Stock market sentiment lexicon acquisition using microblogging data and statistical measures

Nuno Oliveira[a,*], Paulo Cortez[a], Nelson Areal[b]

[a] *ALGORITMI Centre, Department of Information Systems, University of Minho,*
4804-533 Guimarães, Portugal
[b] *Department of Management, University of Minho, 4710-057 Braga, Portugal*

Abstract

Lexicon acquisition is a key issue for sentiment analysis. This paper presents a novel and fast approach for creating stock market lexicons. The approach is based on statistical measures applied over a vast set of labeled messages from StockTwits, which is a specialized stock market microblog. We compare three adaptations of statistical measures, such as pointwise mutual information (PMI), two new complementary statistics and the use of sentiment scores for affirmative and negated contexts. Using StockTwits, we show that the new lexicons are competitive for measuring investor sentiment when compared with six popular lexicons. We also applied a lexicon to easily produce Twitter investor sentiment indicators and analyzed their correlation with survey sentiment indexes. The new microblogging indicators have a moderate correlation with popular Investors Intelligence (II) and American Association of Individual Investors (AAII) indicators. Thus, the new microblogging approach can be used alternatively to traditional survey indicators with advantages (e.g., cheaper creation, higher frequencies).

Keywords: Sentiment analysis; Stock market; Microblogging data

---

1. Introduction

Recently, social media (e.g., Twitter, Facebook, message boards) has enabled a burst of unstructured opinion content that is potentially valuable for diverse decision-making processes [1]. Due to the volume and velocity properties of social media data, human analysis is impracticable and thus sentiment analysis (SA) is used to automatically mine large amounts of opinionated contents in order to summarize the opinions [2]. Several SA approaches apply common supervised classifiers

---

[*] Corresponding author at: R. das Lavouras, 29, 4505-462 Lobão, Portugal. Tel:.+351936607860.

Email addresses: `nunomroliveira@gmail.com` (Nuno Oliveira), `pcortez@dsi.uminho.pt` (Paulo Cortez), `nareal@eeg.uminho.pt` (Nelson Areal)

such as Support Vector Machines [3], Naive Bayes [4] or ensembles [5, 6]. Yet, the utilization of sentiment lexicons allows unsupervised classification of text, relieving the need for arduous manual labeling of text. Moreover, sentiment lexicons permit the creation of important features for supervised SA [7]. The sentiment lexicon is a list of words with a sentiment value (e.g., positive, negative) and it is considered a key element for SA [8]. For example, the sentence "this car is great" can be easily detected as positive if the lexicon has a term "great" with a positive value.

SA is being increasingly used to predict stock market variables [9, 10, 11, 12]. In particular, microblogging data are a useful source for supporting stock market decisions [13, 14]. Users post very frequently and data are readily available at low cost, allowing real-time assessment that can be exploited during the trading day. However, there has been little effort in producing lexicons adapted to the financial domain and microblogs. A financial lexicon was manually built by Loughran and McDonald [15] using text documents extracted from the U.S. Securities and Exchange Commission portal from 1994 to 2008. Mao et al. [16] proposed a procedure to automatically construct a Chinese financial lexicon by exploring a large news corpus classified as positive or negative according to the contemporaneous stock returns. Yet, these lexicons did not consider microblog messages, which is often informal and has character constraints. Furthermore, adopting a manual approach (e.g., [15]) is not feasible in practical terms given the huge effort required to label the large volumes of microblog texts. Moreover, the existing domain independent lexicons (e.g., [17, 18, 19]) may be ineffective for stock market contents. For instance, the word "explosive" is negative in most contexts but it may be positive in financial messages (e.g., "explosive rise").

In this work, we present a novel automated approach for the acquisition of microblog stock market lexicons. Our main contributions are:

i) The adaptation of three statistical measures (e.g., pointwise mutual information) and creation of two new complementary statistics. These measures are applied in labeled messages of the StockTwits microblogging service to calculate a stock market sentiment score. To address negation more efficiently, sentiment scores are also created for affirmative and negated contexts.

ii) The comparison of the resulting stock market lexicons created using the StockTwits test data with six large popular lexical resources: Harvard General Inquirer (GI) [17], opinion lexicon (OL) [2], Macquarie Semantic Orientation Lexicon (MSOL) [20], MPQA subjectivity lexicon (MPQA) [18], SentiWordNet (SWN) 3.0 [19] and financial sentiment dictionaries (FIN) [15].

iii) The assessment of the information content of sentiment indicators produced with a created and a baseline lexicons using a different microblog data source (Twitter). The new Twitter sentiment indicators are correlated with two traditional survey sentiment indicators: Investors Intelligence (II) and American Association of Individual Investors (AAII).

This paper is structured as follows. Section 2 shows related work. Section 3 presents the microblogging data, lexicon methods and sentiment indicators. Section 4 describes the experiments conducted and analyzes the obtained results. Finally, conclusions are drawn in Section 5.

2. Related Work

The sentiment lexicon is considered a key element for SA [8]. The utilization of lexicons permits the execution of effective unsupervised approaches [21, 22] and provides high quality features for supervised SA [23, 7]. Moreover, lexicons can be applied to diverse tasks such as SA, opinion retrieval [24] or opinion question answering and summarization [25], and they can be applied to diverse domains such as stock markets [15], electronic products [2] or the movie industry [26].

The creation of opinion lexicons is an important topic that has been studied for some time under two main approaches: manual and automatic creation. Manual creation is the most labor intensive and expensive approach because it requires experts to manually classify the sentiment value of each term. MPQA subjectivity lexicons [18] and General Inquirer [17] are two important examples of this methodology. Automatic creation requires much less human effort and allows for the faster inclusion of a larger set of lexical items. However, this is often achieved at the expense of accuracy.

There is substantial literature about the automatic construction of lexicons. Many of these studies apply text corpora for this procedure. Hatzivassiloglou and McKeown [27] extracted conjoined adjectives from a large Wall Street Journal corpus and produced a list of adjectives labeled as positive or negative. Using an initial set of adjectives with predetermined orientation labels, they developed a supervised learning algorithm to assign the sentiment polarity. First, they applied a log-linear regression model to determine if each pair of conjoined adjectives had the same or different orientations. Then, a clustering algorithm was used to divide the adjectives into two different positive and negative sets. Wiebe [28] extracted subjective adjectives from corpora by applying a method for clustering words based on distributional similarity [29] and another method to compute polarity and gradability [27]. Turney and Littman [21] tested two co-occurrence measures, Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA), on the AltaVista

Advanced Search engine in order to assign a semantic orientation to each word. The most accurate approach calculated the PMI of each term with pre-classified positive and negative words. A term was considered positive if the sum of its PMI scores with positive words was greater than the total PMI score with negative words, and vice-versa. Qiu et al. [30] explored diverse syntactic relations between opinion words and targets to iteratively extract further opinion words and targets. An initial seed set of opinion words was expanded and opinion targets were collected by continuously identifying terms having those syntactic associations with already extracted terms. Kiritchenko et al. [3] generated lexicons from tweets containing specific hashtag words and emoticons. These symbols were used as signals of the message sentiment (positive or negative). The sentiment score of each term was calculated using the PMI measures with positive and negative messages. These authors also produced distinct scores for negated and non-negated segments to properly obtain the sentiment in these contexts.

Lexical databases and thesaurus are also extensively applied in the creation of opinion lexicons. Kamps et al. [31] calculated the synonymy shortest path on the WordNet database (`wordnet.princeton.edu`) of adjectives to the words "good" and "bad" and determined their sentiment orientation based on these values. Kim and Hovy [32] created a system that automatically extracts holders and sentiment of each opinion about a given topic. This system includes a module for computing word sentiment. Synonymy and antonymy relations from WordNet are applied in this module in order to expand a small set of seed words and to calculate the strength of sentiment polarity. Esuli and Sebastiani [33] applied text classification techniques to the glosses of subjective words in order to determine their sentiment polarity. Mohammad et al. [20] produced a large lexicon using a set of affix patterns and the Macquarie thesaurus. The sentiment of every thesaurus paragraph was classified using a set of positive and negative words collected utilizing affix patterns. Then, each lexical item assumed the most common sentiment label of paragraphs containing the respective term. Baccianella et al. [19] produced the SentiWordNet lexicon by automatically calculating sentiment values to all WordNet synsets. First, these synsets were classified by a group of classifiers trained with pre-classified synsets. The different classification results were combined to generate a sentiment score to each synset. In a second phase, two iterative random-walk procedures were executed for the positivity and negativity values. These processes applied a graph with directed links from synsets included in glosses of other synsets. The random walk phase began with the values created in the previous step and finished when the processes had converged. Neviarouskaya et al. [34] created a lexicon by expanding an initial set of lexicon entries through synonymy, antonymy and hyponymy relations, derivation and compounding.

Other works explore both text corpora and lexical databases. Hu and Liu [2] utilized consumer reviews and Wordnet to select and classify opinion words associated to frequent product attributes. First, they extracted all adjectives included in the sentences of consumer reviews mentioning those product features. The sentiment orientation was assigned according to their semantic association in Wordnet with a seed list of words with known sentiment orientation. Each adjective assumed the same sentiment of synonyms or the inverse polarity of antonyms. The seed list was iteratively expanded with these newly classified words until no further words had antonyms or synonyms in the list. Takamura et al. [35] proposed the utilization of a spin model, where each word had a sentiment polarity (positive or negative), to produce an opinion lexicon. They created a lexical network based on the occurrence of terms in glosses of other terms, the synonymy, antonymy and hypernymy relations in thesaurus and some conjunctive expressions in corpus. Then, the mean-field method was applied on the network to determine the semantic orientations. Lu et al. [36] automatically generated a context-dependent sentiment lexicon by combining the utilization of domain independent lexicons, sentiment ratings of reviews, synonym and antonym relations in Wordnet and linguistic rules.

The utilization of generic lexicons or lexicons associated to other domains may be ineffective to SA on stock market text because sentiment is sensitive to the domain [21]. For example, the verb “underestimate” has often a negative sentiment but an underestimated stock can constitute an opportunity to buy, thus denoting a positive value within the stock market domain. However, the creation of opinion lexicons for the financial domain has been scant. One of the most popular works in this context is by Loughran and McDonald [15], who manually created six word lists (i.e., positive, negative, litigious, uncertainty, modal strong and modal weak) from words occurring in at least 5% of a large collection of 10,000 documents between 1994 and 2008. Also in 2014, Mao et al. [16] presented a procedure to automatically produce a Chinese financial lexicon. A large Chinese news corpus was labeled according to the stock returns. Then, a set of seed words was selected based on the Document Frequency value with news associated with very high or very low returns. The lexicon was expanded by considering the statistical association with seed words and the economic significance of candidate terms. The final lexicon was obtained by an iterative optimization process.

In summary, the majority of the studies applies text corpora (e.g., [27, 28, 2, 35, 30, 3, 16]) and/or existing lexical databases and thesaurus (e.g., [31, 32, 2, 33, 35, 26, 20, 19, 34, 37]). The extraction of opinion words or targets are mainly based on syntactic relations (e.g., [27, 30]), part of speech (e.g., adjectives [27, 28, 2]), co-occurrence with terms (e.g., [2]) and semantic relations in WordNet (e.g., [32, 34]). The calculation of the sentiment polarity or score is mostly performed by statistical measures (e.g., PMI [21, 3]), text classification of glosses (e.g., [33, 20, 19]), semantic associations (e.g., sinonymy, antonymy relations in WordNet [31, 32, 2, 34]), syntactic relations (e.g., [30]) and clustering methods (e.g., [27, 28]). Only one study produced different sentiment scores for affirmative and negated contexts [3], although applied for generic Twitter messages and thus not specifically adjusted to the stock market domain, as performed in this paper. Some of these studies extract and classify opinions words simultaneously by iteratively expanding a pre-labeled seed list by semantic or syntactic relations (e.g., [2, 30]). The newly collected words assume the same or the opposite polarity of the associated term.

Some of these approaches are ineffective for the creation of specialized domain lexicons. Methods based on WordNet or thesaurus produce domain independent lexicons, possibly unadjusted for stock market contents [38]. Therefore, the utilization of methods such as semantic relations in WordNet and text classification of glosses are unsatisfactory for our purpose. The usage of a small group of syntactic relations (e.g., conjunctions [27, 35]) does not allow for the selection of a large set of words. Moreover, extracting only adjectives (e.g., [2, 31, 27]) ignores terms with a strong sentiment, such as "love" and "hate" verbs. In addition, the expansion of a seed list of classified words (e.g., [30]) is less effective than the application of classified text corpora. Since we have classified data exclusively about stock markets, we do not need to restrict the collection to words co-occurring with specific terms (e.g., [2]) and clustering methods (e.g., [27, 28]) become less relevant to assign sentiment polarity.

In this work, we propose a novel automated approach for the creation of microblog stock market lexicons. Three adaptations (e.g., PMI) and two new proposed statistics were applied in a large data set of classified data provided by a microblogging platform exclusively dedicated to stock markets (`stocktwits.com`). We believe this is the largest labeled data set used in the creation of stock market lexicons. Additionally, sentiment scores were created for affirmative and negated contexts in order to address negation properly. Within our knowledge, this work presents the first stock market lexicons with two different context scores for each entry.

3. Material and Methods

3.1. Microblogging Data

Microblogging data are useful for the creation of investor sentiment indicators. The community of investors using these services is growing and becoming more representative. The character limit demands greater objectivity. Microblogging users usually react promptly to events, allowing a near real-time sentiment assessment. Data is quite vast and freely available allowing a more frequent production of indicators than traditional sources, such as extensive surveys. The selection of messages containing cashtags reduces the amount of irrelevant data and permits the creation of sentiment indicators related to particular stocks. A cashtag is composed by a ”$” character and a stock ticker (e.g., $AAPL) and it is usually applied in messages about that stock. In this work, we use StockTwits data to create stock market lexicons and Twitter data in the production of sentiment indicators that are used to correlate with traditional sentiment indicators.

StockTwits is a microblogging service exclusively dedicated to stock market conversations (`stocktwits.com`) that has currently more than 300,000 users. StockTwits users can label their own text messages as “bullish” (optimistic opinion) or “bearish” (pessimistic view). Author supplied sentiment labels are already explored in SA on other social media (e.g., blogs [38]) and topics. In this paper, we explore these labeled messages, as kindly provided by StockTwits from June 2, 2010 to March 31, 2013, in a total of 350,000 posts[1]. We note that such dimension is significantly higher when compared with the majority of works on this topic.

Twitter (`twitter.com`) is the most popular microblogging service. Unlike StockTwits, it is a generic platform. Yet, Twitter users also apply cashtags in stock market conversations. In the creation of microblogging sentiment indicators, we used Twitter for reasons of data availability. Using Twitter REST API (`https://dev.twitter.com/docs/api`) and the R language (or statistical environment), we collected all tweets containing cashtags of all stocks traded in US stock markets from 22$^{\text{nd}}$ of December 2012 to 27$^{\text{th}}$ of March 2015[2].

3.2. Baseline Lexicons

For comparison purposes, we adopt six large and popular lexicons:

---

[1] The total number of StockTwits messages over the same period is nearly 6 millions.

[2] The total number of collected Twitter messages mentioning 3762 different cashtags is approximately 19 millions.

- GI [17] – comprises around 11,000 words (`http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm`). In particular, we used all words of the "Positiv" and "Negativ" attributes.
- OL [2] – contains nearly 6,000 positive and negative terms. The lexicon also contains common social media misspelled words (`http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`).
- MSOL [20] – classifies more than 75,000 n-grams as positive or negative (`http://saifmohammad.com/Lexicons/MSOL-June15-09.txt.zip`).
- MPQA [18] – with around 8,000 entries (`http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`). It contains a list of strong and weak subjective terms. We assigned half of the sentiment score (i.e., 0.5 or -0.5) to weaker terms.
- SWN [19] – with continuous sentiment values to the nearly 117,000 synsets (i.e., group of words that are semantically equivalent in some context) of the WordNet lexical database (`http://sentiwordnet.isti.cnr.it/downloadFile.php`). A word may have multiple scores, since it can belong to diverse synsets. To solve this issue, we averaged the positive and negative values for all (word, POS tag) pairs.
- FIN [15] – contains word lists commonly applied in financial text documents (`http://www3.nd.edu/~mcdonald/Word_Lists.html`). We adopted the terms classified as negative (2349) and positive (354).

3.3. Lexicon Creation

This work applies two different validation approaches. To build a single large lexicon and evaluate its performance, we adopt a holdout split method, where the first 75% StockTwits classified messages are used to create lexicons (training set) and the remaining (most recent) 25% posts are used for evaluation purposes (test set). To perform a robust comparison of the distinct lexicon creation methods, we adopt a realistic rolling window method [39], where the labeled data are split into 20 equally sized parts ordered by time. The first 2/3 messages (training set) of each data window is utilized to create lexicons and the last 1/3 (test set) are applied in the evaluation. After performing the data pre-processing tasks, we selected all items having a minimum number of occurrences in the training set (O_{min}). The removal of non-frequent items is a usual preprocessing task in the creation of lexicons (e.g., [30, 15, 3, 16]). This operation permits the elimination of many orthographic errors. Also, some statistical measures (e.g., PMI) are unsatisfactory estimators of association for infrequent terms [3]. Since the dimension of training data sets for each evaluation procedure is very different, we defined a different minimum number of occurrences (O_{min}) for each evaluation scheme.

The usage of different combinations of statistical measures on training data generates several lexicons. First, we produce three lexicons by applying adaptations of three known statistical measures (Term frequency-inverse document frequency (TF-IDF), Information Gain (IG) and Pointwise Mutual Information (PMI)) to calculate the sentiment score of each selected item. Then, we create three more versions of previous lexicons (i.e., a total of 12 lexicon versions) by utilizing two new complementary statistics ($P_{\text{days}}(l)$ and $M_{\text{assoc}}(l)$). In order to refine the sentiment score, the value obtained by the latter measures is multiplied by the score produced by each adapted measure. Additionally, we tested the calculation of sentiment scores for affirmative and negated contexts. The previously described procedure is used for separate affirmative and negated training data.

3.3.1. Data Pre-Processing

In order to prepare the microblogging data for the lexicon creation, we performed various preprocessing tasks using the R tool [40]:

- substitute all cashtags by a unique term, thus avoiding cashtags to gain a sentiment value related with a particular time period;
- replace numbers by a single tag, since the whole set of distinct numbers is too vast;
- for privacy reasons, all mentions and URL addresses were normalized to "@user" and "URL", respectively;
- exclude messages composed only by cashtags, url links, mentions or punctuation (7,176 messages were removed).

Next, we adopted the Stanford CoreNLP tool [41] to execute common natural language processing operations, such as tokenization, part of speech (POS) tagging and lemmatization.

The holdout split method uses a large training set with 250,000 posts and thousands of distinct terms. Thus, for this evaluation scheme we included only terms with more than $O_{min} = 10$ occurrences in the training data set and excluded all punctuation, resulting in approximately 7,000 unigrams and 27,000 bigrams analyzed. The rolling window method creates lexicons in each one of the 20 data partitions. In this validation scheme, the training set of the partitions is much smaller (about 11,100 messages) and thus we adopted a lower minimum number of occurences, with $O_{min} = 4$ in each training set. Also, we eliminated all punctuation.

3.3.2. Statistical Measures

In this work, we adapt three popular statistical measures and propose two new complementary ones. The former measures were applied to determine the information value of lexical items and thus allow us to discriminate them between "bullish" or "bearish":

1. TF-IDF – often used for textual data representation (e.g., [21]) and that is calculated as:

$$tf(l,d) = \frac{n_{d,l}}{n_D} \tag{1}$$

$$idf(l) = \log \frac{N_d}{N_l + 1} \tag{2}$$

$$tf\text{-}idf(l,d) = tf(l,d) \times idf(l) \tag{3}$$

   where l is a lexical entry, d is a particular document, $n_{d,l}$ is the number of occurrences of l in document d, n_D is the number of lexical items in document d, N_d is the number of documents and N_l is the number of documents containing l. We first created two documents composed by all messages of each class (d_1 – bullish and d_2 – bearish). Then, we executed the *tfidf* function of the textir R package to compute $tf\text{-}idf(l,d)$. To provide a single value that reflects the tendency to a sentiment class, we calculated the sentiment value $S_{\text{TF–IDF}}$ as:

$$S_{\text{TFIDF}}(l) = \frac{tf\text{-}idf(l,d_1) - tf\text{-}idf(l,d_2)}{tf\text{-}idf(l,d_1) + tf\text{-}idf(l,d_2)} \tag{4}$$

   The final sentiment class depends on the $S_{\text{TF–IDF}}(l)$ value: "bullish" if positive, "bearish" if negative or "neutral" if zero.

2. IG – commonly used to access the information value of an attribute (e.g., [42, 43]) and that is computed as:

$$IG(l,c) = \sum_{d\in\{c,\bar{c}\}} \sum_{w\in\{l,\bar{l}\}} p(w,d) \log \frac{p(w,d)}{p(w) \times p(d)} \tag{5}$$

   where c refers to a category (bullish or bearish), $\bar{c}$ means the non-membership in category c and $\bar{l}$ refers to the absence of l. Since there are only 2 categories, $\bar{c}$ of each class corresponds to c of the other class and IG values are equal for both categories. Hence, we propose the slight adaptation:

$$
\begin{aligned}
IG_a(l) = p(l,bl)\log\frac{p(l,bl)}{p(l)\times p(bl)} \\
+ p(\bar{l},br)\log\frac{p(\bar{l},br)}{p(\bar{l})\times p(br)} - p(\bar{l},bl)\log\frac{p(\bar{l},bl)}{p(\bar{l})\times p(bl)} \\
- p(l,br)\log\frac{p(l,br)}{p(l)\times p(br)} \quad (6)
\end{aligned}
$$

where *bl* refers to bullish class and *br* corresponds to bearish category. In this calculation, instead of summing the values of mutual information of all tuples, we add those referring to tuples correlated to bullish class, (l,bl) and $(\bar{l},br)$, and subtract values corresponding to tuples associated to bearish class, (l,br) and $(\bar{l},bl)$. Thus, a positive value indicates a bullish orientation and a negative value means a bearish item. Since very frequent words tend to present very high IG values, we prevent this effect by computing the final sentiment score as:

$$S_{\text{IG}}(l) = \frac{IG_a(l)}{n_l} \quad (7)$$

where n_l is the number of times that term l appears in all texts.

3. PMI – a popular statistic in the development of lexicons (e.g., [21, 3]):

$$PMI(x,y) = \log_2\frac{p(x,y)}{p(x)p(y)} \quad (8)$$

where x and y are words or sets of words, $p(x,y)$ is the probability that they co-occur, and $p(x)$ and $p(y)$ are the probabilities of occurring x and y in the corpus, respectively. PMI will be largely positive if x and y are strongly associated, highly negative if they are complementary and near zero if there is no significant relationship between them. We adapt the sentiment score to include both positive and negative PMI values:

$$S_{\text{PMI}}(l) = PMI(l, bullish) - PMI(l, bearish) \quad (9)$$

where l is a lexical item, *bullish* refers to all bullish messages and *bearish* corresponds to all bearish messages. The sentiment score signal reflects the sentiment orientation.

The three statistical measures were computed to both unigrams (individual words) and bigrams (two sequential terms). We produced one lexicon for each measure, which includes unigrams and bigrams that present a better sentiment score than their constituent terms.

Two novel complementary metrics, $P_{\text{days}}(l)$ and $M_{\text{assoc}}(l)$, are proposed to refine the sentiment score produced by each previously described metric ($S_{\text{TFIDF}}(l)$, $S_{\text{IG}}(l)$ or $S_{\text{PMI}}(l)$). They may increase or decrease the sentiment value calculated by each of the three adapted metrics. We apply the complementary metrics by multiplying them with other metrics (e.g., $S_{\text{TFIDF}}(l)$, $S_{\text{IG}}(l)$ and $S_{\text{PMI}}(l)$). For example, the score of item l produced by the combination of $P_{\text{days}}(l)$, $M_{\text{assoc}}(l)$ and $S_{\text{PMI}}(l)$ is: $P_{\text{days}}(l) \times M_{\text{assoc}}(l) \times S_{\text{PMI}}(l)$.

The $P_{\text{days}}(l)$ statistic calculates, for each lexical item, the percentage of days where the majority of messages mentioning it have the same sentiment polarity of the item. To have a less biased measure favouring the dominant class (i.e., bullish), we multiply the daily number of bearish messages containing the lexical item by the following adjustment value:

$$V_{\text{adj}} = \frac{N_{\text{bull}}}{N_{\text{bear}}} \tag{10}$$

where N_{bull} is the total number of bullish messages in training set and N_{bear} is the total number of bearish messages. We tested the $P_{\text{days}}(l)$ metric to prevent terms appearing in an abnormally high number in few days to have a polluted sentiment score by the predominant opinion in those days. While a low value may indicate the existence of the described situation, a high $P_{\text{days}}(l)$ value means that the lexical item has consistently the same sentiment orientation. Therefore, we expect that this measure may improve sentiment score computation.

Previous measures also do not account for the association of two sets of words: intensifiers (e.g., more, increase, up) and diminishers (e.g., less, decrease, down). Yet, the analysis of these relationships may improve the calculation of sentiment. The presence of diminishers may reverse the sentiment of the following word (e.g., less debt) while intensifiers may reinforce it (e.g., more debt). Thus, previous measures will not be effective in those situations. For instance, the likely presence of "less profit" in a negative message would incorrectly decrease the sentiment score of the positive word "profit" when calculated by former statistical measures. Therefore, we propose the $M_{\text{assoc}}(l)$ metric to address this issue:

$$N_{\text{Int}}(l) = N_{\text{IntBull}}(l) + N_{\text{DimBear}}(l) \tag{11}$$

$$N_{\text{Dim}}(l) = N_{\text{IntBear}} + N_{\text{DimBull}}(l) \tag{12}$$

$$M_{\text{assoc}}(l) = \begin{cases} \frac{N_{\text{Int}}(l)}{N_{\text{Int}}(l) + N_{\text{Dim}}(l) \times \frac{T_{\text{Int}}}{T_{\text{Dim}}}} + 0.5 & \text{if } l \text{ is Bullish} \\ \frac{N_{\text{Dim}}(l)}{N_{\text{Int}}(l) + N_{\text{Dim}}(l) \times \frac{T_{\text{Int}}}{T_{\text{Dim}}}} + 0.5 & \text{if } l \text{ is Bearish} \end{cases} \tag{13}$$

where N denotes the number of occurrences of the lexical item adjoined to: $_{\text{IntBull}}$ – intensifier words (e.g., more profit) in bullish messages; $_{\text{IntBear}}$ – intensifier words in bearish messages; $_{\text{DimBull}}$

– diminisher words (e.g., less profit) in bullish messages; and DimBear – diminisher words in bearish messages. For all analyzed elements, T_{Int} is the sum of $N_{\text{Int}}(l)$ and T_{Dim} is the sum of $N_{\text{Dim}}(l)$.

The $M_{\text{assoc}}(l)$ measure is only used in elements with more than four occurrences adjoined to intensifiers and diminishers. We selected this threshold value because we consider that a lower number would produce many cases of less solid values of association. For instance, it is more likely to happen an excessively high $M_{\text{assoc}}(l)$ value for elements with two occurrences (e.g., $N_{\text{Int}}(l) = 2$ and $N_{\text{Dim}}(l) = 0$ for a bullish term) than with four or more occurrences.

We distinguished the formula for bullish and bearish items because bullish terms shall have higher $N_{\text{Int}}(l)$ values and bearish terms shall produce higher $N_{\text{Dim}}(l)$ values. Since $M_{\text{assoc}}(l) \in [0.5, 1.5]$, a $M_{\text{assoc}}(l)$ value close to 1.5 means that these associations are highly concordant to the assigned sentiment polarity and the absolute sentiment score will increase. A low $M_{\text{assoc}}(l)$ value indicates the opposite, decreasing the absolute score. The intensifiers and diminishers (Table 1) were manually selected. First, we choose a small set of words (e.g., less, more, very) and then added synonyms found in a thesaurus.

Table 1: Intensifiers and Diminishers

Intensifiers	Diminishers
accretion, accrual, addendum, addition, augmentation, boost, expansion, gain, increment, more, plus, proliferation, raise, rise, accelerate, add, aggrandize, amplify, augment, enlarge, escalate, expand, extend, hype, multiply, swell, stoke, supersize, up, accumulate, climb, proliferate, soar, uprise, desire, fancy, prefer, enjoy, relish, admire, adore, esteem, hallow, idolize, revere, venerate, worship, appreciate, love, elevated, escalated, heightened, increased, raised, admiring, applauding, appreciative, approbatory, approving, commendatory, complimentary, friendly, good, positive	abatement, decline, decrease, decrement, depletion, diminishment, diminution, fall, lessening, loss, lowering, reduction, shrinkage, diminish, dwindle, lessen, recede, wane, abate, downsize, lower, minify, reduce, subtract, drop, descend, dip, plunge, dive, sink, slide, abhor, abominate, despise, detest, execrate, loathe, deplore, deprecate, disapprove, disdain, disfavor, dislike, hate, decreased, depressed, dropped, receded, under, down, low, adverse, depreciative, depreciatory, derogatory, disapproving, inappreciative, negative, unappreciative, uncomplimentary, unfavorable, unflattering, unfriendly

3.3.3. Scores for Affirmative and Negative Contexts

The sentiment value of a term may change in different contexts. For instance, negation is a frequent context that can modify the sentiment polarity or intensity of a particular word. While many studies process negation by reverting sentiment polarity (from positive to negative and vice-versa), others argue that sentiment reversion may not be adequate [3]. For example, “frightening” is very negative but “not frightening” often suggests a less intense negative emotion.

To address negation more efficiently, we calculated sentiment scores for negated and affirmative (non-negated) contexts separately [3]. We divided the training data set into an affirmative and a negated corpus. The negated set contains all negated contexts segments and the affirmative set is composed by the remaining segments. The negated contexts are the sentence segments starting with a negation word present in the Christopher Potts' sentiment tutorial (`http://sentiment.christopherpotts.net/lingstruc.html`) and ending with one of the punctuation marks: ',', '.', ':', ';', '!', '?'. Then, we create the stock market lexicon by applying the same procedure utilized for the "general" score (i.e., described in the previous subsection) on each corpus (affirmative and negated), producing two sentiment scores for each item that should be used in the respective context. However, some items do not have sufficient occurrences in each corpus in order to have both sentiment scores calculated. In such situations, we assign its "general" sentiment to the unavailable sentiment context score.

3.4. Lexicon Evaluation

As explained in Section 3.3, we adopt two complementary evaluation procedures that use a time ordered training/test split: a single holdout (75%/25%) and a rolling window (with 20 windows, each with 2/3 for training and 1/3 for testing). The former procedure creates a large lexicon that is publicly made available, while the latter procedure uses much less data to generate each lexicon but it allows us to get several test sets and thus execute statistical significance tests. For both evaluation methods, we performed SA in each test set by applying each lexicon. The message overall sentiment value is computed as the sum of all its lexical scores. When lexicon bigrams are present in the text, we only sum the score of the bigrams and do not account for the score of their individual constituents. The message is classified as "bullish", "bearish" or "neutral" according to the sign of the sum (positive, negative or zero). In SA applying lexicons with affirmative and negated scores, we also identified the affirmative and negated context segments in order to utilize the adequate sentiment score.

The classification measures used were:

- the percentage of correct classifications (CC1);
- the percentage of unclassified messages, i.e., texts with no lexicon items (Unc);
- the percentage of correct classifications excluding unclassified messages (CC2);
- precision for "bullish" ($\mathrm{P}_{\mathrm{Bull}}$) and "bearish" ($\mathrm{P}_{\mathrm{Bear}}$), given by $\frac{TP}{TP+FP}$, where TP denotes the number of true positives and FP the number of false positives;
- recall for"bullish" ($\mathrm{R}_{\mathrm{Bull}}$) and "bearish" ($\mathrm{R}_{\mathrm{Bear}}$), given by $\frac{TP}{TP+FN}$, where FN denotes the number of false negatives;
- F-score for "bullish" ($\mathrm{F1}_{\mathrm{Bull}}$) and "bearish" ($\mathrm{F1}_{\mathrm{Bear}}$), where $F_1 = 2\frac{Precision*Recall}{Precision+Recall}$;

- macro-averaged F-score (F_{Avg}) that averages both F-scores (F1_{Bull},F1_{Bear}).

We assume that the classification is correct when it matches the same sentiment ("bullish" or "bearish") as provided by user who made the post. Under the rolling window scheme, we verified the statistical significance of $CC1$ and F_{Avg} improvements obtained by a specific approach relatively to another one. The parametric paired Student's t-test and the non-parametric Wilcoxon signed rank test were applied to pairs of lexicon creation methods.

3.5. Sentiment Indicators

Some works in the literature argue that sentiment may affect prices. In these studies, sentiment indicators based on indirect measures (e.g., end fund discount, NYSE share turnover) or surveys have informative value in the forecasting of aggregate stock market returns (e.g., [44, 45]) or in the prediction of returns of portfolios formed on diverse attributes (e.g., market value [46, 47], financial distress [46, 47], volatility [44, 46, 47]). Recently, several studies applied computational linguistic methods to textual contents (e.g., microblogs, message boards or newspapers) to extract investor sentiment indicators (e.g., [9, 10, 11, 12, 13, 14, 48, 49, 50]). Some of these papers found that sentiment indicators have predictive value for future market directions (e.g., [10, 11, 12, 13, 48]) and returns (e.g., [49]) and allow profitable trading strategies (e.g., [11, 50]).

In this work, we measured the correlation of microblogging sentiment indicators with two widely applied survey sentiment indicators: the American Association of Individual Investors (AAII) (e.g., [45, 51]) and Investors Intelligence (II) (e.g., [45, 51, 52]). AAII measures the percentage of individual investors who are bullish, bearish, and neutral based on the votes of their members to a poll questioning their sentiment on the stock market for the next six months. AAII values are published online each Thursday morning containing data from previous Thursday until last Wednesday. II analyzes each week over a hundred market newsletters and categorize each author's current opinion about the market as bullish, bearish or correction. The percentage of newsletters classified as bullish, bearish or correction are published every Wednesday and they include the newsletters analyzed until Tuesday. II measures may be more correlated to institutional sentiment than AAII, because many authors are market professionals [51]. AAII and II indicators were collected from Thompson Reuters Datastream (`http://online.thomsonreuters.com/datastream/`).

Twitter sentiment indicators were created by applying SA on the collected Twitter data using two distinct lexicons:

- TWTSML uses the selected stock market lexicon (SML), i.e., the $\text{PMI}_{\text{BiScr}}$ lexicon created using 75% of StockTwits data (Section 3). Affirmative or negated context scores are applied in the respective segments.
- TWTSWN applies the SWN lexicon (the selected baseline lexicon, Section 3).

Six different values are computed for each time period:

$$\text{TWTSML}_{\text{bull,t}} = \text{SML}_{\text{bull,t}}/(\text{SML}_{\text{bull,t}} + \text{SML}_{\text{bear,t}}) \tag{14}$$

$$\text{TWTSML}_{\text{bear,t}} = \text{SML}_{\text{bear,t}}/(\text{SML}_{\text{bull,t}} + \text{SML}_{\text{bear,t}}) \tag{15}$$

$$\text{TWTSML}_{\text{spread,t}} = \text{TWTSML}_{\text{bull,t}} - \text{TWTSML}_{\text{bear,t}} \tag{16}$$

$$\text{TWTSWN}_{\text{bull,t}} = \text{SWN}_{\text{bull,t}}/(\text{SWN}_{\text{bull,t}} + \text{SWN}_{\text{bear,t}}) \tag{17}$$

$$\text{TWTSWN}_{\text{bear,t}} = \text{SWN}_{\text{bear,t}}/(\text{SWN}_{\text{bull,t}} + \text{SWN}_{\text{bear,t}}) \tag{18}$$

$$\text{TWTSWN}_{\text{spread,t}} = \text{TWTSWN}_{\text{bull,t}} - \text{TWTSWN}_{\text{bear,t}} \tag{19}$$

where:

- $\text{SML}_{\text{bull,t}}$ corresponds to the sum of all positive SA scores (i.e., greater than zero) using SML (i.e., $\text{PMI}_{\text{BiScr}}$ lexicon) on all tweets from a given t time period.
- $\text{SML}_{\text{bear,t}}$ corresponds to absolute value of the sum of all negative SA scores (i.e., less than zero) using SML on all tweets for time t.
- $\text{SWN}_{\text{bull,t}}$ corresponds to the sum of all positive SA scores using the selected baseline lexicon (i.e., SWN lexicon) on all tweets for time t.
- $\text{SWN}_{\text{bear,t}}$ corresponds to absolute value of the sum of all negative SA scores using the SWN lexicon on all tweets for time t.

We used survey and Twitter indicators from February 1, 2013 to March 27, 2015. This time period is subsequent to the applied in the creation of the selected stock market lexicon. Since both AAII and II use ratios, we also created bullish and bearish ratios. We decided to use SA scores instead of the number of bullish or bearish messages because we consider that they better indicate the sentiment strength. Additionally, we calculated the bull-bear spread ($\text{TWTSML}_{\text{spread}}$, $\text{TWTSWN}_{\text{spread}}$), a common measure of sentiment (e.g., [51, 52]). Different Twitter sentiment indicators (i.e., $\text{TWTSML}_{\text{bull}}$, $\text{TWTSML}_{\text{bear}}$, $\text{TWTSML}_{\text{spread}}$, $\text{TWTSWN}_{\text{bull}}$, $\text{TWTSWN}_{\text{bear}}$, $\text{TWTSWN}_{\text{spread}}$,) were computed for each survey sentiment indicator corresponding to their time periods. Each Twitter sentiment indicator is correlated to their AAII and II counterparts (e.g., $\mathrm{TWTSML}_{\mathrm{bull}}$ with $\mathrm{AAII}_{\mathrm{bull}}$, $\mathrm{TWTSWN}_{\mathrm{spread}}$ with $\mathrm{II}_{\mathrm{spread}}$).

4. Results

4.1. Lexicon Evaluation

In this section we present the SA results for the tested lexicons. We start by analyzing the use of the proposed statistical measures in the computation of a unique sentiment score for each item. We experimented four different scores for the three main statistical measures (PMI, TFIDF, IG):

- *Scr* corresponds to the value calculated by the main statistical measure (S_{PMI}, S_{TFIDF}, S_{IG});
- *Assoc* is the product of *Scr* and M_{assoc};
- *Days* is the product of *Scr* and P_{days};
- *All* is the product of *Scr*, M_{assoc} and P_{days}.

Table 2 shows all classification metrics for the created lexicons using a unique context score and table 3 indicates which lexicons obtain statistically significant higher CC1 and $\mathrm{F}_{\mathrm{Avg}}$ values compared with other lexicons according to the paired Student's t-test and the Wilcoxon signed rank test. The alternative hypothesis of these tests is that the lexicon in the row has higher results than the lexicon in the column.

The best overall results (e.g., highest CC1, CC2 and $\mathrm{F}_{\mathrm{Avg}}$ values) are obtained by the $\mathrm{PMI}_{\mathrm{All}}$ method, which delivers statistically significant higher CC1 and $\mathrm{F}_{\mathrm{Avg}}$ values than all other lexicons. The complementary metrics proved to be useful because they improved the evaluation results for all main statistical measures. Every lexicon applying M_{assoc} or P_{days} metrics obtain statistically significant higher CC1 and $\mathrm{F}_{\mathrm{Avg}}$ values than their *Scr* counterparts (e.g., $\mathrm{PMI}_{\mathrm{Assoc}}$, $\mathrm{TFIDF}_{\mathrm{All}}$ and $\mathrm{IG}_{\mathrm{Days}}$ are higher than $\mathrm{PMI}_{\mathrm{Scr}}$, $\mathrm{TFIDF}_{\mathrm{Scr}}$ and $\mathrm{IG}_{\mathrm{Scr}}$, respectively). The M_{assoc} measure permits slight gains in both $F1_{\mathrm{Bull}}$ and $F1_{\mathrm{Bear}}$ scores. The P_{days} metric is able to substantially improve $F1_{\mathrm{Bull}}$ values while maintaining the $F1_{\mathrm{Bear}}$ very similar.

Next, we compare the selected $\mathrm{PMI}_{\mathrm{All}}$ with diverse reference lexicons. The SA results obtained by these lexical resources are presented in Tables 4 and 5. $\mathrm{PMI}_{\mathrm{All}}$ based lexicons achieve the best results for all evaluation metrics by a significant margin. For example, sentiment classification using this lexicon obtains a 18.2 ($\mathrm{F}_{\mathrm{Avg}}$), 18.7 (CC2) and 21.4 (CC1) percentage point difference in the holdout split scheme when compared with the baseline resource that has the highest overall results

Table 2: Classification results for the created lexicons with unique context score (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P_{Bull}	R_{Bull}	$F1_{\text{Bull}}$	P_{Bear}	R_{Bear}	$F1_{\text{Bear}}$	F_{Avg}
Panel A: Evaluation results of holdout split method										
PMI_{Scr}	75.2	**0.5**	75.5	88.6	76.5	82.1	51.9	71.5	60.1	71.1
$\text{PMI}_{\text{Assoc}}$	75.6	**0.5**	76.0	88.5	77.4	82.6	52.6	70.6	**60.3**	71.4
PMI_{Days}	**78.8**	**0.5**	79.1	86.0	85.4	85.7	59.5	59.6	59.6	72.6
PMI_{All}	**78.8**	**0.5**	**79.2**	86.0	**85.5**	**85.8**	**59.7**	59.5	59.6	**72.7**
$\text{TFIDF}_{\text{Scr}}$	74.3	**0.5**	74.7	88.6	75.1	81.3	50.6	71.9	59.4	70.4
$\text{TFIDF}_{\text{Assoc}}$	74.8	**0.5**	75.1	88.4	76.2	81.8	51.3	70.8	59.5	70.7
$\text{TFIDF}_{\text{Days}}$	78.4	**0.5**	78.7	85.6	85.4	85.5	58.8	58.3	58.6	72
$\text{TFIDF}_{\text{All}}$	78.5	**0.5**	78.8	85.5	**85.5**	85.5	59.1	58.1	58.6	72.1
IG_{Scr}	70.5	**0.5**	70.8	89.4	68.5	77.5	46.1	**76.3**	57.4	67.5
IG_{Assoc}	71.6	**0.5**	71.9	**89.5**	70.1	78.6	47.3	75.9	58.3	68.4
IG_{Days}	76.0	**0.5**	76.4	87.2	79.5	83.2	53.4	65.9	59.0	71.1
IG_{All}	76.4	**0.5**	76.7	86.9	80.4	83.5	54.1	64.9	59.0	71.3
Panel B: Average evaluation results of rolling window method										
PMI_{Scr}	71.1	**0.5**	71.4	90.0	69.0	78.0	45.7	77.0	56.9	67.4
$\text{PMI}_{\text{Assoc}}$	71.5	**0.5**	71.9	89.9	69.8	78.4	46.2	76.7	57.3	67.9
PMI_{Days}	77.3	**0.5**	77.7	87.4	81.5	84.3	54.1	64.4	58.5	71.4
PMI_{All}	**77.5**	**0.5**	**77.8**	87.3	81.7	84.4	**54.6**	64.2	**58.7**	**71.5**
$\text{TFIDF}_{\text{Scr}}$	71.6	**0.5**	71.9	89.1	70.8	78.8	45.6	73.4	55.8	67.3
$\text{TFIDF}_{\text{Assoc}}$	72.1	**0.5**	72.5	89.0	71.7	79.3	46.3	73.1	56.3	67.8
$\text{TFIDF}_{\text{Days}}$	77.1	**0.5**	77.5	86.4	82.4	84.3	54.0	60.5	56.6	70.5
$\text{TFIDF}_{\text{All}}$	77.3	**0.5**	77.7	86.5	**82.7**	**84.5**	54.5	60.4	56.9	70.7
IG_{Scr}	67.4	**0.5**	67.7	**90.5**	63.8	74.2	42.7	**78.7**	54.3	64.2
IG_{Assoc}	68.0	**0.5**	68.3	**90.5**	64.7	74.9	43.2	78.5	54.6	64.8
IG_{Days}	75.2	**0.5**	75.5	88.4	77.4	82.3	50.6	68.6	57.3	69.8
IG_{All}	75.5	**0.5**	75.8	88.4	77.8	82.6	51.0	68.5	57.6	70.1

Table 3: Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique context scores. The following symbols denote significance at the 5% level: a - paired Student's t t-test for F_{Avg}; b - Wilcoxon signed rank test for F_{Avg}; c - paired Student's t t-test for $CC1$; d - Wilcoxon signed rank test for $CC1$. Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column

	PMI_{Scr}	PMI_{Assoc}	PMI_{Days}	PMI_{All}	$TFIDF_{Scr}$	$TFIDF_{Assoc}$	$TFIDF_{Days}$	$TFIDF_{All}$	IG_{Scr}	IG_{Assoc}	IG_{Days}	IG_{All}
PMI_{Scr}	—								abcd	abcd		
PMI_{Assoc}	abcd	—							abcd	abcd		
PMI_{Days}	abcd	abcd	—		abcd	abcd	ab	ab	abcd	abcd	abcd	abcd
PMI_{All}	abcd	abcd	abcd	—	abcd	abcd	ab	ab	abcd	abcd	abcd	abcd
$TFIDF_{Scr}$					—				abcd	abcd		
$TFIDF_{Assoc}$	cd				abcd	—			abcd	abcd		
$TFIDF_{Days}$	abcd	abcd			abcd	abcd	—		abcd	abcd	acd	cd
$TFIDF_{All}$	abcd	abcd			abcd	abcd	abcd	—	abcd	abcd	abcd	acd
IG_{Scr}									—			
IG_{Assoc}									abcd	—		
IG_{Days}	abcd	abcd			abcd	abcd			abcd	abcd	—	
IG_{All}	abcd	abcd			abcd	abcd			abcd	abcd	abcd	—

(i.e., SWN). All these improvements are statistically significant. In addition, the approximately 20,000 lexical entries that belong to the lexicons created in this work are included in more messages (only 0.5% of the posts are not classified). In contrast, the generic SWN lexicon, which contains a larger number of lexical items (117,000), presents a higher unclassification rate (5.1%). The financial lexicon (FIN) achieves the lowest F_{Avg}, CC1 and CC2 values, despite having the second highest P_{Bull}. The poorer FIN unclassified message performance (66%) confirms that there is a considerable difference between the lexical terms extracted from financial text documents and StockTwit messages. In effect, there are several popular StockTwits terms, such as "bearish", "bullish", "breakout", "put" and "short", that are not present in FIN. Since "bullish" and "bearish" are distinctive terms of stock market terminology, we verified their presence and classification in baseline lexicons. FIN and GI lexicons do not contain these terms and MSOL lexicon incorrectly classifies "bullish" as negative. The remaining lexicons assign the correct classification to these words.

Next, we analyze the differences between the selected large lexicon (PMI_{All}, 20550 items) and baseline lexicons (SWN, 117,000 entries). The lexicons are quite distinct, since only 2695 lexical terms (13% of PMI_{All}) belong to both lexicons. Indeed, the presence of diverse stock market terms in generic opinion lexicons is unlikely [38]. Also, 42% of these common terms (1121) have different sentiment polarities, as shown in Table 6. In particular, Table 6 presents examples of terms associated with: stock price changes (e.g., dip, downside, explosive, outperform, rip, sink);

Table 4: Classification results for the selected lexicon creation method and baseline lexicons (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P_{Bull}	R_{Bull}	$F1_{\text{Bull}}$	P_{Bear}	R_{Bear}	$F1_{\text{Bear}}$	F_{Avg}
Panel A: Evaluation results of holdout split method										
PMI$_{\text{All}}$	**78.8**	**0.5**	**79.2**	**86.0**	**85.5**	**85.8**	**59.7**	**59.5**	**59.6**	**72.7**
FIN	16.8	66.0	49.3	83.5	13.9	23.8	34.3	25.1	29.0	26.4
GI	37.7	25.8	50.8	82.5	36.2	50.3	37.5	42.1	39.7	45.0
MSOL	53.4	1.8	54.3	79.1	58.6	67.3	33.9	38.2	35.9	51.6
MPQA	36.9	37.5	59.0	80.6	40.6	54.0	34.3	26.3	29.8	41.9
OL	31.8	43.0	55.9	82.6	32.7	46.8	37.7	29.4	33.1	39.9
SWN	57.4	5.1	60.5	79.9	59.7	68.3	34.1	50.7	40.7	54.5
Panel B: Average evaluation results of rolling window method										
PMI$_{\text{All}}$	**77.5**	**0.5**	**77.8**	**87.3**	**81.7**	**84.4**	**54.6**	**64.2**	**58.7**	**71.5**
FIN	17.3	63.7	47.8	84.2	13.9	23.8	32.8	27.4	29.3	26.5
GI	37.4	25.8	50.4	82.6	36.1	50.2	35.8	41.1	37.8	44.0
MSOL	52.3	1.2	53.0	80.2	55.8	65.6	32.5	41.9	36.1	50.8
MPQA	39.1	34.7	59.8	81.3	42.6	55.8	35.2	28.4	31.1	43.5
OL	34.1	39.5	56.3	83.5	34.6	48.9	38.2	32.3	34.7	41.8
SWN	58.9	4.4	61.6	80.6	61.4	69.7	33.8	51.1	40.4	55.0

Table 5: Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of PMI$_{\text{All}}$ method with baseline lexicons. The following symbols denote significance at the 5% level: a - paired Student's t t-test for F_{Avg}; b - Wilcoxon signed rank test for F_{Avg}; c - paired Student's t t-test for $CC1$; d - Wilcoxon signed rank test for $CC1$. Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column

	FIN	GI	MSOL	MPQA	OL	SWN
PMI$_{\text{All}}$	abcd	abcd	abcd	abcd	abcd	abcd

stock expectations (e.g., overvalue, underestimate); and stock operations (e.g., long). Under non financial contexts, these terms can suggest different sentiment values. For example, "underestimate" is in general a negative verb but when related with stocks it can suggest an opportunity to buy. These differences highlight the importance of producing specialized stock market lexicons. For demonstration purposes, Figure 1 plots a word cloud of the most interesting PMI$_{\text{All}}$ bullish and bearish terms. Diverse terms with different sentiment polarity or absent from SWN stand out in this figure.

Table 6: Examples of lexical terms with different sentiment value in PMI$_{\text{All}}$ and SWN

Item	**POS tag**	**PMI$_{\text{All}}$**	**SWN**	**Item**	**POS tag**	**PMI$_{\text{All}}$**	**SWN**
careful	adjective	negative	positive	overvalue	verb	negative	positive
dip	noun	positive	negative	outperform	verb	positive	negative
rip	verb	positive	negative	downside	noun	negative	positive
sink	verb	negative	positive	explosive	adjective	positive	negative
long	adjective	positive	negative	underestimate	verb	positive	negative

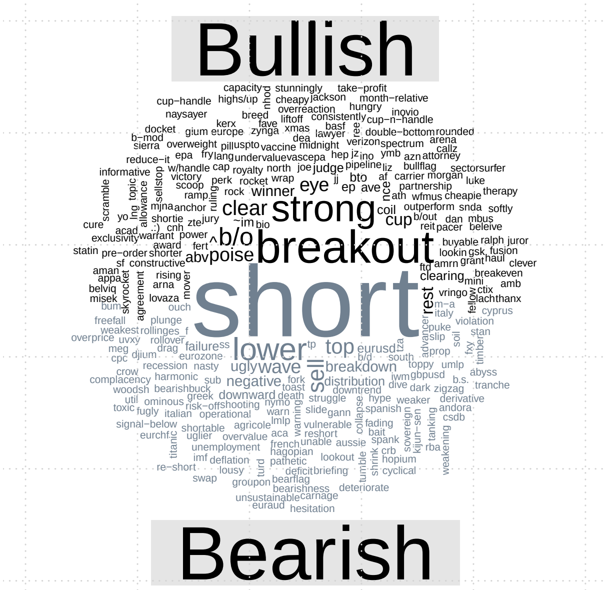

Figure 1: Bullish and bearish word cloud for a stock market lexicon (PMI_{All}).

The utility of affirmative and negated context scores is assessed by comparing PMI_{All}, $\text{TFIDF}_{\text{All}}$ and IG_{All} with the equivalent lexicons containing affirmative and negated context scores ($\text{PMI}_{\text{BiScr}}$, $\text{TFIDF}_{\text{BiScr}}$, IG_{BiScr}). Thus, all sentiment scores apply the complementary metrics (M_{assoc} and P_{days}). Tables 7 and 8 show the evaluation results.

Table 7: Classification results for unique and dual context scores (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P_{Bull}	R_{Bull}	F1_{Bull}	P_{Bear}	R_{Bear}	F1_{Bear}	F_{Avg}
Panel A: Evaluation results of holdout split method										
PMI_{All}	78.8	**0.5**	79.2	86.0	**85.5**	**85.8**	59.7	59.5	59.6	72.7
$\text{TFIDF}_{\text{All}}$	78.5	**0.5**	78.8	85.5	**85.5**	85.5	59.1	58.1	58.6	72.1
IG_{All}	76.4	**0.5**	76.7	86.9	80.4	83.5	54.1	**64.9**	59.0	71.3
$\text{PMI}_{\text{BiScr}}$	**79.0**	**0.5**	**79.3**	86.2	85.4	**85.8**	**59.8**	60.3	**60.1**	**73.0**
$\text{TFIDF}_{\text{BiScr}}$	78.5	**0.5**	78.9	86.0	85.1	85.5	59.0	59.6	59.3	72.4
IG_{BiScr}	76.7	**0.5**	77.0	**87.0**	80.8	83.8	54.7	64.8	59.3	71.5
Panel B: Average evaluation results of rolling window method										
PMI_{All}	77.5	**0.5**	77.8	87.3	81.7	84.4	54.6	64.2	**58.7**	71.5
$\text{TFIDF}_{\text{All}}$	77.3	**0.5**	77.7	86.5	82.7	84.5	54.5	60.4	56.9	70.7
IG_{All}	75.5	**0.5**	75.8	88.4	77.8	82.6	51.0	68.5	57.6	70.1
$\text{PMI}_{\text{BiScr}}$	**78.3**	**0.5**	**78.7**	86.4	**84.2**	**85.2**	**56.8**	60.0	58.1	**71.7**
$\text{TFIDF}_{\text{BiScr}}$	77.3	**0.5**	77.7	86.5	82.6	84.4	54.4	60.7	57.0	70.7
IG_{BiScr}	75.6	**0.5**	76.0	**88.6**	77.8	82.7	51.2	**69.2**	58.0	70.3

The application of affirmative and negated context scores appears to be beneficial. Despite the reduced difference, lexicons having two context scores improve or maintain almost all evaluation results compared to the unique score counterparts. IG_{BiScr} obtains statistically significant higher $CC1$ and F_{Avg} values than IG_{All} and $\text{PMI}_{\text{BiScr}}$ produces statistically significant higher $CC1$ values

Table 8: Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique and two context scores. The following symbols denote significance at the 5% level: a - paired Student's t t-test for F_{Avg}; b - Wilcoxon signed rank test for F_{Avg}; c - paired Student's t t-test for $CC1$; d - Wilcoxon signed rank test for $CC1$. Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column

	PMI_{All}	$\text{TFIDF}_{\text{All}}$	IG_{All}	$\text{PMI}_{\text{BiScr}}$	$\text{TFIDF}_{\text{BiScr}}$	IG_{BiScr}
PMI_{All}	—	ab	abcd		ab	abcd
$\text{TFIDF}_{\text{All}}$		—	acd			cd
IG_{All}			—			
$\text{PMI}_{\text{BiScr}}$	cd	abcd	abcd	—	abcd	abcd
$\text{TFIDF}_{\text{BiScr}}$			acd		—	cd
IG_{BiScr}			abcd			—

than PMI_{All}. Moreover, $\text{PMI}_{\text{BiScr}}$ lexicon has statistically significant higher $CC1$ and F_{Avg} values than all IG and TFIDF lexicons. Figure 2 shows the sentiment scores of all $\text{PMI}_{\text{BiScr}}$ items for both contexts. We can observe that the sentiment reversion in negation is not always appropriate. Indeed, only 41% have their sentiment orientation modified in negated contexts. Moreover, many items have much stronger sentiment value in negated contexts than in affirmative contexts and vice-versa. For example, the term bearish has a -6.634 score in affirmative contexts and just -0.794 in negated contexts. For instance, saying "not bearish" does not signify the same as being bullish. Usually it just means that the opinion is not pessimistic. In the opposite situation, bailout has -5.392 points for negated contexts and only -0.657 for affirmative segments. The refusal of a bailout may imply the business downfall while its application may not necessarily mean a successful future. Negation handling is not a straightforward procedure, it may vary according to each term. Thus, the use of two context scores may be very useful in this matter. The $\text{PMI}_{\text{BiScr}}$ lexicon created using the first 75% labeled messages is available at `https://github.com/nunomroliveira/stock_market_lexicon`.

4.2. Correlation with Survey Sentiment Indicators

To evaluate the relevance of microblogging sentiment indicators created using the stock market lexicon, we assess the association between Twitter sentiment indicators and two popular survey sentiment indicators: AAII and II. A strong correlation may indicate that the microblogging sentiment indicator can be an acceptable alternative or proxy. The correlation calculation uses 112 observations for AAII and 110 observations for II. Table 9 presents the respective Pearson's correlation values.

The obtained results show that Twitter sentiment indicators have a statistical significant moderate correlation with diverse survey sentiment values. Indeed, only the bullish value of AAII is poorly correlated with both Twitter sentiment indicators. II indicators present higher correlation

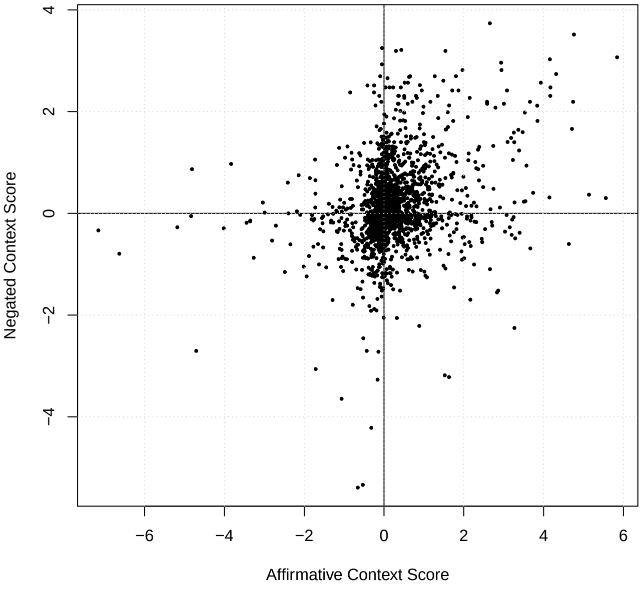

Figure 2: Distribution of sentiment scores for affirmative and negated contexts

Table 9: Pearson's correlation values of Twitter sentiment indicators with survey sentiment indicators ($\star$ – p-value < 0.01, $\diamond$ – p-value < 0.05, best correlation values for each survey sentiment value in **bold**)

Pair	Correlation	Pair	Correlation
($\text{TWTSML}_{\text{bear}}$, $\text{AAII}_{\text{bear}}$)	**0.489***	($\text{TWTSWN}_{\text{bear}}$, $\text{AAII}_{\text{bear}}$)	0.436*
($\text{TWTSML}_{\text{bull}}$, $\text{AAII}_{\text{bull}}$)	**0.233**$^\diamond$	($\text{TWTSWN}_{\text{bull}}$, $\text{AAII}_{\text{bull}}$)	0.220$^\diamond$
($\text{TWTSML}_{\text{spread}}$, $\text{AAII}_{\text{spread}}$)	**0.376***	($\text{TWTSWN}_{\text{spread}}$, $\text{AAII}_{\text{spread}}$)	0.342*
($\text{TWTSML}_{\text{bear}}$, II_{bear})	0.540*	($\text{TWTSWN}_{\text{bear}}$, II_{bear})	**0.628***
($\text{TWTSML}_{\text{bull}}$, II_{bull})	**0.533***	($\text{TWTSWN}_{\text{bull}}$, II_{bull})	0.445*
($\text{TWTSML}_{\text{spread}}$, $\text{II}_{\text{spread}}$)	**0.585***	($\text{TWTSWN}_{\text{spread}}$, $\text{II}_{\text{spread}}$)	0.551*

values than AAII, so it may indicate that Twitter users posting about stock market are informed traders because II is more associated to professional investors than AAII [51]. Moreover, sentiment indicators produced with the selected stock market lexicon (SML, i.e., $\text{PMI}_{\text{BiScr}}$) are more correlated to almost all survey sentiment values than indicators created with the selected baseline lexicon (i.e., SWN). Only II_{bear} is less correlated with TWTSML values than with TWTSWN values.

Sentiment indicators created using automated computational methods have various advantages regarding the traditional sentiment indicators produced from surveys. For instance, the creation of these sentiment indicators is faster and cheaper, permits higher frequencies (e.g., daily) and may be targeted to a more restricted set of stocks (e.g., stock market indices or individual stocks). Therefore, the application of SA in microblogging data may constitute a valuable alternative to the creation of investor sentiment indicators. Additionally, the utilization of stock market lexicons allows an easy and fast unsupervised production of these indicators.

5. Conclusions

With the expansion of social media (e.g., Twitter, message boards), the interest in sentiment analysis (SA) as increased, allowing the summary of opinions from large amounts of opinionated messages and thus support decision-making in several domains, including stock markets. A sentiment lexicon is a crucial resource for SA, enabling an easy and fast unsupervised SA and avoiding the expensive and arduous task of manually labeling data. Moreover, opinion lexicons permit the creation of very informative features for supervised SA. However, there are very few financial lexicons (e.g., [15, 16]) and the existing domain independent lexicons (e.g., [17, 18, 19]) may not be adjusted to the stock market domain.

In this paper, we propose an automated and fast approach to create stock market lexicons for microbloging messages. We employed a large labeled data set of StockTwits messages and tested three adaptations and two novel statistical measures to calculate the sentiment score. Also, we suggest the use of sentiment scores for affirmative and negated contexts in order to improve the difficult task of negation processing. The results on the test data confirmed that these newly created lexicons substantially increase the SA when compared with six reference lexicons. The improvements in evaluation metrics obtained by the created lexicons are statistically significant. Furthermore, the use of the proposed complementary metrics proved to be useful. Lexicons applying any of these measures obtain statistically higher evaluation results in SA than their counterparts that do not use the complementary metrics. Moreover, the utilization of affirmative and negated context scores appears to be beneficial. Lexicons applying these measures improve or maintain almost all evaluation results compared to their counterparts. Some of these improvements are statistically significant. A substantial contribution of this work is to make publicly available a large stock market lexicon with context scores. This is accessable at: `https://github.com/nunomroliveira/stock_market_lexicon`.

Also in this work, we selected a stock market lexicon (SML, i.e., $\mathrm{PMI}_{\mathrm{BiScr}}$) and a baseline lexicon (SWN) to easily generate investor sentiment indicators from Twitter messages holding cashtags of stocks traded in US markets. Twitter based sentiment indicators showed a significant moderate Pearson's correlation with the widely applied AAII and II survey sentiment indicators. Therefore, the Twitter based sentiment indicator can be used as an acceptable proxy for survey sentiment indicators. Moreover, the sentiment indicators created with the proposed lexicon showed higher correlations values than indicators produced with the baseline lexicon in five of the six analyzed survey indicators. A microblogging sentiment indicator presents several advantages when compared with survey sentiment indicators: it is faster and cheaper to produce, it allows higher frequencies (e.g., daily) and it can be adjusted to both stock market indices and individual stocks.

The proposed procedure allows the fast and effortless creation of a lexicon properly adapted to stock market contents. This lexicon may permit an easy and effective unsupervised SA related to the stock market domain, such as the creation of investor sentiment indicators. However, this process requires labeled stock market documents and such data sets are in short supply.

Our results suggest that the proposed microblogging sentiment lexicon approach might be a useful source of information for stock market participants, and this merits future research. For instance, a collective intelligence approach can be used to more easily assign sentiments to unlabeled text and identify stock market terms, thus widening the applicability of the proposed procedure to other stock market message sources (e.g., Twitter) and producing more accurate and comprehensive lexicons. Active learning algorithms may complement this approach by automatically selecting a reduced but more relevant set of text messages for human classification, thus reducing the manual labeling effort. Moreover, it is important to analyze the informative content of microblogging sentiment indicators to forecast stock market behavior. Sentiment indicators created by SA using stock market lexicons can be included in models to predict diverse stock market variables (e.g., returns, trading volume, volatility) in order to assess their predictive ability. In future work, we will also explore other text processing possibilities, such as the inclusion of exclamation points or question marks, which could be relevant in microblogs.

Acknowledgments

This work was supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013. We would like to thank the anonymous reviewers for their helpful suggestions. We also thank StockTwits for the provision of their data.

References

[1] A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments, Decision Support Systems 53 (4) (2012) 675–679.

[2] M. Hu, B. Liu, Mining and summarizing customer reviews, Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04 04 (2) (2004) 168.

[3] S. Kiritchenko, X. Zhu, S. Mohammad, Sentiment Analysis of Short Informal Texts, Journal of Artificial Intelligence Research 50 (2014) 723–762.

[4] H. Saif, Y. He, H. Alani, Semantic sentiment analysis of twitter, in: The Semantic Web–ISWC 2012, Springer, 2012, pp. 508–524.

[5] N. F. da Silva, E. R. Hruschka, E. R. Hruschka, Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 (2014) 170–179.

[6] E. Fersini, E. Messina, F. Pozzi, Sentiment analysis: Bayesian ensemble learning, Decision Support Systems 68 (2014) 26–38.

[7] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, V. Stoyanov, Semeval-2015 task 10: Sentiment analysis in twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, 2015.

[8] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (4) (2013) 82–89.

[9] W. Antweiler, M. Z. Frank, Is all that talk just noise? the information content of internet stock message boards, The Journal of Finance 59 (3) (2004) 1259–1294.

[10] R. P. Schumaker, Y. Zhang, C.-N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems 53 (3) (2012) 458–464.

[11] R. P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: The azfin text system, ACM Transactions on Information Systems (TOIS) 27 (2) (2009) 12.

[12] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, Decision Support Systems 55 (4) (2013) 919–926.

[13] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science 2 (1) (2011) 1–8.

[14] N. Oliveira, P. Cortez, N. Areal, On the predictability of stock market behavior using stocktwits sentiment and posting volume, in: Progress in Artificial Intelligence, Vol. 8154 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 355–365.

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

[16] H. Mao, P. Gao, Y. Wang, J. Bollen, Automatic construction of financial semantic orientation lexicon from large-scale chinese news corpus, in: 7th Financial Risks International Forum, Institut Louis Bachelier, 2014.

[17] P. J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis, Vol. 08, MIT Press, 1966.

[18] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, OpinionFinder : A system for subjectivity analysis, October (October) (2005) 34–35.

[19] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, in: Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10, Vol. 0, European Language Resources Association (ELRA), 2010, pp. 2200–2204.

[20] S. Mohammad, C. Dunne, B. Dorr, Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 2 of EMNLP '09, Association for Computational Linguistics, 2009, pp. 599–608.

[21] P. D. Turney, M. L. Littman, Measuring praise and criticism: Inference of semantic orientation from association, ACM Trans. Inf. Syst. 21 (4) (2003) 315–346.

[22] V. Hatzivassiloglou, J. M. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity, in: Proceedings of the 18th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 2000, pp. 299–305.

[23] Y. Choi, C. Cardie, Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification, in: Proceedings of the 2009 Conference on Empirical Methods

in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics, 2009, pp. 590–598.

[24] I. Ounis, C. Macdonald, I. Soboroff, On the trec blog track., in: ICWSM, 2008.

[25] H. T. Dang, K. Owczarzak, Overview of the tac 2008 opinion question answering and summarization tasks, in: Proc. of the First Text Analysis Conference, 2008.

[26] A. Kennedy, D. Inkpen, Sentiment classification of movie reviews using contextual valence shifters, Computational intelligence 22 (2) (2006) 110–125.

[27] V. Hatzivassiloglou, K. McKeown, Predicting the semantic orientation of adjectives, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics pages (1997) 181.

[28] J. Wiebe, Learning subjective adjectives from corpora, in: AAAI/IAAI, 2000, pp. 735–740.

[29] D. Lin, Automatic retrieval and clustering of similar words, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2, Association for Computational Linguistics, 1998, pp. 768–774.

[30] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Computational Linguistics 37 (1) (2011) 9–27.

[31] J. Kamps, R. Mokken, M. Marx, M. De Rijke, Using WordNet to measure semantic orientation of adjectives, in: Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004, Vol. 4, Citeseer, 2004, pp. 1115–1118.

[32] S.-M. Kim, E. Hovy, Determining the sentiment of opinions, in: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 1367.

[33] A. Esuli, F. Sebastiani, Determining the semantic orientation of terms through gloss classification, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ACM, 2005, pp. 617–624.

[34] A. Neviarouskaya, H. Prendinger, M. Ishizuka, Sentiful: A lexicon for sentiment analysis, Affective Computing, IEEE Transactions on 2 (1) (2011) 22–36.

[35] H. Takamura, T. Inui, M. Okumura, Extracting semantic orientations of words using spin model, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 133–140.

[36] Y. Lu, M. Castellanos, U. Dayal, C. Zhai, Automatic construction of a context-aware sentiment lexicon: an optimization approach, in: Proceedings of the 20th international conference on World wide web, ACM, 2011, pp. 347–356.

[37] D. Tufiş, D. Ştefănescu, Experiments with a differential semantics annotation for wordnet 3.0, Decision Support Systems 53 (4) (2012) 695–703.

[38] D. E. O'Leary, Blog mining-review and extensions:"from each according to his opinion", Decision Support Systems 51 (4) (2011) 821–830.

[39] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems 62 (2014) 22–31.

[40] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).
URL `http://www.R-project.org/`

[41] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology NAACL 03 1 (June) (2003) 173–180.

[42] D. D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task, in: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '92, ACM, New York, NY, USA, 1992, pp. 37–50.

[43] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data (2004).

[44] M. Baker, J. Wurgler, Investor Sentiment in the Stock Market, Journal of Economic Perspectives 21 (2) (2007) 129–151.

[45] K. L. Fisher, M. Statman, Investor Sentiment and Stock Returns, Financial Analysts Journal 56 (2) (2000) 16–23.

[46] M. Baker, J. Wurgler, Investor sentiment and the cross-section of stock returns, The Journal of Finance 61 (4) (2006) 1645–1680.

[47] M. Baker, J. Wurgler, Y. Yuan, Global, local, and contagious investor sentiment, Journal of Financial Economics 104 (2) (2012) 272–287.

[48] C. Oh, O. R. L. Sheng, Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement, in: ICIS 2011 Proceedings, Shanghai, China, 2011.

[49] T. O. Sprenger, A. Tumasjan, P. G. Sandner, I. M. Welpe, Tweets and trades: the information content of stock microblogs, European Financial Management 20 (5) (2014) 926–957.

[50] M. Hagenau, M. Liebmann, D. Neumann, Automated news reading: Stock price prediction based on financial news using context-capturing features, Decision Support Systems 55 (3) (2013) 685–697.

[51] G. W. Brown, M. T. Cliff, Investor sentiment and the near-term stock market, Journal of Empirical Finance 11 (1) (2004) 1–27.

[52] R. Verma, G. Soydemir, The impact of individual and institutional investor sentiment on the market price of risk, The Quarterly Review of Economics and Finance 49 (3) (2009) 1129–1145.