Zuzana JANKOVÁ, PhD (corresponding author)
zuzana.jankova@vutbr.cz
Brno University of Technology, Brno, Czech Republic

Nikoloz KAVELASHVILI, PhD
nikoloz.kavelashvili@iliauni.edu.ge
Ilia State University, Tbilisi, Georgia

Sebastian ESCHENBACH, PhD
sebastian.eschenbach@fh-wien.ac.at
University of Applied Sciences for Management & Communication, Vienna, Austria

Measuring Investor Sentiment in Financial Discourse: How Different Approaches Shape Market Signals

Abstract. *Stock prices are shaped not only by fundamental data but also by investor sentiment, which often deviates from rational decision-making. Given the vast volume of financial texts published by both professional and amateur investors—especially on online financial platforms—sentiment analysis in such unstructured data is essential to understanding their impact on market movements. This study examines the interplay between text data and stock market movements, highlighting the critical role of sentiment extracted from financial news and online news. Existing research has largely relied on general-purpose lexicons or uniform classification techniques, which limits the accuracy of sentiment analysis in financial contexts. To address this gap, we propose a hybrid framework that integrates domain-specific lexicons with advanced machine learning classifiers to improve sentiment extraction from unstructured financial text. Our approach evaluates the impact of lexicon selection on sentiment scores and examines the relationship between classifier choice and prediction accuracy. By improving sentiment analysis methodologies, our findings contribute to the development of more robust stock market forecasting models, strengthen decision-making processes for investors, and increase market efficiency.*

Keywords: *financial social media, investor sentiment, machine learning, StockTwits, sentiment analysis, textual analysis.*

JEL Classification: C39, C55, G17.

Received: 17 June 2025	**Revised: 8 December 2025**	**Accepted: 12 December 2025**

1. Introduction

Financial markets occupy a prominent position in modern society. Successfully predicting the price movements of financial instruments can potentially prevent the harmful effects that an impending financial crisis could have on everyday life. Stock markets in particular exhibit a high degree of chaotic and unexpected movements, and

DOI: 10.24818/18423264/59.4.25.05

their development is the subject of interest of many researchers. As further stated by Hao et al. (2021), the aspiration of every investor is to accurately predict market behaviour aimed at making the best decisions regarding the purchase or sale of stocks seeking to maximise their profits and reduce unexpected risks. This is a difficult task because market behaviour is stochastic, volatile, and influenced by many factors such as the global economy, politics, investor expectations, and others.

The more accurately a system predicts stock price movements, the more profit can be obtained from the prediction model. Stock price forecasting is very popular based on technical and fundamental data analysis alone. However, numerical time series data only contain the event and not the cause of its occurrence. Textual data such as newspaper articles have richer information, and therefore, using textual information, especially in addition to numerical time series data, increases the quality of the input, and better predictions are expected from this type of input, rather than from numerical data alone. Without a doubt, human behaviour is always influenced by its environment. One of the most significant impacts that influence people's behaviour comes from mass media, more specifically from news articles. On the other hand, price movements in stock markets are the result of investors' actions on how they perceive the events surrounding them, including in stock markets. Unstructured data in the form of digitised text is growing rapidly in terms of volume, availability, and relevance for innovation research and beyond. Given the recent advances in social network analysis, it has become possible to include this type of information as a predictive input in forecasting models. Based on the literature review, it can be stated that stock prices are influenced by both published fundamental information and thought processes in the minds of individual market participants (which usually do not arise rationally). These two influences intertwine and operate simultaneously. It would therefore be interesting to investigate whether and what connection there is between texts and stock market movements. Texts can express both fundamental facts (rationality) and people's emotions and opinions (irrationality), and it is possible to examine the influence of investors' (i)rationality on the stock market. For this purpose, it is necessary to design a suitable system that is able to extract sentiment from text data and correctly classify it according to polarity, thereby facilitating their investment decisions. This benefit from investing is important not only for investors themselves, but also for the economy as a whole. This paper argues, based on secondary research, that the use of sentiment analysis to predict stock market movements is not sufficiently developed and that further research is needed, in particular the creation of a suitable model for extracting sentiment from unstructured financial data. This research is therefore focused on the use of hybrid systems to identify possible relationships between textual information and, consequently, the thoughts and opinions of investors extracted from it, and their suitability for use within the financial domain. A critical review of the literature revealed that researchers mostly rely exclusively on one dictionary or use only one method or technique to classify sentiment polarity. Palmer et al. (2022) argue that more intensive development of domain-specific dictionaries is needed, since domain-specific dictionaries are tailored to a specific topic or profession, yet many studies still use general-purpose dictionaries. Sharma & Dutta (2021) state that it is difficult to achieve satisfactory performance in specific and concrete topics and areas. Loughran & McDonald (2016) further add that there is usually no specific dictionary readily available for a particular research context; moreover, its development requires a lot of effort. The development of a framework for creating domain-specific sentiment lexicons and their appropriate selection could play a significant role in the mining of financial texts, respectively. sentiment, which is subsequently used as a predictor of its development (Li et al., 2021a).

In particular, this part of the research on working with text documents and sentiment evaluation plays a key role. An incorrectly chosen dictionary that does not cover the researched issue in detail can provide incorrect outputs that are subsequently worked with, and false conclusions are drawn. The main goal is to classify sentiment scores from unstructured text financial reports and contributions published online using multiple dictionary approaches and machine learning methods. To fulfil these goals, the following research questions will be answered. An important prerequisite for building a classifier is the labelling of input data as positive or negative. Automatic labelling requires the alignment of text data sources with the appropriate dictionaries. As a result, these factors affect the characteristics of the training data set. They determine not only the quality of labelling for positive and negative words, but also the number of word labels in uniform categories. This in turn affects the expected quality of the prediction. The following research question arises: RQ1: What effect does the lexicon of positive and negative words have on sentiment scores?

Predicting stock prices using text mining techniques requires achieving the highest possible prediction performance. This includes the challenge of choosing the right binary classifier. The success of such a classifier largely depends on the quality of the input data. It is necessary to make the right choice of classifier suitable for the specific goal. Therefore, an important research question needs to be asked: RQ2: How is the choice of binary classifier related to the accuracy of the sentiment score calculation with respect to the chosen lexicon?

The paper is organised as follows: Section 1 outlines the relevant research. Section 2 describes the research methodology chosen to meet the objective. Section 3 provides a description of the main research results. Section 4 summarises the findings, discusses the empirical results, and provides final recommendations and answers the research questions.

2. Literature review

Sentiment analysis focuses on detecting sentiment polarity. For this purpose, a number of techniques and methods can be used, which fall into three groups: i) dictionary approach, ii) machine learning, and iii) deep learning. The sentiment lexicon, i.e. the dictionary approach, calculates the sum of positive and negative words in a text (Li et al., 2021b) and traditionally deals with the creation of dictionaries by manually marking the polarity of words. Sekhar et al. (2022) investigates stock volume and price prediction by combining sentiment analysis and technical indicators. They use sentiment analysis from Twitter using the

SentiWordNet dictionary and an LSTM prediction model for sequential data. The results suggest that integrating sentiment data with technical analysis can improve the accuracy of stock market predictions. Bi et al. (2022) use text mining and sentiment analysis to predict stock market performance based on financial news. Sentiment is measured using the Loughran-McDonald dictionary and combined with technical indicators. Machine learning models have shown that positive sentiment correlates with stock price growth, while negative sentiment increases volatility. Adding sentiment to classical predictive models increased prediction accuracy by up to 12%. Kirtac Germano (2024) use large language models (LLMs) such as GPT to analyse sentiment in financial news and social media. The results show that negative sentiment leads to sharper declines in stock prices, while positive news has a milder effect on growth. According to the authors, the combination of LLM and quantitative strategies allows for more effective trading decisions. Basak et al. (2023) evaluate how political and economic changes associated with Brexit affected markets and how the media contributed to shaping public sentiment and investment decisions during this turbulent period. Their results suggest that media sentiment acted as an indicator for investors, who adjusted their portfolios based on these signals, which led to market fluctuations. Fang et al. (2021) found that platforms like Twitter and Reddit were identified as crucial sources of sentiment data, demonstrating how retail investors' collective sentiment impacts market movements. According to the authors, news sentiment was shown to have a measurable influence on stock volatility, especially during periods of high uncertainty or market stress. Chu et al. (2023) investigate the relative impact of investor sentiment during two key periods: the overnight phase (when markets are closed) and the trading-hour period (when markets are open), on forecasting stock market volatility. The research finds that the sentiment generated during the overnight period, when markets are closed, has a stronger predictive relationship with stock market volatility than sentiment observed during trading hours. This suggests that overnight sentiment can act as a precursor to market movements, influencing investor behaviour once the market reopens.

The dictionaries created are of high quality but mostly compiled for a specific professional domain. This implies that dictionaries created for a specific domain use special vocabulary and may not be suitable for other domains. A number of researchers (Rizinski et al., 2024) thus emphasise the domain dependence in sentiment analysis and can lead to incorrect classification. Research by Catelli et al. (2022) revealed that the dictionary approach is preferable to, for example, deep learning methods in the case of limited computational resources or low-resource languages. The literature review also shows that the vast majority of researchers use only one annotated lexicon in sentiment analysis, such as Hao et al. (2021) and many others. Focusing exclusively on one dictionary can significantly distort the resulting sentiment score. An incorrectly determined sentiment score can subsequently distort the impact on stock markets and therefore cause significant losses to investors when implemented into an investment strategy, as it can ultimately indicate incorrect buy and sell signals. For this reason, the paper focuses on multiple dictionaries of positive and negative words (RQ1). Also, from the literature review, some authors claim

(Feuerriegel & Gordon, 2018) that special financial dictionaries tend to more accurately calculate investor sentiment in stock markets. In the afternoon series, the authors rely on various classifiers for sentiment analysis, which play a central role in the correct classification of sentiment. This paper follows a similar approach and examines the impact of the choice of binary classifier on the accuracy of sentiment score calculation (RQ2).

3. Model specification

3.1 Support Vector Machine

Support vector machine (SVM) searches for the best hyperplane among all possible ones. SVM can be used for linear and nonlinear binary classification and more recently for multiple classification and regression. This machine learning approach is based on isolating data by defining a hyperplane with maximum margin. The data in the classifier can be labelled as follows:

$$x_i w + b > 0, c_i = 1$$
$$x_i w + b < 0, c_i = 0$$

where x, w and b can be denoted as feature vector, weight vector and bias of the input, respectively. c_i identifies the input class label. The hyperplane can be calculated by decreasing w, which causes the separation bound to increase. The problem area of classification using SVM is according to Wang & Zhu, (2023):

$$min \frac{1}{2} \|w\|^2$$
$$subject\ to\ c_i(x_i, w + b) \geq 1$$

3.2 Machine Naïve Bayes

The naive Bayes classifier assumes that features are mutually independent and equally important. It is based on probability theory and Bayes' rule can be defined as follows:

$$P(c|t) = \frac{P(c)P(t|c)}{P(t)}$$

where c is a particular class, either negative or positive, t is the classified text, $P(c)$ and $P(t)$ are the prior probabilities of class c and text t. $P(t|c)$ is the probability of the text occurring in a particular class.

3.3 K-nearest neighbour

The k-nearest neighbour classifier classifies data into the same class as the nearest data points. The data is divided into a test dataset containing unclassified data, while the training dataset contains already categorised data. Based on the Euclidean distance d, which determines the difference or similarity or the shortest distance between two data points (x_1, y_1) and $(x_2 . y_2)$ to the formula:

$$d\big((x_1, y_1,), (x_1, y_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

For a specific value of k, k-NN determines a data point and assigns it to the base class with the largest number of data points according to k-neighbours. After calculating the distance, the input x is assigned to the class with the highest probability according to the relationship (Koukaras et al., 2022):

$$P(y = j|X = x) = \frac{1}{K}\sum_{i\in A} I\big(y^{(i)} = j\big)$$

3.4 Decision Trees

Another type of supervised machine learning method is the decision tree, which is referred to as a hierarchical model. A decision tree categorises data through inductive learning from previously known classes. Its tree structure contains leaf nodes representing class labels and internal nodes representing algorithm conditions.

3.5 Wilcoxon test

The Wilcoxon rank-sum test is a nonparametric counterpart of the paired test, which compares the differences in the performance of two models over each data set. The test essentially compares positive and negative differences. The differences are ranked based on their absolute values, and, in case of a match, the average ranks are calculated. Let the difference between the performance scores of the two models on the i-th of the data sets be. Let R^+ be the sum of the ranks for the data sets where the second model outperforms the first, and vice versa, R^-, as noted by Trawiński et al. (2012):

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

Ranks where $d_i = 0$ are divided in half and added to the sum. If T denotes a smaller sum, i.e. $T = \min(R^+, R^-)$, the z-statistic:

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

For larger numbers of data sets, such as larger than 25, it will be approximately normally distributed. The steps of the Wilcoxon test are as follows: Let Y denote the observed value, M_1 denote the predicted value of the first model, and M_2 denote the predictive value of the second model. H_0: two models generate equally accurate results; H_1: one model is more accurate than the other. Then the decision whether to reject the null hypothesis or not is based on the resulting p-value. If the p-value is greater than the chosen significance level, the null hypothesis is not rejected. In this case, a one-sided confidence interval is used.

4. Results and discussion

4.1 Textual data description

Texts can contain primarily factual (e.g., financial results) or opinion (e.g., positive or negative relationship to a company or product) information, or both. In the latter case, we can examine whether the movement of a share price is related to the emotions, moods, and opinions of people trading on stock markets, or whether and how they are influenced by them. For the purposes of this article, the social network Stocktwits is chosen. Stocktwits is the largest social network for investors and traders with more than five million community members and millions of monthly visitors. An extensive dataset is freely available from the authors Jaggi et al. (2021). Figure 1 shows a long-term growing trend in the number of postings of thoughts, opinions, and attitudes regarding the development of stock markets. The largest number of posts was published on February 4, 2020, when an incredible 24,285 posts appeared in one day. The second largest number of 19,676 posts was published on February 5, 2020, the very next day. In general, 2020 can be considered a breakthrough in terms of the number of opinions shared by the platform's users. This can be considered logical, because with the ongoing pandemic situation, expanding computer literacy, and the closure of economies, many users tended to share their opinions on future developments, which were quite unclear. Figure 2 helps visualise the density of post activity. The lighter the colour, the more posts are published on a given day and at a given hour. The greatest activity is evident in the afternoon hours on working or business days. While on the weekend, regardless of the hour, the activity is negligible compared to other days. This is evident from the blue highlighting for both non-business days, when there is no transition to a lighter colour, which indicates a higher volume of posts. It is also necessary to add that the US Eastern Time is considered, when the posts are published.

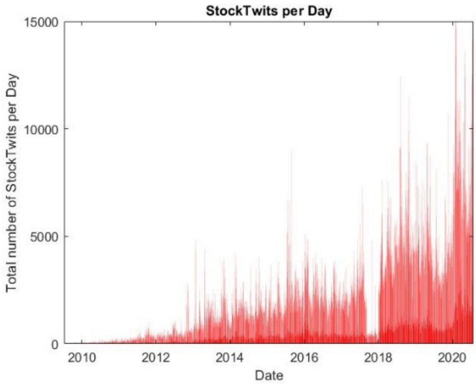

Figure 1. Number of StockTwits in the dataset
Source: Authors' own creation.

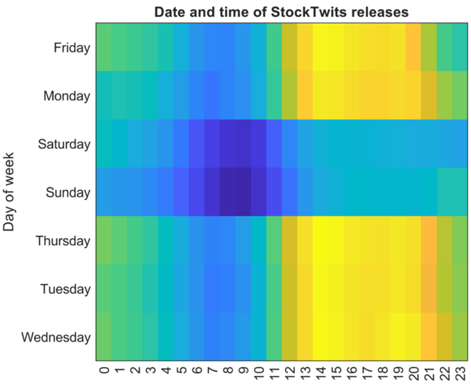

Figure 2. Date and time of posting on StockTwits
Source: Authors' own creation.

Figures 3 and 4 respectively visualise a simplified graph representing the connection between the keywords “Bullish” and “Bearish”, which indicate a bullish and bearish market, respectively. Additionally, as stated in the detailed description of the post published on Stocktwits, these labels are used to detect sentiment indicators that individual users of this platform use to indicate the development of the stock market. The nodes in the graph indicate individual words that are connected via edges and weighted by the cosine distance to the keyword. The value 50 is set for visualisation, which means that the 50 closest word combinations to the keyword are depicted.

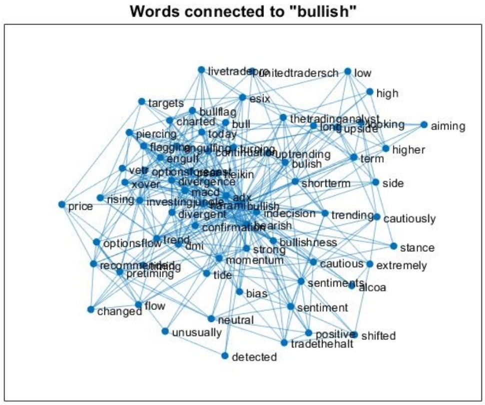
Words connected to "bullish"

Figure 3. Visualisation of words connected to the keyword "bullish"
Source: Authors’ own creation.

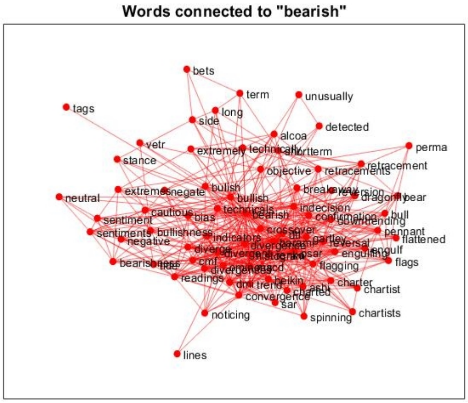
Words connected to "bearish"

Figure 4. Visualisation of words connected with the keyword "bearish"
Source: Authors’ own creation.

According to literature such as Nti et al. (2020), it is recommended to use a ratio of 80% for training and 20% for testing of a machine learning classifier. This ratio is used to define the training and testing sets for validating the statistical model using cross-validation.

4.2 Lexicons of positive and negative words

Several types of lexicon are selected for sentiment analysis, as the selection of an appropriate lexicon is essential for proper sentiment analysis. As a result of the critical review, almost all relevant studies use one to two lexicons for sentiment extraction from text data with subsequent application in the stock market. A total of six lexicons are selected for this research. In terms of general dictionaries, the psychological dictionary Harvard General Inquirer IV-4 (HGIV-4) available at http://www.wjh.harvard.edu/~inquirer/ is selected. The Opinion Lexicon by Hu and Liu (2004) obtained from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. VADER sentiment lexicon by Hutto and Gilbert (2014), which is directly integrated in the used MATLAB software, and AFINN by Nielsen (2011). Given the subsequent application of sentiment with a focus on the financial area, a set of financial dictionaries is also chosen, such as the Loughran-McDonald lexicon by the same-named authors Loughran and McDonald (2011) available at https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20 Lists. Finally, the FinanceSentiment Lexicon is chosen, which is also fully integrated in MATLAB. It is assumed that the last two lexicons should have higher accuracy when applied to text data from financial portals and posts on StockTwits.

4.3 Training and testing classifiers

Several machine learning techniques are selected for training and testing. Specifically, support vector machines, decision trees, naive Bayes, k-nearest neighbours, neural networks, and generalised additive models are selected for classification. These methods were selected based on a critical literature review in Janková (2023).

Support Vectore Machine

First, the support vector machine (SVM) method is chosen for sentiment classification, which classifies word vectors into positive and negative categories. Evaluation metrics are then calculated from the confusion matrix, which are used to compare the selected classification methods. Table 1 shows that the most accurate sentiment classification using SVM is achieved for the Opinion Lexicon dictionary with an accuracy of 96.86%, followed by AFINN and Harvard IV-4 Lexicon with an accuracy of 96.55% and 95.18%, respectively. However, very favourable results were also achieved by the FinanceSentiment Lexicon and VADER dictionaries. Sentiment was classified worst in the special dictionary for the financial area Loughran-McDonald. Another important evaluation indicator is the classification loss, which is a measure of the quality of generalisation. Its interpretation depends on the loss function and the weighting scheme; however, generally better classifiers yield smaller loss values. The Loughran-McDonald lexicon shows an unusually high loss value. Other dictionaries show an acceptable loss of around 3-8%.

Table 1. Evaluation metrics of classification via SVM

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.9552	0.9435	0.9493	0.9686	0.0312
Loughran-McDonald Lexicon	0.2206	0.2376	0.2288	0.2395	0.7599
Harvard IV-4 Lexicon	0.9496	0.9531	0.9514	0.9518	0.0482
FinanceSentiment Lexicon	0.9846	0.9410	0.9623	0.9313	0.0696
VADER	0.9114	0.9061	0.9087	0.9220	0.0782
AFINN	0.9372	0.9728	0.9547	0.9655	0.0325

Source: Authors' processing.

Decision Trees

Decision trees also allow for sentiment polarity classification. Decision trees provide lower accuracy, or higher classification error rate than SVM. Specific numerical data of evaluation metrics can be read from Table 2. The highest accuracy was achieved by the special financial dictionary FinanceSentiment Lexicon with an accuracy value of 83.52%, while on the other hand, the second analysed financial dictionary Loughran-McDonald Lexicon shows a classification accuracy of 20.19%. Other general dictionaries are able to recognise polarity at 73-80%. While the size of the loss for dictionaries ranges from 16-26%.

Table 2. Evaluation metrics for classification through decision trees

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.6517	0.6616	0.6566	0.7899	0.2076
Loughran-McDonald Lexicon	0.2017	0.2092	0.2053	0.2019	0.7981
Harvard IV-4 Lexicon	0.7410	0.7437	0.7423	0.7446	0.2553
FinanceSentiment Lexicon	0.9105	0.9049	0.9077	0.8352	0.1662
VADER	0.6956	0.6837	0.6896	0.7332	0.2676
AFINN	0.6859	0.7661	0.7238	0.7972	0.1951

Source: Authors' processing.

Naive Bayes

After training the model on the training set, the naive Bayes model is tested. Similar to the previous two classifiers, a table of evaluation metrics is created for the naive Bayes; see Table 3. Precision can be interpreted as the number of positive words that were actually classified as positive. In this case, the indicator reaches more than 85%. Recall indicates how many words were correctly captured by the classifier. The F1 metric combines both indicators described above, the closer the value is to 1, the better the polarity classification is achieved. The classification accuracy is very similar to that of the SVM and is around 90%. In the case of Loughran-McDonald, the classification is correct only by 50%. This means that in half of the cases, the naive Bayes was not able to correctly classify the polarity of the words contained in this dictionary.

Table 3. Evaluation metrics for classification via Naïve Bayes

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.8731	0.8864	0.8797	0.9264	0.0726
Loughran-McDonald Lexicon	0.4475	0.5059	0.4749	0.4995	0.5046
Harvard IV-4 Lexicon	0.8885	0.9114	0.8998	0.9018	0.0981
FinanceSentiment Lexicon	0.8673	0.9590	0.9109	0.8489	0.1514
VADER	0.8709	0.8294	0.8496	0.8686	0.1313
AFINN	0.8848	0.9037	0.8942	0.9189	0.0788

Source: Authors' processing.

K-nearest neighbours

Another classification method that is also used in research studies is the k-nearest neighbours' method. However, it can be stated at the outset that this method is not very widespread, unlike the previous ones that are chosen for classification in this work. The indicators are further calculated from the confusion matrix and are entered in Table 4. It can be seen from the table that the k-nearest neighbours classifier shows better performance than decision trees; however, this classifier is not able to overcome SVM and naive Byes. However, it is necessary to state that it does not lag behind these methods much; on the contrary, it provides acceptable performance in terms of high accuracy and low error rate.

Table 4. Evaluation metrics of classification via k-NN

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.8507	0.8529	0.8518	0.9087	0.0902
Loughran-McDonald Lexicon	0.2647	0.2525	0.2585	0.2234	0.7778
Harvard IV-4 Lexicon	0.8849	0.8723	0.8786	0.8786	0.1215
FinanceSentiment Lexicon	0.9383	0.9354	0.9368	0.8874	0.1136
VADER	0.8478	0.8255	0.8365	0.8588	0.1414
AFINN	0.8743	0.9126	0.8930	0.9189	0.078

Source: Authors' processing.

Neural networks

The first fully connected layer of a neural network has connections from the network input, and each subsequent layer has connections from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. Each fully connected layer is followed by an activation function. The final fully connected layer and the subsequent softmax activation function produce the network output, namely the classification score and the predicted sentiment polarity label. The outputs of the neural network-based classifier are shown in Table 5 comparing the performance using evaluation metrics. The generated outputs show that the neural networks perform comparable to those of SVM and naive Bayes. The classification accuracy is around 90% with a very low error rate of around 8%.

Table 5. Evaluation metrics for classification via neural networks

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.9129	0.9107	0.9118	0.9456	0.0538
Loughran-McDonald Lexicon	0.3151	0.3247	0.3198	0.3147	0.6853
Harvard IV-4 Lexicon	0.9137	0.9104	0.9120	0.9125	0.0875
FinanceSentiment Lexicon	0.9506	0.9506	0.9506	0.9121	0.0886
VADER	0.8690	0.8558	0.8623	0.8818	0.1185
AFINN	0.9005	0.9297	0.9149	0.9351	0.0625

Source: Authors' processing.

Generalised Additive Model

Table 6 shows the evaluation metrics of the generalised additive model. The best results in terms of classification accuracy were achieved by Opinion Lexikon and AFINN with the accuracy indicator value of 93.63% and 92.7%, respectively. The worst classification by the model was again the Loughran-McDonald financial dictionary. In general, this method shows very satisfactory results and outperforms decision trees and k-nearest neighbours. This method has not been widely used in previous studies, however, based on the analysis, it can be stated that it is able to compete with popular methods such as SVM or neural networks.

Table 6. Evaluation metrics of classification via a generalised additive model

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexicon	0.8806	0.9100	0.8951	0.9363	0.0626
Loughran-McDonald Lexicon	0.2311	0.2355	0.2333	0.2234	0.7768
Harvard IV-4 Lexicon	0.9101	0.8908	0.9004	0.9000	0.1001
FinanceSentiment Lexicon	0.9753	0.9349	0.9547	0.9176	0.0835
VADER	0.8671	0.8654	0.8662	0.8859	0.1145
AFINN	0.9005	0.9101	0.9053	0.9270	0.0712

Source: Authors' processing.

4.4 Comparison of classifiers

The ROC curve (Receiver Operating Characteristic Curve) is used to evaluate and also graphically represent the performance of classifiers when classifying into two classes, i.e. when evaluating words contained in the analysed dictionaries as positive and negative. Ideally, a ROC curve with perfect discrimination ability would copy the upper left corner of the ROC space. Based on previous calculations, decision trees and k-nearest neighbours appeared to be the weakest classifiers, which is further confirmed by the ROC curve, which for these classifiers deviates significantly from the upper left corner of the ROC space for all analysed lexicons. It can therefore be stated that decision trees and k-nearest neighbours have lower discrimination ability than the other analysed classifiers. The area under the ROC curve is referred to as AUC, which can also be understood as a measure of the quality of the test. AUC is the most common quantitative index describing the ROC curve. The outputs of this indicator are listed in numerical form in Table 7.

Table 7. AUC of all classifiers

	SVM	Decision Tree	naive Bayes	K-nearest neighbours	Neural networks	Generalised Additive Model
Opinion Lexicon	0.9938	0.7463	0.9711	0.8874	0.9819	0.9790
Loughran-McDonald Lexicon	0.2843	0.1669	0.5439	0.2187	0.2888	0.1805
Harvard IV-4 Lexicon	0.9899	0.7350	0.9689	0.8786	0.9560	0.9667
FinanceSentiment Lexicon	0.9316	0.6902	0.8979	0.6439	0.8312	0.9031
VADER	0.9698	0.7046	0.9448	0.8590	0.9080	0.9394
AFINN	0.9923	0.7355	0.9547	0.9185	0.9674	0.9707

Source: Authors' processing.

The accuracy of the classification test to be considered excellent should exceed 0.9, which is the case, for example, for SVM, naive Bayes, neural networks and generalised additive models. Good test accuracy results in the range of 0.7-0.8 were achieved by decision trees and k-nearest neighbours. These results are evident for all selected lexicons except for the already mentioned lexicon by Loughran-McDonald, which is created specifically for the financial domain. In general, it can be stated that the results indicate the superiority of naive Bayes over all other classifiers. Specifically, it achieves an average of 0.25 higher accuracy than decision trees, an average of 0.15 k-nearest neighbours and 0.06 neural networks and generalised additive models. This model achieved almost identical results with SVM, where the quality of binary classification differed on average by only 0.02 in favour of naive Bayes. After training and testing the classifiers, the sentiment score of each word in the text is predicted for the selected text set, and then the average sentiment score is determined for each monitored day. To do this, it is necessary to use a pre-processed text set. Then, words that do not appear in the pre-trained word insertion library are removed from the text. If the classification is correct in the new text, the average sentiment score is then calculated. For each sentence of a text document, words are converted into word vectors, on which a sentiment score is predicted. The score is then transformed using a transformation function, and then the mean sentiment score is calculated. This process is performed on the selected text document for all previously analysed dictionaries of positive and negative words, as well as for all previously analysed classifiers. For an easy visual assessment of the obtained sentiment, a histogram is chosen in which the distribution of the sentiment score is recorded. In Figure 5, you can see individual histograms generated using the histogram function, with each individual image representing a dictionary, and individual histograms within the image indicating the sentiment distribution obtained using six classifiers, which are colour-coded. Figure 5 shows significant differences in the distribution of sentiment scores not only for individual dictionaries, but also for individual classifiers. At first glance, it is obvious that the classification and evaluation of sentiment using SVM and the generalised additive model are almost identical when using all dictionaries. Both classifiers are shown in blue, and the density of the sentiment distribution overlaps. In addition, the Harvard IV-4 dictionary shows a very similar evaluation of a text document using neural networks, naive Bayes, k-nearest neighbours and decision trees, as evidenced by the identical distribution of sentiment scores, since the individual density functions almost overlap. The other dictionaries show a significant bias in the distribution of sentiment scores. This bias is dominant for Opinion Lexikon, AFINN, and the financial dictionaries Loughran-McDonald and FinanceSentiment.

It can be stated that most dictionaries of positive and negative words classified by neural networks, naive Bayes, k-nearest neighbours and decision trees overestimate sentiment scores compared to SVM and generalised additive model classifiers. The only exception is the FinanceSentiment dictionary, which underestimates sentiment scores compared to SVM and generalised additive model. For this reason, it is necessary to adjust the division into positive and negative sentiment and not rely on the general rule that a score greater than 0 indicates positive sentiment, while a score less than 0 indicates negative sentiment. This fact is evident from the distribution of the sentiment score distribution and it is necessary to work with this knowledge further so that sentiment can be correctly applied for subsequent stock market prediction. This means that it is necessary to modify the class labels and set the boundaries for individual classes manually. For example, one could decide that for the VADER dictionary and the SVM classifier, a sentiment score below 0.4 would be considered negative, and a sentiment score above 0.4 would be considered positive. The same would be true for other dictionaries. The goal is to set boundaries that are relevant for each dictionary and each classifier to achieve the highest classification accuracy.

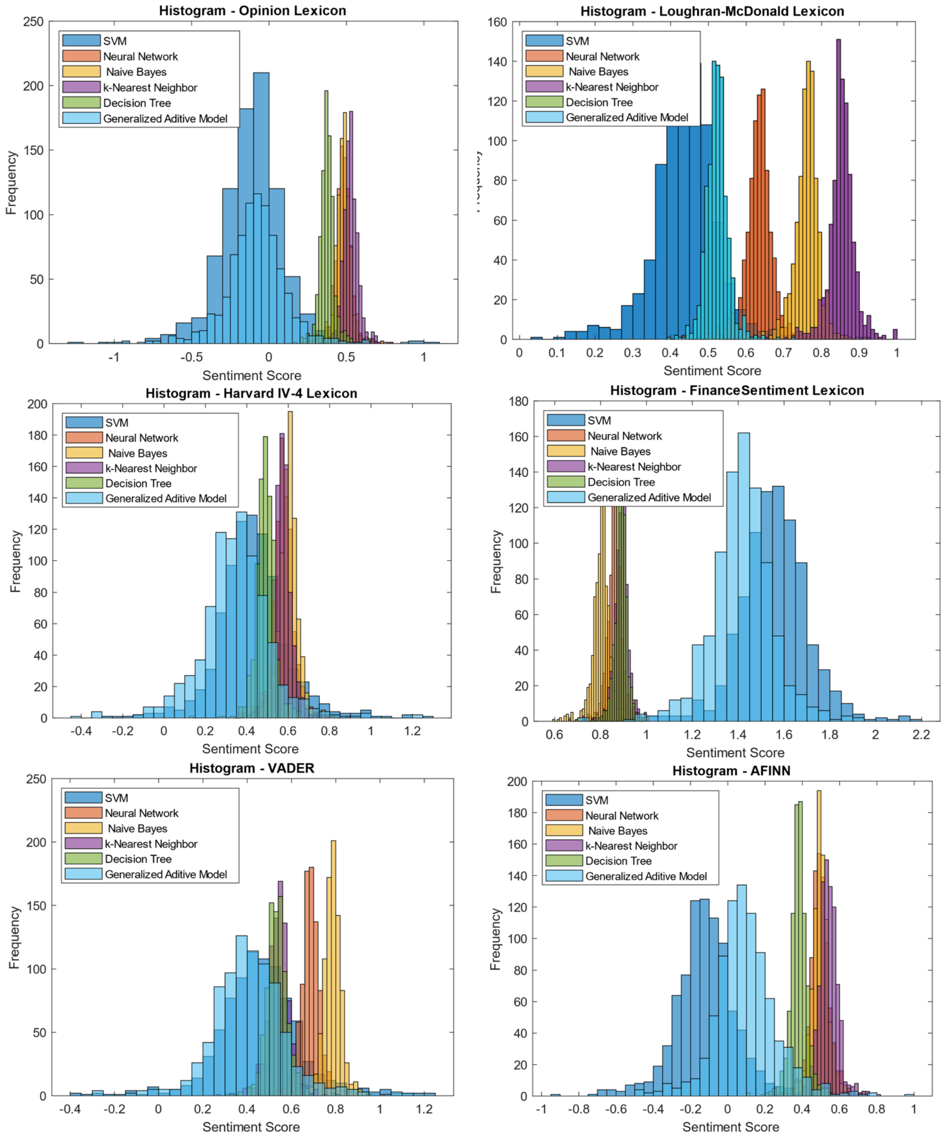

Figure 5. Sentiment score distribution

Source: Authors' own creation.

In addition to using the above-mentioned accuracy metrics to evaluate the effectiveness of individual dictionaries and models, it is necessary to verify whether the differences in performance are statistically significant. In this study, the Wilcoxon test is used. H0: Financial and general vocabulary generate equally accurate sentiment scores; H1: Specialised financial vocabulary generates more accurate sentiment scores than the general vocabulary. A significance level of 0.05 is chosen. Then, the decision whether or not to reject the null hypothesis is based on the resulting p-value. By default, if the p-value is greater than 0.05, the null hypothesis is not rejected. Otherwise, if the p-value is less than 0.05, the null hypothesis will be rejected at the 95% confidence level. The results of the Wilcoxon test on the test data list for each combination of financial and general dictionaries are summarised in Table 8. It is evident from the table that for the training data the p-values are large and higher than the threshold value of 0.05. Therefore, the null hypothesis that financial-specific dictionaries generate more accurate investor sentiment scores from text data cannot be rejected in favour of the alternative hypothesis that financial dictionaries generate more accurate investor sentiment scores from text data. Therefore, there is no significant evidence for a better ability of financial dictionaries to label text data into positive and negative words. Overall, the data presented here provide evidence that financial dictionaries do not generate more accurate values than general dictionaries. Based on the results of the Wilcox test, it can be argued that the FinanceSentiment lexicon is more accurate than the Loughran-McDonald dictionary.

Table 8. Wilcoxon test of individual dictionaries

Lexicon type	Alpha value	P-value	Null hypothesis
Finance Sentiment→Opinion Lexicon	0.0500	0.6582	Not rejected
Finance Sentiment→AFINN	0.0500	0.7101	Not rejected
Finance Sentiment→Loughran-McDonald	0.0500	0.0002	Rejected
Finance Sentiment→VADER	0.0500	0.3394	Not rejected
Finance Sentiment→Harvard	0.0500	0.7133	Not rejected
Loughran-McDonald→Opinion Lexicon	0.0500	1.0000	Not rejected
Loughran-McDonald→AFINN	0.0500	1.0000	Not rejected
Loughran-McDonald→Finance Sentiment	0.0500	0.9998	Not rejected
Loughran-McDonald→VADER	0.0500	0.9992	Not rejected
Loughran-McDonald→Harvard	0.0500	1.0000	Not rejected

Source: Authors' processing.

In the subsequent analysis, only one dictionary, or rather one sentiment score, is used to reduce computational and time requirements. The most accurate dictionary is chosen, namely the Opinion lexicon, which achieved almost 100% accuracy when classified using SVM. However, all the following operations can also be performed with other dictionaries, which is, due to the time constraints of this work, unfeasible, but at least it provides space for subsequent research in this area.

5. Conclusions

This research has shown that sentiment analysis from financial texts, especially those published on online platforms, can provide valuable information for stock market prediction. A hybrid approach combining machine learning with neural networks and multiple lexicons contributed to higher accuracy in sentiment polarity classification. The results confirm that the choice of lexicon and classifier plays a key role in interpreting investor sentiment and its impact on market dynamics. This can be declared by answering the first research question (RQ1): What effect does the lexicon of positive and negative words have on sentiment scores? The sentiment analysis shows significant differences in the distribution of sentiment scores not only for individual dictionaries. At first glance, it is obvious that the classification and evaluation of sentiment using SVM and the generalised additive model are almost identical when using all dictionaries. In addition, the Harvard IV-4 dictionary shows a very similar evaluation of a text document using neural networks, naive Bayes, k-nearest neighbours and decision trees, as evidenced by the identical distribution of sentiment scores. The other dictionaries show a significant bias in the distribution of sentiment scores. This bias is dominant in the Opinion Lexicon, AFINN and the financial dictionaries Loughran-McDonald and FinanceSentiment.

For this reason, it is necessary to adjust the division into positive and negative sentiment and not rely on the general rule that a score greater than 0 indicates positive sentiment, while a score less than 0 indicates negative sentiment. This fact is evident from the distribution of the sentiment score distribution and it is necessary to work with this knowledge further so that sentiment can be correctly applied for subsequent stock market prediction. This means that it is necessary to modify the class labels and set the boundaries for individual classes manually. The goal is to set boundaries that will be relevant for each dictionary and each classifier in order to achieve the highest classification accuracy. Our results contradict the study by Li et al. (2020), who predicted the development of stock prices in Hong Kong using technical indicators and sentiment. They tested four dictionaries, and the best results were achieved by the Loughran–McDonald financial dictionary. Koukaras et al. (2022) using the StockTwits dataset and the TextBlob dictionary shows the best results for SVM with an F-score of 68.7% and an AUC of 53.3%. With the VADER lexicon, SVM and linear regression achieved an F-score of 68%. The authors achieved better results for the Twitter dataset, above 70%. In this context, the hypothesis is answered as to whether a special financial dictionary generates more accurate sentiment scores than a general dictionary. The Wilcoxon non-parametric test is chosen to test this hypothesis. The test results show that the p-values of the financial dictionaries are high for the training data and higher than the threshold value of 0.05. Therefore, the null hypothesis cannot be rejected in favour of the alternative, which claims that specially created financial dictionaries generate more accurate investor sentiment scores from text data. Therefore, there is no significant evidence of a better ability of financial dictionaries to label text data in positive and negative words. Overall, it can be said that the data presented here provide evidence that financial dictionaries do not generate more accurate values than general dictionaries. Based on the results of the Wilcox test, it can be argued that the FinanceSentiment lexicon is more accurate than the Loughran-McDonald dictionary.

Last but not least, the paper also examined the main research question (RQ2): How is the choice of binary classifier related to the accuracy of sentiment score calculation with respect to the chosen lexicon? The accuracy of rating classification as excellent should exceed the accuracy value of 90%, which is the case, for example, for SVM, naive Bayes, neural networks and generalised additive model. Good test accuracy results in the range of 70% to 80% were achieved by decision trees and k-nearest neighbours. These results are evident for all chosen lexicons except for the lexicon from Loughran-McDonald, which is created specifically for the financial domain. In general, it can be stated that the results indicate the superiority of naive Bayes over all other classifiers. Specifically, it achieves an average of 25% higher accuracy than decision trees, an average of 15% for k-nearest neighbours, and 6% for neural networks and generalised additive model. This model achieved almost identical results with SVM, where the quality of binary classification differed on average by only 2% in favour of naive Bayes. However, in the case where the worst-trained dictionary from Loughran-McDonald is not considered. The highest classification accuracy is attributed to the SVM method, which is on average 2% to 3% higher than that of naive Bayes and the generalised additive model. It can be stated that in terms of binary classification, SVM, naive Bayes and the generalised additive model clearly dominate and achieve encouraging results, and these are suitable models for sentiment classification.

Acknowledgements: *This paper was supported by project No. FP-S-25-8736 "Modelling and simulation in business processes" from the Internal Grant Agency at Brno University of Technology and No. 102p17 "Advanced Decision-Making Models for Risk Management in the Digital Economy" funded by the Ministry of Education, Youth and Sports of the Czech Republic within the AKTION Czech Republic–Austria programme.*

References

[1] Basak, G.K., Das, P.K., Marjit, S., Mukherjee, D., Yang, L. (2023), *The British Stock Market, currencies, brexit, and media sentiments: A big data analysis. The North American Journal of Economics and Finance*, 64, 101861, https://doi.org/10.1016/j.najef.2022.101861.

[2] Bi, J., Yang, Z., Yang, Z. (2022), *Stock Market Prediction Based on Financial News Text Mining and Investor Sentiment Recognition. Mathematical Problems in Engineering*, 2022, 1-9, https://doi.org/10.1155/2022/2427389.

[3] Catelli, R., Pelosi, S., Esposito, M. (2022), *Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. Electronics*, 11(3), 374, https://doi.org/10.3390/electronics11030374.

[4] Chu, X., Wan, X., Qiu, J. (2023), *The relative importance of overnight sentiment versus trading-hour sentiment in volatility forecasting. Journal of Behavioral and Experimental Finance*, 39, 100826, https://doi.org/10.1016/j.jbef.2023.100826.

[5] Fang, H., Chung, C. -P., Lu, Y. -C., Lee, Y. -H., Wang, W. -H. (2021), *The impacts of investors' sentiments on stock returns using fintech approaches. International Review of Financial Analysis*, 77, 101858, https://doi.org/10.1016/j.irfa.2021.101858.

[6] Feuerriegel, S., Gordon, J. (2018), *Long-term stock index forecasting based on text mining of regulatory disclosures*, https://doi.org/10.48550/arxiv.1806.09866.

[7] Hao, P. -Y., Kung, C. -F., Chang, C. -Y., Ou, J. -B. (2021), *Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane. Applied Soft Computing*, 98, 106806, https://doi.org/10.1016/j.asoc. 2020.106806.

[8] Hu, M., Liu, B. (2004), *Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, 22-25 August 2004*, 168-177, https://doi.org/10.1145/1014052. 1014073.

[9] Hutto, C.J., Gilbert, E. (2014), *Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media,* 8(1), 216-225, https://doi.org/10.1609/icwsm.v8i1.14550.

[10] Jaggi, M., Mandal, P., Narang, S., Naseem, U., Khushi, M. (2021), *Text Mining of Stocktwits Data for Predicting Stock Prices.* ArXiv, https://doi.org/10.48550/arxiv. 2103.16388.

[11] Janková, Z. (2023), *Critical Review of Text Mining and Sentiment Analysis for Stock Market Prediction. Journal of Business Economics and Management*, 24(1), 177-198, https://doi.org/10.3846/jbem.2023.18805.

[12] Kirtac, K., Germano, G. (2024), *Sentiment trading with large language models. Finance Research Letters*, 62, 105227, https://doi.org/10.1016/j.frl.2024.105227.

[13] Koukaras, P., Nousi, C., Tjortjis, C. (2022), *Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. Telecom*, 3(2), 358-378, https://doi.org/10.3390/telecom3020019.

[14] Li, S., Shi, W., Wang, J., Zhou, H. (2021a), *A Deep Learning-Based Approach to Constructing a Domain Sentiment Lexicon: a Case Study in Financial Distress Prediction. Information Processing & Management*, 58(5), 102673, https://doi.org/10.1016/j.ipm.2021.102673.

[15] Li, M., Chen, L., Zhao, J., Li, Q. (2021b), *Sentiment analysis of Chinese stock reviews based on BERT model. Applied Intelligence*, 51(7), 5016-5024, https://doi.org/10.1007/s10489-020-02101-8.

[16] Li, X., Wu, P., Wang, W. (2020), *Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management*, 57(5), 102212, https://doi.org/10.1016/j.ipm.2020.102212.

[17] Loughran, T., Mcdonald, B. (2011), *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance*, 66(1), 35-65, https://doi.org/10.1111/j.1540-6261.2010.01625.x.

[18] Loughran, T., Mcdonald, B. (2016), *Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research*, 54(4), 1187-1230, https://doi.org/10.1111/1475-679X.12123.

[19] Nielsen, F.Å. (2011), *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. ArXiv, https://doi.org/10.48550/arxiv.1103.2903.

[20] Nti, I.K., Adekoya, A.F., Weyori, B.A. (2020), *Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. Applied Computer Systems*, 25(1), 33-42, https://doi.org/10.2478/acss-2020-0004.

[21] Palmer, M., Roeder, J., Muntermann, J. (2022), *Induction of a sentiment dictionary for financial analyst communication: a data-driven approach balancing machine learning and human intuition. Journal of Business Analytics*, 5(1), 8-28, https://doi.org/10.1080/2573234X.2021.1955022.

[22] Rizinski, M., Peshov, H., Mishev, K., Jovanovik, M., Trajanov, D. (2024), *Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). IEEE Access*, 12, 7170-7198, https://doi.org/10.1109/ACCESS.2024.3349970.

[23] Sekhar, S.R.M., Srinidhi, H., Srinivasa, K.G. (2022), *Forecasting Stock Market Volume Price Using Sentimental and Technical Analysis. Journal of Information Technology Research*, 15(1), 1-13, https://doi.org/10.4018/JITR.299383.

[24] Trawiński, B., Smętek, M., Telec, Z., Lasota, T. (2012), *Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. International Journal of Applied Mathematics and Computer Science*, 22(4), 867-881, https://doi.org/10.2478/v10006-012-0064-z.

[25] Wang, J., Zhu, S. (2023), *A Novel Stock Index Direction Prediction Based on Dual Classifier Coupling and Investor Sentiment Analysis. Cognitive Computation*, 15(3), 1023-1041, https://doi.org/10.1007/s12559-023-10137-4.