Automatic Creation of Stock Market Lexicons for Sentiment Analysis Using StockTwits Data

Nuno Oliveira
ALGORITMI Centre
Dep. of Information Systems
University of Minho
Guimarães, Portugal
nunomroliveira@gmail.com

Paulo Cortez
ALGORITMI Centre
Dep. of Information Systems
University of Minho
Guimarães, Portugal
pcortez@dsi.uminho.pt

Nelson Areal
Department of Management
University of Minho
Braga, Portugal
nareal@eeg.uminho.pt

ABSTRACT

Sentiment analysis has been increasingly applied to the stock market domain. In particular, investor sentiment indicators can be used to model and predict stock market variables. In this context, the quality of the sentiment analysis is highly dependent of the opinion lexicon adopted. However, there is a lack of lexicons adjusted to microblogging stock market data. In this work, we propose an automatic procedure for the creation of such lexicon by exploring a large set of labeled messages from StockTwits, a popular financial microblogging service, and using four statistical measures: adaptations of the known TF-IDF, Information Gain, Class Percentage, and a newly proposed Weighted Class Probability. The obtained lexicons are competitive when compared with a set of six reference lexicons. Moreover, we verified that it is beneficial to use continuous sentiment scores instead of sentiment labels.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; H.3.1 [**Content Analysis and Indexing**]: Dictionaries; I.2.7 [**Natural Language Processing**]: Text analysis; I.5.4 [**Applications**]: Text Processing

General Terms

Economics, Algorithms, Experimentation

Keywords

Sentiment Analysis, Opinion Mining, Stock Market, Lexicon, Microblogging Data, Information Retrieval

1. INTRODUCTION

People's opinions inform diverse decision-making processes. For example, companies are interested in knowing consumers' opinions about their products in order to improve them and to make the best possible marketing decisions. In recent years, social media platforms, and in particular microblogging, have enabled an explosion of unstructured content with opinions regarding a huge variety of topics and have provided a valuable and inexpensive source for opinion analysis about organizations and individuals. However, the analysis of all these data is not feasable without the use of computation. The identification and summarization of important information from vast amounts of data is very challenging for the average person [8] and the current amount of social media data makes human processing impracticable. These limitations can be overcome by Opinion Mining (OM) systems that mine large amounts of opinionated contents and automatically extract and summarize the opinions about a topic [16].

An opinion lexicon is considered the most crucial resource for the majority of OM algorithms [5]. It is composed by a list of opinion words and their respective sentiment value (positive, negative, neutral or a sentiment score). The presence of opinion words in text permits discerning a sentiment orientation. For example, the sentence "the computer is good" can be easily classified as positive if the opinion lexicon contains the word "good" as positive. The utilization of lexicons allows the usage of an easy unsupervised classification of text, thus avoiding the often laborious task of manually labeling training data.

Due to the interest in this area, several opinion lexicons have been proposed for sentiment classification (e.g., [2, 20, 24]). However, these resources may not be appropriate for specific domain contexts and types of messages used. For example, the word "long" can have many sentiment orientations within the Financial domain (e.g., "long debt list", "long Google stocks"). Moreover, due to the small size of microblogging messages (maximum 140 characters), tweets are written differently from larger texts; they are often more direct and resort to abbreviations. Thus, the use of a specialized lexicon, adapted to a particular topic domain and social medium, should potentially lead to better results.

With the rise of Web 2.0 and social media platforms, there has been a recent trend to use OM for forecasting stock market behavior (e.g., [1, 3, 21]). Some authors argue that emotion and mood can impact financial decisions [12]. Thus, investor sentiment can be influential in financial decision-making and their measures can be used to model and predict stock market variables [1, 21, 19, 3, 13, 15, 14]. Despite this interest, the effort in designing opinion lexicons adapted to financial terminology is scant. Loughran and McDon-

IDEAS'14 , July 07 - 09 2014, Porto, Portugal

ACM 978-1-4503-2627-8/14/07 $15.00.
http://dx.doi.org/10.1145/2628194.2628235

ald [10] manually created six word lists utilized in financial documents retrieved from the U.S. Securities and Exchange Commission web site over the years 1994 to 2008. However, such lexicon was not targeted to microblogging text and repeating a similar manual procedure is prohibitive due to the huge laborious effort involved.

In this paper, we propose an alternative and automatic procedure to create a stock market lexicon for microblogging data. In particular, we address the StockTwits social service, which is exclusively dedicated to stock market conversations and used by investors with different levels of expertise. StockTwits has recently implemented an interesting feature that was exploited in this work, where users can classify their own messages as "bullish" or "bearish" providing a valuable and extensive labeled data set for the production of a lexicon. Bullish and bearish are common stock market terms meaning that investors are optimistic or pessimistic respectively. The collected large StockTwits labeled dataset is analyzed by applying four statistical measures: Term Frequency-Inverse Document Frequency (TF-IDF), Information Gain, Class Percentage, and a newly proposed Weighted Class Probability. These measures are adapted in this work to select words with a sentiment value. Finally, the resulting lexicons are compared with a set of six reference lexical resources.

This paper is organized as follows. Section 2 introduces the related work. Section 3 describes the data and methods used in this study. Section 4 presents and discusses the research results. Finally, Section 5 concludes with a summary and discussion of the main results.

2. RELATED WORK

There are diverse methods to create an opinion lexicon. The most labor intensive approach employs a group of experts to manually assign the sentiment value to each lexical entry. General Inquirer [20] and MPQA subjectivity lexicons [24] are examples of two important lexical resources that apply this methodology.

The automatic creation permits including more lexical items with much less human effort at the expense of accuracy. There are various papers describing the automatic construction of lexicons, usually exploring text corpora or dictionaries. Hatzivassiloglou and McKeown [7] extracted adjectives that are associated by conjunctions, such as "and" and "or", with words whose polarity is already known. The sentiment orientation is propagated through these connected words. Turney and Littman [23] infered the semantic orientation of each word from the Pointwise Mutual Information (PMI) with each seed positive and negative term. This work makes use of large text corpora, comprising approximately 350 million web-pages. Kamps et al. [9] extracted adjectives and assigned their sentiment orientation based on the synonymy shortest path on WordNet to the words "bad" and "good". Hu and Liu [8] utilized a large corpus of customer reviews to extract opinion words related to frequent product features. First, Hu and Liu collected all adjectives that appeared in sentences containing frequent product features. Then, they applied Wordnet (`wordnet.princeton.edu`) to determine the sentiment polarity of the opinion words. Terms assumed the sentiment orientation of synonyms or the inverse polarity of antonyms included in a seed set of adjectives with already known semantic orientation. The seed list was continuously expanded with these newly extracted words and the process continued until no further words had synonyms or antonyms in the seed list. Esuli and Sebastiani [4] determined the orientation of subjective terms by applying text classification techniques to term representations of the textual definitions (i.e., glosses) obtained on on-line dictionaries. Mohammad et al. [11] created a broad lexicon using a set of affix patterns and a Roget-like thesaurus. First, a seed set of positive and negative words was collected using a group of affix patterns. Then, those words were applied to classify the sentiment polarity of the terms composing the diverse sets of near-synonymous words of the thesaurus. Baccianella et al. [2] created the SentiWordNet lexicon by automatically assigning sentiment scores to all WordNet synsets. First, they used glosses of a seed set of synsets to train a group of classifiers that subsequently were used to classify all synsets. Then, the diverse classification results were combined to produce a sentiment value to each synset. In a second step, two different iterative random-walk processes were performed for the positivity and negativity measures. These processes used a graph containing directed links from synsets appearing in glosses of other synsets. The random walk step started with the values produced in the previous phase and concluded when the processes had converged. More recently, Qiu et al. [17] proposed a method based on bootstrapping to create an opinion lexicon and extract the opinion targets. The process used an initial seed set of opinion words to iteratively collect further opinion words and targets using a group of syntactic relations linking these items. Additional polarity assignment and target pruning methods were applied to refine results.

In the financial area there has been scarce research on creating opinion lexicons. In the best known work in this topic, Loughran and McDonald [10] analyzed the words that appeared in at least 5% of a large sample of 10-K documents (from the U.S. Securities and Exchange Commission web site) during the years of 1994 to 2008 and produced six word lists corresponding to the sentiments: negative, positive, uncertainty, litigious, modal strong and modal weak.

We argue that many of these methods do not fit our goal. The dictionary-based methods create lexicons that are not sufficiently adjusted to our subject because dictionaries are domain independent. Collocation measures (e.g., PMI) are less effective in microblogging data considering that these messages are very short. The utilization of a reduced set of syntactic relations (e.g., conjunctions) do not allow the extraction of a comprehensive set of words. Additionally, many corpora-based methods only extract adjectives, discarding words with obvious sentiment value (e.g., verbs "love" and "hate"). Therefore, we propose to automatically create a stock market lexicon by exploring a wide data set of classified microblogging data and four statistical measures, as explained in the next section.

3. MATERIALS AND METHODS

3.1 StockTwits Data

Microblogging data has distinguishing characteristics that may benefit the creation of investor sentiment indicators. Investors are increasingly using microblogging services to express their opinion and to share useful information regarding several financial topics. The number of character constraints requires greater objectivity from the author and permits a more accurate linguistic analysis. Users post very

frequently, reacting to events in real-time. This regularity allows a real-time sentiment assessment that can be exploited during the trading day. Additionally, microblogging data is usually abundant and readily available at low cost permitting the creation of sentiment indicators in a more rapid and inexpensive manner than traditional forms (e.g., large-scale surveys). The investor community usually apply a specific term (cashtag) in microblogging conversations related to the respective stock. Cashtags are composed by the stock ticker preceded by the "$" symbol (e.g., $IBM, $GOOG). Concentrating on these messages reduces the amount of irrelevant data.

StockTwits (`stocktwits.com`) is a microblogging platform exclusively dedicated to stock market with more than 200,000 users. Messages are limited to 140 characters and consist of ideas, links, charts and other data. Very recently, since September 2012, users are able to classify their own text messages as "bullish" or "bearish". This data set constitutes a valuable asset to train sentiment analysis methods or to create a stock market lexicon. In this work, we explore StockTwits labeled messages from September 2012 until March 31, 2013. The dimension of this data set, approximately 350,000 messages, is much higher than the classified data sets applied in most studies about mining microblogging data in the finance domain. Moreover, StockTwits data is less noisy when compared with generalist microblogging services (e.g., Twitter) because it is dedicated to the stock market.

3.2 Lexicon Creation

We executed various pre-processing operations on the microblogging data utilized in this study. Using regular expressions in the R tool [18], we performed the following tasks:

- exclude messages composed by charts because they have reduced textual content;
- substitute HTML characters and some contractions, remove URL links, and process punctuation to permit a more effective parsing;
- replace all cashtags by a single tag to prevent some cashtags from having prior sentiment polarity due to its performance in the analyzed period; and
- substitute all numbers by a single tag because the set of referred numbers is very wide.

Then, we performed tokenization, Part of Speech (POS) tagging and lemmatization using Stanford CoreNLP [22]. Additionally, we substituted the POS tags of a small set of words that were frequently mislabeled. For example, the terms "calls", "puts", "futs", "bears" were frequently tagged as verbs when they should be considered nouns.

We applied a time ordered holdout split validation scheme, where the first 75% of StockTwits labeled messages were used as a training set, to create the stock market lexicons, and the remaining and most recent 25% messages were used as a test set, for evaluation purposes. The labeled messages are very unbalanced, with around 75% being bullish and 25% bearish. To avoid a biased sentiment assessment towards the dominant class, we used a random sample of bullish messages in order to select the same number of bearish messages (undersampling method). Approximately 100,000 messages were used in the production of lexicons, after executing all previously described data operations.

Considering the high dimension of training data (several thousand words and about 100,000 messages), we needed to reduce the number of analyzed lexical items. Thus, we included only those with more than 10 occurrences in all messages. This option had the advantage of removing a large number of orthographic errors. Also, we excluded words containing POS tags that we consider to have lower sentiment value (e.g., pronouns). Then, we tested different statistical measures to determine the informative value of lexical items and to classify them as "bullish" or "bearish". We experimented the following measures:

1. Term Frequency-Inverse Document Frequency (TF-IDF) based: TF-IDF is a common text mining statistic used to determine the importance of a term to classify a document. The value increases with the number of occurrences of the item in messages of a class (''bullish" or "bearish") but decreases with the frequency of the item in all messages. TF-IDF can be calculated as follows:

$$tf(l,c) = \frac{n_{c,l}}{n_M} \tag{1}$$

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

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

where l is the lexical item, $c \in \{c_1, c_2\}$ (where $c_1 =$ "bullish" and $c_2 =$ "bearish") is the class label, $n_{c,l}$ is the number of occurrences of the lexical item l in class c, n_M is the total number of lexical items in all messages and N_l is the number of classes that contains the lexical item l. We applied function *tfidf* of the R package *textir* to retrieve $tf\text{-}idf(l,c)$ values.

In this study, for each lexical item, we use the difference between the TF-IDF value for bullish and bearish classes, in order to have a single value that reflects the tendency to a sentiment class. Thus, the continuous sentiment score $S_{\text{TF-IDF}}$ is given by:

$$S_{\text{TF-IDF}}(l) = tf\text{-}idf(l,c_1) - tf\text{-}idf(l,c_2) \tag{4}$$

The assigned sentiment label is: "bullish" if $S_{\text{TF-IDF}}(l) > 0$; "bearish" if $S_{\text{TF-IDF}}(l) < 0$; else it is "neutral".

2. Information Gain (IG): IG is a statistical measure often applied in Information Theory to determine the informative value of a variable. In this case, IG is used to measure the relevance of lexical item l:

$$\begin{aligned} IG(l) = &- \sum_{i\in\{1,2\}} p(c_i)\log p(c_i) \\ &+ p(l) \sum_{i\in\{1,2\}} p(c_i|l)\log p(c_i|l) \\ &+ p(\bar{t}) \sum_{i\in\{1,2\}} p(c_i|\bar{l})\log p(c_i|\bar{l}) \end{aligned} \tag{5}$$

where $p(c_i) = \frac{n_c}{n}$ is the probability for class c_i (n_c is the number of messages of class c_i and n is the total number of messages), $p(l)$ is the probability for lexical item l (computed similarly to $p(c_i)$), $p(c_i|l)$ is the probability of class c_i for documents with lexical item

l and $\bar{l}$ denotes the set of documents without lexical item l. We calculated these values using function *information.gain* of the R package *FSelector*.

IG values inform about term relevancy but not about the sentiment orientation. Thus, the sentiment label is obtained by checking what is the most common class within messages that contain the lexical term l. And the continuous sentiment score is computed as $S_{\text{IG}} = IG(l)$, if the sentiment label is c_1 ("bullish"); or $S_{\text{IG}} = -IG(l)$, if the sentiment label is c_2 ("bearish").

3. Class Percentage (CP): CP is defined as the percentage of the number of occurrences of a lexical item in messages of a class relative to the total number of occurrences of that item in all messages. It is computed as follows:

$$cp(l, c) = \frac{n_{c,l}}{n_l} \tag{6}$$

where n_l is the number of occurrences of the lexical item l in all messages. The sentiment label is given by $\arg\max(cp(l, c_i))_{c_i \in \{c_1, c_2\}}$, i.e., the class that produces the highest $cp(l, c)$ value. Similarly to IG, the continuous sentiment score is computed as $S_{\text{CP}} = cp(l, c_1)$, if the sentiment label is "bullish" or $S_{\text{CP}} = -cp(l, c_2)$, if the sentiment label is "bearish".

4. Weighted Class Probability (WCP): in this study we propose this measure intending to increase the weight of more frequent items. The formula is as follows:

$$wcp(l) = \frac{2n_{c_1,l} - n_l}{n_l} \times \log(1 + |2n_{c_1,l} - n_l|) \tag{7}$$

The continuous sentiment score is $S_{\text{WCP}} = wcp(l)$. Similarly to TF-IDF, the assigned sentiment label is: "bullish" if $S_{\text{WCP}}(l) > 0$; "bearish" if $S_{\text{WCP}}(l) < 0$; else it is "neutral".

We applied these statistical measures to two types of lexical items:

- unigrams, composed by all individual words and the respective POS tag; and
- bigrams, corresponding to all two contiguous words.

For each metric, we created one lexicon composed by unigrams and the bigrams with higher informative value than their words individually. Finally, all lexical items were sorted in decreasing order by the absolute continuous sentiment score. This ordering is useful when we need to select subsets of items from the full lexicon.

3.3 Evaluation

To assess the relevance of the created lexicons, we compared the results of sentiment analysis performed on the test set using these lexicons and the following lexical resources:

- Harvard General Inquirer (GI) [20] - This resource comprises more than 11,000 words classified in 182 categories. These categories come from four sources: the Harvard IV-4 dictionary; the Lasswell value dictionary; categories recently constructed, and "marker" categories containing syntactic and semantic markers.

  We exploited this resource by utilizing all words of the "positive" and "negative" categories.

- Opinion Lexicon (OL) [8] - This lexicon contains two lists of positive and negative opinion words for English, including misspelled words that appear frequently in social media contents. It comprises nearly 6,000 words.
- Macquarie Semantic Orientation Lexicon (MSOL) [11] - This resource classifies more than 75,000 n-grams, as positive or negative.
- MPQA Subjectivity Lexicon (MPQA) [24] - this lexicon is part of OpinionFinder, a system that identifies various aspects of subjectivity (e.g., sources of opinion, sentiment expressions) in text. MPQA Subjectivity Lexicon has more than 8000 entries and contains attributes such as POS tag, prior polarity and subjectivity type. In the latter attribute, a word is classified as *strongsubj* if it is subjective in most contexts and it is considered *weaksubj* if it only has certain subjective usages.
- SentiWordNet (SWN) 3.0 [2] - it is a lexical resource that assigns continuous sentiment scores to each synset of WordNet [6]. A synset is a group of words or expressions that are semantically equivalent in some context. SentiWordNet has more than 117,000 entries, corresponding to the number of WordNet synsets. Each word may have multiple scores because it can belong to diverse synsets of Wordnet. In this paper, we used the average positivity and negativity score for each pair (word, POS tag) because we did not disambiguate the various synsets.
- Financial Sentiment Dictionaries (FIN) [10] - it contains 6 word lists commonly applied in financial text documents. The lists are: negative (2349 words), positive (354 words), uncertainty (297 words), litigious (886 words), modal strong (19 words), modal weak (26 words). In this study we only utilized the negative and positive word lists because the other lists do not have a clear sentiment polarity.

We used a bag of words approach in the sentiment analysis performed to evaluate all baseline lexicons as well as the created lexicons. A message is considered to be: "bullish" – if the total sentiment value of all lexical items is positive; or "bearish" – if the total is negative; and "neutral" – if the total is zero. When processing sentiment labels, we substitute positive or bullish labels by 1 and negative or bearish labels by -1. In MPQA lexicon, we also assigned half of the sentiment score to *weaksubj* words.

The evaluation measures applied in this study were:

- Global classification accuracy (Acc1): percentage of messages correctly classified when considering the full data set;
- Unclassified messages (Uncl): percentage of messages that do not contain any lexicon entry; and
- Classification accuracy (Acc2): similar to Acc1 except that the unclassified messages are not considered in the computation.
- For each label, bullish and bearish, we calculated:
  - Precision (Pre_{Bull}, Pre_{Bear}) – measures the proportion of true positives relative to the number of true and false positives;

– Recall (Rec_{Bull}, Rec_{Bear}) – measures the proportion of true positives relative to the number of true positives and false negatives (also known as Sensitivity); and

– $F_1 = 2\frac{Precision*Recall}{Precision+Recall}$ score (F1_{Bull}, F1_{Bear}) – which considers both Precision and Recall under a single measure.

4. RESULTS

In this section we present the results of sentiment analysis on the test set (last 25% of all messages) using different lexicons. We applied the baseline lexicons GI, OL, MSOL, MPQA, SWN and FIN and the created new lexicons using all four statistical measures. Regarding the latter lexicons, we tested them with 10 different dimension sizes, with 100%, 90%, 80%, ..., 20% and 10% of its total lexical items, discarding those with inferior absolute sentiment score. The intention is to verify which lexical items are unimportant for sentiment classification and should be excluded from the lexicon.

We start the analysis by considering the reference lexicons (Table 1). Considering all evaluation metrics, the best results are in general obtained by SWN. When compared with MSOL (second best Acc1 value), this lexicon gets higher Acc1 and Acc2 values. Also, the precision and recall values are better for SWN except for Pre_{Bear}, where both lexicons achieve the same result. This SWN performance can be explained by its greater scope, as it includes more than 117,000 lexical items, and also by the adopted continuous sentiment scores. The financial lexicon (FIN) produces the highest Pre_{Bull}. However, it also achieves the lowest Acc1 and Acc2 values. The lack of coverage of this lexicon, which presents the highest number of unclassified messages (66%), can be explained by the different nature of the 10-K documents when compared with the StockTwits messages. For instance, there are several StockTwits frequent terms, such as "bearish", "bullish", "short", "put" and "breakout", that do not appear in FIN.

Next, we analyze the created lexicons. Table 2 shows all evaluation metrics for the best baseline lexicon, FIN, overall best Acc1 lexicon (full WCP lexicon) and overall best Acc2 lexicon (top 10% CP lexicon). Lexicons created in this work clearly outperform the baseline lexicons. Sentiment classification using continuous sentiment scores of full WCP lexicon produced 77% accuracy for both Acc1 and Acc2, resulting in a 17 (Acc2) and 20 (Acc1) percentage point difference when compared with SWN. Moreover, the 8172 lexical items that compose the WCP lexicon are present in more messages (only 1% of the messages are not classified). While using a much higher number of lexical items (117,000), SWN lexicon cannot classify a larger portion of messages (6%). The top 10% of items of CP lexicon produces higher levels of precision. Sentiment scores of these unigrams and bigrams permit correctly classifying 82% of messages containing them, with 93% precision in positive messages and 71% in negative messages. However, this resource has low recall values.

Figures 1 and 2 compare all created lexicons by presenting the Acc1 and Acc2 values using the sentiment labels and continuous sentiment scores. One important outcome is that the usage of continuous sentiment scores appears to be beneficial. They achieve better results than sentiment labels in almost every evaluation measure. These values correspond to the informativeness of each item. Therefore, those lexical items more correlated to a specific class will have higher scores than items less associated. This differentiation seems to improve sentiment classification.

Regarding the four statistical measures tested, IG presents constant results that do not exceed 62% for Acc1 and 71% for Acc2 because only 7% of lexical items have IG coefficients different than zero. TFIDF also produces results with small variations. Top 30% TFIDF items present very similar results when compared with the full TFIDF lexicon. Lexicon dimension is more significant for WCP and CP measures. Larger lexicons obtain higher Acc1 results, so the utilization of less valued items seems to be useful for sentiment classification. Nevertheless, sentiment scores of top 80% WCP lexicon produces almost identical Acc1 values to the full WCP lexicon and substantially higher than other measures. The greater range of values obtained by this measure may explain its superiority when applying continuous sentiment scores. For demonstration purposes, Figure 3 shows a word cloud of the most WCP valued words. The CP statistical measure is particularly good for creating lexical resources with higher precision. Lexicons composed by the 10% and 20% most CP valued lexical items have significantly higher Acc2 values for both sentiment labels and sentiment scores. Frequent items are less weighted using the CP measure than applying other measures. Thus, it may indicate that some infrequent words have great informational value.

To better understand the difference in performance between lexicons produced in this work and the baseline lexicons, we particularly detail the differences between the lexicons WCP and SWN. These are the lexical resources that achieved the best Acc1 values among their counterparts, the created and baseline lexicons, respectively. The WCP lexicon has 8172 items from which only 1885 belong to the SWN. Some common stock market words and bigrams not included in SWN are shown in Table 3. Moreover, 733 lexical items have different sentiment polarities in both lexicons. Table 4 presents some examples of items with different sentiment orientation. These words are usually associated with stock performance. Some words are related to stock price movements (e.g., downside, rip, dip, explosive, sink, jump, explode, outperform), others are associated to expectations and considerations about stocks (e.g., underestimate, overvalue, cheap, exhaustion, caution, cautious, careful, steady) and other items refer to stock operations (e.g., long, pick). Thus, these words may transmit different sentiments than they would convey in everyday situations. For instance, "explosive" is usually related to accidents but in stock market conversations, it is associated with huge rises in stock prices. Also, the verb "underestimate" generally has negative feelings but underestimated stocks means a good chance to buy and profit later.

Figure 4 compares the number of occurrences of four types of lexical items:

- items included in both lexicons but with different polarities (Df_sent);
- items that have identical sentiment polarity in SWN and WCP (Eq_sent);
- SWN items not present in WCP (SWN_sent); and
- WCP items not present in SWN (WCP_sent).

Table 1: Sentiment analysis results using baseline lexicons (in %, best values in bold)

Lexicon	Acc1	Uncl	Acc2	Pre_{Bull}	Rec_{Bull}	$F1_{Bull}$	Pre_{Bear}	Rec_{Bear}	$F1_{Bear}$
FIN	17	66	49	**83**	14	24	35	26	30
GI	38	26	51	82	36	50	**38**	42	40
MSOL	53	**2**	54	79	58	67	34	39	36
MPQA	37	38	59	80	40	53	35	27	30
OL	32	43	56	82	32	46	**38**	29	33
SWN	**57**	6	**60**	80	**59**	**68**	34	**51**	**41**

Table 2: Comparison of two baseline lexicons and two created lexicons (in %, best values in bold)

Lexicon	Acc1	Uncl	Acc2	Pre_{Bull}	Rec_{Bull}	$F1_{Bull}$	Pre_{Bear}	Rec_{Bear}	$F1_{Bear}$
FIN	17	66	49	83	14	24	35	26	30
SWN	57	6	60	80	59	68	34	51	41
100% WCP	**77**	**1**	77	86	**82**	**84**	56	**63**	**59**
10% CP	17	80	**82**	**93**	13	23	**71**	27	39

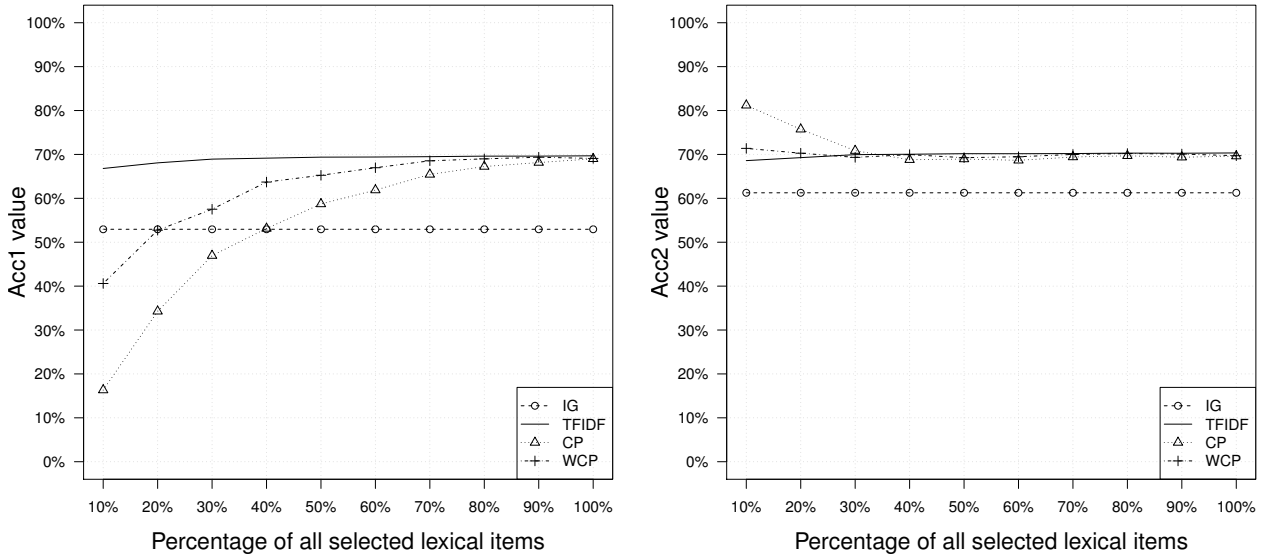

Figure 1: Comparison of created lexicons using sentiment labels (left – Acc1 values; right – Acc2 values).

Table 3: Examples of common lexical terms not included in SWN

Lexical terms
recession, bear flag, watch list, addition, sell signal, negative divergence, rise wedge, open sell, break out, below, break down, breakout, new low, current stop, double top, upside, look weak, good volume

The 6287 items that constitute WCP_sent have a substantially higher number of occurrences than the 100,000 items composing SWN_sent. Moreover, the 733 lexical items with distinct sentiment polarity in both lexicons present a significant number of occurrences. Therefore, domain independent lexicons may not be adjusted to classify stock market sentiment because they do not contain influent lexical items and do not label them correctly. The utilization of stock market lexicons should minimize these errors.

5. CONCLUSIONS

An opinion lexicon is a crucial and useful resource that can be employed to perform unsupervised sentiment classification and avoid the exhaustive task of manually labeling data. Although there is a set of large and popular opinion lexicons (e.g., [2, 20, 24]), these resources are generally domain independent. Thus, they may not be adapted to the specific terminology and semantics of stock market contents. However, sentiment analysis applied to social media

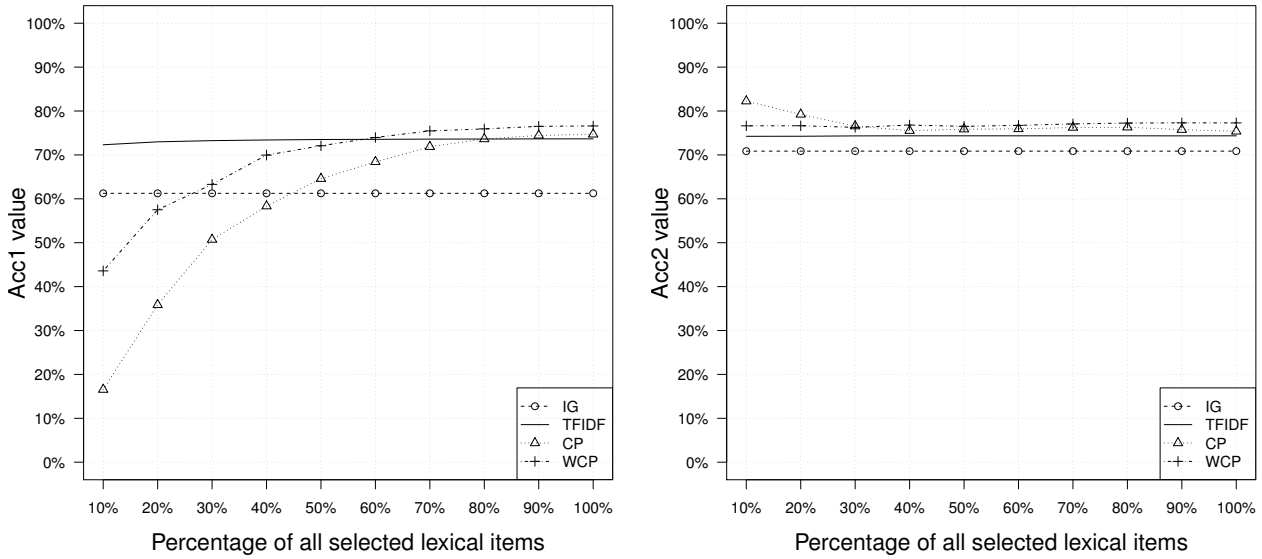

Figure 2: Comparison of created lexicons using continuous sentiment scores (left – Acc1 values; right – Acc2 values).

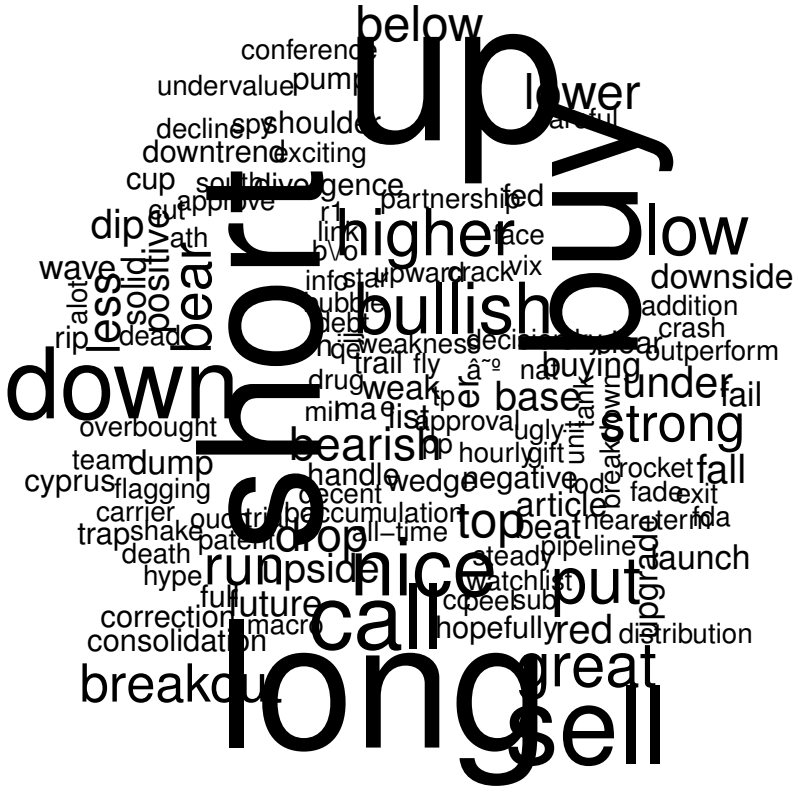

Figure 3: Word cloud of WCP lexicon.

Table 4: Lexical items with different sentiment polarity

Item	POS tag	WCP	SWN
downside	Noun	Negative	Positive
careful	Adjective	Negative	Positive
overvalue	Verb	Negative	Positive
dip	Noun	Positive	Negative
rip	Verb	Positive	Negative
long	Adjective	Positive	Negative
caution	Noun	Negative	Positive
steady	Adjective	Positive	Negative
exhaustion	Noun	Negative	Positive
explosive	Adjective	Positive	Negative
outperform	Verb	Positive	Negative
cautious	Adjective	Negative	Positive
sink	Verb	Negative	Positive
jump	Noun	Positive	Negative
pick	Verb	Positive	Negative
cheap	Adjective	Positive	Negative
explode	Verb	Positive	Negative
underestimate	Verb	Positive	Negative

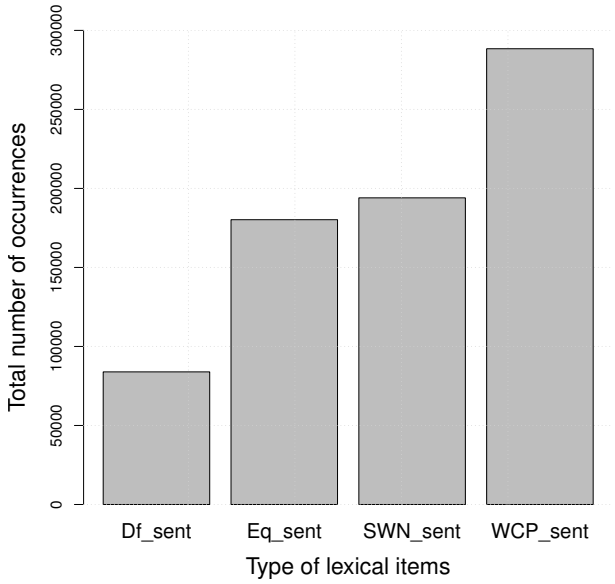

Figure 4: Number of occurrences of Df_sent, Eq_sent, SWN_sent and WCP_sent lexical items.

has become increasingly important in the stock market context. For instance, investor sentiment indicators have been applied to model and predict stock market variables [1, 3, 13, 19, 21].

The main purpose of this study was to create an opinion lexicon adjusted to the stock market domain. We explored a large data set of StockTwits labeled messages and proposed four statistical measures to select lexical items and assign them sentiment values. To verify the relevance of the created lexicons, we performed sentiment analysis on a test dataset using the created lexicons and compared these results against the ones obtained using six reference lexical resources.

Results show that lexicons created in this work permit significantly improving sentiment classification relative to baseline lexicons (e.g., with an improvement of 17 and 20 percentage points). For instance, full WCP lexicon achieves 77% accuracy, while the most accurate baseline lexicon (i.e., SWN) only obtains 57%. Despite the large number of lexical items that SWN contains, there is a considerable number of occurrences of items that are either not included or misclassified in this lexicon. Moreover, we confirmed that it is beneficial to use continuous sentiment scores instead of sentiment labels. Although the majority of reference lexicons applies sentiment labels, the utilization of sentiment scores produced the best results in this work. Therefore, we consider that the application of these stock market lexicons may contribute to improve and facilitate the creation of investor sentiment indicators applied to microblogging data. In the future, we intend to use the created lexicons to predict useful stock market variables.

6. ACKNOWLEDGMENTS

We wish to thank StockTwits for kindly providing their data. This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

7. REFERENCES

[1] W. Antweiler and M. Frank. Is All That Talk Just Noise? The Information Content of Interest Stock Message Boards. *Journal of Finance*, 59(3):1259, 2004.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, volume 0, pages 2200–2204. European Language Resources Association (ELRA), 2010.

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

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

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

[6] C. Fellbaum. *WordNet: An Electronic Lexical Database*, volume 71 of *Language, Speech, and Communication*. MIT Press, 1998.

[7] V. Hatzivassiloglou and 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:181, 1997.

[8] M. Hu and 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):168, 2004.

[9] J. Kamps, R. Mokken, M. Marx, and 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*, volume 4, pages 1115–1118. Citeseer, 2004.

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

[11] S. Mohammad, C. Dunne, and 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*, volume 2 of *EMNLP '09*, pages 599–608. Association for Computational Linguistics, 2009.

[12] J. R. Nofsinger. Social Mood and Financial Economics Social Mood and Financial Economics. *Journal of Behavioral Finance*, 6(3):144–160, 2005.

[13] C. Oh and 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.

[14] N. Oliveira, P. Cortez, and N. Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In L. Correia, L. P. Reis, and J. Cascalho, editors, *Progress in Artificial Intelligence - 16th Portuguese Conference on Artificial Intelligence (EPIA)*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer, 2013.

[15] N. Oliveira, P. Cortez, and N. Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In D. Camacho, R. Akerkar, and M. D. Rodríguez-Moreno, editors, *3rd International Conference on Web Intelligence, Mining and Semantics (WIMS '13)*, page 31. ACM, 2013.

[16] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(2):1–135, 2008.

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

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[19] T. O. Sprenger and I. M. Welpe. Tweets and Trades: The Information Content of Stock Microblogs. *Social Science Research Network Working Paper Series*, pages 1–89, 2010.

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

[21] P. C. Tetlock. Giving Content to Investor Sentiment : The Role of Media in the Stock Market. *Journal of Finance*, 62(3):1139–1168, 2007.

[22] K. Toutanova, D. Klein, C. D. Manning, and 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):173–180, 2003.

[23] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information System*, 21(4):315–346, 2003.

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