Sentiment Analysis in Finance: A Comparative Study of Lexicon-Based and Model-Based Methods

Wai Jean Koh

MSc in Artificial Intelligence
The University of Bath
2023

Sentiment Analysis in Finance: A Comparative Study of Lexicon-Based and Model-Based Methods

Sentiment Analysis in Finance: A Comparative Study of Lexicon-Based and Model-Based Methods

Submitted by: Wai Jean Koh

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Artificial Intelligence in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

This paper conducts a comparative analysis of sentiment analysis algorithms, focusing on lexicon-based and model-based approaches, within the context of predicting stock price movements. It involves an extensive literature review exploring the development of these two model categories and the implementation of state-of-the-art methods. The study employs intraday stock price data and news articles related to S&P500 companies to evaluate accuracy across various holding periods. While model-based methods theoretically outperform lexicon-based ones, prior studies yielded inconclusive results, primarily due to the lack of labelled training data. This research highlights a recent breakthrough in model-based approaches, particularly the FinBERT model, which overcomes previous limitations, demonstrating superior predictive performance over lexicon-based methods.

The study further examines two versions of the FinBERT model, v1 and v2, noting substantial differences in their performance. FinBERT v2 exhibits superior accuracy and precision, with a balanced performance across positive and negative sentiment predictions. Backtesting across different time intervals reveals that a 30-minute window optimally captures the influence of news articles on market dynamics. The research also investigates the impact of classification thresholds on model performance, finding modest effects on evaluation metrics. FinBERT's effectiveness stems from its deep understanding of financial language, adaptability to contextual nuances, and nuanced sentiment predictions due to extensive pre-training, distinguishing it from lexicon-based methods constrained by predefined sentiment lists.

However, challenges persist regarding time and cost efficiency in employing FinBERT. Its resource-intensive training and deployment demand substantial computational resources and costs, potentially limiting accessibility. Additionally, interpretability remains a concern, given FinBERT's complex architecture, hindering transparent explanations and feature importance assessments. These limitations underscore the importance of considering practical deployment constraints when utilizing FinBERT for financial decision-making.

Contents

List of Figures

Chapter 1

Introduction

The impact of news sentiment on stock price reaction has been a subject of considerable research in the financial literature. Positive news, such as favorable earnings reports or news of new product launches, can lead investors to perceive a company as performing well, which may drive demand for the company's stock and increase its price. Conversely, negative news, such as product recalls or earnings decreases, can undermine investor confidence and lead to a sell-off and a subsequent drop in the stock price. The rapid dissemination of news and the degree of its impact on the market can also play a role in stock price reactions. Consequently, investors and traders closely monitor news sentiment to make informed decisions about buying and selling stocks. Overall, news sentiment has emerged as a key driver of stock price reactions (Costola et al., 2023), and understanding its dynamics can offer valuable insights into stock market behavior.

The growing recognition of the importance of news sentiment in financial markets, as well as the potential benefits of predicting news sentiment, has motivated a significant body of literature to study sentiment analysis techniques. Sentiment analysis is a branch of natural language processing that involves the use of computational methods to extract subjective information, such as opinions, emotions, and attitudes, from text data. Researchers have developed a range of sentiment analysis techniques, which can broadly be split into lexicon-based methods and model-based methods. These techniques differ in their approach to sentiment analysis, with the former relying on manually curated dictionaries of sentiment words and phrases, while the latter uses statistical models to learn sentiment from large text datasets.

Despite the increasing popularity and importance of sentiment analysis in various fields, limited research has been done on the comparative analysis between lexicon-based and model-based sentiment analysis methods. One reason for this is that sentiment analysis is a relatively new and rapidly evolving field, and the development of new techniques and methods has outpaced the research on their comparative analysis. Another reason is the lack of standardization in sentiment analysis evaluation and benchmark datasets, which makes it difficult to compare the performance of different methods. This lack of standardization makes it challenging to conduct a fair and comprehensive comparative analysis of different approaches. Furthermore, there is a lack of consensus on the appropriate evaluation metrics for sentiment analysis, as different studies employ different metrics depending on whether they frame the problem as a regression or classification analysis (Kraus & Feuerriegel, 2017).

Model-based methods are expected to perform better than lexicon-based methods in financial sentiment analysis for several reasons. Firstly, financial texts contain specific jargon and terminology that may not be captured accurately by general-purpose lexicons (Araci, 2019). Model-based methods can be trained on financial-specific data, allowing them to learn the specific language and terminology used in the financial domain and adjust their sentiment analysis accordingly. Secondly, financial texts often contain more complex and nuanced language than general-purpose text, making it more challenging to accurately determine sentiment. Model-based methods are better suited to capture the context and nuances of language, allowing them to capture more subtle variations in sentiment that may be missed by lexicon-based methods.

Among the earlier research that has been conducted in the financial domain, it has been reported that lexicon-based methods have outperformed model-based approaches (Mohan et al., 2019; Sohangir et al., 2018). However, these findings were observed in studies conducted prior to the introduction of FinBERT, which has demonstrated superior performance and has become widely recognized for its effectiveness in capturing financial language nuances and domain-specific sentiment. Researchers have recognized the advantages of leveraging pre-trained models, such as FinBERT, which addresses the traditional machine learning problem of requiring large amounts of labeled data. Therefore, while earlier studies may have indicated the superiority of lexicon-based methods over model-based approaches, it is crucial to consider the evolution of model-based approaches in the financial domain.

As sentiment analysis becomes increasingly important in the financial field, there is a need for more comprehensive and standardized research on the comparative analysis of different methods to guide the selection of appropriate sentiment analysis methods for different applications. Hence, the research objective of this paper is to conduct a thorough evaluation of state-of-the-art algorithms in lexicon-based and model-based methods for sentiment analysis. The objective is to compare the performance of these two approaches in analysing financial news articles, with a focus on identifying which approach is more effective in identifying sentiment in the financial domain. The evaluation will consider various metrics, such as accuracy and precision, to assess the effectiveness of the algorithms in capturing the correct sentiment. The thesis also aims to investigate the limitations of these methods and identify areas for further research to improve their effectiveness.

This paper is organized into distinct chapters. Chapter 2 delves into an in-depth literature survey, tracing the development of both model categories. Chapter 3 outlines the dataset collection process and the methodologies adopted for performance assessment. Chapter 4 presents the outcomes and engages in a comprehensive discussion of the observed results. Finally, Chapter 5 encapsulates the conclusions drawn and hints at potential directions for future research.

Chapter 2

Literature Survey

Sentiment analysis in the financial domain involves the use of natural language processing techniques to extract sentiment from financial text data, such as news articles, social media posts, and financial reports. The aim is to determine the positive or negative sentiment associated with a particular entity, event, or product in the financial market. The sentiment analysis techniques can broadly be categorized into lexicon-based or model-based methods.

2.1 Lexicon-based method

Lexicon-based methods for sentiment analysis in the financial domain rely on sentiment lexicons that are specifically designed for financial text data. These lexicons contain a list of words and phrases that are associated with positive or negative sentiment in the financial domain. Several well-known sentiment lexicons are SentiWordNet (Baccianella et al., 2010), Textblob (Loria, 2018) and VADER (Hutto & Gilbert, 2014).

SentiWordNet is a publicly available lexical resource that assigns sentiment scores to each word in WordNet, a large lexical database of English. The sentiment scores in SentiWordNet are based on three categories: positivity, negativity, and objectivity. Each word in WordNet is assigned a score between 0 and 1 for each of these three categories, with a higher score indicating a stronger association with that category. To determine the sentiment of a text using SentiWordNet, the sentiment scores of each word in the text are aggregated to obtain an overall sentiment score. One common approach is to calculate the average positivity and negativity scores of all words in the text, and then calculate the sentiment score as the difference between the two scores.

On the positive side, SentiWordNet provides predefined sentiment scores for a wide range of English words, allowing for convenient sentiment analysis without the need for manual annotation. It enables fine-grained sentiment analysis at the word level and incorporates semantic information through synsets, capturing contextual meaning. However, SentiWordNet has limitations. Its coverage may be limited, resulting in missing sentiment information for specific terms. It lacks explicit consideration of negation and intensifiers, which can affect the accuracy of sentiment analysis. Additionally, its fixed sentiment scores may not account for evolving language and sentiment changes, potentially affecting its effectiveness in different domains or over time.

Textblob is a popular Python library that uses a lexicon-based method to perform sentiment analysis. When analyzing a piece of text, TextBlob first assigns a polarity score to each word in the text, based on a pre-defined polarity lexicon that contains words and their corresponding polarity scores. Next, TextBlob calculates the overall polarity score of the text by aggregating the individual word scores and taking the average. It then classifies the text as positive, negative, or neutral, based on the polarity score. Moreover, TextBlob also uses a set of rules to determine the polarity of a sentence. These rules take into account factors such as negation and intensifiers, and can help to improve the accuracy of the sentiment analysis. However, TextBlob's sentiment analysis accuracy may be limited compared to more advanced models, as it relies on simple heuristics and lacks extensive consideration of contextual or domain-specific information. Customization options are also limited, preventing users from fine-tuning the sentiment analysis model or incorporating domain-specific knowledge.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based sentiment analysis tool that assigns a sentiment score to a piece of text by calculating the positive, negative, and neutral word densities in the text. The tool uses a sentiment lexicon that has been manually curated and labeled for sentiment intensity by 10 independent and trained human raters. VADER also applies 5 heuristics based on grammatical and syntactical rules such as punctuation, capitalization, degree modifier, contrastive conjunction “but” and negation to handle amplification and negation in the texts. The tool outputs a sentiment score ranging from -1 (most negative) to 1 (most positive), as well as a compound score that represents the overall sentiment intensity of the text. An important advantage of VADER is that its lexicon can be customized to include domain-specific terms and their associated sentiment scores. By tailoring the lexicon to the financial domain, VADER can provide more relevant and precise sentiment analysis results.

Despite not being domain-specific, general-purpose sentiment analysis tools have shown promising performance in the financial domain. (Gupta et al., 2022) applied Textblob to classify tweet data and combined it with historical stock price data before feeding them into the LSTM algorithm to predict the future price. They found that the accuracy of the model has improved by more than 5 percentage points when compared with the start-of-the-art. On the other hand, (Long et al., 2023) highlighted the need to design a unique lexicon containing specific terms frequently used on the r/WallStreetBets subreddit when using VADER because certain financial words are excluded in the VADER lexicon. Applying VADER directly to the finance field may lead to a high misclassification of the Reddit sentiment. Therefore, general-purpose tools may not be appropriate for sentiment analysis in the financial domain, as they may not be able to capture domain-specific sentiment expressions.

Loughran and McDonald's Financial Sentiment Word Lists (Loughran & McDonald, 2011) is a lexicon-based method tailored for sentiment analysis in the financial domain. It consists of two word lists: the Loughran-McDonald Positive Word List (LM-Positive) and the Loughran-McDonald Negative Word List (LM-Negative), which contain words commonly used in financial news articles to express positive or negative sentiment, respectively. The lists were constructed by analysing a large corpus of financial text data, such as 10-K filings, earnings announcements, and news articles, and each word was manually labelled as either positive or negative based on the context in which it was used. The lists have been shown to have high precision and recall in identifying sentiment in financial texts, and have been widely used in academic and industry research. To use the lists, each word in a text is compared to the words in the LM-Positive and LM-Negative lists. If a word appears in the LM-Positive list, it is assigned a positive sentiment score, and if it appears in the LM-Negative list, it is assigned a negative sentiment score. The sentiment score for the text is then calculated as the difference between the total positive and negative scores. While the LM-Positive and LM-Negative lists have been shown to be effective in identifying sentiment in financial texts, they may not capture all possible expressions of sentiment in financial texts, as new words and phrases may emerge over time. Therefore, the authors had been actively maintaining the word list by adding commonly appearing words to the list. [1]

However, some industries may require a more specific lexicon than Loughran’s word list. For example, (Shah et al., 2018) had to create a dictionary [2] that contains words which are specific to the pharmaceutical industry such as “fda approval” as these phrases might indicate whether the companies have received drug approval from the United States Food and Drug Administration’s (FDA). The process of creating a custom dictionary requires significant domain expertise and does not scale well.

(Turner et al., 2021) presented a probabilistic approach to generate financial sentiment lexicon in an automated way. They collected a list of earnings conference call logs and assigned each document as positive or negative based on the adjusted

[1] https://sraf.nd.edu/loughranmcdonald-master-dictionary/

[2] https://github.com/queensbamlab/NewsSentiments/blob/master/dict.csv

stock price change during the trading day. Then, they extracted a list of words from the document and calculated the sentiment score based on the number of positive and negative documents the word belongs to. The method outperformed SentiWordNet, Loughran and McDonald's lexicon and Henry's lexicon (Henry, 2008) in classifying document sentiment. However, the authors stated that they did a random split, rather than a time series split, when they split the earnings call data to a training set to build the sentiment lexicon and a test set for evaluation. Given that the earnings call data span 10 years from 2008 to 2018, the correct approach would be to split the training set before a fixed time, say 2016, and test it on data after 2016 so the lexicon wouldn't contain any word that appears after it was created. A random split implied that the lexicon might contain words in the future, which gave it an unfair advantage over the other lexicons. Furthermore, I have not been able to locate the lexicon used by the authors despite conducting an extensive online search. It is possible that the lexicon has not been made available online or that it is not yet accessible to the public. Hence, we have excluded this study from our experiment.

Lexicon-based methods are easy to implement and computationally efficient, and they do not require large amounts of labelled data for training. They can be effective in predicting sentiment in the financial domain, particularly when using lexicons that are tailored to the financial language. However, they have limitations in handling sarcasm, irony, and other forms of figurative language, and they may not perform well in detecting subtle nuances of sentiment.

2.2 Model-based method

Model-based methods for sentiment analysis involve the use of machine learning algorithms to identify patterns and learn from labelled training data to predict the sentiment of new, unlabelled text. These methods use various types of models such as naive Bayes, support vector machines, decision trees, and deep learning models, among others. The process involves training the model on a large dataset of labelled text, which is then used to classify new text based on the learned patterns and features. Model-based methods can be highly effective in identifying complex patterns and nuances in text, and can be trained specifically for a particular domain or task, making them potentially more accurate than lexicon-based methods. However, they require a large amount of labelled data for training and may be computationally intensive and resource-intensive. It can be subdivided into three groups: machine learning, deep learning and pretrained models.

2.2.1 Machine learning

An important step in the machine learning method is text featurization, also known as text representation or feature extraction. The goal of text featurization is to convert raw text data into a structured format that machine learning algorithms can process. In this section, we will discuss some popular text featurization methods.

Bag-of-Words (BoW) is a widely used approach in text featurization. The basic idea of the BoW approach is to represent a text document as a collection of words, and then to use the frequency distribution of those words as a feature vector for classification. In other words, the BoW approach simply counts the frequency of each word in a text document, and then uses those counts as a feature vector to represent the document. The resulting feature vector is then used as input to a machine learning algorithm for classification. The BoW approach is simple and effective, but it has some limitations, such as ignoring the order of words in a text document and not capturing the meaning of phrases or idiomatic expressions.

Term Frequency-Inverse Document Frequency (TF-IDF) is another popular approach that improves upon the BoW method by taking into account the importance of words in the document. The idea behind TF-IDF is to weight the importance of a word in a document by measuring its frequency in the document and inversely proportional to its frequency in the corpus. The resulting weight reflects how important the word is to the document relative to the corpus. TF-IDF has been shown to be effective in many text classification tasks, including sentiment analysis.

In financial sentiment analysis, bigram and trigram are frequently used text featurization methods to capture more context and improve accuracy compared to unigram. Bigram refers to the combination of two adjacent words in a sentence, while trigram refers to the combination of three adjacent words. By considering the combination of words, bigram and trigram capture more context and provide a better representation of the meaning of a sentence. This is particularly important in financial sentiment analysis because a single word may have different meanings in different contexts, and the correct sentiment can depend on the combination of words. For example, the bigram "strong buy" or trigram "earnings per share" can carry a sentiment that cannot be captured by the individual words alone. By using bigram and trigram in addition to unigram, the sentiment analysis model can take into account the context of the words and improve accuracy.

When using bigram and trigram featurization methods, the number of features can become very large, making the data matrix sparse and computationally expensive to process. Latent Semantic Analysis (LSA) can help to address this issue by reducing the dimensionality of the feature space. It does this by creating a lower-dimensional representation of the original feature matrix using singular value decomposition (SVD). The resulting matrix retains most of the important information from the original matrix while significantly reducing its size, making it more manageable for subsequent analysis. Hence, LSA can help to overcome the dimensionality issue of bigram and trigram featurization methods.

Numerous studies have been conducted on the use of machine learning methods for sentiment analysis in financial domains, and many have reported promising results. (Gupta & Chen, 2020) did an extensive study on using different combination of text featurization methods (BoW, TF-IDF, bigram, trigram LSA) and machine learning algorithms (Naïve Bayes, Logistic Regression, Support Vector Machine). They showed that the best combination (Logistic Regression and TF-IDF) achieved an accuracy level between 75% and 85% when conducting sentiment analysis on StockTwists data for five US companies. While lexicon-based methods may require separate dictionaries for social media data and news articles, machine learning-based methods can work reasonably well on different text data provided that we have sufficient training data to train the machine learning algorithm. For example, (Kalyani et al., 2016) transformed news articles on major news aggregators (Google News, Yahoo Finance, Reuters) to TF-IDF vectors and achieved an accuracy of more than 80% in identifying the correct sentiment.

One disadvantage of traditional machine learning models compared to deep learning models is that they treat each word or phrase as a distinct feature rather than interpreting them as a sequence of words. This can result in a loss of contextual information and can make it difficult to capture the underlying meaning and nuances of language. Moreover, traditional machine learning models may struggle with handling large amounts of unstructured data, which can require a significant amount of pre-processing and feature engineering.

2.2.2 Deep learning

Word embedding is a text featurization technique that represents words in a vector space based on their meaning and semantic relationships. In the context of financial sentiment analysis, the word embedding model can learn from a large corpus of financial documents and generates a high-dimensional vector representation for each word. The advantage of using word embedding is that it can capture the relationship between words, making it possible to capture the nuances of the financial domain. With word embedding, the model can identify financial terms that have similar meanings and can better differentiate between positive and negative sentiments. Aggregating word embeddings for a document involves combining the embeddings of individual words in the document into a single vector representation. One common approach for this is the "average word embedding" method. In this method, the embeddings of each word in the document are added together, and the resulting vector is divided by the total number of words in the document. This generates a single vector representation for the entire document, which can be used as input for machine learning or deep learning models.

There are various approaches available for generating word embeddings, and one of the popular methods is Word2vec (Mikolov et al., 2013). Word2vec works by training a neural network on a large corpus of text. The network takes as input a sequence of words and tries to predict the next word in the sequence. During training, the weights of the neural network are updated to minimize the difference between the predicted and actual next words. The weights of the network form the word embeddings. This approach has been shown to be effective in financial sentiment analysis. For example, (Pagolu et al., 2016) utilized Word2vec to extract features from human annotated tweets about Microsoft. The sum of 300 dimensional vectors of all words in a tweet are then passed into a Random Forest algorithm and they achieved an accuracy of 70% on the sentiment classification task. However, aggregating word embeddings into a sentence or document-level representation might result in the loss of some nuances of the individual words, as the context and relationships between words may not be fully captured by the aggregation method.

A Recurrent Neural Network (RNN) is a type of neural network that can process sequential data by maintaining an internal state, also known as a "memory". RNNs are designed to handle input data with temporal dependencies, such as text or speech. Unlike traditional feedforward neural networks, which treat each input as independent and do not retain any memory of previous inputs, RNNs use their internal state to capture information from past inputs, allowing them to model sequential data more accurately. One advantage of RNNs over word embedding is their ability to model sequences of variable lengths. Word embedding methods such as Word2Vec generate a fixed-length vector representation for each word in the vocabulary. However, in natural language processing, sequences can vary in length and meaning. RNNs are capable of processing variable-length sequences and can adapt their internal state to the length of the input sequence, making them more flexible in modeling language data. Additionally, RNNs can capture the context of words within a sentence or document, which can be important in sentiment analysis tasks. For example, the sentiment of a sentence can be influenced by the presence of negation words such as "not" or "but". RNNs can learn to capture such contextual information and use it to improve sentiment analysis accuracy.

However, RNNs can suffer from the vanishing gradient problem, where the gradients used in backpropagation become very small as they propagate through multiple time steps, leading to slow convergence or even no convergence at all. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can address the vanishing gradient problem in traditional RNNs. LSTM networks use a memory cell to store information over a long period of time, and three gates to control the flow of information into and out of the cell: the input gate, the output gate, and the forget gate. The input gate controls how much new information is added to the cell, the forget gate controls how much old information is discarded from the cell, and the output gate controls how much information is output from the cell. By using these gates, the LSTM can selectively remember or forget information from previous time steps, allowing it to learn long-term dependencies more effectively than traditional RNNs. This makes it well-suited for tasks such as natural language processing, where long-term dependencies between words are often crucial for understanding the meaning of a sentence or document. Numerous studies (Gite et al., 2021; Kraus & Feuerriegel, 2017) have demonstrated the effectiveness of LSTM in financial sentiment analysis.

One disadvantage of LSTM is that it requires a large amount of training data to learn the relationships between words and their contexts. This can be challenging in some domains where the data may be limited or expensive to collect. Pretrained models, on the other hand, are trained on large datasets and can be fine-tuned for a specific task with smaller amounts of data. This can be advantageous in situations where data is limited or where the cost of collecting data is high. Pretrained models can also reduce the need for extensive hyperparameter tuning and can help to overcome the overfitting problem. Another disadvantage of LSTM is that it can be computationally expensive and slow to train, especially on large datasets. Pretrained models, on the other hand, are already trained and can be readily used for inference, which can save time and computational resources.

2.2.3 Pretrained model

Pretrained model addresses the traditional machine learning problem of requiring large amounts of labelled data by leveraging the power of transfer learning. Transfer learning is a technique that allows a model to learn from a large dataset in one domain (pre-training) and then apply the learned knowledge to a different domain with a smaller labeled dataset (fine-tuning). For example, BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model based on the transformer architecture that can be fine-tuned for various NLP tasks, including sentiment analysis (Devlin et al., 2018).

The BERT model is pre-trained on a large corpus of text data using two unsupervised tasks: masked language modeling and next sentence prediction. In masked language modeling, a certain percentage of words in the input sentence are randomly masked, and the model is trained to predict the masked words based on the context of the surrounding words. In next sentence prediction, the model is trained to predict whether a given sentence follows the previous sentence in the text corpus. This pre-training process benefits from the availability of large-scale unlabelled financial data, which is relatively easier to obtain compared to annotated labelled data. Once the pre-training phase is complete, the BERT model can be fine-tuned for various downstream NLP tasks by adding a task-specific output layer and training the model on task-specific labelled data. In the case of financial sentiment analysis, fine-tuning allows it to adapt and specialize in capturing sentiment specifically in the financial domain. This transfer learning approach enables BERT to leverage the pre-trained knowledge and generalize well to the target task, even with a limited amount of labelled data.

The advantage of BERT over other pre-trained models is its ability to understand the context of words in a sentence, as it uses a transformer architecture that allows for bidirectional processing of input sequences. This allows the model to capture the relationship between words in a sentence and understand the meaning of the sentence as a whole. Moreover, BERT has reduced the need for large amounts of labelled training data, which can be time-consuming and expensive to obtain.

Several studies have attempted to train FinBERT, a pre-trained language model fine-tuned specifically for financial sentiment analysis, by further pretraining BERT on financial corpus.

(Araci, 2019; Yang et al., 2020) conducted a study where the BERT model was pretrained on the Reuter's TRC2-financial dataset, consisting of 1.8 million news articles published by Reuters between 2008 and 2010. This pretrained BERT model was then fine-tuned using a labeled sentiment dataset, and various techniques were employed to address the issue of catastrophic forgetting during the fine-tuning process. Comparisons were made between FinBERT and two other pre-trained language models, namely ELMo and ULMFit, and FinBERT exhibited superior performance, surpassing the state-of-the-art results by a significant margin. However, Araci noted that FinBERT has limitations when it comes to determining sentiment in the absence of words that indicate direction, such as "increase" or "decrease." In such cases, FinBERT may struggle to make the determination and may predict a neutral sentiment instead. Nevertheless, a significant finding of the study is that FinBERT was able to outperform previous model-based methods even with a relatively small training set of only 500 examples. This highlights the advantage of FinBERT in mitigating the requirement for large, labeled datasets, which has been a challenge for previous model-based approaches.

Yang et al. (2020) have undertaken the training of a FinBERT model, and we posit that its performance is likely to surpass that of Araci's version for several reasons. Firstly, Yang et al. employed a broader range of financial data sources, including corporate reports, earnings conference call transcripts, and analyst reports, for pretraining the BERT model. The incorporation of a longer history of financial data may also enhance the model's understanding of financial language and sentiment. Secondly, they developed a custom vocabulary specifically tailored to the financial corpus, departing from the utilization of the original BERT vocabulary. This custom vocabulary is expected to capture the specific terminologies and nuances prevalent in financial texts more accurately. While Yang et al. demonstrated the advantages of FinBERT over a generic domain BERT model, they did not provide a direct comparison with Araci's version. Consequently, this paper aims to conduct a comparative analysis between these two models to ascertain their relative performance in financial sentiment analysis.

As highlighted in the work of Araci (2019), the objective of financial sentiment analysis extends beyond sentiment classification alone. Its ultimate purpose lies in providing support for informed financial decision-making. Consequently, it becomes crucial to assess not only the accuracy of sentiment classification but also the predictive capacity of sentiment analysis in relation to stock market movements. In light of this, we also conducted a comprehensive review of several studies that have utilized FinBERT for the prediction of stock prices.

(Sidogi et al., 2021) integrated the news headline sentiment feature extracted from FinBERT into an LSTM model, demonstrating a significant enhancement in its predictive performance for intraday stock price prediction of Amazon. Similarly, (Halder, 2022) adopted a comparable approach, combining FinBERT and LSTM to forecast the closing price of the NASDAQ-100 index. Both studies revealed that the incorporation of the FinBERT sentiment feature within the LSTM model yielded superior results compared to baseline models. However, it is important to note that the baseline models employed in these studies solely utilized historical stock price data as input, without incorporating news sentiment features. Consequently, it remains uncertain whether the observed superiority of the proposed models stems solely from the inclusion of additional news sentiment features or from the inherent capabilities of the FinBERT model itself. To ensure a fair and accurate comparison between FinBERT and other sentiment analysis models, this paper aims to employ the sentiment feature in alternative models as well, thereby allowing for an isolated assessment of FinBERT's performance relative to other sentiment analysis models in the prediction of stock prices.

2.3 Comparison of lexicon-based and model-based method

While some research has been conducted on the comparative analysis between lexicon-based and model-based sentiment analysis methods, it is noteworthy that the majority of these studies have focused on general domains rather than specifically in the finance domain. The existing literature has primarily explored the performance and limitations of these methods in generic sentiment analysis tasks using datasets from other domains, such as social media posts (Kolchyna et al., 2015) and product reviews (Nguyen et al., 2018).

Despite the theoretical advantages of model-based methods over lexicon-based methods in sentiment analysis, empirical results have shown mixed results, and in some cases, lexicon-based methods have outperformed model-based methods in financial sentiment analysis. For example, (Mohan et al., 2019) compared lexicon-based method from NLTK libraries with LSTM to predict sentiment on news articles and found that, on average, the model-based method performed slightly worse in terms of MAPE. Furthermore, (Sohangir et al., 2018) performed the experiment on StockTwits data and showed that lexicon-based methods outperformed machine learning-based methods in terms of area under ROC curves. Additionally, lexicon-based methods can be faster and less resource-intensive than model-based methods, which is important for real-time financial sentiment analysis.

On the other hand, (Alostad & Davulcu, 2015) showed that applying a logistic regression classifier with a chi-squared feature selection algorithm on unigram features from news articles yields the best result in predicting hourly directional stock price movement. They concluded that adding extracted sentiment features with the Loughran-McDonald sentiment lexicon does not lead to a statistically significant improvement in accuracy. However, this is not a fair comparison between lexicon-based and model-based methods because they did not evaluate the effectiveness of the Loughran-McDonald sentiment lexicon as a standalone method, but rather combined it with the model-based method and observed if it brought marginal improvement over the baseline accuracy. The lexicon-based method could be equally effective but not more effective when used in conjunction with the model-based method. Overall, while model-based methods are theoretically expected to outperform lexicon-based methods in financial sentiment analysis, empirical results have shown that the effectiveness of sentiment analysis methods varies, and this warrants further investigation.

Our literature review indicates that the existing studies have primarily focused on exploring the effectiveness of either lexicon-based or model-based approaches individually, without directly comparing their performance on financial sentiment analysis tasks. This can be attributed to the following factor. Firstly, the availability of comprehensive and domain-specific lexicons for financial sentiment analysis is limited. Constructing a lexicon that accurately captures the nuances and terminology specific to the financial domain is a challenging task. These challenges necessitate the development of specialized lexicons that accurately capture the sentiment associated with financial language, which can be resource-intensive and time-consuming.(Long et al., 2023; Shah et al., 2018). On the other hand, model-based methods, such as deep learning models, have gained significant popularity in recent years due to their ability to automatically learn and capture complex patterns in textual data. These models have shown promising results in various NLP tasks, including sentiment analysis. Moreover, the rapid advancements in deep learning techniques and the availability of large-scale pre-trained models, such as BERT and GPT, have further shifted the focus towards model-based methods. These models provide a more flexible and versatile approach as they can capture contextual information and learn from vast amounts of unlabelled data. Consequently, researchers have been more inclined towards exploring and refining model-based methods rather than extensively studying lexicon-based approaches.

Nevertheless, it is important to recognize the value of lexicon-based methods, especially in domains like finance where domain-specific knowledge is vital. Therefore, it is crucial to bridge this research gap and conduct comparative analyses between lexicon-based and model-based methods specifically in the finance domain. Such studies can provide insights into the strengths and weaknesses of each approach, and their suitability for sentiment analysis in finance, thereby enabling more informed and effective sentiment analysis techniques in financial applications.

Chapter 3

Methods

To assess the performance of the lexicon-based and model-based methods for sentiment analysis in the financial domain, an experimental study was conducted on news articles using state-of-the-art lexicon-based and model-based methods. The study aimed to analyze the stock price reaction following the publication of news articles within different time windows.

3.1 Dataset selection

The aim is to collect a comprehensive dataset of financial news articles covering the S&P500 companies. The reason is that the top 500 companies by market cap in the US should have more news coverage compared to smaller companies and this should give us sufficient data to analyse the impact of news on stock prices.

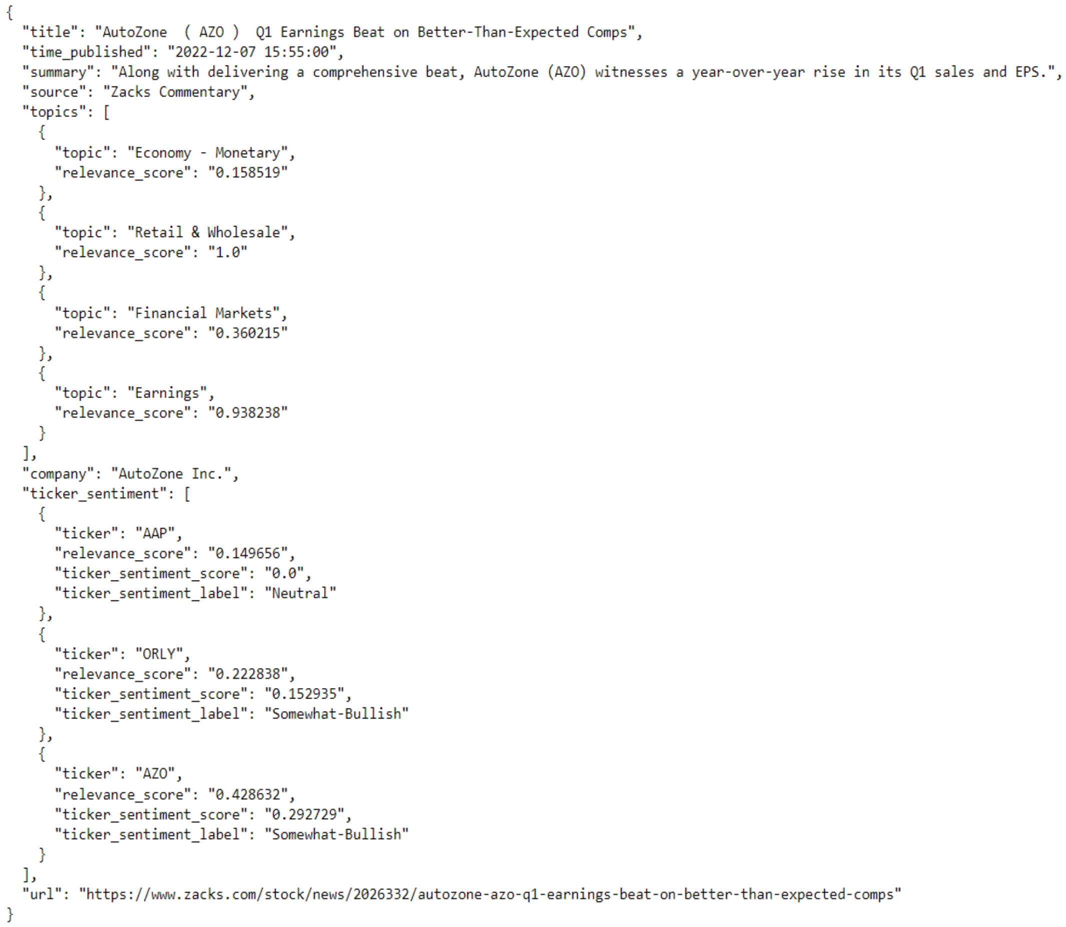
Our first attempt was to utilize the New York Times API. [3] However, it's not a scalable solution to retrieve the news articles of all S&P500 companies as we must first use the Semantic API to search for the organization name in The New York Times controlled vocabulary before using the Article Search API and its filter query to search for articles that are associated with the organization. Manual work was required to verify the correct organization name was the intended S&P500 company. Moreover, it doesn't cover news articles published by other news organization.

Our second attempt was to collect the news articles from Alpha Vantage. [4] Given a stock symbol, the API returns news data from over 50 major financial news outlets around the world with 1 year of history. The dataset includes the publication time,

[3] https://developer.nytimes.com/apis

[4] https://www.alphavantage.co/documentation/#news-sentiment

title and summary of the news articles and relevance score of the news to the stock (see Figure 1). Using the API, we retrieved all news articles related to S&P 500 companies and published between March 1, 2022, to March 31, 2023. To ensure relevance to our research objective, we applied a filtering criterion to select articles specifically focused on earnings-related topics, giving us a total of 245,719 articles.

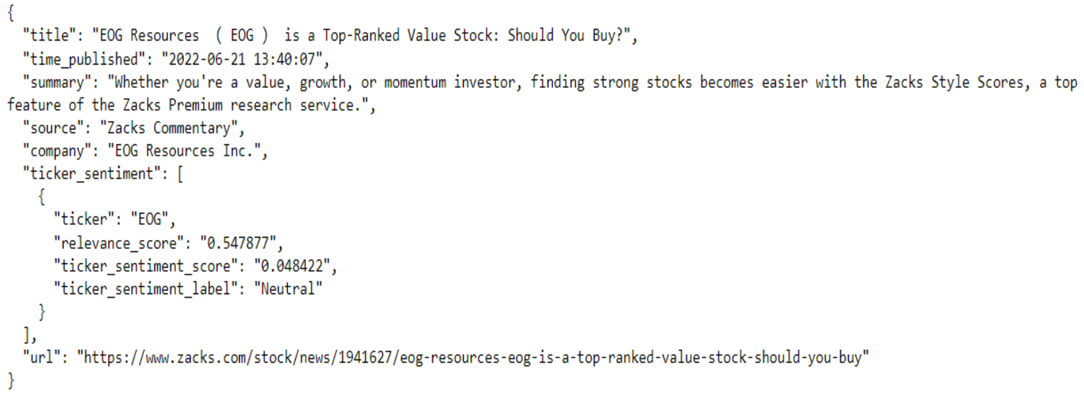


```
{
  "title": "AutoZone  ( AZO )  Q1 Earnings Beat on Better-Than-Expected Comps",
  "time_published": "2022-12-07 15:55:00",
  "summary": "Along with delivering a comprehensive beat, AutoZone (AZO) witnesses a year-over-year rise in its Q1 sales and EPS.",
  "source": "Zacks Commentary",
  "topics": [
    {
      "topic": "Economy - Monetary",
      "relevance_score": "0.158519"
    },
    {
      "topic": "Retail & Wholesale",
      "relevance_score": "1.0"
    },
    {
      "topic": "Financial Markets",
      "relevance_score": "0.360215"
    },
    {
      "topic": "Earnings",
      "relevance_score": "0.938238"
    }
  ],
  "company": "AutoZone Inc.",
  "ticker_sentiment": [
    {
      "ticker": "AAP",
      "relevance_score": "0.149656",
      "ticker_sentiment_score": "0.0",
      "ticker_sentiment_label": "Neutral"
    },
    {
      "ticker": "ORLY",
      "relevance_score": "0.222838",
      "ticker_sentiment_score": "0.152935",
      "ticker_sentiment_label": "Somewhat-Bullish"
    },
    {
      "ticker": "AZO",
      "relevance_score": "0.428632",
      "ticker_sentiment_score": "0.292729",
      "ticker_sentiment_label": "Somewhat-Bullish"
    }
  ],
  "url": "https://www.zacks.com/stock/news/2026332/autozone-azo-q1-earnings-beat-on-better-than-expected-comps"
}
```

Figure 1: An example of a news article pulled from the Alpha Vantage's News API. An article comes with title, summary, and corresponding metadata such as publication time, URL, and relevance score of each company mentioned in the article.

To mitigate the presence of irrelevant articles, a relevance threshold was applied to filter out those with minimal relevance to the tickers of interest. Specifically, articles with a relevance score below 0.5 were excluded from the dataset.

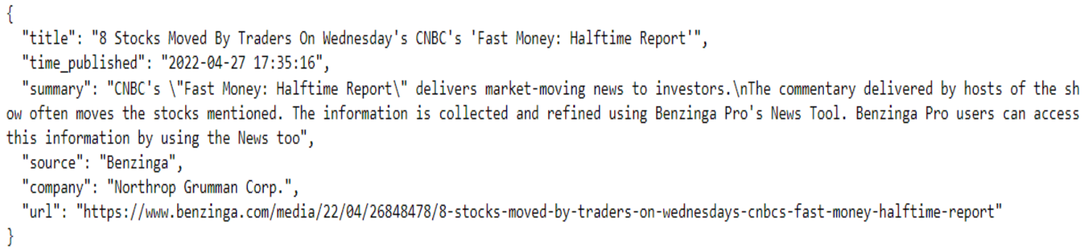
To prepare the input for the sentiment analysis model, a text column was created. The default approach involved using the article summary as the input. However, specific rules were applied to handle certain edge cases. If the summary was deemed uninformative (see Figure 2), alternative approaches were adopted. In such cases, either the article title alone was used as the input, or the title and summary were combined to provide additional context.



```
{
  "title": "EOG Resources  ( EOG )  is a Top-Ranked Value Stock: Should You Buy?",
  "time_published": "2022-06-21 13:40:07",
  "summary": "Whether you're a value, growth, or momentum investor, finding strong stocks becomes easier with the Zacks Style Scores, a top
feature of the Zacks Premium research service.",
  "source": "Zacks Commentary",
  "company": "EOG Resources Inc.",
  "ticker_sentiment": [
    {
      "ticker": "EOG",
      "relevance_score": "0.547877",
      "ticker_sentiment_score": "0.048422",
      "ticker_sentiment_label": "Neutral"
    }
  ],
  "url": "https://www.zacks.com/stock/news/1941627/eog-resources-eog-is-a-top-ranked-value-stock-should-you-buy"
}
```

Figure 2: An example of a news article with uninformative summary, even though it's highly relevant to the company (a high relevance score of 0.54). The title indicates that this article expresses a favourable view of the company.

In situations where both the summary and title are deemed uninformative, as depicted in Figure 3, the most optimal solution would be to scrape the corresponding website for relevant information. However, due to time limitations, developing a custom scraper for each website was considered beyond the scope of this project. Therefore, this approach was not pursued, but it can be considered as a potential avenue for future research. As a result of this filtering process, the total number of articles was reduced from an initial count of 245,719 to 33,318 articles that met the relevance criteria and were deemed more pertinent to the study.



```
{
  "title": "8 Stocks Moved By Traders On Wednesday's CNBC's 'Fast Money: Halftime Report'",
  "time_published": "2022-04-27 17:35:16",
  "summary": "CNBC's \"Fast Money: Halftime Report\" delivers market-moving news to investors.\nThe commentary delivered by hosts of the sh
ow often moves the stocks mentioned. The information is collected and refined using Benzinga Pro's News Tool. Benzinga Pro users can access
this information by using the News too",
  "source": "Benzinga",
  "company": "Northrop Grumman Corp.",
  "url": "https://www.benzinga.com/media/22/04/26848478/8-stocks-moved-by-traders-on-wednesdays-cnbcs-fast-money-halftime-report"
}
```

Figure 3: An example of a news article where both the summary and the title are uninformative. We excluded news articles of this nature from our analysis due to their lack of textual content suitable for sentiment extraction.

Alpha Vantage also offers intraday time series stock price data with 2 years of history from the Securities Information Processor (SIP) market-aggregated data. [5] We obtained the split/dividend-adjusted intraday data from this endpoint. While the API offers 1-minute interval price data, we opted for the 5-minute interval data due to memory constraints. The stock price data consists of Open, High, Low, Close (OHLC) prices and Volume for each 5-minute period, covering extended trading hours where applicable (e.g., 4am to 8pm Eastern Time for the US market).

3.2 Lexicon-based method

Our lexicon-based method utilized the well-established Loughran and McDonald's Financial Sentiment Word Lists (LM Words Lists). However, LM Words Lists are based on simple word count and do not explicitly handle negation and amplification. The lists assign positive and negative sentiment scores to words based on their association with positive or negative sentiments in financial contexts. The absence of explicit negation and amplification handling is the main limitation, as they can significantly affect sentiment polarity in text.

To overcome this limitation and capture more nuanced sentiment information, we incorporate LM Words Lists into the VADER sentiment analysis framework. This can be done by adding the financial-specific terms and their associated sentiment scores to the VADER lexicon. By combining VADER and LM Words Lists, we can apply more sophisticated grammatical and syntactical rules to handle punctuation, capitalization, degree modifier, contrastive conjunction “but” and negation.

On the other hand, lexical features from VADER are more adapted to social media contexts and applying it directly to financial text might not be appropriate. For example, certain words from accounting language are expected to appear in both positive and negative contexts (e.g. limited, liability). Removing the default word lists from VADER will help avoid labelling words that are typically not negative in a financial context as negative. Adopting LM Words Lists will allow for a more domain-specific sentiment analysis in financial contexts, leveraging the specialized sentiment knowledge provided by financial domain experts.

3.3 Model-based method

The model-based approach utilized the state-of-the-art sentiment analysis model, FinBERT, which had undergone pre-training on an extensive financial text dataset and exhibited superior performance compared to other transfer learning models.

[5] https://www.alphavantage.co/documentation/#intraday-extended

There are two versions of FinBERT that will be compared in this study. For clarity, the version developed by Araci will be referred to as FinBERT V1[6], while the version by Yang et al. will be referred to as FinBERT V2[7]. Both models are publicly available on the Hugging Face platform.

Initially, attempts were made to load the model from the transformer library and perform predictions using local machines. However, this approach encountered memory errors when conducting bulk predictions. Therefore, the decision was made to utilize the Inference API, which runs the inference process on the Hugging Face infrastructure and provides predictions in JSON format. While this approach requires adherence to the API rate limit to prevent resource overload, it reduces the computational resources required, albeit at the cost of longer prediction times for all news articles.

3.4 Time Window

Several studies have explored the optimal time window for new market information to manifest in stock prices. Gidofalvi and Elkan (2001) conducted a seminal investigation in this area, observing a significant correlation between news articles and stock price behavior within a 20-minute window after the release time of the news article. It is important to acknowledge that this experiment was conducted two decades ago, and since then, the dynamics of the stock market may have evolved. Nevertheless, this study provides a valuable starting point for determining an appropriate time window. Consequently, in this research, we have defined various time windows to capture the stock price reactions following the publication of news articles. These time windows encompass intraday reactions, such as 30 minutes and 1 hour, as well as longer-term effects spanning multiple days, including 1 day and 5 days.

3.5 Stock Price Return

The granularity of news article publication time is at the second level, whereas the stock price data is available in 5-minute intervals. To ensure consistency in aligning the news article with the corresponding stock price, we adopt the following approach. We consider the latest trading price just before the news article is published as the initial price, denoted as P0. For instance, if a news article is published at 9:07 AM, we select the price recorded at 9:05 AM as P0. Next, we define a holding period, such as 30 minutes, during which we aim to evaluate the stock price reaction. To determine the price at the end of this holding period,

[6] https://huggingface.co/ProsusAI/finbert

[7] https://huggingface.co/yiyanghkust/finbert-tone

denoted as P1, we consider the first available price after the holding period is over. For example, if the holding period is 30 minutes, we would consider the price at 9:40 AM as P1, representing the first available price after 9:37 AM. We then calculate the percentage return within the specified time window, which can be expressed as (P1/P0) - 1.

3.5.1 Missing Price

To address the issue of missing prices during trading hours (4 am to 8 pm Eastern Time) when no trading activity occurs within a 5-minute interval, we employ a forward fill approach. In this approach, if there is a missing price for a particular 5-minute interval, we utilize the price from the last available time as a substitute for the missing price. By propagating the last known price forward, we ensure that there are no gaps in the price data and maintain a consistent time series. The forward fill approach leverages the assumption that adjacent data points in a time series are often correlated. By filling in missing prices with the last available price, the approach preserves the temporal ordering and ensures a smooth transition between adjacent time intervals. This method also allows us to handle missing prices effectively and ensure a continuous stream of price data for the desired time intervals.

3.5.2 Out-of-Hours Publication

The inclusion of news articles published outside of trading hours, such as weekends or non-trading hours on weekdays, is crucial to maintain a comprehensive analysis and ensure that significant events are not excluded due to publication time. To address this, our methodology involves utilizing the price at the end of the last trading day as P0 and the first available price after the holding period as P1 for news published outside of trading hours.

Using the price at the end of the last trading day as P0 allows for a relevant and consistent baseline for evaluating the stock price reaction. This reference point takes into account the market context prior to the publication of the news article. Subsequently, selecting the first available price after the holding period as P1 ensures that the analysis captures the immediate impact of the news article on subsequent trading activity. This approach enables a standardized and consistent evaluation of stock price reactions across different news articles, regardless of when they were published.

3.6 Classification Labels

3.6.1 Lexicon-based Method

To convert sentiment scores into categorical labels, we employed the VADER ruleset, adhering to the following guidelines. If the sentiment score surpassed the threshold of 0.05, we categorized it as positive. Conversely, if the score fell below -0.05, we classified it as negative. Scores within the range between the positive and negative thresholds were designated as neutral. The justification for applying these specific rules lies in their alignment with established conventions in the use of VADER. The selected thresholds aim to capture sentiment intensity that is sufficiently distinct from neutrality, enhancing the granularity and accuracy of the sentiment classification process.

3.6.2 Model-based Method

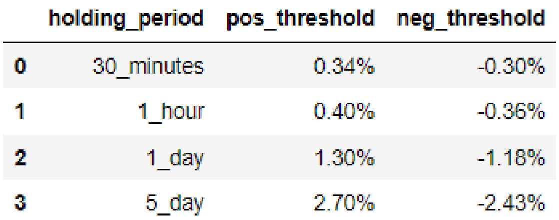
In contrast to the sentiment score generated by lexicon-based models, FinBERT produces a probability distribution encompassing positive (p1), negative (p2), and neutral (p3) labels, where the sum of these probabilities equals 1. However, relying solely on the label associated with the highest probability may not yield optimal results as it fails to consider the model's certainty in its prediction. For instance, if the probability distribution is nearly uniformly distributed (e.g., p1=0.4, p2=0.3, and p3=0.3), the model exhibits low confidence regarding the article's positivity. Consequently, a threshold is introduced to classify articles based on their respective labels. The classification threshold is regarded as a tunable parameter within our framework. Initially, we have set this threshold to 0.8. In Chapter 4 of this study, we will conduct hyperparameter tuning procedures to identify the optimal threshold value that maximizes predictive performance.

Articles with a probability exceeding the threshold are categorized accordingly. Any article where the probability of all labels falls below 0.8 indicates a high degree of uncertainty in the predicted label by the model, warranting a classification of neutral. The justification for employing this approach stems from the aim of incorporating the model's certainty into the classification process. By setting a threshold, we introduce a level of confidence that enhances the reliability of the sentiment classification. It prevents misclassifications based on uncertain predictions and promotes the inclusion of articles where the model exhibits a high degree of confidence in its sentiment assignment.

3.6.3 Stock Price Return

The classification of stock price percentage returns into positive, negative, and neutral labels requires careful consideration to account for the varying magnitudes of returns across different time windows. A uniform threshold for all time windows is inadequate as it fails to capture the disparity in return magnitudes between shorter and longer durations. To address this issue, we adopt a percentile-based approach that takes into account the distribution of returns within each time window. Specifically, we set the 25th percentile of the return distribution as the negative threshold and the 75th percentile as the positive threshold. By using percentiles, we ensure that the classification thresholds are tailored to the specific characteristics of each time window.

This approach is justified by the need to accurately classify returns while considering the differences in price movements across various time frames. For example, the average absolute return within a 30-minute window may differ significantly from that within a 5-day window. Utilizing percentiles allows us to dynamically adjust the thresholds based on the inherent characteristics of each time window. Consequently, a higher threshold is applied to longer time windows to account for the typically larger price movements observed within those periods. By employing percentile-based thresholds, we align the classification process with the relative magnitude of returns within each time window. Figure 4 shows the positive and negative threshold applied for each holding period. This approach ensures that a comparable proportion of returns in each window are assigned to positive, negative, and neutral labels, capturing the nuanced nature of stock price movements across different time horizons.



	holding_period	pos_threshold	neg_threshold
0	30_minutes	0.34%	-0.30%
1	1_hour	0.40%	-0.36%
2	1_day	1.30%	-1.18%
3	5_day	2.70%	-2.43%

Figure 4: Positive and negative threshold used for each holding period.

3.7 Performance Evaluation

In order to simplify the multi-class classification problem, we employed a binary classification approach for both the feature (sentiment label) and target variable (percentage return label). To begin, we focused solely on articles associated with positive or negative returns, as our objective did not involve predicting neutral return events. Subsequently, these selected articles were subjected to sentiment analysis algorithms capable of classifying them into three categories: positive, neutral, and negative. Articles classified as neutral by the sentiment analysis algorithm were excluded from further analysis, as no actionable decisions were to be based on neutral predictions. The adoption of a binary classification approach allowed us to focus exclusively on articles associated with discernible positive or negative returns, filtering out the neutral category for practical decision-making purposes.

Our primary objective was to assess the ability of the sentiment analysis algorithm to accurately predict the direction of return within the specified time window. Performance evaluation of the sentiment analysis algorithm involved comparing its predictions against the actual positive or negative returns observed. Various performance metrics such as accuracy and precision were employed to assess the effectiveness of both lexicon-based and model-based methods.

In predicting stock price returns, our focus was on prioritizing precision over recall. The emphasis on precision stems from the greater significance placed on the accuracy and reliability of predictions when the model indicates a positive or negative stock price return. Conversely, the measurement of capturing all possible positive or negative returns within our dataset, reflected by recall, held less importance. The justification for prioritizing precision lies in the aim of obtaining precise and trustworthy predictions to inform decision-making processes. In financial contexts, the accuracy of predicting stock price movements is crucial for investors and stakeholders who rely on reliable information to guide their actions. By emphasizing precision, we prioritize the minimization of false positives and the provision of accurate signals when the model predicts a positive or negative stock price return.

Chapter 4

Results

4.1 Accuracy and Precision

Looking at Figure 5, FinBERT V2 demonstrates superior accuracy, positive precision and negative precision compared to both the lexicon-based method and FinBERT V1 when we compute the mean metric across various holding periods.

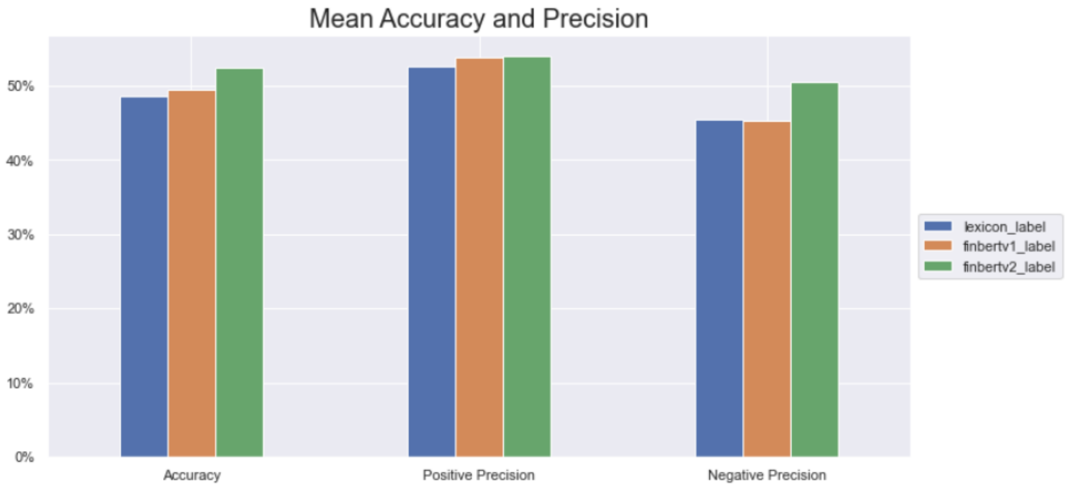


Figure 5: The plot shows that FinBERT V2 outperforms lexicon-based method and FinBERT V1 in terms of average accuracy, positive precision and negative precision. The mean metric is computed across different holding periods.

4.1.1 Accuracy

A qualitative examination of error analysis unveils several factors contributing to the enhanced performance of FinBERT’s accuracy. Firstly, unlike lexicon-based methods, FinBERT is pre-trained on a substantial corpus of financial text data, enabling it to understand the intricate nuances, jargon, and context prevalent in financial language. This contextual understanding allows it to accurately decipher the sentiment behind complex financial terms and expressions, which lexicon-based methods may struggle to capture effectively.

> *Example 1: Moderna took a nearly $500 million hit on write-downs for vaccine inventory that has expired or is expected to expire before it can be used.*

The term "write-down," absent from the negative word list, leads the lexicon-based model to render a neutral prediction. In contrast, FinBERT, a product of diverse financial text pre-training, accurately derives a negative sentiment prediction.

Secondly, financial news articles can have complex sentence structures that require an understanding of relationships between various parts of a sentence. FinBERT's deep learning approach excels in capturing these intricate dependencies, enabling it to grasp the sentiment within complex sentence formations.

> *Example 2: Inflation could interfere with Hasbro's new game plan to boost profits and sales.*

A lexicon-based approach, guided by a straightforward word list association, ascribes a neutral sentiment prediction to this sentence, due to the inclusion of the terms "interfere" and "boost" within the lexicon's negative and positive categories, respectively. It fails to effectively discern the nuanced sentiment conveyed by the sentence. In contrast, FinBERT, harnessing its contextual understanding, accurately predicts the negative sentiment.

4.1.2 Precision

In our analysis, it becomes evident that all three models display a bias towards predicting positive labels with greater precision than negative labels. This observation holds particular significance as it was not previously identified during the literature review stage, nor was it explicitly cautioned by the original developers of FinBERT. However, we note that the gap between the precision scores for positive and negative sentiment predictions is narrower for FinBERT v2 in comparison to both FinBERT v1 and the lexicon-based method.

One plausible explanation for this phenomenon may be rooted in the composition of the financial text data used for the pretraining of FinBERT. It is conceivable that a substantial proportion of the text corpus used for pretraining predominantly comprises positive sentiment texts. Consequently, FinBERT may be more effective in recognizing and interpreting patterns associated with positive sentiment expressions, thereby leading to the observed precision bias in favor of positive labels in its predictions.

On the other hand, the bias within the lexicon-based method may stem from the structural composition of the pre-defined lexicon itself. Notably, the Loughran and McDonald's Financial Sentiment Word Lists exhibit an overrepresentation of negative terms, which surpasses the frequency of positive terms by a ratio of 6.8 to 1. This skewed distribution inherently inclines the lexicon-based method towards predicting negative sentiment labels, thereby contributing to a diminished precision in its predictions.

4.2 Holding Period

Looking at Figure 6, across all three models, there is a gradual decrease in accuracy as the holding period extends. The highest level of accuracy is observed when the holding period is set at 30 minutes, suggesting that this duration might represent an optimal window for news dissemination and participant response in the market. However, it is noteworthy that the accuracy of all three models drops below 50% when the holding period is extended to 1 day and 5 days. This suggests that a longer time frame could introduce noise that masks the influence of news events on stock price movements.

This observation substantiates the findings of Gidofalvi and Elkan (2001), whose experiment demonstrated a substantial correlation between news articles and subsequent stock price behavior within a 20-minute window following the release of the news article. Despite the fact that this investigation was conducted approximately two decades ago, it is noteworthy that the pace at which news articles disseminate information into the stock market has not exhibited significant evolution since that time.

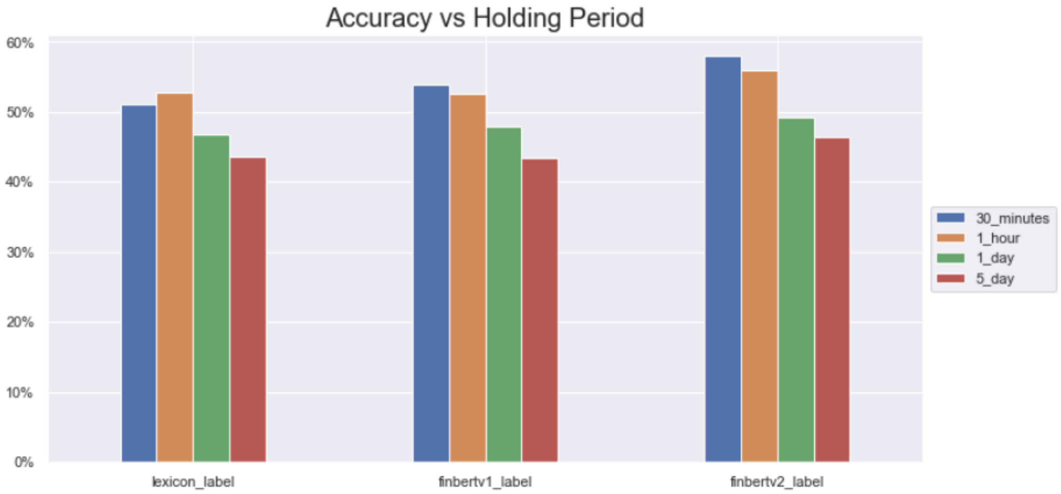


Figure 6: The plot shows the effect of holding period on accuracy for different models. In general, there is a decline in accuracy as the duration of the holding period is extended.

4.3 Classification Threshold

As previously mentioned, the FinBERT model generates a probability distribution across positive, negative, and neutral sentiment labels, necessitating the application of a classification threshold to map these probabilities into discrete classification labels. Prior to this section, a default threshold of 0.8 was employed for both FinBERT V1 and V2. In this section, we explore the impact of varying these classification thresholds on various evaluation metrics. To conduct this analysis, we selected the FinBERT V2 model in conjunction with a 30-minute holding period, a combination that has exhibited superior performance in our study thus far.

Upon reviewing Figure 7, we note that as the classification threshold increases, there is a slight corresponding improvement in accuracy. This trend aligns with intuitive expectations, as higher-confidence predictions by the model would logically contribute to enhanced overall accuracy. However, it is important to emphasize that while this effect of the classification threshold on accuracy is observable, it remains relatively modest. Furthermore, the precision metric exhibits minimal fluctuation across varying threshold values. This suggests that the model maintains consistent precision levels regardless of the specific classification threshold employed.

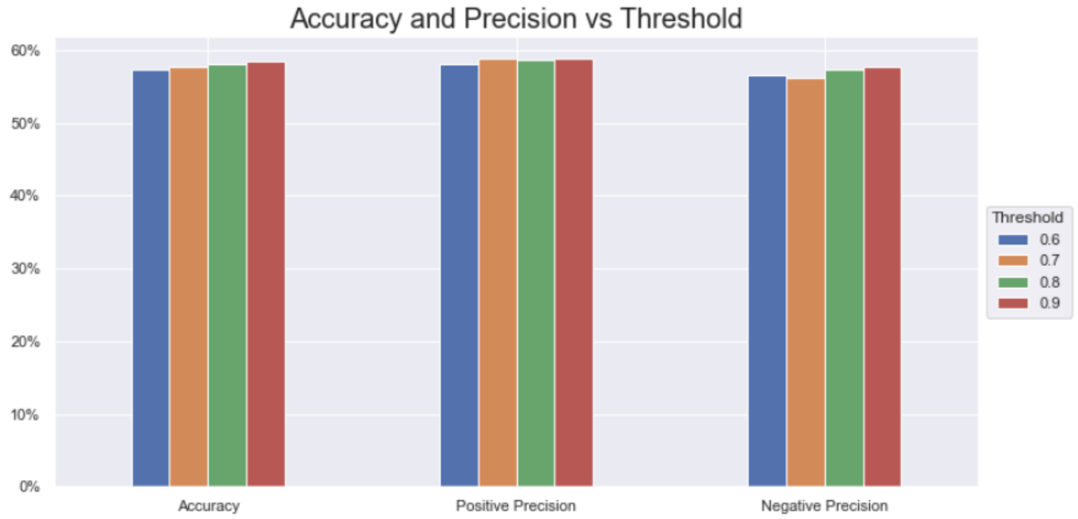


Figure 7: The plot shows the effect of changing Finbert V2 threshold on accuracy and precision for 30-minute holding period.

4.4 Time and cost

In our experiment, a notable observation emerged wherein the lexicon-based method displayed a distinctive advantage in terms of time and cost efficiency for both implementation and deployment, when compared with the FinBERT model. The lexicon-based approach involves a streamlined process that primarily relies on predefined dictionaries, thus requiring minimal computational resources and relatively less time for implementation. This efficiency not only reduces the operational overhead but also simplifies the deployment process, making it an attractive choice particularly for real-time or large-scale applications.

While the accuracy of the lexicon-based method might be comparatively lower due to its inherent limitations in capturing nuanced sentiments, its efficiency in terms of computational requirements and implementation timelines might outweigh this drawback. This is especially relevant in scenarios where rapid decision-making is essential, such as high-frequency trading or time-sensitive market analysis. However, as hardware technology continues to advance, particularly in the field of graphics processing units (GPUs), it has the potential to mitigate the time and cost constraints associated with training and deploying sophisticated models like FinBERT, making them more accessible and practical for various applications.

4.5 Interpretability

Within our error analysis, we have encountered a significant hurdle in grasping the rationale behind FinBERT's predictions. This obstacle to interpretability is primarily rooted in the complex nature of FinBERT's deep learning architecture, characterized by an intricate network of layers and an extensive array of interconnected neurons. These neural networks process textual information in a highly nonlinear and convoluted manner, making it exceptionally challenging to delineate how individual input features contribute to the model's predictions.

Moreover, FinBERT represents words within high-dimensional vector spaces, wherein each dimension corresponds to a learned feature. The comprehension of these high-dimensional embeddings and the collective influence they exert on sentiment predictions is far from straightforward. Consequently, FinBERT does not readily offer feature importance scores or provide clear explanations for its predictions. It lacks built-in mechanisms for attributing specific sentiment classifications to particular words or phrases within the input text. This opacity in interpretability is a significant drawback of the FinBERT model.

In contrast, lexicon-based methods offer transparency in their decision-making process. They use predefined word lists to classify sentiment, allowing for straightforward inspection and debugging. Users can easily track which words contribute to a positive or negative sentiment classification. This transparency enables the lexicon-based method to be refined by adding missing terms. However, this manual refinement process can lead to overfitting, where the model becomes too tailored to the training data and performs poorly on new data. Additionally, the lexicon-based approach is less scalable since it necessitates human intervention to maintain and improve the sentiment lexicon.

Conclusion

5.1 Contribution

The primary objective of this research paper is to conduct a comparative analysis between state-of-the-art algorithms in lexicon-based and model-based methods for sentiment analysis. We conducted a thorough literature survey to research the development of both model categories and implemented the state-of-the-art methods within each domain. Intraday stock price data and news articles of S&P500 companies were then collected, processed, and synthesized. The performance evaluation encompassed the prediction of stock price movements across varying holding periods. While model-based methods are expected to outperform lexicon-based methods in theory, prior comparative studies yielded inconclusive outcomes, mainly because model-based methods require large amount of labelled training data to be useful in practice. Our investigation indicates that the recent advancement in model-based approach, specifically FinBERT, has successfully surmounted the constraints previously associated with model-based methodologies, exhibiting superior predictive performance when compared with lexicon-based counterparts.

In our extensive analysis, we scrutinized 2 different versions of FinBERT model and identified noteworthy distinctions in their performance characteristics. Notably, when comparing FinBERT v1 and v2, the latter demonstrated superior accuracy and precision. While differences in precision scores were observed between positive and negative sentiment predictions, these distinctions were less pronounced for FinBERT v2. This finding indicates that FinBERT v2 offers a more balanced and consistent performance, showing no significant bias in favour of either positive or negative sentiment labels.

Furthermore, we extended our investigation by conducting backtesting across diverse time intervals, aiming to discern the optimal temporal window for news articles to exert their influence on the financial market. Our empirical findings revealed that a 30-minute interval serves as the optimal timeframe within which the impact of news articles becomes discernible within the market dynamics. This temporal choice aligns with the notion that financial markets respond swiftly to new information, and this swift reaction is most evident within the initial half-hour following the dissemination of relevant news.

Additionally, we delved into an examination of the classification threshold, a parameter critical for converting FinBERT's probabilistic outputs into definitive sentiment labels. Our investigation into the impact of varying classification thresholds on evaluation metrics revealed that alterations in this parameter yielded only modest effects on the overall model performance. This outcome suggests that FinBERT's predictive stability remains relatively robust across a range of classification thresholds, reinforcing its consistency as an analytical tool in sentiment analysis tasks.

FinBERT's effectiveness can be attributed to its ability to grasp the complexities of financial language, adjust to changing contexts, and make nuanced sentiment predictions based on its extensive pre-training. This provides a distinct advantage over lexicon-based methods, which are often constrained by predefined sentiment lists and lack the depth of comprehension that FinBERT's advanced deep learning framework offers. As our findings demonstrate that FinBERT has a predictive accuracy better than random chance for news articles, investors could potentially utilize this insight for developing trading strategies in the financial market.

Despite its notable advancements in accuracy and precision, FinBERT, like many sophisticated deep learning models, is not without its inherent drawbacks. The primary concerns revolve around issues related to time and cost efficiency as well as interpretability, both of which are critical factors in the practical deployment of sentiment analysis models within financial domains.

Firstly, the time and cost efficiency of FinBERT poses a noteworthy challenge. The training and fine-tuning of complex neural network models like FinBERT typically demand substantial computational resources, including high-performance hardware and extensive training times. These resource-intensive requirements translate into elevated costs, making it less accessible for smaller organizations or researchers with limited budgets. Additionally, the deployment of such models in real-time applications can be computationally burdensome, which might hinder their adoption in scenarios where quick decision-making is paramount. Thus, while FinBERT excels in predictive performance, these resource-related constraints could limit its practicality in certain operational contexts.

Secondly, interpretability remains a substantial concern with FinBERT. Its effectiveness is largely attributed to its complex deep learning architecture, consisting of numerous layers and thousands of neurons. While this complexity contributes to its superior performance, it also renders the model a "black box" in terms of interpretability. Understanding how FinBERT arrives at specific predictions is a formidable challenge, as the intricate interplay of features and the high-dimensional embeddings employed make it difficult to discern the underlying rationale. Consequently, FinBERT does not readily provide transparent explanations or feature importance scores, impeding the ability to pinpoint the precise textual cues or linguistic elements that influence its sentiment predictions. In an industry where decision-makers often require transparent insights to inform financial strategies, this lack of interpretability poses a significant limitation.

5.2 Further work

5.2.1 Data

A prospective avenue for enhancement would involve the acquisition of news article content through the utilization of URL links, as alluded to in Section 3.1. This stems from the recognition that certain summaries within the dataset may not entirely capture the substantive essence of the corresponding news articles. This augmentation could potentially enhance the accuracy and depth of the sentiment analysis by providing a more comprehensive understanding of the news articles' content.

5.2.2 Methodology

Financial prediction using news articles involves two primary steps: analyzing the sentiment of news content and using this sentiment to predict stock price movements. In our methodology, we consider the stock price movement as the ground truth for the sentiment label of news articles. However, it's plausible that while the sentiment analysis might be precise, its applicability to accurate stock price prediction could be limited. With adequate resources and manpower, an alternative approach would be manually annotating sentiment for each news article, enabling a more isolated assessment of these two prediction phases.

5.2.3 FinBERT

In our literature review, we have identified the existence of a third iteration of FinBERT (Liu et al., 2021). This version introduces a novel approach, involving the construction of six pre-training tasks, concurrently applied to both general language corpora and specialized financial domain corpora. The objective is to augment FinBERT's proficiency in comprehending language intricacies and capturing nuanced semantic information. Regrettably, despite our diligent efforts, we have been unable to locate the pre-trained model or access the source code pertaining to this version of FinBERT through online resources. This absence of access has constrained our ability to perform a comparative analysis between this third iteration and the preceding two. Nevertheless, this intriguing development presents a promising trajectory for potential future research efforts. Subsequent investigations could pivot towards evaluating and contrasting the performance of these distinct FinBERT versions, thereby shedding light on the enhancements and divergences intrinsic to each.

5.2.4 Generative AI

An emerging avenue for research involves the utilization of Generative AI for the dual tasks of sentiment analysis and stock price prediction. A notable advantage of this approach lies in its potential for enhanced interpretability. Unlike conventional models such as FinBERT, where comprehending the rationale behind specific predictions can prove challenging, Generative AI offers a distinctive advantage by enabling explicit inquiry into the decision-making process. This feature empowers researchers to seek detailed explanations for predictions, facilitating the identification and rectification of any flawed assumptions or reasoning that may have influenced the outcome. Consequently, the integration of Generative AI into sentiment analysis and stock price prediction could contribute to a more transparent and accountable predictive framework.

BibliographyReferences

Alostad, H., & Davulcu, H. (2015). Directional prediction of stock prices using breaking news on twitter. 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Lrec,

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in International Business and Finance*, *64*, 101881.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., & Pandey, N. (2021). Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, *7*, e340.

Gupta, I., Madan, T. K., Singh, S., & Singh, A. K. (2022). HiSA-SMFM: historical and sentiment analysis based stock market forecasting model. *arXiv preprint arXiv:2203.08143*.

Gupta, R., & Chen, M. (2020). Sentiment analysis for stock price prediction. 2020 IEEE conference on multimedia information processing and retrieval (MIPR),

Halder, S. (2022). FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis. *arXiv preprint arXiv:2211.07392*.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363-407.

Hutto, C., & 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,

Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*.

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38-48.

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence,

Long, S., Lucey, B., Xie, Y., & Yarovaya, L. (2023). “I just like the stock”: The role of Reddit sentiment in the GameStop share rally. *Financial Review*, *58*(1), 19-37.

Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*(8).

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.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. 2019 IEEE fifth international conference on big data computing service and applications (BigDataService),

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, *1*(4), 7.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 international conference on signal processing, communication, power and embedded system (SCOPES),

Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the effects of news sentiments on the stock market. 2018 IEEE International Conference on Big Data (Big Data),

Sidogi, T., Mbuvha, R., & Marwala, T. (2021). Stock price prediction using sentiment analysis. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC),

Sohangir, S., Petty, N., & Wang, D. (2018). Financial sentiment lexicon analysis. 2018 IEEE 12th international conference on semantic computing (ICSC),

Turner, Z., Labille, K., & Gauch, S. (2021). Lexicon-based sentiment analysis for stock movement prediction. *Journal of Construction Materials*, *2*(3), 3-5.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.