specific lexicon

Three categories of words are defined in the domain specific lexicon used in this paper. The overall distribution of words in the lexicon is given in Table 2.

The first category of words is related to performance indicators. The performance indicator words are categorized as lagging and leading indicators. Lagging indicators reflect the results of firms activity e.g. improvement or decline in sales, market share, operating profit, operating cost, orders, inventory turns etc.

The leading indicators signal future events and are precursors to the firms future performance. Some of the common examples of leading indicators are: #new stores, #employee recruitments, #employee reductions/layoffs, #new customers, % increase in productivity, % increase in production capacity, #new contracts won or awarded, and % increase in plant utilization.

**Table 2** Distribution of words in the dictionary

Category	Type of word (tags)	No. of words (% of all entries)
Performance Indicators	Lagging Indicator words (LagInd)	67 (2.29%)
	Leading indicator words (LeadInd)	70 (2.39%)
Directionality	Down (DOWN)	53 (1.81%)
	Up (UP)	51 (1.74%)
Finance-specific sentiment words	Negative (NEG)	2337 (79.73%)
	Positive (POS)	353 (12.04%)

The specific words used in the dictionary for leading indicators include only the key terms like store, employee, customer, productivity, efficiency. The words that signify the direction of movement of these terms (i.e. increase, decrease, reduction, layoff etc.) are maintained separately as part of the directionality category that is described next.

The second category of words defined in the lexicon is related to the directionality of leading/lagging indicators. Directionality is the word (or n-gram) describing direction of events [25]. Examples include increase, improve, skyrocket, decrease, and plummet. For building the directionality related words, the words defined under specific categories (namely, rise, fall, increase, and decrease) in the Harvard GI lexicon [33] were used as seed words. The words were manually reviewed to remove words that have different meaning in financial context. A few additional words that signify directionality of lagging/leading indicator terms were included e.g. layoff (#employees laid off), terminated (#staff terminated), and awarded (#new contracts awarded). The final number of words under the directionality category was about 100.

The third, and final, category of words used in the lexicon is the domain-specific lexicon, LM dictionary [22]. The polarity bearing words defined in the LM dictionary were also used in this study.

### 3.2 Text tagging using lexicon

The input to the proposed method is a set of financial text sentences. Each financial text sentence is parsed using NLTK Parts-Of-Speech (POS) tagger [4]. The sentences are then tagged or labeled based on the occurrence of different categories of words (i.e. performance indicators, directionality, interaction of performance indicators & directionality, and sentiment words). The occurrence patterns of different categories of words are analyzed using several heuristic rules. Fig. 1 provides a couple of illustrative examples of parsed and tagged text sentences. Fig. 1 gives the raw text sentence, its corresponding POS tags and the lexicon tagged words. The lexicons used for tagging are the ones described in section 3.1. The complete grammar and the heuristic rules used for parsing and labeling financial text sentences are given in Appendix.

The output of this step is the conversion of text sentences into tagged sequence of words. The specific tags used are: LagInd, LeadInd, UP, DOWN, POS, NEG, LagInd::UP, LagInd::DOWN, LeadInd::UP, and LeadInd::DOWN. The last four

tags are the interactions between performance indicators and directionality. All other words in the original sentence are discarded. The above process is repeated for each and every sentence in the review collection. The resulting collection of tagged words is then used for polarity classification.

<p>Net sales rose by 30 %</p> <p>JJ NNS VBD IN CD NN</p> <p>LagInd:UP</p>	<p>Following the increase , Huhtamaki's number of shares outstanding is 105 mn</p> <p>VBG DT NN , NNP NN IN NNS JJ VBZ CD NN</p> <p>UP POS</p>
<p>Operating profit for the year 2007 declined by 15 %</p> <p>NN NN IN DT NN CD VBD IN CD NN</p> <p>LagInd:DOWN</p>	<p>Aspokem posted an operating profit of 2.7 mln</p> <p>NNP VBD DT NN NN IN CD NNS</p> <p>LagInd</p>
<p>Finnish Outokumpu Technology has been awarded several new grinding technology contracts</p> <p>JJ NNP NNP VBZ VBN VBN JJ JJ VBG NN NNS</p> <p>LeadInd:UP</p>	
<p>Sports equipment sales was good due to the prolonged winter season</p> <p>NNS NN NNS VBD JJ JJ TO DT JJ NN NN</p> <p>LagInd POS NEG</p>	

**Fig. 1** Illustration of text sentences tagged with POS tags and lexicon related tags

### 3.3 Polarity classifier model

The proposed method aims to classify the financial text as positive, neutral or negative. The classifier proposed in this paper uses the concept of association rule mining [1]. Some of the common associative classifiers used in the literature include: CBA [20], CMAR [18], LB [27] and ART [3]. The proposed approach utilizes a variation of the CMAR method for sentiment classification. The key steps of the associative classifier proposed in this paper are as follows:



### 3.3.1 Prepare transaction data

The input data required for rule mining is the set of transactions with items. The tags identified in the earlier step are used as items for rule mining. In addition, the class label of the sentence is appended to the set of items. For example, the transaction data for the illustrative example in Figure 1 is shown in Table 3.

**Table 3** A sample tagged transaction database

Transaction #	Type of word (tags)
1	LagInd::UP, positive
2	UP, POS, neutral
3	LagInd::DOWN, negative
4	LagInd, neutral
5	LeadInd::UP, positive
6	LagInd, POS, NEG, neutral

### 3.3.2 Build a classifier model

The model building involves frequent itemset mining, association rule generation, and rule ordering.

Frequent itemset and rule mining. The frequent itemsets are mined using Apriori algorithm [1]. Association rules are mined from the frequent itemsets with the constraint that the consequent of the rule contains one item and is a class (positive, negative, or neutral class). Antecedent of the rule can contain any number of items. That is, rules of the form:  $X \rightarrow c$ , are mined from the generated frequent itemsets. A sample set of association rules mined from the transactions in Table 3 is given in Figure 2.

LagInd  $\rightarrow$  neutral (33.33%, 100%)  
 POS  $\rightarrow$  neutral (33.33%, 100%)  
 UP, POS  $\rightarrow$  neutral (16.67%, 100%)  
 LagInd::DOWN  $\rightarrow$  negative (16.67%, 100%)  
 LagInd::UP  $\rightarrow$  positive (16.67%, 100%)  
 LeadInd::UP  $\rightarrow$  positive (16.67%, 100%)  
 UP  $\rightarrow$  neutral (16.67%, 100%)

**Fig. 2** Sample Association Rules

Rule ordering. The generated association rules are ordered according to decreasing precedence based on their confidence, support and antecedent length. This heuristic is used in the literature to give precedence to more confident rules [18,20] while making predictions.

### 3.4 Predict sentiment

The rules mined in the previous step are used as a model for predicting the polarity of new sentences. The complete pseudo code for polarity prediction is given in Algorithm 1.

Given a new financial text sentence, the polarity prediction works as follows: First, the new text sentence is tagged with performance indicators, directionality and finance-specific sentiment words. The tagged words are then looked up in the rule base (i.e. the set of rules generated by the polarity classifier model). It is to be noted that each rule in the rule base is of the form: Premise (one or more tag words)  $\rightarrow$  Conclusion (the class label). The tagged words are matched against each and every rule in the rule base (steps 5-19 of Algorithm 1). If a rule match is found, a matching score is assigned based on the confidence of the rule (the higher the confidence, the more important the rule). The scores are grouped based on class labels, as multiple rules with different class labels are likely to match the tagged words. The process of grouping scores based on class labels is similar to the approach followed by associative classifiers in the literature [18]. The above process is repeated for each and every rule in the rule base and the scores are accumulated. At the end of the iteration, the algorithm produces a score for each class label. The class with the highest score is treated as the final predicted class (step 20 of Algorithm 1).

Let us illustrate the above prediction procedure with the following example:

*Olvi expects market share to increase in the first quarter of 2010*

The above sentence after parsing and dictionary lookup will be tagged as LagInd::UP. The class label is then determined by looking up the tag in the set of mined rules. The only matching rule available in Figure 2 for the tagged sentence is LagInd::UP  $\rightarrow$  positive (16.67%, 100%). The scores are, therefore, generated only for the 'positive' class label. The most appropriate class label for the sentence is predicted as "positive". If there are rule matches across class labels, then scores for each class label are generated (refer to steps 5 to 19 of Algorithm 1). The final class label is predicted as the one with highest score. If no rule matches are found during the search process, then the sentence is assigned a default prediction value of "neutral".

The polarity classifier model and prediction method described above is generic and can be used for multi-class problems (positive, neutral, negative sentiments). However, in the financial sentiment analysis context, the total number of training examples in each class is likely to be highly unbalanced. That is, one is likely to encounter more number of neutral sentences than positive/negative sentences in financial texts. This class imbalance can also be observed in the annotated corpus used by researchers in the past literature. For example, the number of neutral examples was well over 53%; and negative examples was a small fraction of about 12-14% in the earlier works [11, 25].

In order to effectively address the class imbalance problem, we propose a hierarchical classifier, named Hierarchical Sentiment Classifier (HSC). The proposed method (HSC) first classifies a given sentence as polarized or neutral. Subsequently, the polarized sentences are classified as either positive or negative. The underlying classifier model and prediction methods used in HSC were described earlier (sec-

**Algorithm 1** Pseudocode for polarity prediction

---

**Input:** *inputText*, financial text sentence  
*ruleBase*, mined association rules

**Output:** *result*, the predicted class

```

1: result={} ▷ a dictionary that holds confidence values for each class
2: tags ← Parse and tag the text sentence #section 3.2
3: ▷ iterate through each rule and match the tags
4: ▷ each rule, premise → class, contains support and confidence values
5: for r in ruleBase do
6:   ▷ check if complete set of tags match
7:   if r.premise==tags then
8:     ▷ accumulate the score for each class in result
9:     result[r.class] = result[r.class] + r.confidence
10:  else
11:    ▷ match each tag individually
12:    for t in tags do
13:      if r.premise==t then
14:        ▷ one of the tags match
15:        result[r.class] = result[r.class] + r.confidence
16:      end if
17:    end for
18:  end if
19: end for
20: return class with the highest average confidence

```

---

tion 3). A comparative evaluation of the hierarchical versus the multi-class and the standard one-against-one classifier is presented in the experimental results section.

#### 4 Experimental Evaluation

The publicly available financial phrase bank dataset [25] is used for the experimental analysis. This dataset is a reasonably large annotated corpus of about 5000 sentences. The sentences are marked as positive, neutral or negative by the annotators. Four different datasets were prepared by the authors based on the level of agreement amongst the annotators. These datasets are labelled as DS100 (100% agreement amongst annotators), DS75 (>75% agreement), DS66 (>66% agreement) and DS50 (>50% agreement). The total number of examples and distribution of class labels on each of the datasets is given in Table 4.

**Table 4** Dataset characteristics

Dataset	Number of sentences (%positive, %neutral, %negative)
DS100	2259 (13.4, 61.4, 25.2)
DS75	3448 (12.2, 62.1, 25.7)
DS66	4211 (12.2, 60.1, 27.7)
DS50	4840 (12.5, 59.4, 28.2)

The proposed classifier model is evaluated on standard classifier evaluation measures [9] such as Precision (P), Recall (R), F-measure (F) and Accuracy (A).

To ensure robustness of the results, all the experiments were conducted using a stratified 10-fold cross-validation.

The proposed method (HSC) is compared against the most commonly used domain specific dictionary (LM [22]) based method (W-Loughran), and the most recent LPS model [25]. The dictionary based method is too simple to capture all intricacies of financial text. The LPS model is more complex requiring transformation of text using domain dictionary and inferencing of complex sentence structures using SVM based approach. HSC, on the other hand, is a relatively simple approach that transforms the text into a set of tags and applies rule mining on the identified tags to predict sentiment. One of the major advantages of the proposed method is that the rationale for prediction can be clearly explained as it is a rule based approach.

The rule mining requires two configurable parameters namely, minimum support (minsup) and minimum confidence (minconf). We use a default minsup and minconf value of 0.5% and 60% respectively. A sensitivity analysis of the parameter value changes on sentiment prediction are presented in section 4.4.

The results of our initial experiments are presented in Table 5. The best F-measure and accuracy values are marked in bold. The results reveal that the proposed method is significantly better compared to other state-of-the-art methods in all of the datasets studied. The accuracy values for the neutral classes alone are the same for both LPS and HSC models. It is to be noted that there are large number of neutral examples in the dataset (about 60%, refer to Table 4) and both the models are able to learn well from these examples and achieve a very high accuracy value of over 93%.

**Table 5** Comparative evaluation of HSC against other methods

Dataset	Measure	W-Loughran			LPS			HSC		
		pos	neut	neg	pos	neg	neut	pos	neg	neut
DS100	P	0.56	0.64	0.36	0.74	0.85	0.84	0.83	0.93	0.86
	R	0.13	0.91	0.16	0.74	0.87	0.79	0.82	0.93	0.81
	F	0.20	0.75	0.22	0.74	0.86	0.81	<b>0.83</b>	<b>0.93</b>	<b>0.83</b>
	A	0.76	0.63	0.85	0.87	0.83	<u><b>0.95</b></u>	<b>0.91</b>	<b>0.92</b>	<u><b>0.95</b></u>
DS75	P	0.60	0.65	0.38	0.69	0.83	0.76	0.77	0.90	0.83
	R	0.17	0.90	0.20	0.66	0.84	0.80	0.75	0.90	0.78
	F	0.26	0.76	0.26	0.67	0.83	0.78	<b>0.76</b>	<b>0.90</b>	<b>0.80</b>
	A	0.76	0.64	0.86	0.84	0.79	<u><b>0.95</b></u>	<b>0.88</b>	<b>0.88</b>	<u><b>0.95</b></u>
DS66	P	0.61	0.63	0.40	0.60	0.86	0.73	0.75	0.86	0.80
	R	0.18	0.89	0.21	0.82	0.71	0.77	0.70	0.88	0.75
	F	0.28	0.74	0.28	0.69	0.77	0.75	<b>0.72</b>	<b>0.87</b>	<b>0.77</b>
	A	0.74	0.62	0.87	0.80	0.75	<u><b>0.94</b></u>	<b>0.85</b>	<b>0.84</b>	<u><b>0.94</b></u>
DS50	P	0.57	0.62	0.41	0.65	0.76	0.72	0.72	0.84	0.79
	R	0.19	0.88	0.22	0.54	0.81	0.77	0.66	0.85	0.71
	F	0.28	0.73	0.29	0.59	0.78	0.74	<b>0.69</b>	<b>0.85</b>	<b>0.75</b>
	A	0.73	0.61	0.86	0.79	0.74	<u><b>0.93</b></u>	<b>0.82</b>	<b>0.81</b>	<u><b>0.93</b></u>

The best F & A cases are marked in **bold** and the tie cases are underlined.

#### 4.1 Study of the effect of performance indicators

In the next set of experiments, we study the influence of performance indicators and financial sentiment words on sentiment prediction. Three different scenarios were considered for the analysis: (1) using lagging indicators (along with directionality), (2) using both lagging and leading indicators (along with directionality), and (3) using all of the tags including financial sentiment words (baseline case). The analysis results are presented in Table 6. The best F-measure and accuracy values are marked in bold. In addition, the values that are tied are underlined.

The results present a few interesting insights, when it is analysed in relation to the characteristics of the annotated datasets. First, the results for DS100 dataset using only the lagging indicators is almost close to that of the baseline case. The performance difference between the two cases widens as we navigate down the table (DS75, DS66 and DS50). Second, the performance results improve as the leading indicators are included in the model. Further performance improvements are observed when all of the tags are used. One can observe a correlation between inter-annotator agreement and the use of indicators. A model that uses only the lagging indicators works quite well when the inter-annotator agreement is 100%. The second model that use both lagging and leading indicators perform better, even when the inter-annotator agreement declines to 75%. Finally, the third model works best, even when the inter-annotator agreement declines to a very low value of 50%. These results imply that human annotators, experienced in finance domain, have less agreement when a financial text uses leading indicators to describe company’s performance. This is in line with our expectation, as the relationship amongst the leading indicators, investor sentiment and the firms future performance are often unclear. One can also find evidence from the finance and accounting literature on the lack of clear relationship between leading indicators and firm performance [12]. Let us consider the following two sentences:

1. VDW combined with LXE devices enhances productivity, enabling workers to use a single device to perform voice, scanning and keyboard functions.
2. Last year, UPM cut production, closed mills in Finland and slashed 700 jobs.

The first sentence refers to worker productivity improvement (a leading indicator) and the second sentence refers to cut in production and jobs (leading indicators). Both of these sentences do not have 100% agreement and are annotated by 75% of the annotators as polarized i.e. the first and second sentences are marked respectively as positive and negative.

The financial texts also contain general positive opinions. For example, let us consider the following sentence:

”He believes that the soy-oats have a good chance of entering the UK market”

The above sentence contains neither lagging nor leading indicators but expresses a positive outlook. A high level of disagreement amongst annotators was observed and only 50% of the annotators have marked the above sentence as positive. Another sentence that has been rated as positive by 50% of the annotators was: ”The companys strength is its Apetit brand”.

From the foregoing discussions and the experimental results, it is clear that financial texts are interpreted differently based on the nature of the indicators and

**Table 6** Influence of performance indicators on sentiment prediction

Dataset	Measure	HSC: LagInd Only			HSC: LagInd, LeadInd			HSC: All		
		pos	neut	neg	pos	neg	neut	pos	neg	neut
DS100	P	0.89	0.91	0.87	0.89	0.92	0.87	0.83	0.93	0.86
	R	0.77	0.96	0.79	0.80	0.96	0.78	0.82	0.93	0.81
	F	0.82	0.93	0.82	<b>0.84</b>	<b>0.94</b>	0.82	0.83	0.93	<b>0.83</b>
DS75	A	0.91	0.91	<b>0.95</b>	<b>0.92</b>	<b>0.92</b>	<b>0.95</b>	0.91	<b>0.92</b>	<b>0.95</b>
	P	0.82	0.85	0.86	0.82	0.88	0.84	0.77	0.90	0.83
	R	0.65	0.94	0.68	0.70	0.94	0.76	0.75	0.90	0.78
DS66	F	0.73	0.89	0.75	<b>0.76</b>	<b>0.91</b>	<b>0.80</b>	<b>0.76</b>	0.90	<b>0.80</b>
	A	0.87	0.86	0.94	<b>0.89</b>	<b>0.88</b>	<b>0.95</b>	0.88	<b>0.88</b>	<b>0.95</b>
	P	0.79	0.80	0.81	0.79	0.84	0.81	0.75	0.86	0.80
DS50	R	0.58	0.93	0.57	0.63	0.91	0.72	0.70	0.88	0.75
	F	0.67	0.86	0.67	0.70	<b>0.87</b>	0.76	<b>0.72</b>	<b>0.87</b>	<b>0.77</b>
	A	0.84	0.81	0.93	<b>0.85</b>	<b>0.84</b>	<b>0.94</b>	<b>0.85</b>	<b>0.84</b>	<b>0.94</b>
DS50	P	0.75	0.77	0.82	0.75	0.81	0.80	0.72	0.84	0.79
	R	0.54	0.91	0.51	0.59	0.90	0.69	0.66	0.85	0.71
	F	0.62	0.83	0.63	0.66	<b>0.85</b>	0.74	<b>0.69</b>	<b>0.85</b>	<b>0.75</b>
	A	0.82	0.78	0.92	<b>0.83</b>	0.80	<b>0.93</b>	0.82	<b>0.81</b>	<b>0.93</b>

sentiment words. It is to be noted that sentiment analysis in other domains (such as movies, music) utilizes general sentiment lexicon words to predict sentiments. In the financial sentiment analysis literature, similar approach has been borrowed with a refinement of dictionary words to finance domain. This paper presents a new perspective and suggests the use of multiple levels of analysis (lagging indicator, leading indicator, domain specific lexicon words) to improve the quality of financial sentiment analysis. The experimental results clearly demonstrate the utility of such a multi-level sentiment analysis.

The multi-level approach to financial sentiment analysis can be very useful in building models to suit specific application requirements. For example, an investor might prefer a model built using lagging (or lagging and leading) indicators over a more accurate baseline model. The former model is more likely to help the investor make the best investment decisions, even though it offers relatively lower accuracy.

#### 4.2 Analysis of alternate classifier models

In the financial sentiment analysis context, the number of examples in each of the classes is likely to be highly unbalanced. Therefore, a hierarchical sentiment classifier was used in this paper. Table 7 presents an analysis of hierarchical versus multi-class and one-against-one versions of the associative classifier. A comparison of multi-class versus hierarchical classifier reveals that the multi-class version performs marginally better for DS100 dataset. However, the performance degrades considerably for other datasets. For example, the F-measure values for DS50 dataset is 64% and 69% respectively for multi-class and hierarchical versions.

A comparison of one-against-one versus hierarchical classifier suggests that the hierarchical classifier performs better. However, the margin of difference is not very significant, except for a few cases (e.g. D75 dataset, negative class values of 74% versus 80%). Overall, the comparative evaluation of different variants of the classifier models did not indicate statistically significant differences in the ob-

served results. Future work could consider detailed evaluation of alternate classifier models on more diverse datasets to better understand the trade-offs involved.

**Table 7** Performance analysis of One-against-One, Multi-class and HSC classifiers

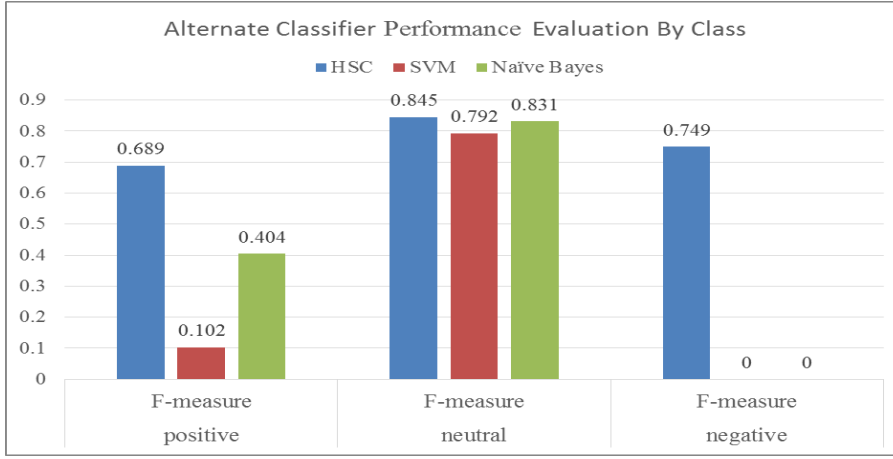
Dataset	Measure	One-against-one			Multi-class			HSC		
		pos	neut	neg	pos	neg	neut	pos	neg	neut
DS100	P	0.84	0.92	0.87	0.85	0.94	0.84	0.83	0.93	0.86
	R	0.82	0.94	0.76	0.81	0.92	0.89	0.82	0.93	0.81
	F	<b>0.83</b>	<b>0.93</b>	0.81	<b>0.83</b>	<b>0.93</b>	<b>0.86</b>	<b>0.83</b>	<b>0.93</b>	0.83
	A	<b>0.91</b>	0.91	0.94	<b>0.91</b>	<b>0.92</b>	<b>0.95</b>	<b>0.91</b>	<b>0.92</b>	<b>0.95</b>
DS75	P	0.78	0.88	0.85	0.78	0.89	0.83	0.77	0.90	0.83
	R	0.75	0.92	0.66	0.71	0.90	0.81	0.75	0.90	0.78
	F	<b>0.76</b>	0.89	0.74	0.74	0.89	<b>0.82</b>	<b>0.76</b>	<b>0.90</b>	0.80
	A	<b>0.88</b>	0.86	0.94	0.87	0.87	<b>0.95</b>	0.87	<b>0.88</b>	<b>0.95</b>
DS66	P	0.76	0.85	0.81	0.75	0.84	0.80	0.75	0.86	0.80
	R	0.69	0.89	0.70	0.64	0.88	0.76	0.70	0.88	0.75
	F	<b>0.73</b>	<b>0.87</b>	0.75	0.69	0.86	<b>0.78</b>	0.72	<b>0.87</b>	0.77
	A	<b>0.85</b>	0.83	<b>0.94</b>	0.84	0.82	<b>0.94</b>	<b>0.85</b>	<b>0.84</b>	<b>0.94</b>
DS50	P	0.73	0.82	0.80	0.78	0.80	0.79	0.72	0.84	0.79
	R	0.65	0.87	0.67	0.54	0.89	0.73	0.66	0.85	0.71
	F	<b>0.69</b>	0.84	0.73	0.64	0.84	<b>0.76</b>	<b>0.69</b>	<b>0.85</b>	0.75
	A	<b>0.83</b>	0.80	0.93	0.81	0.79	<b>0.94</b>	0.82	<b>0.81</b>	0.93

#### 4.3 Comparative evaluation of HSC and machine learning models

In the recent literature on financial sentiment analysis, machine learning approaches like Naive Bayes (NB) and Support Vector Machines (SVM) have been explored. These approaches utilize bag-of-words or n-grams or sentiment words as features for building the model. Such models generate large number of features and pose scalability issues. Besides, the models based on bag-of-words (or n-grams) are shown to be less predictive in the past research studies [19,32].

We analyze the use of machine learning models with parsimonious set of features (tags) introduced in this paper. A term document matrix was built using the tags as features. The generated term document matrix was then weighted with the help of standard Term frequency-Inverse document frequency (TF-IDF) measure. The standard python library (sklearn) was utilized for building NB and SVM models. For SVM, the default parameter values of C,  $\gamma$  and RBF kernel were used for the experimentation.

The results of our experiments are shown in Fig. 3. The results reveal that HSC perform better compared to other models. SVM and NB models produced a zero F-measure values for the negative class. This can be attributed to the models inability to learn from the parsimonious set of features. The naive bayes method was found to perform better than SVM for the positive and neutral classes. It is evident from the results that the standard machine learning models produced poor results compared to HSC. These results clearly demonstrate that an associative classifier (HSC) is quite effective in predicting sentiments even with parsimonious set of features.



**Fig. 3** Classifier Performance for DS50 dataset

#### 4.4 Effect of changes in rule mining parameters

The association rule mining requires two parameters to be specified, namely, minimum support and minimum confidence. The minimum confidence parameter influences the quality of rules generated and plays a critical role in sentiment prediction. We study the influence of the choice of minimum confidence values on the precision and recall values of the classifier. The results of our investigation are presented in Fig. 4.

The results show the trade-off involved between precision and recall. An increase in confidence value is associated with increase in precision for positive and negative classes. But, the precision increase at higher confidence values comes at the cost of lower recall. The observed relationship is quite intuitive and is in line with our expectations. One can choose an appropriate value of confidence based on the level of precision and recall desired. Alternately, one can choose an optimal threshold value that maximizes F-measure. It is to be noted that F-measure is the harmonic mean of precision and recall.

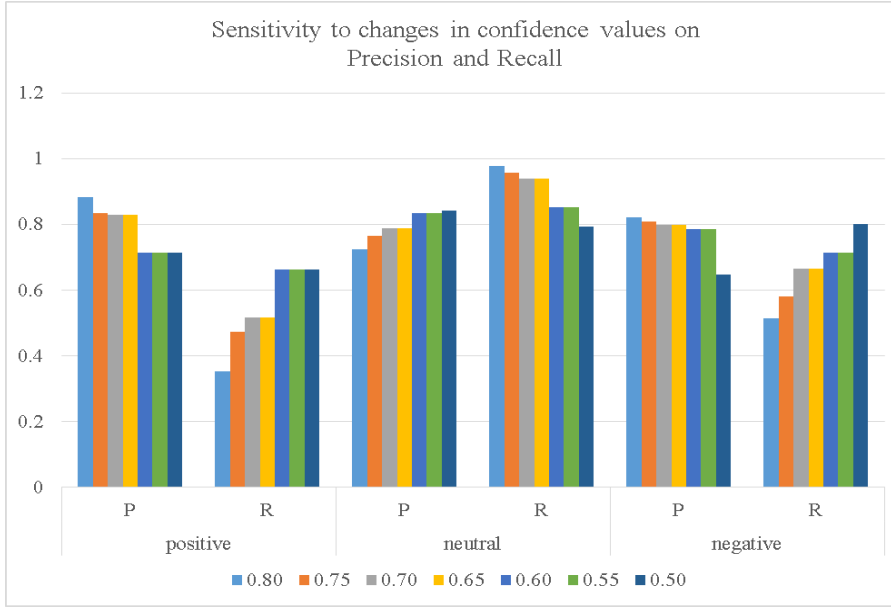
#### 4.5 Study of the use of performance indicator - directionality reversal

In the next set of experiments, we analyze the influence of directionality reversals. Let us consider the following two sentences to understand the need for directionality reversals.

1. *Turnover* rose to EUR 21mn from EUR 17 mn.
2. Unit *costs* for flight operations fell by 6.4 percent.

Both of the above sentences belong to the positive class. The above two sentences will be tagged respectively as LagInd::UP and LagInd::DOWN. It is reasonable to assume that one is likely to encounter more number of LagInd::UP cases in positive class and more number of LagInd::DOWN cases in negative class. Hence,





**Fig. 4** Performance analysis of confidence parameter changes for DS50 dataset

an associative classifier (HSC) is most likely to predict the class for the second example as negative.

In the literature, Malo et al [25] use a lexicon to pre-define directional dependence for each financial term. The baseline associative classifier presented in this paper does not use such directional dependencies. In order to understand the influence of directional dependencies on classifier performance, a directionality reversal is applied as a post-processing operation. That is, after the initial tagging, a lookup is made for specific indicators (like operating cost, operating loss, expenses) and the directionality is reversed. For the illustrative example in sentence 2 above, the tag is reversed from LagInd::DOWN to LagInd::UP as lower cost is considered as a positive outcome.

The results of the directionality reversal experiments are presented in Table 8. The results show that there is a general improvement in classifier performance for both positive and negative classes. This is in line with our expectation as the positive and negative example cases are correctly tagged with directionality reversals. The observed improvement is marginally better compared to the baseline case but the results are statistically insignificant. These results demonstrate that the proposed model (HSC) offers better predictive results even when the directionality reversal words are not used.

An alternate way to handle the directional dependence is to tag the identified indicators along with the business terms. For example, while tagging sentence 'Operational cost rose by 30%', the term 'operational cost' can be tagged along with 'LagInd::UP'. The association rule mining can then be applied on the transactions that contain both performance indicators and the associated terms. The main challenge with this approach is that the support of individual terms (like cost, expenses) will be very low and is unlikely to satisfy the minimum support

**Table 8** Experimental analysis of PI and Directionality reversal

Dataset	Measure	HSC			PI-Directionality reversal		
		pos	neut	neg	pos	neg	neut
DS100	P	0.83	0.93	0.86	0.83	0.93	0.89
	R	0.82	0.93	0.81	0.85	0.93	0.80
	F	0.83	<b>0.93</b>	0.83	<b>0.84</b>	<b>0.93</b>	<b>0.84</b>
	A	0.91	<b>0.92</b>	0.95	<b>0.92</b>	<b>0.92</b>	<b>0.96</b>
DS75	P	0.77	0.90	0.83	0.77	0.90	0.85
	R	0.75	0.90	0.78	0.77	0.90	0.77
	F	0.76	<b>0.90</b>	0.80	<b>0.77</b>	<b>0.90</b>	<b>0.81</b>
	A	<b>0.88</b>	<b>0.88</b>	<b>0.95</b>	<b>0.88</b>	<b>0.88</b>	<b>0.95</b>
DS66	P	0.75	0.86	0.80	0.75	0.86	0.82
	R	0.70	0.88	0.75	0.71	0.88	0.74
	F	0.72	<b>0.87</b>	0.77	<b>0.73</b>	<b>0.87</b>	<b>0.78</b>
	A	<b>0.85</b>	<b>0.84</b>	0.94	<b>0.85</b>	<b>0.84</b>	<b>0.95</b>
DS50	P	0.72	0.84	0.79	0.71	0.84	0.81
	R	0.66	0.85	0.71	0.68	0.85	0.71
	F	0.69	<b>0.85</b>	<b>0.75</b>	<b>0.70</b>	<b>0.85</b>	<b>0.75</b>
	A	0.82	<b>0.81</b>	0.93	<b>0.83</b>	<b>0.81</b>	<b>0.94</b>

constraints or produce models with very poor accuracy (if very low minimum support value is used). In such cases, the use of more complex variable support rule mining approaches can be explored.

#### 4.6 Discussion

This paper proposed a hierarchical sentiment classifier using performance indicators for financial sentiment prediction. The proposed approach was found to be quite effective achieving an accuracy of over 80% and F-measure of over 68% in all of the datasets studied.

The findings of this paper have implications for both theory and practice. From a theoretical perspective, the paper presents a novel approach to financial sentiment analysis using performance indicators. The presented approach is more closely aligned with the way humans interpret financial texts while making investment decisions. The associative sentiment classifier generates easily explainable predictions compared to other methods available in the literature. Our rigorous experimental evaluation clearly demonstrate the utility of the approach over the state-of-the-art methods. Furthermore, the results observed in Table 6 clearly reveals the reasons for differences amongst annotators with the help of performance indicators (lagging and leading indicators).

The paper presents new perspectives on the influence of performance indicators and financial sentiment lexicon words on sentiment prediction. The experimental results reveal that it is useful to consider financial sentiment analysis at multiple levels (using lagging indicator, combination of lagging and leading indicators, and all indicators along with sentiment words). An investor, analyst or financial institution can benefit by choosing an appropriate financial sentiment model to suit their specific application requirements e.g. track overall market outlook, make investment decision and so on.

The transparency and replicability of results is often a challenge in the financial sentiment analysis literature [24] due to the proprietary nature of datasets, dictionaries and advanced sentence parsing rules. To ensure transparency and replicability of the results presented, the detailed grammar and sentence parsing steps used for tagging are provided in Appendix.

## 5 Conclusion

This paper examined the use of performance indicators for predicting sentiments from financial texts. It presented a hierarchical classifier that uses the concept of association rules. The effectiveness of the classifier was demonstrated through rigorous experimental analysis on a benchmark financial dataset.

A study of the influence of performance indicators on sentiment prediction revealed interesting insights. The presence of varying levels of influence of lagging indicators, leading indicators and sentiment words on sentiment prediction were observed for different datasets. The results are clearly in alignment with the way humans interpret financial texts and make decisions.

As part of our future work, we plan to explore several interesting extensions. First, a more fine-grained analysis of performance indicators is likely to reveal interesting insights. For instance, a balanced scorecard approach [13] analyses company's performance from multiple dimensions, namely, financial, customers, internal business processes and learning & growth. Performance indicators can be categorized across each of these dimensions and the influence of each of the dimensions on financial sentiment can be explored. Second, advanced variable support rule mining methods can be studied to capture the directional dependencies in financial texts without using a lexicon. Third, the proposed method is not tuned to handle sentences that have mix of positive and negative orientations on performance indicators. One can investigate the use of utility mining [21] approaches to give different weights to different indicators. Such an approach allows one to capture varying influence of indicators and further improve sentiment prediction.

The findings of this study are likely to be useful for financial institutions in building superior sentiment analysis models. The investors or financial analysts can make better investment decisions using qualitative financial texts.

## References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: *Acm sigmod record*, pp. 207–216. ACM (1993)
2. Antweiler, W., Frank, M.Z.: Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance* **59**(3), 1259–1294 (2004)
3. Berzal, F., Cubero, J.C., Sánchez, D., Serrano, J.M.: Art: A hybrid classification model. *Machine Learning* **54**(1), 67–92 (2004)
4. Bird, S.: Nltk: the natural language toolkit. In: *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72. Association for Computational Linguistics (2006)
5. Blitzer, J., Blitzer, J., Dredze, M., Dredze, M., Pereira, F., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Annual Meeting-Association for Computational Linguistics* **45**(1), 440 (2007)
6. Cambria, E., Olsher, D., Rajagopal, D.: Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, pp. 1515–1521. AAAI Press (2014)

7. Dang, Y., Zhang, Y., Chen, H.: A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems* **25**(4), 46–53 (2010)
8. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: *Proceedings of LREC*, vol. 6, pp. 417–422. Citeseer (2006)
9. Fawcett, T.: An introduction to roc analysis. *Pattern recognition letters* **27**(8), 861–874 (2006)
10. Ferguson, N., Philip, D., Lam, H., Guo, J.: Media Content and Stock Returns: The Predictive Power of Press. *Multinational Finance Journal* **19**(1), 1–31 (2014)
11. Huang, A.H., Y.zang, A., Zheng, R.: Evidence on the information content of text in analyst reports. *Accounting Review* **89**(6), 2151–2180 (2014)
12. Ittner, C.D., Larcker, D.F.: Are nonfinancial measures leading indicators of financial performance? an analysis of customer satisfaction. *Journal of accounting research* **36**, 1–35 (1998)
13. Kaplan, R.S., Norton, D.P.: Linking the balanced scorecard to strategy. *California Management Review* **39**(1), 53–79 (1996)
14. Kearney, C., Liu, S.: Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* **33**(Cc), 171–185 (2014)
15. Li, F.: The information content of forward- looking statements in corporate filings-A naive bayesian machine learning approach. *Journal of Accounting Research* **48**(5), 1049–1102 (2010)
16. Li, Q., Chen, Y., Jiang, L.L., Li, P., Chen, H.: A tensor-based information framework for predicting the stock market. *ACM Transactions on Information Systems (TOIS)* **34**(2), 11 (2016)
17. Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., Chen, Y.: The effect of news and public mood on stock movements. *Information Sciences* **278**, 826–840 (2014)
18. Li, W.L.W., Han, J.H.J., Pei, J.P.J.: CMAR: Accurate and efficient classification based on multiple class-association rules. *Proceedings 2001 IEEE International Conference on Data Mining* pp. 369–376 (2001)
19. Li, X., Xie, H., Chen, L., Wang, J., Deng, X.: News impact on stock price return via sentiment analysis. *Knowledge-Based Systems* **69**, 14–23 (2014)
20. Liu, B., Hsu, W., Ma, Y., Ma, B.: Integrating Classification and Association Rule Mining. *Knowledge Discovery and Data Mining* pp. 80–86 (1998). DOI 10.1.1.48.8380
21. Liu, M., Qu, J.: Mining high utility itemsets without candidate generation. In: *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 55–64. ACM (2012)
22. Loughran, T., McDonald, B.: When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance* **66**(1), 35–65 (2011)
23. Loughran, T., McDonald, B.: The Use of Word Lists in Textual Analysis. *Journal of Behavioral Finance* **16**(1), 1–11 (2015)
24. Loughran, T., McDonald, B.: Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* (2016)
25. Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P.: Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* **65**(4), 782–796 (2014)
26. Man, Y., Yuanxin, O., Hao, S.: Investigating association rules for sentiment classification of web reviews. *Journal of Intelligent & Fuzzy Systems* **27**(4), 2055–2065 (2014)
27. Meretakakis, D., Wüthrich, B.: Extending naive bayes classifiers using long itemsets. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 165–174. ACM (1999)
28. Mo, S.Y.K., Liu, A., Yang, S.Y.: News sentiment to market impact and its feedback effect. *Environment Systems and Decisions* pp. 1–9 (2016)
29. Moilanen, K., Pulman, S., Zhang, Y.: Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression. *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010) at the 19th European Conference on Artificial Intelligence (ECAI 2010)* pp. 36–43 (2010)
30. O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P.P., Gurrin, C., Smeaton, A.F., O'Hare, N.: Topic-Dependent Sentiment Analysis of Financial Blogs. *International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement* pp. 9–16 (2009)

31. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* pp. 79–86 (2002)
32. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)* **27**(2), 12 (2009)
33. Stone, P.J., Bales, R.F., Namenwirth, J.Z., Ogilvie, D.M.: The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as Unit of Information. *Computers in Behavioral Science* **7**(4), 484–498 (1962)
34. Tetlock, P.C.: Giving Content to Investor Sentiment : The Role of Media in the Stock Market Published by : Wiley for the American Finance Association Stable URL : <http://www.jstor.org/stable/4622297> The Role of Media in the Stock Market Giving Content to Investor Sentim. *The Journal of Finance* **62**(3), 1139–1168 (2016)
35. Tetlock, P.C., Saar-Tsechansky, M., MacSkassy, S.: More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance* **63**(3), 1437–1467 (2008)
36. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* **63**(1), 163–173 (2012)
37. Turney, P.D.: Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 417–424 (2002)
38. Van De Kauter, M., Breesch, D., Hoste, V.: Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications* **42**(11), 4999–5010 (2015)
39. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* **39**(2-3), 165–210 (2005)
40. Yang, C.C., Tang, X., Wong, Y., Wei, C.P.: Understanding online consumer review opinions with sentiment analysis using machine learning. *Pacific Asia Journal of the Association for Information Systems* **2**(3) (2010)

## Appendix

### A. Parsing sentences using the NLTK toolkit

#### Grammar

```

JJ : {< JJ.* > *}
VB : {< VB.* >}
NP : {( < NNS|NN > ) *}
NPP : {< NNP|NNPS >}
RB : {< RB.* >}
NPJJ : {
  ((( < NP|NPP > + < IN > < .* > * < , > ) | ( < JJ > * < NP|NPP > +
    < VB > < NPP > * < .* > * < , > ) ) |
    ( < RB > | < DT > < NP > | < VB > < TO > ) ) |
    < JJ > * < NP > ( < IN > < DT > * < JJ > * < NPP > * < NP > * ) * < VB > |
    < JJ > + < NP > < VB > |
    < NP|NPP > + ( < ( > < .* > * < < > ) * ( ( < IN > < JJ > * < NP > )
      ( < IN > < CD > ) * ) * < VB > + |
      < NP > < .* > + ( < RB > | < JJ > ) |
      < NP|NPP > + ( < IN > < NP|NPP > ) * < .* > * < VB >
      ( < DT > < JJ > ) * |
      < NP > < VB > ( < RB > | ( < TO > < DT > < NP > ) ) |
      < VB > < NP|NPP > * < POS > < JJ > * < NP > |
      < VB > < PRP.* > < JJ > * < NP > < IN > |
      < VB > < TO > * < DT > * < JJ > * < NP > < IN > * < NP > * |
      < NP|NPP > + ( < IN > < DT > * < RB > * < JJ > * < NP|NPP > ) *
      < RB > * ( < VB > < JJ > < NP > ) * < VB >
      ( < DT > < CD > < NP > ) * |
      ( < JJ > ) * < NP|NPP > + < .* > * ( < , > < .* > * < , > ) * < NP >
    )
}

```

#### Steps:

1. Parse the sentence with a regular expression parser using the grammar specified above
2. If a match for 'NPJJ' tree pattern is found:
  - (a) Traverse the subtree to get the combination of NP (potential performance indicator) and JJ/RB/VB (potential directionality word). Look up the dictionary for the matching indicator and directionality word.
  - (b) If the combination of NP and JJ/RB/VB are not found, check for presence of individual words (either performance indicator or directionality word) in the dictionary
  - (c) Tag the sentence based on the identified matches.

### B. Parsing numeric values to determine directionality

#### Preconditions:

1. If a sentence has not been tagged with combination of performance indicators and directionality using the parse rules specified in Section A above.
2. The sentence contains terms like compared to, versus, down from, up from  
*#common sentences where one is likely to observe multiple numeric values without the use of directional words.*

Example sentence: *Operating profit* margin was 8.3%, compared to 11.8%  
a year earlier.

Expected tag output: LagInd::DOWN

### Grammar

```

JJ : {< JJ.* > *}
VB : {< VB.* >}
NP : {( < NNS|NN > ) *}
NPP : {< NNP|NNPS >}
RB : {< RB.* >}
CD : {< CD >}
NPJJ : {
    < NP|NPP > + ( < ( > < .* > * < > ) > ) * ( < IN > < DT > * < RB > *
        < JJ > * < NP|NPP > ) *
        < RB > * ( < VB > < JJ > < NP > ) * < VB > ( < DT > < CD > < NP > ) *
        < NP|NPP > * < CD > * < .* > * < CD > * |
    < NP|NPP > < IN > < NP|NPP > < CD > < .* > * < , > < VB >
        < IN > < NP|NPP > < CD >
}

```

### Steps:

1. Parse the sentence with a regular expression parser using the grammar specified above
2. If a match for 'NPJJ' tree pattern is found:
  - (a) Traverse the subtree to get the combination of NP, CD, CD to extract the performance indicator and numeric values. The numeric values are analysed to determine the directionality (UP/DOWN)
  - (b) Tag the sentence based on the identified matches.