Cognitive Computation
https://doi.org/10.1007/s12559-021-09833-w

Automatically Building Financial Sentiment Lexicons While Accounting for Negation

Thomas Bos[1] · Flavius Frasincar[1]

Received: 18 June 2020 / Accepted: 12 January 2021

Abstract
Financial investors make trades based on available information. Previous research has proved that microblogs are a useful source for supporting stock market decisions. However, the financial domain lacks specific sentiment lexicons that could be utilized to extract the sentiment from these microblogs. In this research, we investigate automatic approaches that can be used to build financial sentiment lexicons. We introduce weighted versions of the Pointwise Mutual Information approaches to build sentiment lexicons automatically. Furthermore, existing sentiment lexicons often neglect negation while building the sentiment lexicons. In this research, we also propose two methods (Negated Word and Flip Sentiment) to extend the sentiment building approaches to take into account negation when constructing a sentiment lexicon. We build the financial sentiment lexicons by leveraging 200,000 messages from StockTwits. We evaluate the constructed financial sentiment lexicons in two different sentiment classification tasks (unsupervised and supervised). In addition, the created financial sentiment lexicons are compared with each other and with other existing sentiment lexicons. The best performing financial sentiment lexicon is built by combining our Weighted Normalized Pointwise Mutual Information approach with the Negated Word approach. It outperforms all the other sentiment lexicons in the two sentiment classification tasks. In the unsupervised sentiment classification task, it has, on average, a balanced accuracy of 69.4%, and in the supervised setting, a balanced accuracy of 75.1%. Moreover, the various sentiment classification tasks confirm that the sentiment lexicons could be improved by taking into account negation while building the sentiment lexicons. The improvement could be made by using one of the proposed methods to incorporate negation in the sentiment lexicon construction process.

Keywords Sentiment analysis · Financial sentiment lexicon · Negation · Weighted pointwise mutual information · Neural networks

Introduction

The introduction of the Web made it possible and easy for users to express their opinions online. The number of online opinions has snowballed over the past years and is still growing [1]. Processing and analyzing online opinions have emerged as an important task for organizations and researchers since they contain valuable information. Manually identifying the sentiment of opinions and summarizing opinions is very challenging and impracticable [2]. Consequently, there is a rising demand for approaches that overcome the drawbacks of manually processing opinions. Sentiment mining approaches are computational approaches that automatically obtain the sentiment of an opinion [3]. Sentiment lexicons play a key role in these approaches since most of them use a sentiment lexicon [4]. They can be constructed manually or automatically. Manually creating these lexicons ensures their high quality because they are made by language and domain experts. However, this process is time-cons uming, and those experts may not always be available. Consequently, the coverage of the manually built sentiment lexicon is low. These disadvantages turned the focus to automatically building sentiment lexicons. Sentiment lexicons can be constructed for the general domain or a specific domain, such as the financial domain. Building a domain-specific sentiment lexicon is more challenging since words could have domain-specific meanings and sentiments.

✉ Flavius Frasincar
frasincar@ese.eur.nl

Thomas Bos
tja.bos@hotmail.com

[1] Erasmus Universiteit, Rotterdam, Netherlands

Published online: 11 February 2021

In this paper, we focus on building sentiment lexicons for the financial domain. Financial investors make trades based on available information. Some of this information is made available by social media. Previous research has proved that social media messages and news articles are useful sources for supporting stock market decisions [5, 6]. Consequently, sentiment analysis is being increasingly used to predict stock market variables [7]. For example, Malandri et al. [8] use a financial sentiment lexicon to predict the best asset allocation. Xing et al. [9] use sentiment analysis to create market views. These market views are integrated into an asset allocation method. Picasso et al. [10] and Weng et al. [11] use, among other things, sentiment analysis on news articles to forecast stock prices. Although the interest in sentiment analysis in the stock market is rising, the domain lacks good sentiment lexicons. In the past, manually made financial sentiment lexicons, like the sentiment lexicon made by Loughran and McDonald [12], are not always performing well compared to automatically built financial sentiment lexicons [13, 14].

In this research, we investigate existing automatic approaches that can be used to build financial sentiment lexicons. Furthermore, we investigate how they can be extended to account for negation while building a financial sentiment lexicon. These solutions are all focused on building the sentiment lexicons without any domain or language knowledge. This kind of approach is also known as an a priori approach. We use three different types of a priori approaches to create sentiment lexicons for the financial domain, namely probability-based, information retrieval-based, and sentiment-aware word embedding-based approaches. The financial sentiment lexicons are built by using messages from StockTwits, which is a financial microblogging platform. The messages are marked as either *bullish* or *bearish*. In the financial domain, *bullish* indicates positive sentiment, and *bearish* indicates negative sentiment. Hereafter, we use the terms *bullish* and positive interchangeably. In addition, we also use the terms *bearish* and negative interchangeably. Moreover, we do not consider the sentiment class neutral for financial corpora in this research due to this class's ambiguity. However, it is still possible that words in the sentiment lexicon end up having a sentiment strength of zero, i.e., a neutral sentiment orientation. We define the sentiment orientation as the sign of the sentiment strength.

After building the financial sentiment lexicons, we evaluate these lexicons by classifying financial messages. We compare the financial sentiment lexicons and other general and financial sentiment lexicons created by other researchers in two different settings. We evaluate the sentiment classification in an unsupervised and supervised setting. For the evaluation part, we use three different financial corpora. The financial corpora consist of messages from StockTwits, financial-related tweets from Twitter, and financial headlines. The different classification (unsupervised and supervised) tasks show us that the probability-based approaches outperform the information retrieval-based and sentiment-aware word embedding-based approaches. Moreover, the proposed weighted versions of the Pointwise Mutual Information (PMI) approaches outperform other researchers' general and financial sentiment lexicons in all the sentiment classification tasks. Furthermore, we notice that accounting for negation while building the sentiment lexicons leads to better performing sentiment lexicons, which other approaches neglect when building them.

The main contributions of this paper are as follows:

- We propose weighted versions of the PMI approaches. The sentiment lexicons built by these weighted approaches outperform other lexicons in different sentiment classification tasks in the financial domain;
- We discuss how to deal with negation in sentences, and we show how the sentiment lexicon building approaches could be extended to account for negation when determining the sentiment orientation and strength of a word. We propose two different methods, namely the Negated Word (NW) approach and the Flip Sentiment (FS) approach.

The remainder of this paper is structured as follows. In the next section, the “Related Work” section, we review the literature that is relevant to our research. The related work is followed by a description of the implementation of the various approaches that are used to automatically build financial sentiment lexicons in the “Methodology” section. The process of building the financial sentiment lexicons and the performed evaluation of these are described in the "Results" section. In the "Conclusion" section, we provide concluding remarks and suggest future research directions.

Related Work

Sentiment lexicons play a crucial role in the sentiment analysis approaches since most of the existing sentiment mining approaches use a sentiment lexicon [4]. There are multiple ways to create a sentiment lexicon. They can be divided into two main categories: manual and automatic approaches. Moreover, the latter category can also be divided into two subcategories: dictionary-based and corpus-based approaches [15, 16].

The first category, manual approaches, consists of sentiment lexicons that are entirely made by hand. These approaches are the most labor-intensive and expensive approaches because they require domain and language experts to manually assign sentiment orientations and sentiment strengths to words and phrases. Consequently, these sentiment lexicons are of high quality. On the other hand, they are time-consuming to build, hard to maintain, and not immune to the evolution of words and their sentiment orientation. Moreover, the coverage of the manually built sentiment lexicon is low. The Harvard General Inquirer [17] and the MPQA subjectivity sentiment lexicon [18] are great examples of manually built sentiment lexicons. The Harvard General Inquirer is an extensive collection of words containing syntactic, semantic, and pragmatic information of part-of-speech tagged words. The Harvard General Inquirer also includes whether a word can be classified as a word with either a positive or negative sentiment orientation. The MPQA subjectivity sentiment lexicon has the same structure as the Harvard General Inquirer, but it also contains the subjectivity strength of a word or phrase. The subjectivity strength could be strong if the word or phrase has a strong meaning, like “excellent,” or the subjectivity strength could be weak if it has a weak meaning, like “fine.”

For the financial domain, the manually made lexicons by Loughran and McDonald [12] and Jegadeesh and Wu [19] are the best known manually built sentiment lexicons. Loughran and McDonald [12] made use of 10-K documents from the U.S. Securities and Exchange Commission. They built six lexicons named after the sentiment they represent: positive, negative, uncertainty, litigious, modal strong, and modal weak. Jegadeesh and Wu [19] also worked with the 10-K documents from the U.S. Securities and Exchange Commission. However, they focused on the importance of assigning a weight to the words in the sentiment lexicon.

The second category, the dict ionary-based approaches, consists of approaches that exploit semantic relations, such as synonyms and antonyms, between words. Most of the approaches start with a small set of seed words. This set of seed words consists of a small group of words for which the sentiment orientation is already known. The small set of seed words is expanded by looking up the seed words’ synonyms and antonyms in a dictionary [20]. An example of a dictionary is the online (semantic) lexical resource WordNet [21]. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by utilizing conceptual-semantic and lexical relations. An example of a synset is the synset of the word “stock market.” This synset contains the synonyms “stock exchange” and “securities market.” Using a dictionary-based approach, one starts defining the set of seed words to build a sentiment lexicon. Thereafter, the process continues expanding the seed set by searching for synonyms and antonyms of the words that are contained in the seed set.

The third category, corpus-based approaches, consists of approaches that extract the sentiment lexicon’s words from a corpus or corpora. These approaches could also use a list of seed words, but the list is expanded using corpora instead of a dictionary. An advantage of using corpus-based approaches is that they use the fact that these corpora contain domain-specific knowledge. This domain-specific knowledge gives words a domain-specific sentiment orientation. There are multiple types of approaches in the category of corpus-based approaches. We point out the studies that are most related to our work. The first set of studies uses unsupervised techniques, such as information-theoretic techniques and other statistical measurements. The first significant work that uses these techniques is the work of Turney [22]. The author applies PMI and information retrieval measurements to estimate the semantic orientation of words or phrases. Later, other information-theoretic techniques and statistical measurements were used [13, 14, 23, 24]. These works show that approaches using information-theoretic techniques and statistical measurements belong to the state-of-the-art approaches to create domain-specific sentiment lexicons. Next to the unsupervised techniques, many studies use supervised techniques that create sentiment lexicons. Li and Shah [25], Tang et al. [26], and Wang and Xia [27] learn word embeddings by using a neural network to capture both the syntactic structure and semantics of a word. The approach of Vo and Zhang [28] consists of a simple neural network that learns the sentiment orientation of a word. The neural network learns the sentiment orientations of words by optimizing the accuracy of predicting the sentiment orientation of messages. The authors show that building sentiment lexicons by optimizing predictions improves the sentiment lexicon’s accuracy compared to sentiment lexicons built by counting-based methods. Recently, there is an upcoming interest in methods that adapt existing lexicons to a specific domain. An example of such an approach is the work of Xing et al. [29]. The authors introduce a cognitive-inspired approach that uses the wrongly predicted sentences to adjust the polarity scores. The newly constructed sentiment lexicons achieve higher accuracies in the sentiment classification tasks than the original sentiment lexicons.

Furthermore, other approaches make use of both a dictionary and a corpus. The dictionary-based approaches usually do not give domain or context-dependent meanings to words. In addition, employing a corpus-based approach makes it hard to find a large set of opinion words if the corpus is not large. The disadvantages of both types of approaches can be tackled by combining these types [16]. An example of a study that combines both types of approaches is the work of Hu and Liu [2]. Hu and Liu [2] start by extracting adjectives from corpora, which are, in this case, consumer reviews. Thereafter, the authors assign a sentiment orientation to these adjectives based on the known sentiment orientation of a list of original seed adjectives. The list of seed adjectives is iteratively expanded by using the seed adjectives’ semantic relations in WordNet.

This way, it contains both domain-specific adjectives obtained from the corpus and general adjectives, which are the original seed adjectives.

Methodology

In this section, we discuss the methodology we use to create financial sentiment lexicons and evaluate the created financial sentiment lexicons. We start by discussing the probability-based approaches, information retrieval-based approaches, and the sentiment-aware word embedding-based approach. Thereafter, we elaborate on how to account for negation while building a sentiment lexicon. Last, we discuss the methods we use to evaluate the quality of the created financial sentiment lexicons.

Financial Sentiment Lexicon Approaches

In this section, we dive deeper into the different approaches we use to create financial sentiment lexicons. Before we dive deeper into these approaches, we introduce some general notation in Table 1.

Probability-Based Approaches

The probability-based approaches are focused on the probabilities of a sentiment class given a word, i.e., the probabilities of a word being positive and negative. The different probabilities are obtained by counting the occurrences in a training set. Hence, we also refer to this type of approach as counting-based approaches. We start with the Bayes' Theorem Benchmark (BTB) approach, which is the most intuitive approach. The BTB approach makes use of Bayes' theorem and is focused on counting the frequencies of words. Thereafter, we continue with the PMI approach, which is similar to the BTB approach. However, the PMI approach is focused on counting the frequency of messages.

Table 1 General definitions and notations

Notation	Definition
c	Sentiment class (positive or negative)
$\|C\|$	The cardinality of the set of all sentiment classes
pos	Positive sentiment class, i.e., *bullish*
neg	Negative sentiment class, i.e., *bearish*
M	Total number of all the messages
M_c	Number of messages that have a label of sentiment class c
M_{pos}	Number of the positive labeled messages
M_{neg}	Number of the negative labeled messages
M_w	Number of messages that contain word w
$M_{w,c}$	Number of messages that contain word w and are labeled with the sentiment class c
m	Message in the corpus, $m \in M$
w	Word $w \in m$
$\mathrm{SS}_x(w)$	Sentiment strength of w computed by approach x

In this table, we state the most common and important abbreviations for the considered sentiment lexicon building approaches

Bayes' Theorem Benchmark. Our first approach is defined by Labille et al. [23]. It is derived from the Bayes' theorem introduced by Bayes and Price [30]. We define the sentiment strength of word w, computed by the BTB approach, $\mathrm{SS}_{\mathrm{BTB}}(w)$, as the difference between the probability of being positive, $p(pos|w)$, and the probability of being negative, $p(neg|w)$. The $\mathrm{SS}_{\mathrm{BTB}}(w)$ is stated in Eq. 1.

$$\begin{aligned}\mathrm{SS}_{\mathrm{BTB}}(w) &= p(pos|w) - p(neg|w),\\ &= \frac{\sum_{m\in M_{pos}} n_{wm}}{\sum_{m\in M} n_{wm}} - \frac{\sum_{m\in M_{neg}} n_{wm}}{\sum_{m\in M} n_{wm}},\end{aligned} \tag{1}$$

where n_{xy} denotes the number of word(s) x in the set y. The probabilities $p(pos|w)$ and $p(neg|w)$ can be interpreted as counting the number of times word w appears in messages with that specific sentiment class, divided by the total appearances of word w in all messages.

Pointwise Mutual Information. PMI measures the association between two words or sets of words. The PMI measurement was derived by Church and Hanks [31] from Fano's original definition of mutual information [32]. In this research, we follow the works of Turney [22] and Oliveira et al. [14] to suit the PMI measure to the needs of sentiment analysis. However, the interpretation of the PMI measure slightly differs from the interpretation of the works mentioned above. We interpret it as counting the frequency of messages instead of counting the frequency of words. The sentiment strength $\mathrm{SS}_{\mathrm{PMI}}(w)$ is defined as follows:

$$\begin{aligned}\mathrm{SS}_{\mathrm{PMI}}(w) &= \mathrm{PMI}(w, pos) - \mathrm{PMI}(w, neg),\\ &= \log_2\left(\frac{M_{w,pos} \times M}{M_w \times M_{pos}}\right) - \log_2\left(\frac{M_{w,neg} \times M}{M_w \times M_{neg}}\right).\end{aligned} \tag{2}$$

There are two significant drawbacks of the PMI approach, as defined in Eq. 2. The first drawback is that we could come across a word that only appears in messages that belong to one of the two sentiment classes. Consequently, the logarithm's inner term in the PMI measure of the other sentiment class becomes equal to zero. Since the logarithm is undefined for zero, we are unable to compute the corresponding PMI measure. We tackle this problem by setting the PMI measure of the corresponding sentiment class to be equal to zero, as has been suggested by Bouma [33].

Table 2 Example 1 drawback PMI - Part 1

Word	M_w	$M_{w,pos}$	$M_{w,neg}$	$\mathrm{PMI}(w, pos)$	$\mathrm{PMI}(w, neg)$	$\mathrm{NPMI}(w, pos)$	$\mathrm{NPMI}(w, neg)$
w_1	5	5	0	1	0	1	0
w_2	6	5	1	0.74	-1.58	0.74	-0.48
w_3	3	2	1	0.41	-0.58	0.18	-0.18
w_4	6	4	2	0.41	-0.58	0.31	-0.25
w_5	4	4	0	1	0	0.76	0

In this table, we compute the PMI and NPMI values for the positive and negative class of the words w_1, w_2, w_3, w_4, and w_5

To illustrate the second drawback, we look at an example. In this example, we assume that $M = 10$, $M_{pos} = 5$, and $M_{neg} = 5$. Further details of this example are stated in Table 2. If we look at w_2, we see that it occurs in all the positive messages and only in one negative message. Furthermore, it has a $\mathrm{PMI}(w, pos)$ value of 0.74 and a $\mathrm{PMI}(w, neg)$ value of -1.58. If we compare the absolute values of both PMI measures, we notice that the value of $\mathrm{PMI}(w, neg)$ is more than twice as large as the $\mathrm{PMI}(w, pos)$ value. Consequently, the influence of $\mathrm{PMI}(w, neg)$ on $\mathrm{SS}_{\mathrm{PMI}}(w)$ is not in line with the occurrence of w_2 in the negative message compared to the occurrences of w_2 in the positive messages. One expects that the value of $\mathrm{PMI}(w, pos)$ would be larger than the value of $\mathrm{PMI}(w, neg)$ for w_2 and thus have a larger influence on $\mathrm{SS}_{\mathrm{PMI}}(w)$. To tackle this problem, Bouma [33] suggests normalizing the PMI measure. The maximum value of the Normalized PMI (NPMI) measure is equal to one, which only occurs if a word solely appears in messages of a specific sentiment class.

Next, we compute $\mathrm{NPMI}(w, pos)$ and $\mathrm{NPMI}(w, neg)$ for w_2. The NPMI measure values are 0.74 and -0.48, respectively. These values are in line with the values that one expects given $M_{w,pos}$ and $M_{w,neg}$ for w_2. However, there is also a disadvantage of using the NPMI measure. To illustrate this disadvantage, we take a look at w_3 and w_4 from Table 2. Both words occur in the same ratio in positive and negative messages, namely 2:1 and 4:2, respectively. Since they have the same ratio, one could intuitively expect that they have the same PMI and NPMI values. Nevertheless, this holds only for the PMI measure and not for the NPMI measure.

To choose between using the PMI measure and the NPMI measure, we look at the sentiment strengths of w_1, w_2, w_3, and w_4, which are displayed in Table 3. In our example, w_1 solely occurs in positive messages. Consequently, we expect w_1 to have the highest sentiment strength. We use the sentiment strength of w_1 as our benchmark to compare the sentiment strengths of w_2, w_3, and w_4. First, we take a look at the $\mathrm{SS}_{\mathrm{PMI}}(w)$ values. The $\mathrm{SS}_{\mathrm{PMI}}(w)$ value for w_2 is larger than the $\mathrm{SS}_{\mathrm{PMI}}(w)$ value of w_1 and does not reflect that w_2 is also found in negative messages. Furthermore, the $\mathrm{SS}_{\mathrm{PMI}}(w)$ values for w_3 and w_4 are approximately equal to one, which is also unwanted since w_3 and w_4 also occur in negative messages. Therefore, we do not want to use $\mathrm{SS}_{\mathrm{PMI}}(w)$ to compute the sentiment strengths in our sentiment lexicon. Next, we compute the $\mathrm{SS}_{\mathrm{NPMI}}(w)$ values for all the words. The $\mathrm{SS}_{\mathrm{NPMI}}(w)$ is defined as follows:

$$\begin{aligned}\mathrm{SS}_{\mathrm{NPMI}}(w) &= \mathrm{NPMI}(w, pos) - \mathrm{NPMI}(w, neg), \\ &= \frac{\log_2\left(\frac{M_{w,pos} \times M}{M_w \times M_{pos}}\right)}{-\log_2\left(\frac{M_{w,pos}}{M}\right)} - \frac{\log_2\left(\frac{M_{w,neg} \times M}{M_w \times M_{neg}}\right)}{-\log_2\left(\frac{M_{w,neg}}{M}\right)}.\end{aligned} \tag{3}$$

Now, we see in Table 3 that only for w_2, the $\mathrm{SS}_{\mathrm{NPMI}}(w)$ value of 1.21 is not in line with our expectations. Therefore, we decided not to use $\mathrm{SS}_{\mathrm{NPMI}}(w)$ for our sentiment lexicon creation. We propose to use weighted versions of the $\mathrm{SS}_{\mathrm{PMI}}(w)$ and $\mathrm{SS}_{\mathrm{NPMI}}(w)$, which is one of this paper's contributions. We weigh the (N)PMI values by the ratio of occurrence in messages with the specific sentiment class. The

Table 3 Example 1 drawback PMI - Part 2

Word	M_w	$M_{w,pos}$	$M_{w,neg}$	$\mathrm{SS}_{\mathrm{PMI}}(w)$	$\mathrm{SS}_{\mathrm{NPMI}}(w)$	$\mathrm{SS}_{\text{W-PMI}}(w)$	SS_{W-NPMI}
w_1	5	5	0	1	1	1	1
w_2	6	5	1	2.32	1.21	0.88	0.69
w_3	3	2	1	1	0.35	0.47	0.18
w_4	6	4	2	1	0.57	0.47	0.29
w_5	4	4	0	1	0.76	1	0.76

In this table, we compute the $\mathrm{SS}_{\mathrm{PMI}}(w)$, $\mathrm{SS}_{\mathrm{NPMI}}(w)$, $\mathrm{SS}_{\text{W-PMI}}(w)$, and $\mathrm{SS}_{\text{W-NPMI}}(w)$ of the words w_1, w_2, w_3, w_4, and w_5

Weighted PMI (W-PMI) sentiment strength, $\mathrm{SS}_{\text{W-PMI}}(w)$, is computed as follows:

$$\begin{aligned}\mathrm{SS}_{\text{W-PMI}}(w) =& \frac{M_{w,pos}}{M_w} \times \mathrm{PMI}(w, pos) - \frac{M_{w,neg}}{M_w} \times \mathrm{PMI}(w, neg),\\ =& \frac{M_{w,pos}}{M_w} \times \log_2\left(\frac{M_{w,pos} \times M}{M_w \times M_{pos}}\right)\\ &- \frac{M_{w,neg}}{M_w} \times \log_2\left(\frac{M_{w,neg} \times M}{M_w \times M_{neg}}\right).\end{aligned} \tag{4}$$

The Weighted NPMI (W-NPMI) sentiment strength, $\mathrm{SS}_{\text{W-NPMI}}(w)$, is computed as follows:

$$\begin{aligned}\mathrm{SS}_{\text{W-NPMI}}(w) =& \frac{M_{w,pos}}{M_w} \times \mathrm{NPMI}(w, pos)\\ &- \frac{M_{w,neg}}{M_w} \times \mathrm{NPMI}(w, neg),\\ =& \frac{M_{w,pos}}{M_w} \times \frac{\log_2\left(\frac{M_{w,pos}\times M}{M_w \times M_{pos}}\right)}{-\log_2\left(\frac{M_{w,pos}}{M}\right)}\\ &- \frac{M_{w,neg}}{M_w} \times \frac{\log_2\left(\frac{M_{w,neg}\times M}{M_w \times M_{neg}}\right)}{-\log_2\left(\frac{M_{w,neg}}{M}\right)}.\end{aligned} \tag{5}$$

After defining the weighted versions of $\mathrm{SS}_{\text{PMI}}(w)$ and $\mathrm{SS}_{\text{NPMI}}(w)$, we compute $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ for words w_1, w_2, w_3, and w_4. The $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ are displayed in Table 3. One can see that all the $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ values for words w_2, w_3, and w_4 are smaller than the $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ values of w_1. The weighted versions give us desired sentiment strength values. Furthermore, we obtained different values of $\mathrm{SS}_{\text{W-NPMI}}(w)$ for w_3 and w_4, even though the words have the same ratio between $M_{w,pos}$ and $M_{w,neg}$. On the one hand, one could argue that having the same ratio between $M_{w,pos}$ and $M_{w,neg}$ should result in an equal sentiment strength, which is the case for $\mathrm{SS}_{\text{W-PMI}}(w)$. On the other hand, one could argue that the sentiment strengths, in this case, should not be equal because one should also take into account the ratio between $M_{w,c}$ and M, which is not the same for w_3 and w_4.

Similar to the discussion between $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ above, one could argue that w_1 and w_5 should have equal sentiment strengths, which is the case for $\mathrm{SS}_{\text{W-PMI}}(w)$. The argument is that both words appear only in positive messages and, therefore, should have the same sentiment strength. However, one could again argue that one should take into account the relation between $M_{w,c}$ and M. Since w_1 occurs more often in the positive messages (i.e., higher $M_{w,pos}$), one could argue that you are more certain about the sentiment orientation and sentiment strength of w_1 compared to w_5. Therefore, w_1 and w_5 should have different sentiment strengths. In this research, we use both $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ to compute the sentiment strengths for our sentiment lexicon.

Last, using $\mathrm{SS}_{\text{W-PMI}}(w)$ to compute the sentiment strengths has a small disadvantage in the case of a word that is hugely unevenly distributed over $M_{w,pos}$, and $M_{w,neg}$. We illustrate this disadvantage with an example, which is displayed in Table 4. In this example, M is equal to 40, and M_c is equal to 20. In the case of w_7, the $\mathrm{SS}_{\text{W-PMI}}(w)$ value is slightly larger than the $\mathrm{SS}_{\text{W-PMI}}(w)$ value of w_6, which is unwanted since w_6, in contrast to w_7, only appears in positive messages. Therefore, we suggest clamping the sentiment strength, such that it is between -1 and 1. These values are the minimum and maximum sentiment strengths in the cases of having a word that only occurs in either positive or negative messages. We clamp of the sentiment strength with the following equation:

$$\mathrm{SS}_{\text{W-PMI}}(w) = \max(\min(\mathrm{SS}_{\text{W-PMI}}(w), 1), -1). \tag{6}$$

Information Retrieval-Based Approach

In general, there exist many information retrieval techniques. One of the most popular information retrieval techniques is the Term Frequency-Inverse Document Frequency (TF-IDF) statistic proposed by Salton and Buckley [34]. The statistic reflects how important a specific term t is to a document d in a corpus. Wang and Zhang [35] introduced the Term Frequency-Inverse Category Frequency (TF-ICF) statistic, a similar statistic to the TF-IDF statistic but designed explicitly for categories instead of documents. The intuition behind the ICF term: the more categories in which word w occurs, the less discrimination power word w has. Next to the TF-ICF measure, Wang and Zhang [35] propose an extension of the TF-ICF measure, namely the Inverse Category Frequency-based (ICF) measure. This ICF-based measure combines the TF-ICF measure and the Relevance Frequency (RF) measure introduced by Lan et al. [36].

To define a sentiment score based on the information retrieval measure, we follow the work of Oliveira et al. [14].

Table 4 Example 2 drawback $\mathrm{SS}_{\text{W-PMI}}(w)$

Word	M_w	$M_{w,pos}$	$M_{w,neg}$	$\mathrm{SS}_{\text{W-PMI}}(w)$	SS_{W-NPMI}
w_6	19	19	0	1	0.93
w_7	20	19	1	1.05	0.85

In this table, we compute the $\mathrm{SS}_{\text{W-PMI}}(w)$ and $\mathrm{SS}_{\text{W-NPMI}}(w)$ of the words w_6 and w_7

Oliveira et al. propose the following equation to compute the sentiment strength for word w using the TF-IDF measure:

$$SS_{\text{TF-IDF}}(w) = \frac{\text{TF-IDF}(w, pos) - \text{TF-IDF}(w, neg)}{\text{TF-IDF}(w, pos) + \text{TF-IDF}(w, neg)}. \tag{7}$$

In our case, we adjust Eq. 7 to the following equation:

$$\begin{aligned} SS_{\text{ICF}}(w) &= \frac{\text{ICF}(w, pos, neg) - \text{ICF}(w, neg, pos)}{\text{ICF}(w, pos, neg) + \text{ICF}(w, neg, pos)}, \\ &= \frac{\dfrac{\left(tf_{w,pos}\times\log_2\left(2+\frac{M_{w,pos}}{\max(1,M_{w,neg})}\times\frac{|C|}{cf_w}\right)\right.}{\left.-tf_{w,neg}\times\log_2\left(2+\frac{M_{w,neg}}{\max(1,M_{w,pos})}\times\frac{|C|}{cf_w}\right)\right)}}{\dfrac{\left(tf_{w,pos}\times\log_2\left(2+\frac{M_{w,pos}}{\max(1,M_{w,neg})}\times\frac{|C|}{cf_w}\right)\right.}{\left.+tf_{w,neg}\times\log_2\left(2+\frac{M_{w,neg}}{\max(1,M_{w,pos})}\times\frac{|C|}{cf_w}\right)\right)}} \end{aligned} \tag{8}$$

where $tf_{w,c}$ is the number of times word w occurs across all messages with sentiment class c; $|C|$ is the cardinality of the set of all sentiment classes, i.e., the number of sentiment classes, which equals two in our case; cf_w is the number of sentiment classes that contain word w.

Sentiment-Aware Word Embedding-Based Approach

The final type of approach makes use of sentiment-aware word embeddings. Word embeddings represent words or phrases that are mapped to vectors with real numbers. Words with similar contexts appear closer to each other than words that do not have a similar context. The algorithms that create the word embeddings use a large corpus to capture and process the words' semantic and syntactic contexts. Popular algorithms that create word embeddings are word2vec [37] and GloVe [38]. In the field of sentiment analysis, there is a demand for word embeddings that also contain the sentiment of the words. However, standard word embedding creating algorithms cannot always capture the sentiment successfully in the word embeddings [27]. Consequently, one cannot utilize these general word embeddings and should focus on sentiment-aware word embeddings, which also contain the sentiment of words.

In our research, we use the Simple Neural Network (SNN) approach of Vo and Zhang [28] to construct the sentiment-aware word embeddings. We start by defining the words as word embeddings. A word w takes the form of $[n, p]$, where n stands for the negative sentiment value of the word and p for the positive sentiment value. The positive and negative sentiment values are the weights of the neural network obtained after training the neural network. We refer to the work of Vo and Zhang [28] for further details about the neural network that we use to train the sentiment-aware word embeddings. We compute the sentiment strength $SS_{\text{SNN}}(w)$ of word w by simply subtracting n from p.

Adjustments for Negation

Taking negation into account when performing sentiment analysis could improve determining the messages' sentiment orientation [39]. For example, the sentiment of the sentence "It is looking good." is the opposite of the sentiment of the sentence "It is not looking good.", while the sentences are word-wise very similar to each other. The challenge of detecting the negation consists of two parts: negation cue detection and negation scope detection [40]. The negation cue is the negation keyword that indicates that there is a negation in a sentence. We can distinguish two types of negation cues: explicit and implicit negation cues [41]. Explicit negation cues are negation words, such as "not" and "never," which affect the following words and change their meaning.

On the other hand, we have implicit negation cues, such as "dislike" and "hopeless." The implicit negation cues can be recognized by their affixes and suffixes, such as "dis-," "im-," and "-less," and their negation affects only these single words. In this research, the implicit negation cues are treated as ordinary words, and therefore, they automatically receive their own sentiment orientation and strength in our sentiment lexicon. Hence, we only pay extra attention to the explicit negation cues and leave the implicit negation cues as future research. In this work, we focus on the explicit negation cues as defined by Jia et al. [40] and Councill et al. [41]. In Table 5, we state all the explicit negation cues that we use in this research. Since we are dealing with microblogging

Table 5 Explicit negation cues

no	not	never	less	without
barely	hardly	rarely	scarcely	seldom
lack	lacking	lacks	neither	nor
rather	nobody	none	nothing	nowhere
no longer	no more	no way	no where	by no means
at no time	not (...) anymore	ain't	aren't	can't
cannot	couldn't	didn't	doesn't	don't
hasn't	haven't	isn't	mightn't	mustn't
needn't	shan't	shouldn't	wasn't	weren't
wont	wouldnt	aint	arent	cant
couldnt	didnt	doesnt	dont	hasnt
havent	isnt	mightnt	mustnt	neednt
shant	shouldnt	wasnt	werent	wont
wouldnt				

In this table, we state all the explicit negation cues that we use in this research messages, we also take into account abbreviations of the explicit negation cues, such as “isnt” and “cant.”

After detecting a negation cue, we must still tackle the challenge of detecting the negation scope. The negation scope is the set of words affected by the negation cue, and the sentiment orientation is inverted. There exist many approaches to detect the negation scope. They vary from simply setting a fixed window as the negation scope [40] to using machine learning approaches to determine the negation scope [42]. We follow the work of Hogenboom et al. [39] and consider the two words following the negation cue as the negation scope. The authors show that this is a simple and effective approach to use in sentiment classification. Finally, after determining the negation cues and negation scope, we treat the explicit negation cues as stopwords and remove them from the messages.

We propose two approaches to account for the negated words in the negation scope while computing the sentiment orientation and strength. The first approach creates two entries for a word in the sentiment lexicon, one for the original word and one for its negated version. We refer to this approach as the NW approach. We transform the negated word *w* to “NOT_*w*” and give it a separate entry in the sentiment lexicon. If we look at the example sentence, “It is not looking good.”, then we create a new entry for the negated version of “good,” namely “NOT_good.” Now, an own sentiment orientation and strength are assigned to “NOT_good.” An advantage of this approach is that we only have to change the words in the negation scope to their negated versions. In addition, we do not have to change any input for the previously described sentiment lexicon building approaches because the negated words get separate entries.

Consequently, if we come across a negated word in the sentiment classification task, we do not pay particular attention to the negated word. Nevertheless, this approach has a disadvantage. The negated version of a word may receive the same sentiment orientation as the original version, which is possible due to the low number of occurrences of the negated version of the word. We could tackle this by setting a threshold of minimal occurrences before the negated version of a word is included in the sentiment lexicon. However, this results in a loss of information since we do not add these negated versions to the sentiment lexicon.

The second approach considers the negated words to have a sentiment orientation that is the opposite of the message’s sentiment class. Thus, we consider the negated words in a message with a positive sentiment class as negative words and the negated words in a message with a negative sentiment class as positive words. In other words, we flip the sentiment orientation of the negated words. We refer to this approach as the FS approach. To clarify the FS approach further, we look at the example sentence: “It is not looking good.”. The message’s sentiment class is negative, but since “good” is in the negation scope, we consider good” to be the opposite of negative, i.e., positive. The probability-based and information retrieval-based sentiment lexicon building approaches specifically rely on the number of occurrences of word w in sentiment class c (= $\sum_{m \in M_c} n_{wm}$ and $tf_{w,c}$) and on the number of messages of sentiment class c in which word w occurs (= $M_{w,c}$).

Consequently, we adjust the values of $\sum_{m \in M_c} n_{wm}$, $tf_{w,c}$, and $M_{w,c}$ for the negated word w. If we have the word “good” from the example sentence, then we adjust $M_{w,pos}$ by adding one message to $M_{w,pos}$, and $M_{w,neg}$ by subtracting one message from $M_{w,neg}$. For the sentiment-aware word embedding-based approach, we treat the negated scope(s) of a message as a separate message, which has the opposite sentiment class of the original message. Furthermore, the second approach tackles the disadvantage of having the same sentiment orientation by having a single entry for each word.

In Table 6, we state all the automatically sentiment lexicon building approaches discussed in this paper. We select the five most advanced approaches (per type and per category), which we use for evaluation. For each approach, we create three different versions. The first version is made without adjusting for negation, i.e., the benchmark sentiment lexicon. The second and third versions of the sentiment lexicon are created with accounting for negation using the two previously described approaches. Hence, we construct in total fifteen financial sentiment lexicons. We compare the different financial sentiment lexicons made while accounting for negation to their benchmarks to analyze whether there is a significant difference in performance on the sentiment classification tasks.

Evaluation

We evaluate the created financial sentiment lexicons in different ways. We first discuss the supervised and unsupervised classification evaluation of these lexicons. The evaluation is done internally by comparing the built financial sentiment lexicons with each other and externally by comparing them with different existing lexicons. The external

Table 6 All sentiment lexicon building approaches

Type of approach	Name of approach
Probability-based	**BTB**, PMI, NPMI, **W-PMI**, and **W-NPMI**
Information retrieval-based	TF-IDF, TF-ICF, and **ICF**
Sentiment-aware word embedding-based	**SNN**

In this table, we display the advanced approaches that we consider in our evaluation in bold comparison is made with the following general and financial sentiment lexicons:

- **Harvard General Inquirer Lexicons (GI)** - The Harvard General Inquirer [17] contains a positive and a negative lexicon, which is originally constructed by the Harvard IV dictionary. Since the words lack a sentiment strength, we assign a value of +1 to the words in the positive lexicon and a value of −1 to the words in the negative lexicon.
- **MPQA Subjectivity Lexicon** - The MPQA Subjectivity Lexicon has been manually built by Wilson et al. [18]. We use the prior polarity of words as the sentiment orientation. The prior polarity can either be positive, negative, neutral, or both positive and negative. An example of a word that is both positive and negative is the word "demand." In this research, we only use the words with either a positive, neutral, or negative prior polarity. We assign a value of 0 to the words that have neutral prior polarity.
- **Hu and Liu Lexicons (HL)** - Hu and Liu [2] built two lexicons, a positive and a negative lexicon. Similar to the GI lexicon, we assign a value of +1 to the words in the positive lexicon and a value of −1 to the words in the negative lexicon. The words' subjectivity strength indicates whether the meaning of a word is either strong or weak. Similar to Oliveira et al. [14], we use the words' subjectivity strength to adjust the weak words' sentiment strength to +0.5 or −0.5. In the case of a strong word, we keep the sentiment strengths of +1 and −1.
- **NRC Hashtag Sentiment Lexicon (NRC-H)** - Mohammed et al. [43] created the first sentiment lexicon using the PMI measure, as described in the "Probability-Based Approaches"section. The PMI measure was applied to the words of 775,000 tweets, which were marked as either positive or negative by their hashtags. The authors used positive hashtags, such as #good, and negative hashtags, such as #bad, to identify the tweet's sentiment orientation. In our research, we use the sentiment lexicon that consists of unigrams.
- **NRC Emoticon Sentiment Lexicon (NRC-E)** - The second sentiment lexicon generated by Mohammed et al. [43] is constructed by applying the PMI measure on the sentiment140 corpus [51]. The tweets in the corpus were classified as either positive or negative based on the emoticon(s) in the tweet.
- **VADER Sentiment Lexicon** - Ten individual raters rated more than 7,500 words to create the VADER sentiment lexicon [44]. The raters rated the words on a scale of [−4, +4]. Thereafter, the average of these ten ratings is taken as the sentiment strength of a word.
- **Loughran and McDonald Lexicons (LM)** - Loughran and McDonald [12] constructed six lexicons out of financial 10-K documents. The lexicons are named after the sentiment they represent. These lexicons only contain words and do not contain any specific sentiment orientations or strengths. In this research, we only use the positive and negative lexicons because it is unclear which sentiment orientation we should assign to the other lexicons. We assign a positive sentiment orientation and a sentiment strength of +1 to the words in the positive lexicon. In addition, we assign a negative sentiment strength and a sentiment strength of −1 to the words in the negative lexicon.
- **SenticNet 6.0 Lexicon** - Cambria et al. [45] introduced an approach that combines both symbolic and subsymbolic models and leverages their strengths. In this research, we make use of the sixth version of the SenticNet knowledge base.
- **Stock Market Sentiment Lexicon (SM)** - Oliveira et al. [14] generated a financial sentiment lexicon using the PMI measure. This sentiment lexicon was constructed by leveraging messages from StockTwits. The SM sentiment lexicon is the only external sentiment lexicon that considered negation. The authors account for negation by dividing the messages of StockTwits into an affirmative corpus and a negated corpus. Thereafter, they learn two separate sentiment strengths for each word, one without negation and one with negation.

Sentiment Classification Evaluation

In the sentiment classification evaluation, we use the obtained sentiment lexicons to classify unseen messages as either positive or negative. The evaluation is done internally by comparing the created financial sentiment lexicons and externally by comparing them with the earlier mentioned lexicons constructed by other researchers. The comparisons are made in a supervised and unsupervised manner. In the comparisons, we use different metrics, which are all based on the well-known confusion matrix.

In the unsupervised setting, the sentiment lexicon may be unable to classify a message as either positive or negative due to the insufficient coverage of the sentiment lexicon or because the sentiment strengths cancel each other out. Consequently, we can distinguish two groups of test messages in all the unseen messages (A). The first group consists of the unclassified messages (U), and the second group consists of the classified messages (C). Based on this differentiation, we compute the following evaluation metrics:

Overall Accuracy (ACC1): The overall percentage of correctly classified messages.

$$= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{A}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{U}+\mathrm{C}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{U}+\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}};$$

Unclassified (UNCL):The percentage of unclassified messages due the insufficient coverage of the sentiment lexicon.

$$= \frac{\mathrm{U}}{\mathrm{A}} = \frac{\mathrm{U}}{\mathrm{U}+\mathrm{C}};$$

Classification Accuracy (ACC2): The percentage of correctly classified messages after adjusting for the unclassified messages.

$$= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{C}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}};$$

Balanced Accuracy (BA): Balanced accuracy of the classified messages.

$$= \frac{\mathrm{TP}\times(\mathrm{TN}+\mathrm{FP})+\mathrm{TN}\times(\mathrm{TP}+\mathrm{FN})}{2\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})};$$

F$_1$Positive (F$_1$Pos): The F$_1$measure for the pos. sentiment class.

$$= \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}};$$

F$_1$Negative (F$_1$Neg): The F$_1$-measure for the neg. sentiment class.

$$= \frac{2\times\mathrm{TN}}{2\times\mathrm{TN}+\mathrm{FN}+\mathrm{FP}};$$

Macro F$_1$ (Macro F$_1$) :The macro F$_1$measure.

$$= \frac{\mathrm{F_1Pos}+\mathrm{F_1Neg}}{2}.$$

In this research, we are mainly focusing on the balanced accuracy and the macro F$_1$ metric because they are combinations of the other metrics. The balanced accuracy combines the true positive rate and the true negative rate. The macro F$_1$ measure combines the recall and precision metrics. Last, we also account for negation in the sentiment classification tasks. Here, we again follow the work of Hogenboom et al. [39] and define the negation scope as the two words following the negation cue. If we come across a negated word in a sentiment classification task that has not a separate entry in the sentiment lexicon, we flip the sentiment orientation of the word and maintain the sentiment strength.

Unsupervised Classification. In the unsupervised setting, we look up the messages' words in the sentiment lexicon and take the sum of all the individual words' sentiment strengths to obtain an overall sentiment score of the message. In the case of a word that is not stated in the sentiment lexicon, the word's sentiment strength is set to zero such that it does not influence the overall sentiment score. If the overall sentiment score of a message is positive, then we classify the message as positive. On the other hand, if the overall sentiment score is negative, we classify the message as negative.

Supervised Classification. Next to the unsupervised evaluation of the financial sentiment lexicons, we also evaluate the financial sentiment lexicon in a supervised manner. We start by extracting some pre-defined sentiment lexicon features on the test dataset, as defined by Zhu et al. [46]. The sentiment lexicon features are as follows:

- The number of words in message *m* that have a sentiment strength in the sentiment lexicon;
- The total sentiment value of message *m*, which is computed by taking the sum over the sentiment strengths of all the words in *m*;
- The largest sentiment strength of the words in message *m*;
- The sum of sentiment strengths of the words in message *m* that have a positive sentiment orientation;
- The sum of sentiment strengths of the words in message *m* that have a negative sentiment orientation;

– The sentiment strength of the last word in message m.

We use the sentiment lexicon features as input for the supervised sentiment classification. We train a linear classifier with LibLinear [47]. A linear classifier works well on a large number of features, and it supports interpretability. We perform a grid search on the accuracy to tune the type of classifier and the hyperparameter α, representing the cost of constraints violation on the five-fold cross-validation. As described by Fan et al. [47], we consider six different types of multi-class classifiers.

Results

In this section, we discuss the created financial sentiment lexicons and their evaluation. First, we give more details about constructing the financial sentiment lexicons and provide more insight into them. Thereafter, we look at the performance of the obtained financial sentiment lexicons in various sentiment classification tasks. The constructed financial sentiment lexicons and the R and Python implementation codes used to produce these are made available at https://github.com/ThomasJABos/Financial-Sentiment-Lexicons-Negation.

Building Financial Sentiment Lexicons

In this research, we make use of three datasets. The first dataset is used to construct the sentiment lexicons and test the sentiment lexicons. The second and third dataset are solely meant as complementary datasets for the sentiment classification tasks. The first dataset consists of collected messages from StockTwits. We received permission from StockTwits to use their database to collect these messages. StockTwits users can mark their messages as *bullish* or *bearish*. We set the overall sentiment values of the messages to +1 and -1, respectively. We collect 10,000 *bullish* and 10,000 *bearish* messages for each month in the year 2019. Hence, we collect a total of 240,000 messages. An advantage of collecting messages each month is that topics differ monthly, which results in a richer vocabulary of words. In addition, the advantage of having an equal number of messages in each sentiment class is that there are words that only occur in messages that belong to one of the two sentiment classes. Both advantages lead to increased coverage of the financial sentiment lexicons. The second and third dataset are made available by Cortis et al. [48]. The second dataset is the microblogging dataset, and the third dataset is the financial headlines dataset. We state an overview of the number of messages in each dataset in Table 7.

Before we can use the StockTwits messages to construct the financial sentiment lexicons, we undertake some preprocessing steps to clean the messages. We start by removing URLs, user mentions, and cashtags. One reason for removing cashtags is to prevent that cashtags get labeled with a sentiment orientation and strength related to the time period. We also remove punctuation, emoticons, and emojis to ease the (pre)processing steps. The emoticons and emojis could be indicators of sentiment, but these are outside the scope of this research. In addition, intentional spelling mistakes, such as “boreddd” and “cooool,” could also carry a sentiment. In this study, we do not correct these intentional spelling mistakes. Furthermore, we remove simple stopwords from the messages. Stopwords are words that often do not provide any additional information or insight [49]. Examples of stopwords are “a” and “the.” We use the stopwords from the list of stopwords introduced by Feinerer et al. [50].

Finally, we process the numbers from the messages. The numbers in the messages could be classified into three categories. The first category contains numbers that are prior to a percentage sign, %. This category contains all the percentage increases and decreases. We replace the percentage increases, e.g., +15%, with *posperc*, and the percentage decreases, e.g., -18%, with *negperc*. The second category consists of numbers that indicate increases or decreases, but a percentage sign does not follow these numbers. We replace the increases, e.g., +15, with *posnum* and the decreases, e.g., -18, with *negnum*. The reason for replacing the increases and decreases with a tag is that we want to prevent that single numbers would receive a sentiment orientation and strength. By replacing them with tags, we still maintain the sentimental value of the numbers. The last category consists of single numbers without a sign, i.e., without a + or −. An example of a phrase that does not contain a sign is the following phrase: “selling at 50.2”. The number does not have a meaning without knowing the context, which is, in this case, the stock price of a particular stock. We remove the numbers from the third category.

Table 7 Number of messages in the datasets

Dataset	Purpose	Positive messages	Negative messages	Total messages
StockTwits	Training	100,000	100,000	200,000
StockTwits	Testing	20,000	20,000	40,000
Microblogging	Testing	1,067	562	1,629
Headlines	Testing	623	436	1,059

In this table, we state the number of messages in each dataset

The financial sentiment lexicons are constructed using 200,000 messages from StockTwits. The training set of messages consists of 100,000 messages with a positive sentiment class and 100,000 messages with a negative sentiment class. In this research, we focus on building financial sentiment lexicons that do only contain unigrams, i.e., single words. The reason for focusing only on unigrams is that computation time significantly increases if we also consider *n*-grams, i.e., sets of *n* words as entries for our financial sentiment lexicon. We only consider unigrams that occur at least five times in our dataset. We refer to the lexicons built without accounting for negation as the original financial sentiment lexicons. Next, we refer to the lexicons constructed using the first negation approach, which focuses on creating separated entries for the negated words, as NW financial sentiment lexicons. Last, the FS financial sentiment lexicons are the lexicons created with the second negation approach, which is based on the principle that the negated words belong to the opposite sentiment class of the message.

Sentiment Classification Evaluation

We start by performing the sentiment classification in an unsupervised setting. In the unsupervised setting, we evaluate the sentiment lexicons of each category on the three test datasets. We start with the StockTwits test dataset. Thereafter, we evaluate the microblogging dataset, and finally, we discuss the financial headlines dataset. The unsupervised setting is followed by the sentiment classification in a supervised setting. In the supervised setting, we need to train the linear classifiers and test the sentiment lexicons using a test set. In order to have a well-trained classifier and at the same time have enough test messages remaining, we need a sufficiently large test set. Therefore, we only evaluate the financial sentiment lexicons on the StockTwits test dataset in the supervised setting.

Unsupervised Sentiment Classification Evaluation

Table 8 shows the evaluation metrics of the fifteen financial sentiment lexicons on the StockTwits test dataset. The sentiment lexicons built using the BTB, W-PMI, W-NPMI, and ICF approaches all have similar values for the evaluation metrics. However, the sentiment lexicons constructed using the SNN approach have dissimilar values for the evaluation metrics compared to the other approaches. Looking at the balanced accuracy and the macro F_1 measure, we see that the sentiment lexicons of the BTB, W-PMI, and W-NPMI approaches, the probability-based approaches, perform slightly better than the sentiment lexicons of the ICF approach. In the category of original sentiment lexicons, we notice that the BTB and W-NPMI sentiment lexicons have a balanced accuracy of 73.2% and a macro F_1 measure of 72.7%.

Table 8 Evaluation metrics of the financial sentiment lexicons in unsupervised sentiment classification on the StockTwits dataset

Original Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	73.1	**0.1**	**73.2**	**73.2**	76.3	**69.0**	**72.7**
W-PMI	73.0	**0.1**	73.1	73.1	76.2	**69.0**	72.6
W-NPMI	**73.2**	**0.1**	**73.2**	**73.2**	**76.5**	68.9	**72.7**
ICF	72.6	**0.1**	72.6	72.6	76.0	68.1	72.1
SNN	54.1	**0.1**	54.2	54.2	19.1	68.0	43.5
NW Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	**73.4**	**0.1**	**73.5**	**73.5**	**76.7**	69.2	**73.0**
W-PMI	**73.4**	**0.1**	**73.5**	**73.5**	**76.7**	**69.3**	**73.0**
W-NPMI	73.3	**0.1**	73.3	73.3	**76.7**	68.8	72.8
ICF	72.9	**0.1**	73.0	73.0	76.4	68.3	72.4
SNN	54.5	**0.1**	54.6	54.6	19.0	68.4	43.7
FS Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	**73.0**	**0.1**	**73.0**	**73.0**	**76.4**	**68.6**	**72.5**
W-PMI	72.9	**0.1**	72.9	72.9	76.3	68.5	72.4
W-NPMI	72.8	**0.1**	72.8	72.8	**76.4**	68.1	72.2
ICF	72.5	**0.1**	72.6	72.6	76.1	67.7	71.9
SNN	49.4	**0.1**	49.4	49.4	21.3	62.8	42.0

In this table, the best metrics per category are stated in bold. The overall best metrics across all the lexicons are underlined

If we look at the NW category's sentiment lexicons, we notice that all the evaluation metrics of the BTB, W-PMI, W-NPMI, and ICF sentiment lexicons have been improved. Hence, accounting for negation while building the sentiment lexicons pays off. Finally, looking at the FS lexicons, we notice that they perform slightly worse than the other two categories' sentiment lexicons. Overall, based on the balanced accuracy and the macro F_1 measure, the BTB and W-PMI NW sentiment lexicons perform the best on the StockTwits dataset. The BTB and W-PMI NW sentiment lexicons have a balanced accuracy of 73.5% and a macro F_1 measure of 73.0%.

Table 9 shows the evaluation metrics of the financial sentiment lexicons on the microblogging dataset. Similar to Table 8, we notice that the sentiment lexicons of the SNN approach have dissimilar evaluation metrics compared to the sentiment lexicons of the other approaches. Furthermore, in all the categories, the sentiment lexicons created using the W-NPMI approach have the highest values compared to the other sentiment lexicons. Moreover, the W-NPMI sentiment lexicon of the FS category overall has the highest values for the evaluation metrics with a balanced accuracy of 72.5% and a macro F_1 measure of 73.5%. In addition, we notice that the balanced accuracy and macro F_1 measure of all approaches are similar or higher for the sentiment lexicons that account for negation. Looking at the probability-based approaches, we notice that the FS approach leads to a higher balanced accuracy and macro F_1 measure compared to the NW approach. However, the opposite is true for the ICF sentiment lexicons.

In Table 10, one finds the evaluation metrics of the financial sentiment lexicons on the headlines dataset. We notice again that the values of the SNN sentiment lexicons' evaluation metrics are dissimilar compared to the other sentiment lexicons. In the category with the original sentiment lexicons, we see that the BTB sentiment lexicon slightly outperforms the other lexicons in this category. In the second category, the NW sentiment lexicons, the W-NPMI sentiment lexicon slightly outperforms the other lexicons based on the balanced accuracy and macro F_1 metric. The W-PMI sentiment lexicon of the FS category performs slightly better than the other sentiment lexicons that belong to this category. Based on the balanced accuracy and the macro F_1 measure, we select the W-PMI sentiment lexicon of the FS category as our best performing sentiment lexicon on the headlines dataset. This sentiment lexicon has a balanced accuracy of 62.3% and a macro F_1 metric of 62.1%. In addition, we notice that accounting for negation while constructing the financial sentiment lexicons leads to an increase in the F_1 Pos measure. Looking at the balanced accuracy and the macro F_1 measure, we see that the FS lexicons outperform

Table 9 Evaluation metrics of the financial sentiment lexicons in unsupervised sentiment classification on the microblogging dataset

Original Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	74.1	1.4	75.1	70.6	81.8	60.9	71.3
W-PMI	74.0	1.3	74.9	70.6	81.6	60.9	71.2
W-NPMI	**75.4**	1.3	**76.4**	**72.0**	**82.7**	**62.9**	**72.8**
ICF	74.0	1.3	75.0	70.5	81.7	60.7	71.2
SNN	44.6	**1.2**	45.2	56.9	31.1	54.5	42.8
NW Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	74.8	1.3	75.8	71.2	82.3	61.6	72.0
W-PMI	74.1	**1.2**	75.0	70.5	81.7	60.7	71.2
W-NPMI	**75.8**	**1.2**	**76.8**	**72.1**	**83.1**	**62.9**	**73.0**
ICF	75.0	**1.2**	75.9	71.1	82.5	61.4	71.9
SNN	46.8	**1.2**	47.4	58.3	36.4	55.2	45.8
FS Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	75.3	1.4	76.3	71.4	82.8	61.9	72.4
W-PMI	75.1	1.3	76.1	71.4	82.5	61.9	72.2
W-NPMI	**76.2**	1.3	**77.2**	**72.5**	**83.4**	**63.6**	**73.5**
ICF	74.6	1.3	75.6	70.5	82.3	60.6	71.4
SNN	41.9	**1.2**	42.4	49.4	37.8	46.5	42.1

In this table, the best metrics per category are stated in bold. The overall best metrics across all the lexicons are underlined

Table 10 Evaluation metrics of the financial sentiment lexicons in unsupervised sentiment classification on the headlines dataset

Original Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	**65.3**	**0.0**	**65.3**	**61.6**	73.6	49.3	**61.4**
W-PMI	64.5	0.2	64.6	61.0	73.0	48.8	60.9
W-NPMI	**65.3**	**0.0**	**65.3**	61.5	**73.7**	48.9	61.3
ICF	64.6	**0.0**	64.6	60.9	73.2	48.0	60.6
SNN	47.0	**0.0**	47.0	54.1	23.7	**59.4**	41.6
NW Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	65.0	0.2	65.1	61.1	73.8	47.7	60.7
W-PMI	65.3	0.2	65.5	61.6	73.9	48.8	61.4
W-NPMI	**66.3**	0.2	**66.4**	**62.1**	**75.2**	48.0	**61.6**
ICF	64.8	0.2	64.9	61.0	73.6	47.5	60.6
SNN	48.3	**0.0**	48.3	53.5	35.2	**56.9**	46.1
FS Sentiment Lexicons							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
BTB	65.7	**0.0**	65.7	61.9	74.2	49.1	61.6
W-PMI	**66.2**	**0.0**	**66.2**	**62.3**	74.6	**49.6**	**62.1**
W-NPMI	66.0	**0.0**	66.0	61.6	**75.0**	46.9	61.0
ICF	65.8	**0.0**	65.8	61.9	74.4	48.7	61.5
SNN	47.6	**0.0**	47.6	49.3	47.1	48.1	47.6

In this table, the best metrics per category are stated in bold. The overall best metrics across all the lexicons are underlined

the original sentiment lexicons, except for the W-NPMI and SNN sentiment lexicons. Furthermore, we notice that the W-NPMI NW sentiment lexicon performs better than the other two categories' W-NPMI sentiment lexicons.

Overall, we notice that the probability-based approaches perform relatively better than the information retrieval-based and the sentiment-aware word embedding-based approaches. Moreover, we notice that our introduced weighted versions of the PMI approaches perform better than the other approaches. Furthermore, we observe that the quality could be improved by accounting for negation while building the sentiment lexicons. Looking at the balanced accuracy and the macro F_1 measure, we observe over the three test datasets that the SNN approach could be improved with the NW approach. The other four approaches could be improved by either the NW or FS approach, depending on the test dataset.

After selecting the best financial sentiment lexicon for each test set, we compare these financial sentiment lexicons with the external sentiment lexicons mentioned in the "Evaluation" section. In Table 11, one finds the evaluation metrics of the external lexicons. We notice that the manually made sentiment lexicons, such as the GI and LM lexicons, struggle with classifying the messages as either positive or negative. The LM sentiment lexicon is, on average, unable to classify approximately 70% of the test messages. The high percentages of unclassified messages confirm the disadvantage of the low coverage of manually made sentiment lexicons, as discussed in the "Related Work" section. Therefore, we only consider the external sentiment lexicons with a low and similar percentage of unclassified messages as the best financial sentiment lexicons. We consider the NRC-H lexicon [43] and the SM sentiment lexicon [14]. In addition, we notice that the evaluation metrics for the NRC-H sentiment lexicon are lower than the evaluation metrics of the SM sentiment lexicon and the best financial sentiment lexicon. This result is in line with our expectations because the NRC-H sentiment lexicon is, in contrast to the other two lexicons, not explicitly constructed for the financial domain.

Similar to the previous unsupervised comparisons, we focus on the balanced accuracy and the macro F_1. We see that the BTB and W-PMI NW sentiment lexicons outperform both considered external sentiment lexicons on the StockTwits dataset with a balanced accuracy of 73.5% and a macro F_1 measure of 73.0%. In the case of the microblogging dataset, the W-NPMI FS lexicon slightly outperforms the NRC-H sentiment lexicon and the SM sentiment lexicon. However, in the case of the headlines dataset, the SM sentiment lexicon slightly outperforms the W-PMI FS lexicon. Overall, we notice that the newly introduced W-PMI and W-NPMI sentiment lexicons, which are built while accounting for negation, perform very well, both internally and externally, compared to other sentiment lexicons.

Table 11 Evaluation metrics of the external sentiment lexicons in unsupervised sentiment classification on the test datasets

StockTwits							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
GI	36.4	38.7	59.3	59.3	60.4	58.2	59.3
MPQA	37.5	33.9	56.8	56.3	63.2	47.6	55.4
HL	34.2	46.1	63.4	63.4	63.1	63.8	63.4
NRC-H	55.0	0.6	55.4	55.4	53.0	57.5	55.2
NRC-E	39.9	22.5	51.5	52.0	42.1	58.3	50.2
VADER	39.3	34.9	60.4	60.5	64.3	55.4	59.9
LM	19.7	68.7	63.0	63.2	62.1	63.8	63.0
SenticNet	48.1	13.9	55.9	55.8	62.6	46.2	54.4
SM	65.3	2.0	66.7	66.6	70.5	61.6	66.1
BTB NW	**73.4**	**0.1**	**73.5**	**73.5**	**76.7**	69.2	**73.0**
W-PMI NW	**73.4**	**0.1**	**73.5**	**73.5**	**76.7**	**69.3**	**73.0**
Microblogging							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
GI	31.7	55.4	71.1	69.4	77.8	58.7	68.2
MPQA	37.8	45.9	69.7	68.2	76.2	58.3	67.3
HL	35.7	54.2	78.0	77.5	82.0	71.8	76.9
NRC-H	57.7	1.8	58.8	58.2	65.8	48.1	57.0
NRC-E	31.7	28.5	44.3	53.2	40.1	48.0	44.0
VADER	35.5	52.7	75.2	73.3	80.0	67.2	73.6
LM	20.3	73.5	76.4	77.8	78.4	74.0	76.2
SenticNet	54.3	22.3	70.0	64.2	78.4	50.9	64.6
SM	74.1	2.3	75.8	72.0	82.1	62.9	72.5
W-NPMI FS	**76.2**	**1.3**	**77.2**	**72.5**	**83.4**	**63.6**	**73.5**
Headlines							
Lexicon	ACC1	UNCL	ACC2	BA	F_1Pos	F_1Neg	Macro F_1
GI	36.8	47.4	70.0	70.1	72.2	67.4	69.8
MPQA	37.6	49.0	73.7	73.6	75.5	71.6	73.6
HL	31.0	58.8	75.2	74.7	71.6	78.0	74.8
NRC-H	56.0	**0.0**	56.0	55.5	60.8	49.8	55.3
NRC-E	48.1	7.2	51.8	54.5	48.9	54.3	51.6
VADER	46.8	35.6	72.7	72.4	75.4	69.4	72.4
LM	27.3	66.9	82.6	78.6	72.9	87.2	80.0
SenticNet	59.2	4.9	62.3	58.0	72.2	41.4	56.8
SM	**66.3**	0.2	**66.4**	**63.6**	73.6	**53.7**	**63.7**
W-PMI FS	66.2	**0.0**	66.2	62.3	**74.6**	49.6	62.1

In this table, the best metrics of the considered sentiment lexicons per test set are stated in bold

Supervised Sentiment Classification Evaluation

Next to the unsupervised sentiment classification, we also perform supervised sentiment classification. We use a linear classifier introduced by Fan et al. [47] for the supervised sentiment classification. First, we extract the six sentiment lexicon features for each message in the StockTwits test dataset, as described in the "Sentiment Classification Evaluation" section. Hereafter, we split the StockTwits test set into an 80% training set to train the linear classifier and a 20% test set to evaluate the sentiment lexicons using a fixed seed. The training set consists of 16,000 positive and 16,000 negative messages. The test set consists of 4,000 messages and 4,000 negative messages.

We perform a grid search on the accuracy to tune the type of classifier and the hyperparameter α on the five-fold cross-validation of the training set. We consider the six different types of multi-class classifiers described by Fan et al. [47]. We let the values for c vary from 0.0001 to 1000. The optimal classifier is the L2-regularized logistic regression, which is also the same type of classifier, as Tang et al. used [26].

Table 12 Evaluation metrics of the financial sentiment lexicons in supervised sentiment classification on the StockTwits test set

Original Sentiment Lexicons					
Lexicon	ACC	BA	F_1Pos	F_1Neg	Macro F_1
BTB	**74.4**	**74.4**	**75.4**	73.2	74.3
W-PMI	74.3	74.3	75.3	73.2	74.3
W-NPMI	**74.4**	**74.4**	75.3	**73.4**	**74.4**
ICF	**74.4**	**74.4**	**75.4**	73.3	**74.4**
SNN	66.1	66.1	66.1	66.2	66.1
NW Sentiment Lexicons					
Lexicon	ACC	BA	F_1Pos	F_1Neg	Macro F_1
BTB	74.9	74.9	76.0	73.8	74.9
W-PMI	75.0	75.0	76.0	73.9	74.9
W-NPMI	**75.1**	**75.1**	**76.1**	**74.1**	**75.1**
ICF	74.8	74.8	75.8	73.7	74.8
SNN	67.8	67.8	66.8	68.7	67.8
FS Sentiment Lexicons					
Lexicon	ACC	BA	F_1Pos	F_1Neg	Macro F_1
BTB	74.5	74.5	**75.6**	73.3	**74.5**
W-PMI	74.4	74.4	75.5	73.3	74.4
W-NPMI	**74.6**	**74.6**	**75.6**	**73.5**	**74.5**
ICF	74.4	74.4	75.4	73.2	74.3
SNN	50.8	50.8	50.2	51.4	50.8

In this table, the best metrics per category are stated in bold. The overall best metrics across all the lexicons are underlined

In Table 12, one finds the evaluation metrics of the financial sentiment lexicons on the StockTwits test set. In general, we see that the lexicons of the BTB, W-PMI, W-NPMI, and ICF approaches score similarly on the evaluation metrics across all three categories. This result is similar to the results we obtained in the unsupervised setting. The evaluation metrics of the SNN approach of the original and NW categories are more similar to the evaluation metrics of the other approaches. This similarity was not the case in the unsupervised sentiment classification. This difference indicates that the SNN sentiment lexicons' sentiment strengths are more suited for supervised sentiment classification compared to the unsupervised sentiment classification. A possible explanation for the SNN sentiment lexicons' lower performance in both sentiment classification tasks could be the number of training messages used in the neural network to compute the sentiment strengths. Neural networks tend to perform better when leveraging large datasets.

Similar to the unsupervised setting, we focus on the balanced accuracy and the macro F_1 measure. We notice that the W-NPMI and ICF sentiment lexicons slightly outperform the BTB and W-PMI sentiment lexicons in the original category. In the negation categories, the W-NPMI sentiment lexicons are slightly outperforming the other sentiment lexicons. In addition, by looking at the balanced accuracy and macro F_1 measure, we notice that the sentiment lexicons' quality could be improved by accounting for negation while building the sentiment lexicons. The balanced accuracy and macro F_1 measure of the NW sentiment lexicons are higher than the balanced accuracy and macro F_1 measure of the two other categories' sentiment lexicons. The balanced accuracy and macro F_1 measure of the FS lexicons are similar to the balanced accuracy and macro F_1 measure of the original sentiment lexicons. Overall, the W-NPMI NW sentiment lexicon performs slightly better than the other sentiment lexicons with a balanced accuracy of 75.1% and a macro F_1 measure of 75.1%.

Table 13 Evaluation metrics of the external sentiment lexicons and the best financial sentiment lexicon in supervised sentiment classification on the StockTwits test set

Lexicon	ACC	BA	F_1Pos	F_1Neg	Macro F_1
GI	55.6	55.6	60.0	50.1	55.0
MPQA	55.9	55.9	52.5	58.8	55.6
HL	57.7	57.7	63.3	50.3	56.8
NRC-H	56.7	56.7	55.4	57.9	56.7
NRC-E	53.1	53.1	47.3	57.7	52.5
VADER	57.2	57.2	64.5	46.2	55.3
LM	55.0	55.0	33.7	65.9	49.8
SenticNet	55.3	55.3	52.4	57.9	55.1
SM	66.5	66.5	66.8	66.3	66.5
W-NPMI NW	**75.1**	**75.1**	**76.1**	**74.1**	**75.1**

In this table, the best metrics are stated in bold

After selecting the best financial sentiment lexicon for the StockTwits test set in the supervised sentiment classification setting, we compare this financial sentiment lexicon with the external sentiment lexicons mentioned in the "Evaluation" section. Table 13 shows the external sentiment lexicons' evaluation metrics and the best financial sentiment lexicon on the StockTwits test set.

In general, the balanced accuracy and macro F_1 measure of all the external sentiment lexicons are very similar, except for the SM lexicon. Based on the balanced accuracy and the macro F_1 measure, the SM lexicon is the best performing external sentiment lexicon with a balanced accuracy of 66.5% and a macro F_1 metric of 66.5%. However, the W-NPMI NW sentiment lexicon has significantly higher values for the balanced accuracy and the macro F_1 measure compared to all the external sentiment lexicons.

Conclusion

The financial domain is currently lacking specific sentiment lexicons. In this research, we discuss several approaches to build financial sentiment lexicons automatically. We introduce two new approaches to automatically build these sentiment lexicons, namely the W-PMI and W-NPMI approach. Furthermore, we propose two different methods to account for negation while building the sentiment lexicons. The first method, the NW approach, creates a separate entry in the lexicon for the word's negated version. The second method, the FS approach, considers the negated word to have a sentiment orientation that is the opposite of the message's sentiment. This way, the method corrects for the negation without creating a new entry for the negated version of the word in the sentiment lexicon. We evaluate the constructed sentiment lexicons in two different sentiment classification tasks by comparing them with each other and with external sentiment lexicons created by other researchers.

The first sentiment classification task is done by evaluating the sentiment lexicons in an unsupervised setting across three different test sets. The test sets consist of StockTwits messages, microblogging messages, and financial headlines. In this unsupervised setting, we focused on the balanced accuracy and the macro F_1 measure. We noticed that the probability-based approaches achieved relatively higher metrics compared to the other types of approaches. Furthermore, we noticed that the sentiment lexicons achieve higher scores for the evaluation metrics if they account for either of the two proposed negation approaches while building the sentiment lexicons. Moreover, we noticed that the W-PMI and W-NPMI sentiment lexicons outperform all the internal and external sentiment lexicons in the unsupervised sentiment classification task.

In the second sentiment classification task, we evaluate the financial sentiment lexicons in a supervised setting. Again, we noticed that the quality of the sentiment lexicons could be improved by accounting for negation while building them. The W-NPMI NW sentiment lexicon slightly outperforms the other financial sentiment lexicons. Moreover, we observed that the W-NPMI NW sentiment lexicon has significantly higher scores for the evaluation measures compared to all the external sentiment lexicons. Furthermore, we noticed that the financial sentiment lexicons that take into account negation by using the NW approach achieve higher scores for the evaluation measures on the test set.

In the considered sentiment classification tasks (unsupervised and supervised), the probability-based approaches outperformed the other types of approaches. We compared building the financial sentiment lexicons while accounting for negation using the NW approach and FS approach to the baseline, not accounting for negation. The sentiment lexicons can be improved by accounting for negation while building the sentiment lexicons using either the proposed NW approach or the FS approach. In general, the financial sentiment lexicon obtained using the proposed W-NPMI approach and the NW approach performs best.

The constructed financial sentiment lexicons could be further improved in different ways. In this research, we focused on the explicit negation cues, such as "not" and "never." A possible future research direction is to focus, next to the explicit negation cues, on the implicit negation cues, such as "dislike" and "hopeless." Furthermore, the financial sentiment lexicons could be improved by taking into account intensifiers, such as "really" and "very," and downtoners, such as "hardly" and "slightly," while constructing the sentiment lexicons. In addition, the sentiment lexicons could be refined by taking into account emoticons and emojis, which are becoming more and more popular in microblogs [52, 53]. Last, we plan to apply the introduced W-PMI and W-NPMI sentiment building approaches together with accounting for negation on other domains, such as the consumer product domain.

Acknowledgements We want to thank StockTwits for their help and for providing access to their back-up database.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors

References

1. Liu B. Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press; 2015.
2. Hu M, Liu B. Mining and summarizing customer reviews. In: 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 2004. p. 168–177.
3. Cambria E, Hussain A, Havasi C, Eckl C. Sentic computing: exploitation of common sense for the development of emotion-sensitive systems. Development of Multimodal Interfaces: Active Listening, and Synchrony, Lecture Notes in Computer Science. Springer, Cambria, Erik and Li, Yang and Xing. Soujanya and Kwok, Kenneth: Frank Z. and Poria. 2010. vol. 5967, p. 148–156.
4. Feldman R. Techniques and applications for sentiment analysis. In: Communications of the ACM. 2013. vol. 56, p. 82–89.
5. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. In: Journal of Computational Science. Elsevier 2011. vol. 2, p. 1–8.
6. Montoyo A, MartíNez-Barco P, Balahur A. Subjectivity and sentiment analysis: an overview of the current state of the area

and envisaged developments. In: Decis Support Syst. Elsevier 2012. vol. 53 (4), p. 675–679.

7. Xing FZ, Cambria E, Welsch RE. Natural language based financial forecasting: a survey. Artif Intell Rev. 2018;50(1):49–73.
8. Malandri L, Xing FZ, Orsenigo C, Vercellis C, Cambria E. Public mood-driven asset allocation: the importance of financial sentiment in portfolio management. Cogn Comput. 2018;10(6):1167–76.
9. Xing FZ, Cambria E, Welsch RE. Intelligent asset allocation via market sentiment views. IEEE Comput Intell Mag. 2018;13:25–34.
10. Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications. 2019;135:60–70.
11. Weng B, Lu L, Wang X, Megahed FM, Martinez W. Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems with Applications. 2018;112:258–73.
12. Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. In: J Financ, vol. 66, p. 35–65. Wiley Online Library (2011)
13. Oliveira N, Cortez P, Areal N. Automatic creation of stock market lexicons for sentiment analysis using stocktwits data. In: 18th International Database Engineering & Applications Symposium. ACM 2014. p. 115–123.
14. Oliveira N, Cortez P, Areal N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. In: Decis Support Syst. Elsevier 2016. vol. 85, p. 62–73.
15. Al-Twairesh N, Al-Khalifa H, AlSalman A. AraSenTi: Large-scale twitter-specific arabic sentiment lexicons. In: 54th Annual Meeting of the Association for Computational Linguistics ACL. 2016. vol. 1, p. 697–705.
16. Liu B. Sentiment analysis and opinion mining. Morgan & Claypool Publishers (2012)
17. Stone PJ, Dunphy DC, Smith MS. The General Inquirer: a computer approach to content analysis, vol. 8. MIT Press; 1966.
18. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing. ACL 2005. p. 347–354.
19. Jegadeesh N, Wu D. Word power: a new approach for content analysis. In: J Financ Econ. Elsevier 2013. vol. 110, p. 712–729.
20. Heerschop B, Hogenboom A, Frasincar F. Sentiment lexicon creation from lexical resources. In: 14th International Conference on Business Information Systems (BIS 2011), Lecture Notes in Business Information Processing. Springer 2011. vol. 87, p. 185–196.
21. Miller GA. WordNet: a lexical database for english. In: Communications of the ACM. 1995. vol. 38, p. 39–41.
22. Turney PD. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: 40th Annual Meeting on Association for Computational Linguistics. ACL 2002. p. 417–424.
23. Labille K, Alfarhood S, Gauch S. Estimating sentiment via probability and information theory. In: 8th International Conference on Knowledge Discovery and Information Retrieval KDIR. SciTePress 2016. p. 121–129.
24. Labille K, Gauch S, Alfarhood S. Creating domain-specific sentiment lexicons via text mining. In: 6th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining. WISDOM 2017.
25. Li Q, Shah S. Learning stock market sentiment lexicon and sentiment-oriented word vector from Stocktwits. In: 21st Conference on Computational Natural Language Learning. CoNLL 2017, p. 301–310.
26. Tang D, Wei F, Qin B, Zhou M, Liu T. Building large-scale twitter-specific sentiment lexicon : a representation learning approach. In: 25th International Conference on Computational Linguistics. ACL 2014, p. 172–182.
27. Wang L, Xia R. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In: 2017 Conference on Empirical Methods in Natural Language Processing. ACL 2017. p. 502–510.
28. Vo DT, Zhang Y. Don't count, predict! An automatic approach to learning sentiment lexicons for short text. In: 54th Annual Meeting of the Association for Computational Linguistics. ACL 2016. vol. 2, p. 219–224.
29. Xing FZ, Pallucchini F, Cambria E. Cognitive-inspired domain adaptation of sentiment lexicons. Inf Process Manag. 2019;56(3):554–64.
30. Bayes T, Price R. An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. In: Philosophical Transactions of the Royal Society of London. 1763. vol. 53, p. 370–418.
31. Church KW, Hanks P. Word association norms, mutual information, and lexicography. In: 27th Annual Meeting on Association for Computational Linguistics. ACL. 1989. vol. 16, p. 22–29.
32. Fano RM. Transmission of information: a statistical theory of communication. In: Am J Phys. MIT Press. 1961. vol. 29, p. 793–794.
33. Bouma G. Normalized (pointwise) mutual information in collocation extraction. From Form to Meaning: Processing Texts Automatically. 2009. p. 31–40.
34. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. In: Inf Process Manag. Pergamon Press, Inc. 1988. vol. 24, p. 513–523.
35. Wang D, Zhang H. Inverse-category-frequency based supervised term weighting schemes for text categorization. J Inf Sci Eng. 2013;29:209–25.
36. Lan M, Tan CL, Su J, Lu Y. Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans Pattern Anal Mach Intell. 2009;31:721–35.
37. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations. ICLR 2013, p. 1–12.
38. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: 2014 Empirical Methods in Natural Language Processing. ACL 2014. p. 1532–1543.
39. Hogenboom A, Van Iterson P, Heerschop B, Frasincar F, Kaymak U. Determining negation scope and strength in sentiment analysis. In: IEEE International Conference on Systems, Man, and Cybernetics. 2011. p. 2589–2594.
40. Jia L, Yu C, Meng W. The effect of negation on sentiment analysis and retrieval effectiveness. In: 18th ACM Conference on Information and Knowledge Management. CIKM 2009. p. 1827–1830.
41. Councill IG, McDonald R, Velikovich L. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In: Workshop on Negation and Speculation in Natural Language Processing. ACL. 2010. p. 51–59.
42. Morante R, Liekens A, Daelemans W. Learning the scope of negation in biomedical texts. In: 2008 Conference on Empirical Methods in Natural Language Processing. ACL 2008. p. 715–724.
43. Mohammad SM, Kiritchenko S, Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: 7th International Workshop on Semantic Evaluation. ACL. 2013.
44. Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: 8th International AAAI Conference on Weblogs and Social Media. ICWSM 2014. p. 216–225.
45. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. SenticNet 6: ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: 29th ACM International Conference on

Information & Knowledge Management. ACM CIKM. 2020. p. 105–114.

46. Zhu X, Kiritchenko S, Mohammad S. NRC-Canada-2014: recent improvements in the sentiment analysis of tweets. In: 8th International Workshop on Semantic Evaluation. ACL 2014. p. 443–447.
47. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LibLineaR: a library for large linear classification. In: J Mach Learn Res. JMLR.org. 2008. vol. 9, p. 1871–1874.
48. Cortis K, Freitas A, Daudert T, Huerlimann M, Zarrouk M, Handschuh S, Davis B. SemEval-2017 task 5: fine-grained sentiment analysis on financial microblogs and news. In: 11th International Workshop on Semantic Evaluation. ACL. 2017. p. 519–535.
49. Kwartler T. Text mining in practice with R. John Wiley & Sons; 2017.
50. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. J Stat Softw. 2008;25(5):1–54.
51. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. In: Final Projects from CS224N for Spring 2008/2009 at the Stanford Natural Language Processing Group. Stanford University 2009. p. 1–6.
52. Hogenboom A, Bal D, Frasincar F, Bal M, De Jong F. Kaymak U. Exploiting emoticons in polarity classification of text. In: Journal of Web Engineering (JWE). Rinton Press 2015, vol. 14 (1 & 2), p. 22–40.
53. Hogenboom A, Bal D, Frasincar F, Bal M, de Jong F, Kaymak U. Exploiting emoticons in sentiment analysis. In: 28th Symposium on Applied Computing. SAC 2013, p. 703–710.