Article

JIS

Journal of Information Science
2025, Vol. 51(6) 1443–1456
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01655515231171361
journals.sagepub.com/home/jis
Sage

SA-MAIS: Hybrid automatic sentiment analyser for stock market

Bruno Taborda
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal; Centre for Informatics and Systems of the University of Coimbra (CISUC), Portugal

Ana Maria de Almeida
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Portugal; Centre for Informatics and Systems of the University of Coimbra (CISUC), Portugal

José Carlos Dias
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal; Business Research Unit (BRU-IUL), Portugal

Fernando Batista
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal; INESC-ID Lisboa, Portugal

Ricardo Ribeiro
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal; INESC-ID Lisboa, Portugal

Abstract

Sentiment analysis of stock-related tweets is a challenging task, not only due to the specificity of the domain but also because of the short nature of the texts. This work proposes SA-MAIS, a two-step lightweight methodology, specially adapted to perform sentiment analysis in domain-constrained short-text messages. To tackle the issue of domain specificity, based on word frequency, the most relevant words are automatically extracted from the new domain and then manually tagged to update an existing domain-specific sentiment lexicon. The sentiment classification is then performed by combining the updated domain-specific lexicon with VADER sentiment analysis, a well-known and widely used sentiment analysis tool. The proposed method is compared with other well-known and widely used sentiment analysis tools, including transformer-based models, such as BERTweet, Twitter-roBERTa and FinBERT, on a domain-specific corpus of stock market-related tweets comprising 1 million messages. The experimental results show that the proposed approach largely surpasses the performance of the other sentiment analysis tools, reaching an overall accuracy of 72.0%. The achieved results highlight the advantage of using a hybrid method that combines domain-specific lexicons with existing generalist tools for the inference of textual sentiment in domain-specific short-text messages.

Keywords

Sentiment analysis; sentiment classification; sentiment lexicon; stock market; tweets

1. Introduction

The social network Twitter was created in 2006 and is nowadays widely used, with around 500 million tweets per day, covering all kinds of content. As such, Twitter becomes a popular social network when one wants to analyse the expression of sentiment in textual data [1–3]. Due to the tweet's maximum number of 280 characters, authors need to express opinions straight to the point. When tweets originate in specific interest areas, it is usual to resort to the use of technical

Corresponding author:
Bruno Taborda, ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon 1649-026, Portugal.
Email: Bruno_Taborda@iscte-iul.pt

jargon and polysemic terms, with the meaning understood by the tweet's particular context domain. Thus, currently, sentiment analysis of domain-specific tweets is considered a challenging task [4].

Sentiment analysis, a natural language processing (NLP)–related task, employs several methods and tools to automatically detect relevant information from a text to determine the prevalent sentiment or opinion it expresses [5]. The most common approach for sentiment analysis or polarity detection consists of classifying the sentiment towards something as positive, neutral or negative [6]. To perform sentiment analysis, and since this is a highly domain-dependent task, domain-specific knowledge is critical for obtaining a good classification performance [7]. Most of the existent research focusing on textual sentiment analysis adopts tools built for generic domains [8,9] which results in lower performance, given that sentiment analysis is a domain-dependent task, and domain-specific words are not taken into account and may express different sentiments for different domains. In fact, only very few approaches are specifically designed for financial and stock-market domains [10–12]. The stand-alone usage of a domain-specific dictionary presents its own shortcomings, given that it is created based on a specific and relatively small data set. To overcome these shortcomings, a few works attempt to combine general and domain-specific dictionaries to provide a better sentiment classifier [13–15]. Despite the wide usage of general domain, domain-specific or hybrid lexicons, we did not identify any other research that creates a specific and up-to-date dictionary to evaluate the sentiment of tweets related to stock markets.

This study investigates: (a) if sentiment analysis for stock-market tweets can be improved using a specifically automatically constructed stock-market lexicon from Twitter and (b) how is sentiment analysis of stock-market tweets' performance affected by the use of hybrid lexicons compared with general approaches. Moreover, we present an enhancement for the actual reference financial dictionary, the Loughran–McDonald dictionary (LMcD) [10]. This updated lexicon, LMcD20, contains additional words not previously included in the lexicon and is automatically drawn from a corpus of recent stock market–related tweets. We also extend the existing body of knowledge by proposing SA-MAIS: a hybrid approach for sentiment analysis of a domain-specific text that integrates general dictionaries with an up-to-date domain-specific sentiment lexicon. This methodology positively compares with existing tools. Our experiments show that this innovative approach achieves better results when compared with pre-trained models or stand-alone usage of generalist or specialised dictionaries. SA-MAIS is available in GitHub,[1] and the data set used to validate SA-MAIS and enhance LMcD is published in IEEE DataPort.[2]

The article is organised as follows: after the introduction, a brief review of the most relevant literature is presented in section 2. Section 3 presents the objectives and the research questions. Section 4 presents the data set and SA-MAIS system's architecture. Section 5 details the domain-specific dictionary enhancement. Section 6 describes SA-MAIS implementation. Section 7 highlights theoretical and practical implications as well as the limitations of this work. Section 8 discusses possible directions for further research and presents the main conclusions of this work.

2. Literature review

Sentiment analysis, also named opinion analysis [16], uses NLP and text analysis to explore sentiment's valence in textual data (e.g. documents, tweets and direct messages). In general, sentiment analysis intends to evaluate the objectivity or subjectivity of a text and classifies it as positive or negative, and thus, this type of classification is considered a binary problem [17,18]. Saif et al. [19] created SentiCircle, a semantic sentiment representation of words. It captures the contextual data, based on the occurrence of tweets and updates the sentiment based on the contextual semantics.

Nofsinger [20] concluded that, due to the nature of stocks, the stock market is directly impacted by social mood, and this behaviour helps to predict 'financial and economic activity'. Sul et al. [21] collected tweets where stock symbols (like AAPL for Apple Inc.) of S&P 500 companies were referred to and classified the sentiment in each tweet as positive or negative. The authors were able to show that their sentiment analysis was related to the firm's stock returns. They also demonstrated that users with many followers directly impact the same day's returns, while users with fewer followers impact future returns (10 days returns). In a nutshell, most of the previous works support the claim that public mood and sentiment expressed in word-of-mouth (WOM) impact stock-market prices. Bollen et al. [22] created a system that correlated Dow Jones Industrial Average with Twitter feeds. Chandra Pandey et al. [23] proposed a new clustering method to evaluate the sentiment of tweets. The proposed method outperforms five of the most well-known algorithms. In more recent work, Song et al. [4] describe a technique that combines supervised and unsupervised learning and uses a new text representation model named Word2PLTS for short-text sentiment analysis. This model is based on probabilistic linguistic term sets that fully describe the possibilities for the sentiment polarity of the word.

Hu et al. [24] proposed an approach to summarise the task of manually verifying customer reviews by reviewing the features related to the products in the data set. The authors created a domain-specific lexicon for customer reviews as an attempt to increase the performance of their lexicon. Loughran et al. [10] identified that previous approaches classified financial texts incorrectly. Thus, the authors created a domain-specific dictionary to classify the sentiment of financial texts more accurately. However, Li et al. [25] implemented a ‘generic stock price prediction framework’ using Harvard IV-4 dictionary and the Loughran–McDonald financial sentiment dictionary [10] for sentiment analysis. The proposed generic framework was tested using 5 years of historical data on prices and news on the Hong Kong Stock Exchange. The authors concluded that models focusing only on positive and negative sentiment classes do not provide good predictions. Junqué et al. [26] used articles from all major Flemish newspapers between 2007 and March 2012. The authors concluded that sentiment analysis using Pattern for Python [27], which is a python package with multiple functionalities including natural language processing, underperformed when compared with Bag-Of-Words or market technical indicators. More recently, Oliveira et al. [28] noted that there is a lack of financial lexicons adjusted to micro-blogging stock markets. The authors proposed a new automatic procedure to create a lexicon based on the StockTwits[3] data set.

Li et al. [29] combined technical indicators with news articles as an attempt to predict Hong Kong stock prices. Four dictionaries were used to perform sentiment analysis on the news articles, namely, Harvard IV-4 Dictionary, Loughran–McDonald financial sentiment dictionary [10], SentiWordNet 3.0 [30] and SenticNet 5 [31]. The authors concluded that Loughran–McDonald financial sentiment dictionary outperformed the remaining dictionaries.

In relation to the specificity of a tweet’s textual message, several authors use Twitter hashtags (i.e. *#fail*, *#iloveit*), emoticons (i.e.:*)* and:\/) to evaluate the sentiment of tweets [32–34]. Using lexicons, Kiritchenko et al. [35] created a supervised statistical text classification tool that analyses the sentiment of short-text messages (e.g. Twitter or SMS), named by the authors as ‘message-level task’ and also analysed the sentiment of a word or a phrase contained in a message, which they called ‘term-level task’. The lexicons were generated automatically using tweets, hashtags and emotions. The authors concluded that the usage of automatically generated lexicons improves performance by 6.5%. In terms of specific financial lexicons, Oliveira et al. [28] used term frequency–inverse document frequency (TD-IDF), information gain, class percentage and weighted class probability to create a lexicon based on StockTwits. Li et al. [36] used cluster sentiment classifiers by applying the TD-IDF weighting method. The authors concluded that this approach outperformed WordNet. WKWSCI Sentiment Lexicon [37] was created to evaluate Amazon reviews and was compared with six available lexicons: Hu & Liu Opinion Lexicon, Multi-perspective Question Answering (MPQA) Subjectivity Lexicon, General Inquirer, National Research Council Canada (NRC) Word-Sentiment Association Lexicon and Semantic Orientation Calculator (SO-CAL) lexicon. The authors concluded that WKWSCI Sentiment Lexicon best fits the non-reviews text. Hassan et al. [38] created a sentiment analysis tool to evaluate an Altmetrics data set. The authors concluded that support vector machine (SVM) with uni-gram outperformed logistic regression and Naïve Bayes. Sohangir et al. [39] compared lexicon-based tools with machine learning techniques to perform sentiment analysis in StockTwits. The authors concluded that VADER [9], which is a general sentiment analyser, outperformed the remaining lexicon-based tools under the experiment, namely, SentiWordNet and TextBlob. In addition, VADER outperformed machine learning techniques such as Logistic Regression, Linear SVM and Naive Bayes classification.

Devlin et al. [40] proposed in 2018 a new pre-trained model called BERT which uses bidirectional encoding representations from transformers. One of the main advantages of this model comes from its flexibility, given that we can fine-tune a BERT model for a specific task. Thus, to create new models derived from BERT, one only output layer needs to be added. Quoc Nguyen et al. [41] proposed a BERT tweet for sentiment analysis (BERTweet), a model based on BERT trained with more than 40,000 tweets. FinBERT was proposed by Araci [42] based on BERT, with the model made to specialise in financial news. FinBERT was trained with news from Reuters and financial phrase bank data sets. Li et al. [43] used FinBERT to analyse news headlines and compared it with other long short-term memory (LSTM) models. The authors concluded that FinBERT outperforms other models. roBERTa Twitter sentiment analyser (Twitter-roBERTa) is a model based on BERT, trained over more than 58 million tweets and fine-tuned with TweetEval benchmark proposed by Barbieri et al. [44].

3. Objectives

Our goal is to provide a proof of concept through a case study: the evaluation of the sentiment of tweets directly related to stock markets.

Following the existing literature, most of the tools that try to predict the market’s behaviour use either generalist approaches or domain-specific dictionaries to quantify the sentiment of tweets or other important sources of data (i.e. news). In the particular case of stock market–related tweets, to the best of our knowledge, there are no domain-specific dictionaries for sentiment analysis. The following section presents a new simple method for upgrading an existing financial dictionary. Regarding generalist tools, we have chosen to use VADER, TextBlob and Stanza [8] and the state-of-the-art of specifically trained language models BERTweet [41], Twitter-roBERTa and FinBERT [42]. These models have been part of the most recent studies in sentiment analysis, not only in financial domains [39] but also in generic domains [45].

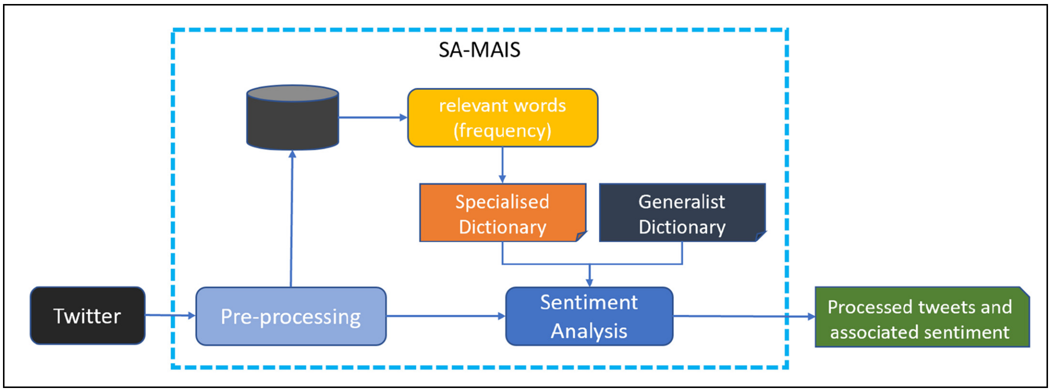

Figure 1. SA-MAIS system architecture.

As such, the present work addresses the following research questions:

RQ1. Is the performance of generalist sentiment analysis tools sound when analysis of stock-market tweets is intended?
RQ2. Can sentiment analysis for stock-market tweets be improved by using specifically constructed stock-market lexicons?
RQ3. Can sentiment analysis of stock-market tweets be improved by integrating a general analyser with a domain-specific lexicon?

To answer these questions, we show that a methodology that seeks to unveil the polarity of a specific domain's short-textual messages must incorporate up-to-date domain knowledge.

4. Methodology

During the analysis of our data set, it became evident that some words like 'breakthrough' and 'barred' were not adequately classified by generalist sentiment analysers (GSAs). In the past few years, either domain-specific dictionaries or lexicons have been explored for sentiment analysis. Different applications found that if the data set is specific enough on a particular topic, such as finance, a domain-specific dictionary may improve the results. We address the problem by proposing a new approach, SA-MAIS, a sentiment analyser that differs from previous tools because it combines a generalist tool and a domain-specific dictionary. SA-MAIS system's architecture is depicted in Figure 1. The methods and definitions created for the system's implementation are detailed later in section 6.

The proposed methodology follows the commonly established framework for mining sentiment in tweets: data collection, pre-processing (removal of numbers, emails, hashtags and hyperlinks), performing sentiment classification and validating the model results.

To validate SA-MAIS as a short-text sentiment analyser, we explore and compare six well-known generalist sentiment classifiers: TextBlob, Stanza, VADER, BERTweet, Twitter-roBERTa and FinBERT. The first one, TextBlob,[4] is a pre-trained Python library for NLP that returns two values for sentiment analysis: the text's polarity and its subjectivity. The latter is a measure for the level of lack of objectivity and thus abstract and subject to individual perception and opinion, whereas polarity expresses the tweet's overall sentiment and is evaluated in the range $[-1.0, 1.0]$, with -1.0 being the most negative sentiment and 1.0 indicating an utterly positive one. Stanza [8] is a toolkit created in 2020 at Stanford University. This classifier was trained using 112 data sets to analyse text in multiple languages (English, German and Chinese). The polarity values differ from those returned by TextBlob since Stanza outputs the values 0, 1 and 2, representing a negative, neutral and positive sentiment, respectively. Hutto and Gilbert [9] developed VADER, which can be considered an embedded library of NLTK [46]. The authors describe VADER as a 'simple rule-based model for general sentiment analysis'. VADER was compared with different sentiment/opinion lexicons: Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN) and Word-Sense Disambiguation (WSD) using WordNet. The authors concluded that VADER outperformed all the lexicons when classifying social media text [9].

Table 1. Number of tweets and companies per sector in the data set.

Sector	Number of companies	Number of tweets
Communication services	5	234,940
Consumer cyclical	2	112,519
Consumer defensive	3	19,875
Energy	2	16,102
Financial services	5	106,000
Healthcare	4	39,126
Index	1	163,801
Technology	4	235,637

Table 2. Manual annotated data set distribution.

Sector	Number of companies	Number of tweets
Communication services	5	536
Consumer cyclical	2	265
Consumer defensive	3	40
Energy	2	34
Financial services	5	240
Healthcare	4	88
Index	1	361
Technology	4	536

BERTweet is a model created by Quoc Nguyen et al. [41] that was trained with SemEval 2017 corpus (around 40,000 tweets). The model classifies the sentiment in three different classes: POS, NEG and NEU, respectively, positive, negative and neutral. Twitter-roBERTa was trained on approximately 58 million tweets and fine-tuned for sentiment analysis with the TweetEval benchmark [44]. The model produces three labels to classify the sentiment, 0 that represents negative sentiment, 1 that represents neutral sentiment and 2 that represents positive sentiment. FinBERT was proposed by Araci [42], and it is specialised in financial news. This model was fine-tuned using the Financial Phrasebank by Malo et al. [47]. The model produces three labels as output: positive, negative and neutral.

At this point, it is important to notice that, to the extent of our knowledge, (1) there are no manually annotated stock-market tweet data set for the three possible sentiment classes: positive, neutral and negative. Moreover, (2) there are no benchmark data sets with a large enough amount of tweets related to S&P500 and its firms.

To this end, we have collected and filtered about 928,000 stock-market tweets, between 9 April 2020 and 16 July 2020, concerning the top 25 companies with higher volume in S&P500 index stock symbol (cash tag), $SPX and #stock. The time window was used to reduce possible time/seasonal patterns (i.e. the uptrend of the sentiment) that could impact the experiments and, consequently, the results. Table 1 shows the distribution of the companies and tweets per economic sector.

Creating an annotated data set for a domain-specific task is time-consuming and is subject to a high degree of subjectivity by the annotators. For example, Li et al. [48] used two annotated data sets to propose a new sentiment analysis of user reviews using deep learning models. Both data sets had the validation set, respectively, with 2210 and 802 reviews. Mowlaei et al. [49] proposed a new aspect-based sentiment analyser. To validate this model, the authors used a data set containing 367 positive reviews and 267 negative reviews, making a total of 634 reviews. We did not identify any manually annotated tweets data set specialised in the stock-market domain. Therefore, we have manually annotated a random sample of 2100 tweets. Table 2 shows the distribution of the companies and tweets per economic sector for the manually annotated tweets. The tweets were manually classified using the three available sentiment values: positive, neutral and negative. The annotation was performed using two independent annotators, which are experts in stock markets, as an attempt to reduce subjectivity while labelling the data set. Cohen's kappa [50] statistic was used to quantify the inter-annotator agreement. This measure is in the range of $[-1.0, 1.0]$, where 1.0 means a complete agreement between annotators and -1.0 means no agreement at all. The manual annotation performed for this data set has the kappa value of 0.88, representing an almost perfect agreement between annotators [50].

Table 3. LMcD20 – positive and negative words added to Loughran–McDonald dictionary.

Positive words	Negative words
Invest	Sell
Investing	Selling
Buy	Short
Buying	Bear
Long	Bearish
Bull	Red
Bullish	Lows
Upper	Down
Green	Low
Highs	Lower
Higher	Risk
Open	Sold
Added	Bottom
Uptrend	
Upwards	
Up	
High	
Top	
Profit	
Rally	
Profits	
Rise	
Revenue	

To keep SA-MAIS as much up-to-date as possible based on the word frequency, we created another data set with almost 1 million tweets, providing large-scale quality data for the analysis. Both data sets have been made public available [51].

5. Lexicon improvements

As previously pointed out, Loughran and McDonald created sentiment lists based on the most probably interpretation of a word in a business context, resulting in two dictionary lists that contain 354 positive and 2329 negative words [10]. The dictionary lists from now on will be named the LMcD, which is nowadays a financial and accounting dictionary of reference [52]. However, it was constructed based on financial accounting texts to enhance sentiment analysis in this specific domain.

This section describes the changes introduced into the LMcD lexicon in order to improve its representative power for stock markets' analysis (section 5.1) and highlights the results of the newly enhanced dictionary (section 5.2).

5.1. LMcD20: domain-specific dictionary enhancement

A deeper exploration of the second stock-tweet data set (not annotated data set) highlighted that some specific and frequent stock market–related words (like 'breakthrough' and 'barred') were not included in the generalised sentiment analysis tools, leading to the need to use domain-specific dictionaries. However, a more in-depth exploration of the LMcD also revealed that some of the words used currently in financial tweets were still not present in the LMcD. We have also noted that words used to express opinions on Twitter are subject to subjective interpretation and vary over time. Therefore, a Twitter domain-specific dictionary cannot be static but must be adjusted over time based on real and up-to-date content.

In order to solve this problem, we have used stock-tweet data containing more recent data as a means to improve our lexicon. A sample of positive and negative words not included in LMcD was selected between the most frequent 500 words extracted from the large-scale data set. From these, 23 words expressed a positive sentiment (such as 'buy' or 'bull'), and 13 words expressed a negative sentiment (like 'short' or 'bear'). The authors manually selected words expressing sentiment, resulting in 36 finance-related new words (Table 3) added to the LMcD dictionary, thus creating LMcD20.

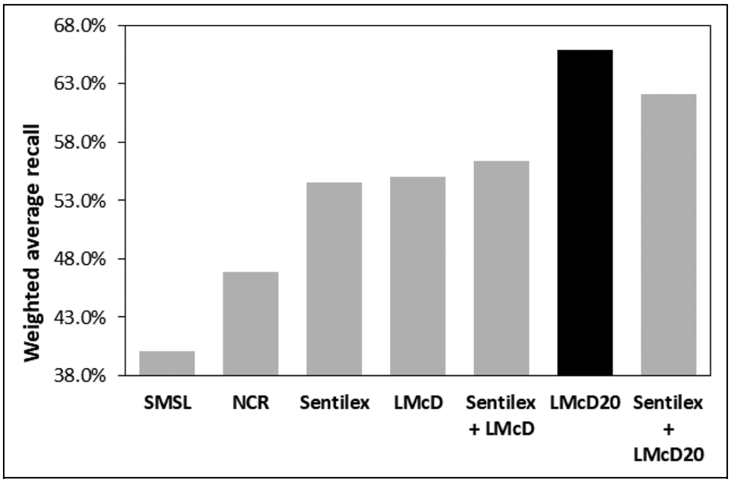

Figure 2. Comparing lexicons – weighted average recall (WAR).

5.2. Comparing different lexicons

In terms of model evaluation, that is, validation of the results of positive, negative or neutral classification of words, and due to the slight imbalance between the negative class and the remaining ones, the metric selected to compare the experiments is the weighted average recall (WAR). This function weights the recall of each class by the number of samples from that class. WAR is defined by equation (1), where TP_p, TP_n and TP_t are the true positives associated with the positive (p), the negative (n) and the neutral (t) sentiment classes, respectively, and N represents the total number of tweets

$$\mathrm{WAR} = \frac{TP_p + TP_n + TP_t}{N} \tag{1}$$

As an initial experiment, a comparison between domain-specific lexicons was performed. In terms of domain-specific dictionaries, Hu et al. [24] (Sentilex) is one of the most well-known domain-specific lexicons and was created for analysis of customer reviews. Oliveira et al. [11] (stock-market sentiment lexicon (SMSL)) and Loughran and McDonald [10] (LMcD) are the examples of financial dictionaries created to be used in specific economic contexts. Mohammad and Turney [53] (NRC) is one of the most well-known domain-specific lexicons incorporating the sentiment polarity and the emotions in the same lexicon for a crowd-sourcing scenario. Loughran and McDonald created the LMcD dictionary in 2011 to evaluate financial reports from companies. LMcD contains two subsets of words: the negative words' subset, which includes 2329 words, and the positive words' subset, which contains 354 words [10]. Many words related to financial markets can be found in the dictionary, but unlike tweets, financial reports are very well structured and carefully written.

The SMSL dictionary was created in 2016 with the primary objective of evaluating StockTwits sentiment [11]. This dictionary contained 20,551 uni-grams and bi-grams and was generated automatically based on a StockTwits sample using a statistical approach. Nonetheless, the validation of SMSL was performed using a selection of 5000 StockTwits classified by the authors of the posts but excluded the neutral sentiment. Each n-gram $(n = 1, 2)$ has a different weight depending on the sentence's context at that moment, being positive or negative. The SMSL was created automatically based on a data set from StockTwits, but unlike tweets, StockTwits' main focus is on financial markets.

Figure 2 shows the results of sentiment analysis using each domain-specific dictionary. It is possible to observe that SMSL shows the worst WAR (40.0%). One of the reasons for SMSL results may be that it was automatically generated using StockTwits. This platform is different from Twitter and the language employed is much more targeted, very specific and directed mainly to financial markets readers. A second reason is that SMSL uses uni-grams and bi-grams generated automatically with the StockTwits data set. Notice that the two best domain-specific lexicons were combined, namely, Sentilex and LMcD. Based on Figure 2, Sentilex combined with LMcD outperformed the LMcD by 1.4 p.p. Comparing the LMcD against the remaining domain-specific lexicons (NRC and Sentilex), LMcD outperformed both by 8 p.p. and 0.5 p.p., respectively.

As an attempt to improve the overall WAR, LMcD20 was created as previously described in section 5.1. Noteworthy, the stand-alone use of LMcD20 achieved the best performance (65.9%) compared with the remaining dictionaries. In particular, LMcD20 shows an improvement of 11 p.p. over LMcD. There was a decrease in the WAR when combining Sentilex and LMcD20. The main reason for this decrease is that LMcD20 is domain-specific for stock-market tweets and Sentilex is domain-specific for customer reviews.

Albeit the results of LMcD20 can be considered satisfactory, a careful analysis of the results reveals that a uni-gram words dictionary, such as LMcD20, still has limitations in tweets sentiment classification. Therefore, we felt the need to combine a GSA with this domain-specific dictionary, something that was achieved by implementing SA-MAIS as described in section 6.

6. SA-MAIS: a hybrid method for the analysis of stock-market tweets

This section details the implementation of SA-MAIS. Section 6.1 outlines how the GSAs and domain-specific dictionaries are combined, and section 6.2 highlights the results achieved with this implementation.

6.1. Classifying tweets

To achieve the best performance, SA-MAIS combines a generalist sentiment classifier, such as TextBlob library, Stanza toolkit, VADER, BERTweet, Twitter-roBERTa, FinBERT and a domain-specific lexicon (like SMSL, NRC, Sentilex, LMcD or the enhanced LMcD20 dictionary). The technique relies upon using both tools for textual analysis and integrating the resultant classification using a convex combination.

6.1.1. GSA component. As previously mentioned, TextBlob returns the polarity expressing the tweet's overall sentiment: a value in the interval $[-1.0, 1.0]$, with -1.0 expressing an entirely negative sentiment and 1.0 a totally positive one. VADER also reports polarity in the same range of values, but Stanza results are 0, 1 or 2, representing a negative, neutral or positive sentiment, respectively. The three deep learning models based on BERT used in this article, namely, BERTweet, Twitter-roBERTa and FinBERT, output three categories representing the sentiment analysed in the text. Despite the labels being different as mentioned before, all of them tend to represent positive, negative and neutral sentiments.

This means that, independently of the tool being used, the sentiment polarity given by the GSA component of SA-MAIS is a value $P_0 \in [-1.0, 1.0]$. Therefore, while the value returned (*pol*) by the generalist analysers TextBlob and VADER is taken by its facial value, in case Stanza is to be used, its output is converted into -1, 0 or 1, indicating a negative, a neutral or a positive sentiment, respectively. Similar to Stanza conversion, BERTweet, Twitter-roBERTa and FinBERT outputs are converted into -1, 0 or 1, indicating a negative, a neutral or a positive sentiment, respectively. Equation (2) defines the GSA polarity value for SA-MAIS, where *pol* represents the sentiment of a given GSA and BERT variations represents the models used in this research (BERTweet, Twitter-roBERTa and FinBERT)

$$P_0 = \begin{cases} pol - 1, & \text{if } pol \text{ is Stanza' soutput} \\ -1, & \text{if } pol \text{ is negative using a BERT} - \text{based model} \\ 0, & \text{if } pol \text{ is neutral using a BERT} - \text{based model} \\ 1, & \text{if } pol \text{ is positive using a BERT} - \text{based model} \\ pol, & \text{otherwise} \end{cases} \tag{2}$$

6.1.2. Domain-specific dictionary component. The domain-specific component of SA-MAIS classifies each tweet by comparing its content with a domain-specific dictionary at the word level. Let W_{pos} be the number of matches in the set of positive words, and W_{neg} be the number of matches in the set of negative words. The output of this component is computed by equation (3)

$$P_1 = \begin{cases} \frac{(W_{pos} - W_{neg})}{(W_{pos} + W_{neg})}, & \text{if}(W_{pos} + W_{neg}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In case none of the tweet words matches the domain-specific dictionary, the output of the domain-specific dictionary is zero.

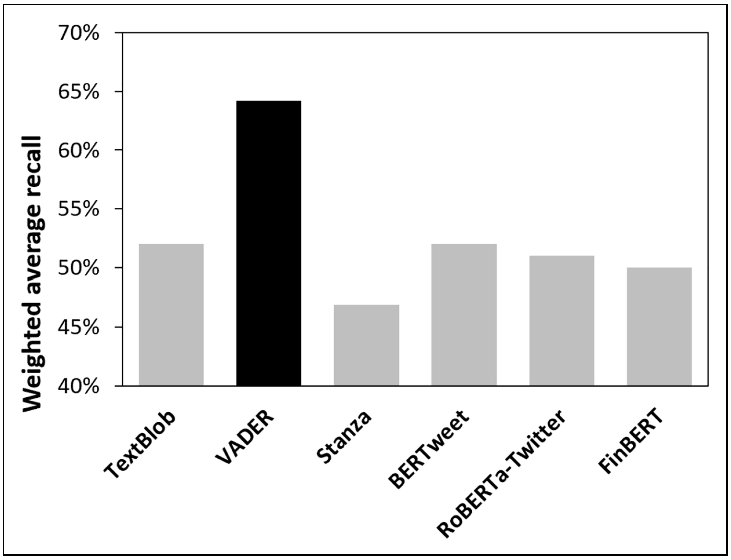

Figure 3. Comparing the performance of the generalist sentiment analysis tools on our data set.

6.1.3. SA-MAIS sentiment value integrator. The output of SA-MAIS is a linear convex combination of the two previous components, given by equation (4), where P_0 is the polarity of the GSA, P_1 is the output of the domain-specific lexicon, and λ is a parameter that defines the importance of the domain-specific lexicon in the final result. As $\lambda \in [0.0, 1.0]$, the polarity's range of values is not altered

$$P = (1 - \lambda) \times P_0 + \lambda \times P_1 \tag{4}$$

Finally, equation (5) defines the categorical output of SA-MAIS. Each tweet is classified as negative, neutral or positive, according to the value of P given by equation (4). The parameter β is dependent on the GSA being used: $\beta = 0.0$ when using either TextBlob, Stanza, BERTweet, Twitter-roBERTa or FinBERT and $\beta = 0.05$ when using VADER, since its authors consider a value ranging from -0.05 to 0.05 as being a neutral value

$$C = \begin{pmatrix} \text{negative}, & -1.0 \leq P < \beta \\ \text{neutral}, & -\beta \leq P \leq \beta \\ \text{positive}, & \beta < P \leq 1.0 \end{pmatrix} \tag{5}$$

6.2. Results

The first experiment is meant to establish the baseline and consists of the evaluation of the performance of TextBlob, Stanza, VADER, BERTweet, Twitter-roBERTa and FinBERT as stand-alone analysers on the annotated data set. This means that the experiment is performed by setting the parameter $\lambda = 0$ in equation (4), and therefore, the domain-specific dictionaries are not used.

As shown in Figure 3, VADER clearly outperformed the remaining tools, achieving approximately a WAR of 64.0% versus 52.0% from BERTweet, with the latter being the model that achieves the highest WAR for the remaining models. Thus, from the compared tools, VADER stands out as the most suitable classifier for stock-market tweet sentiment analysis.

Comparing the classification confusion matrices of TextBlob and VADER (Tables 4 and 5), it is possible to see that TextBlob has a higher failure rate in the positive sentiment class than when predicting a neutral or a negative sentiment class. On the contrary, although VADER shows a more homogeneous confusion matrix, it is possible to see that it fails mainly in classifying negative sentiment tweets. Regarding BERTweet's confusion matrix (Table 6), the model fails mostly in the positive and negative sentiment classes, classifying many of these tweets as neutral. Therefore, it seems that the model is either missing context or words with domain-specific polarity to achieve a proper classification for the financial stock-market tweets.

Table 4. TextBlob – confusion matrix.

	Prediction		
	Negative	Neutral	Positive
Actual			
Negative	140	212	168
Neutral	61	509	226
Positive	66	278	440

Table 5. VADER – confusion matrix.

	Prediction		
	Negative	Neutral	Positive
Actual			
Negative	266	113	141
Neutral	54	514	228
Positive	51	165	568

Table 6. BERTweet – confusion matrix.

	Prediction		
	Negative	Neutral	Positive
Actual			
Negative	180	328	9
Neutral	44	723	39
Positive	45	540	192

With the previous results in mind, SA-MAIS integrated approach was evaluated using VADER as the general component provider combined with LMcD20, the lexicon that achieved the best results in the experiments in section 5.2.

As it can be observed in Figure 4, the WAR of sentiment analysis classification has increased from 65.9% with LMcD20 and 64.0% with VADER up to 71.8% with VADER + LMcD20. Thus, there is an overall increase in 6 p.p. compared with the stand-alone LMcD20 and 8 p.p. with the stand-alone VADER. Notice that the best result was achieved at $\lambda = 0.5$, which means that both components have an equal share of involvement in the final tweet's sentiment classification.

The overall metric values for SA-MAIS using VADER plus LMcD20 with $\lambda = 0.5$ are shown in Table 7, which details the behaviour of SA-MAIS regarding each of the sentiment classes by displaying the most common evaluation metrics (precision, recall and F_1-score). Regarding recall, the statistics show that the most effective classification is for the positive class, whose value is 87.1%. The negative sentiment class shows a recall of 71.4%, while the neutral sentiment class presents the lowest recall value (57.8%). The lowest precision value, 64.6%, is achieved for the positive class, while the neutral sentiment class is the one presenting the highest precision, with 82.9%. According to the previous results, the classes displaying the best F_1-score are the positive and negative sentiment classes having 74.2% and 73.6%, respectively. Overall, the weighted average F_1-score of SA-MAIS is 71.7%.

To better illustrate what can be a positive, neutral and negative tweet, Table 8 shows an example of a correct prediction of SA-MAIS for each of these sentiment classes.

7. Contributions and practical implications

This work proposes a hybrid parametric approach for the analysis of sentiment polarity of short-text messages, combining a general sentiment analyser with an up-to-date domain-specific lexicon termed SA-MAIS.

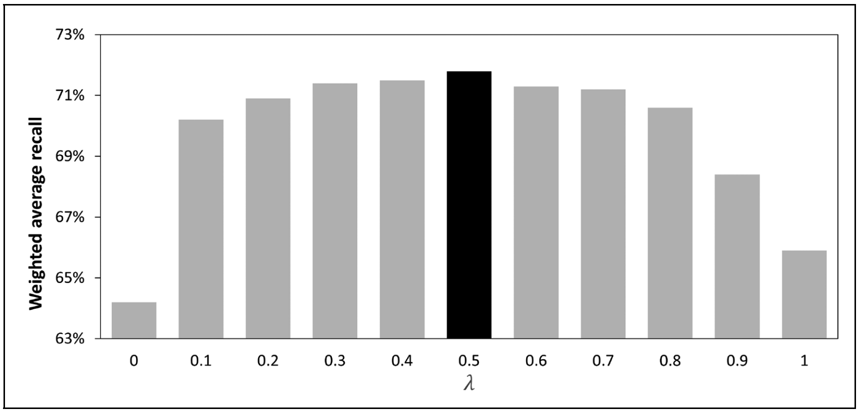

Figure 4. Comparing the performance of LMcD20 and VADER for different values of λ.

Table 7. Metrics for each sentiment class using VADER + LMcD20 with λ = 0.5.

	Precision	Recall	F_1-score
Negative	0.76	0.71	0.74
Neutral	0.83	0.58	0.68
Positive	0.65	0.87	0.74
Weighted average	0.74	0.72	0.72

Table 8. Examples of correct SA-MAIS prediction.

Prediction	Tweet
Positive	Analysts are bullish, expecting the $SPX to rally to above 3200 within 12 months
Neutral	TOS frozen for anyone else? #es_f $spx $spy #options #futures #cl_f $forex
Negative	$AAPL a pause after an epic run up, $BAC upper shadow inside a bearish candle ...

The study offers novel insights into domain-specific sentiment analysis of short-text social media. It describes a new methodological approach for timely analysis and shows that (a) a simple yet effective way of incorporating up-to-date vocabulary from domain-specific short text provides added value for classification tasks and that (b) the usage of this enhanced lexicon improves existing general sentiment analysis, providing a more accurate tool for analysis of textual sentiment in a specific technical domain language. In particular, we describe a proof-of-concept of this methodology by applying it to stock-market tweets analysis and prediction. Despite the focus of this work on stock markets, the general approach may be applied in different domains, whether financial or not, by substituting the specific dictionary and using tweets from the chosen domain area for enhancing the new dictionary.

Second, it contributes to machine learning and text mining research by providing a novel annotated stock market–related corpus to benchmark new approaches and techniques. Third, by comparing the performance of several existing generalist tools, it shows that the latter, on their own, are mostly inadequate for accurate and precise classification of sentiment for stock market–related tweets.

8. Conclusion

A new sentiment analyser method, SA-MAIS, using a framework based on the controlled integration of a GSA and a domain-specific dictionary has been presented. This system combines the well-known GSA VADER with a domain-specific lexicon, LMcD20, updated with the more recent lexical trends. An enhanced version of the LMcD financial lexicon, named LMcD20, that incorporates newer and up-to-date specific finance-related words automatically retrieved from stock-market tweets was also presented.

The SA-MAIS hybrid combination of generalist and domain-specific analyses was comprehensively tested using six popular GSAs: TextBlob, Stanza [8], VADER [9] and the specifically trained state-of-the-art models BERTweet [41], Twitter-roBERTa and FinBERT [42] together with the four existing specialised financial dictionaries: LMcD financial sentiment dictionary [10], the SMSL [11], Sentilex [24] and NRC [53]. As a proof of concept, after running several experiments, it was possible to conclude that the novel-enhanced dictionary LMcD20 shows an increase in WAR results of about six percentage points for the Twitter stock market–related corpus. Furthermore, the SA-MAIS implementation using the integration of VADER with LMcD20 improves the former results over all the possible classification classes – positive, negative and neutral. These results indicate that SA-MAIS can be used as a tool in more elaborate systems for market evolution prediction as it outperforms the state-of-art in terms of NLP models using deep learning.

Finally, all the experiments were conducted using a novel annotated corpus publicly available at https://github.com/taborda11/SAMAIS. In terms of further directions for research, this study inevitably presents some limitations. Ideally, the 2100 annotated documents data set should be extended, since further manual annotation of tweets would not only allow for the enrichment of the annotated corpus but would be an important asset for future enhancement of LMcD20, eventually leading to an increase in SA-MAIS sentiment classification results. Second, more complex or purposeful dictionaries, possibly representing relations between words and additional linguistic information could be pivotal for improving the results presented here. Third, performing online tests using SA-MAIS could improve this tool's performance and expand its scope.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UIDB/50021/2020 and UIDB/00315/2020.

Notes

1. https://github.com/taborda11/SAMAIS.
2. https://ieee-dataport.org/open-access/stock-market-tweets-data.
3. https://stocktwits.com.
4. https://textblob.readthedocs.io.

ORCID iD

Bruno Taborda https://orcid.org/0000-0002-3564-8662

References

[1] Giachanou A and Crestani F. Like it or not: a survey of Twitter sentiment analysis methods, 2016, https://dl.acm.org/doi/abs/10.1145/2938640

[2] Wagh R and Punde P. Survey on sentiment analysis using Twitter dataset. In: *Proceedings of the 2nd international conference on electronics, communication and aerospace technology (ICECA)*, 2018, pp. 208–211. Institute of Electrical and Electronics Engineers. DOI: 10.1109/ICECA.2018.8474783.

[3] Chaturvedi I, Cambria E, Welsch RE et al. Distinguishing between facts and opinions for sentiment analysis: survey and challenges. *Inform Fus* 2018; 44: 65–77.

[4] Song C, Wang XK, Cheng Pf et al. SACPC: a framework based on probabilistic linguistic terms for short text sentiment analysis. *Knowl Based Syst* 2020; 194: 105572.

[5] Liu B. Sentiment analysis and subjectivity, 2010, https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf

[6] Dave K, Lawrence S and Pennock DM. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: *Proceedings of the 12th international conference on World Wide Web (WWW)*, 2003, pp. 519–528. DOI: 10.1145/775152.775226.

[7] Khan MT, Durrani M, Ali A et al. Sentiment analysis and the complex natural language. *Complex Adapt Syst Model* 2016; 4(1): 2.

[8] Qi P, Zhang Y, Zhang Y et al. Stanza: a Python natural language processing toolkit for many human languages. In: *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

[9] Hutto CJ and Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the 8th international conference on weblogs and social media (ICWSM)*, 2014, May, pp. 216–225.Ann Arbor, Michigan USA: PKP Publishing Services Network.

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

[11] Oliveira N, Cortez P and Areal N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decis Support Syst* 2016; 85: 62–73, https://www.sciencedirect.com/science/article/pii/S0167923616300240

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

[13] Cho H, Kim S, Lee J et al. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowl Based Syst* 2014; 71: 61–71.

[14] Agarwal B, Mittal N, Bansal P et al. Sentiment analysis using common-sense and context information. *Comput Intell Neurosci* 2015; 2015: 30.

[15] Asghar MZ, Kundi FM, Ahmad S et al. T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. *Expert Systems* 2018; 35(1): e12233.

[16] Pang B and Lee L. Opinion mining and sentiment analysis. *Found Trends Inform Retr* 2008; 2(1–2): 1–135, http://www.nowpublishers.com/article/Details/INR-011

[17] Pang B, Lee L and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on empirical methods in natural language processing-EMNLP '02*, pp. 79–86, vol. 10. Morristown, NJ: Association for Computational Linguistics. DOI: 10.3115/1118693.1118704.

[18] Turney PD. Thumbs up or thumbs down? In: *Proceedings of the 40th annual meeting on association for computational linguistics-ACL '02*, p. 417. Morristown, NJ: Association for Computational Linguistics. DOI: 10.3115/1073083.1073153.

[19] Saif H, He Y, Fernandez M et al. Contextual semantics for sentiment analysis of Twitter. *Inform Process Manag* 2016; 52(1): 5–19.

[20] Nofsinger JR. Social mood and financial economics. *J Behav Finance* 2005; 6(3): 144–160.

[21] Sul HK, Dennis AR and Yuan L. Trading on Twitter: the financial information content of emotion in social media. In: *Proceedings of the annual Hawaii international conference on system sciences*, 2014, pp. 806–815. DOI: 10.1109/HICSS.2014.107.

[22] Bollen J, Mao H and Pepe A. Modeling public mood and emotion: blog and news sentiment and socio-economic phenomena. In: *Proceedings of the fifth international AAAI conference on weblogs and social media*, pp. 692–699, https://www.sciencedirect.com/science/article/abs/pii/S0167739X17323750

[23] Chandra Pandey A, Singh Rajpoot D and Saraswat M. Twitter sentiment analysis using hybrid cuckoo search method. *Inform Process Manag* 2017; 53(4): 764–779.

[24] Hu M and Liu B. Mining and summarizing customer reviews. In: *KDD-2004-proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 168–177. New York: Association for Computing Machinery. DOI: 10.1145/1014052.1014073.

[25] Li X, Xie H, Chen L et al. News impact on stock price return via sentiment analysis. *Knowl Based Syst* 2014; 69(1): 14–23.

[26] Junqué De Fortuny E, De Smedt T, Martens D et al. Evaluating and understanding text-based stock price prediction models. *Inform Process Manag* 2014; 50(2): 426–441.

[27] De Smedt T and Daelemans W. Pattern for Python. *J Mach Learn Res* 2012; 13(66): 2063–2067, https://dl.acm.org/doi/10.5555/2503308.2343710

[28] Oliveira N, Cortez P and Areal N. Automatic creation of stock market lexicons for sentiment analysis using StockTwits data. In: *IDEAS '14: Proceedings of the 18th international database engineering & applications symposium*, pp. 115–123. New York: Association for Computing Machinery. DOI: 10.1145/2628194.2628235.

[29] Li X, Wu P and Wang W. Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong. *Inform Process Manag* 2020; 57(5): 102212.

[30] Baccianella S, Esuli A and Sebastiani F. SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the 7th international conference on language resources and evaluation (LREC)*, 2010, pp. 2200–2204. DOI: 10.1.1.61.7217.

[31] Cambria E, Poria S, Hazarika D et al. SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In: *Proceedings of the thirty-second AAAI conference on artificial intelligence*, vol. 32, pp. 1–8, www.aaai.org

[32] Davidov D, Tsur O and Rappoport A. Enhanced sentiment learning using Twitter hashtags and smileys. In: *Coling 2010: posters*, 2010, August, pp. 241–249. Beijing, China: Coling 2010 Organizing Committee.

[33] Kouloumpis E, Wilson T and Moore J. Twitter sentiment analysis: the good the bad and the OMG! In: *Fifth international AAAI conference on weblogs and social media*, 2011, https://ojs.aaai.org/index.php/ICWSM/article/view/14185

[34] Mohammad SM. #Emotional tweets. In: *Conference on lexical and computational semantics*, pp. 246–255, http://www.ark.cs.cmu.edu/GeoText

[35] Kiritchenko S, Zhu X and Mohammad SM. Sentiment analysis of short informal texts. *J Artif Intell Res* 2014; 50: 723–762, https://jair.org/index.php/jair/article/view/10896

[36] Li G and Liu F. Application of a clustering method on sentiment analysis. *J Inform Sci* 2012; 38(2): 127–139.

[37] Khoo CS and Johnkhan SB. Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. *J Inform Sci* 2017; 44(4): 491–511.

[38] Hassan SU, Saleem A, Soroya SH et al. Sentiment analysis of tweets through Altmetrics: a machine learning approach. *J Inform Sci* 2020; 47(6): 712–726.

[39] Sohangir S, Petty N and Wang D. Financial sentiment lexicon analysis. In: *Proceedings-12th IEEE international conference on semantic computing (ICSC)*, 2018, January, pp. 286–289, vol. 2018. Institute of Electrical and Electronics Engineers. DOI: 10.1109/ICSC.2018.00052.

[40] Devlin J, Chang MW, Lee K et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: *2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2 June–7 June 2019, pp. 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

[41] Quoc Nguyen D, Vu T, Tuan Nguyen A et al. BERTweet: a pre-trained language model for English tweets. In: *Proceedings of EMNLP 2020: system demonstrations*, pp. 1–6. DOI: 10.48550/arxiv.2005.10200.

[42] Araci D. *FinBERT: financial sentiment analysis with pre-trained language models*. PhD Thesis, University of Amsterdam, Amsterdam, 2019.

[43] Li JB, Lin SY, Leu FY et al. Stock price trend prediction using LSTM and sentiment analysis on news headlines. In: *BWCCA 2022: advances on broad-band wireless computing, communication and applications*. Springer, Cham, pp. 282–291. DOI: 10.1007/978-3-031-20029-8_27.

[44] Barbieri F, Camacho-Collados J, Neves L et al. TweetEval: unified benchmark and comparative evaluation for tweet classification. In: *Findings of the association for computational linguistics: EMNLP* 2020, 2020, pp. 1644–1650. DOI: 10.48550/arxiv.2010.12421.

[45] Shelar A and Huang CY. Sentiment analysis of Twitter data. In: *Proceedings-2018 international conference on computational science and computational intelligence (CSCI)*, 2018, pp. 1301–1302. Institute of Electrical and Electronics Engineers. DOI: 10.1109/CSCI46756.2018.00252.

[46] Bird S, Loper E and Klein E. *Natural language processing with Python*. O'Reilly Media, 2009. Sebastopol, CA

[47] Malo P, Sinha A, Korhonen P et al. Good debt or bad debt: detecting semantic orientations in economic texts. *J Assoc Inf Sci Technol* 2014; 65(4): 782–796, https://www.researchgate.net/publication/251231107_Good_Debt_or_Bad_Debt_Detecting_Semantic_Orientations_in_Economic_Texts

[48] Li W, Zhu L, Shi Y et al. User reviews: sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. *Appl Soft Comput J* 2020; 94: 106435.

[49] Mowlaei ME, Saniee Abadeh M and Keshavarz H. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Syst Appl* 2020; 148: 113234.

[50] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20(1): 37–46.

[51] Taborda B, de Almeida A, Carlos Dias J et al. Stock market tweets data, 2021, DOI: 10.21227/g8vy-5w61.

[52] Loughran T and McDonald B. Textual analysis in accounting and finance: a survey. *J Account Res* 2016; 54(4): 1187–1230.

[53] Mohammad SM and Turney PD. Crowdsourcing a word-emotion association lexicon. *Comput Intell* 2013; 29(3): 436–465, http://arxiv.org/abs/1308.6297