JOHANNES KEPLER
UNIVERSITY LINZ

Author
Miguel Azuero, BSc

Submission
Institute of Business Informatics - Data & Knowledge Engineering

Thesis Supervisor
Assoz.-Prof. Mag. Dr. Christoph Schütz

Assistant Thesis Supervisor
Simon Staudinger, MSc

Month Year
June 2024

Topic Modeling and Sentiment Analysis for Predicting the Direction of Movement of the S&P 500 Index Based on Financial News

Master's Thesis

to confer the academic degree of

Master of Science

in the Master's Program

Economic and Business Analytics

**JOHANNES KEPLER
UNIVERSITY LINZ**
Altenberger Straße 69
4040 Linz, Austria
jku.at

Abstract

Predicting stock prices and determining the optimal moments to buy or sell stocks is a long-standing challenge for investors. Advances in natural language processing (NLP) allow for extracting valuable insights from unstructured text, and diverse studies have used news articles to predict the stock market, employing techniques such as lexicon-based sentiment analysis and topic modeling through Latent Dirichlet Allocation (LDA). These traditional approaches, however, do not consider the semantic relationships among words. Language models that use text embedding techniques, such as BERT, have gained popularity in the NLP field for their ability to consider the context of words.

This thesis evaluates the use of BERT-based topic modeling and sentiment analysis of financial news in the context of training a classifier to predict the direction of movement of the S&P 500 index. On the one hand, this thesis evaluates BERT-based models that consider semantic relationships among words, specifically FinBERT and BERTopic, in conjunction with various classification algorithms, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, among others. On the other hand, to provide a benchmark, the method is applied with the same classification algorithms using traditional techniques for sentiment analysis and topic modeling that do not consider word context. The benchmark sentiment analysis relies on a lexicon-based approach utilizing the Loughran and McDonald dictionary, while the topic modeling employs Latent Dirichlet Allocation (LDA).

The comparison between the BERT-based method and the selected benchmark involves evaluating the accuracy, precision, sensitivity, and other classification metrics. Furthermore, the research explores the influence of several factors on prediction outcomes, including the size and frequency of training the topic model and the impact of utilizing only the headline versus the full article. The results indicate that BERT-based methods marginally outperform traditional approaches in predicting stock price direction. However, it has become apparent that relying solely on sentiment information and topic models derived from financial news may not suffice for accurately forecasting the S&P 500 index's direction.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
BERT	Bidirectional Encoder Representations from Transformers
BERTopic	BERT model pre-trained for Topic Modeling
FinBERT	BERT model pre-trained for sentiment analysis on financial text
GBC	Gradient Boosting Classifier Model
JST	Joint Sentiment-Topic Model
LDA	Latent Dirichlet Allocation
LM	Language Model
NLP	Natural Language Processing
SVC	Support Vector Machine Classifier
SVM	Support Vector Machine

1. Introduction

Stock prices are hard to predict because they depend on a wide variety of factors, including political and economic developments but also investors’ sentiment. Still, predicting stock prices and consequently the optimal moment to buy or sell stocks has been described as “one of investors’ most persistent endeavors” [1, p. 118].

In financial theory, the concept of market efficiency states that investors base their investment decisions under the premise that security prices incorporate and "fully reflect" all available information, thereby defining such a market as "efficient” [2, p. 383] . Examples of such information include financial statements but may also be news reports. Under the assumption of market efficiency, the analysis of publicly available information should thus allow for the prediction of stock prices, but analyzing and interpreting the information in a systematic way is challenging. First, collecting all relevant information is impossible in practice. Second, the relevant information is often only available in natural language text, i.e. in an unstructured form.

Birz and Lott [3] investigated how macroeconomic news influences the U.S. stock market, finding evidence that macroeconomic news affects the stock market. Ranganathan and Brown [4] observed a positive correlation between announcements of enterprise resource planning (ERP) systems adoption and the consequent stock returns of the involved firms. The assembled data of ERP announcements included newspapers and wired report services. Furthermore, Birz and Lott suggest that not just the existence of the news itself, but the investors’ interpretation of the news affects stock prices. Sidogi et al. [5] also found correlation between the movement of stock prices and the publication of financial news.

Recent advances in natural language processing (NLP) allow for the automatic extraction of useful insights from unstructured natural language text. Examples of such insights include a sentiment score, which quantifies the sentiment of a text, i.e., whether the text expresses a positive or negative opinion, and lists of topics or common themes in text documents. The consideration of insights extracted through NLP from textual information (e.g., news, tweets, blogs) may improve the accuracy of stock price predictions [6] [7].

This thesis investigates the use of NLP technologies to extract insights from financial news to predict the evolution of stock prices. In particular, the thesis investigates the use of language models to uncover topics or common themes in financial news and analyze the sentiment of financial news, which then serves as the input for training a model for predicting the direction of the movement of the S&P 500 stock price index.

The remainder of this chapter comprises the problem statement, a description of the objectives, and a summary of the outline of the thesis.

1.1. Problem Statement

News sentiment is frequently employed as a proxy for assessing investor sentiment, providing insights into society's confidence in financial markets [8, p. 158]. There have been different studies that consider the effects of news articles' sentiment on the stock market [5] [6] [7] [8] [9] [10]. Medhat et al. [11] conducted a survey of different sentiment analysis articles and determined that by the time of their research (2014), lexicon-based approaches were most common. Lexicon-based approaches consist of using a dictionary/lexicon that gives a score to specific words and calculates the sentiment based on the count of positive and negative words. Still in 2023, Leippold [12] asserts that in the finance literature, lexicon-based approaches are the most used for sentiment analysis, with examples including the Loughran and McDonald dictionary [13], which is a finance-specific dictionary. This dictionary is further explained in the literature review. According to Araci [14], the problem with such word counting approaches is their limitation to analyze deeper semantic meaning of a given text.

Topic modeling has been used in stock market prediction [15] [16] [17]. Topic modeling aims at extracting common topics from a corpus of documents and associating topics with new documents. According to Maier et al. [18, p. 94], among the different statistical algorithms used for topic modeling, Latent Dirichlet Allocation (LDA) stands out as a widely employed and general model. For example, Yono et al. [16] proposed the supervised Joint Sentiment Topic (sJST) model, which uses LDA together with sentiment analysis of news on currency exchange in the context of predicting the movement of foreign exchange markets.

According to Grootendorst [19] conventional topic models, such as LDA consider a document text as a *bag of words*, but through these representations of *bag of words*, they do not consider semantic relationships among words. Egger and Yu [20] evaluated the performance of different topic modeling techniques and provided advantages and disadvantages of the different techniques. Egger and Yu [20] stated that the problem with LDA assumption of topic independence means that it relies only on the frequency of the common occurrence of words. Consequently, no relationships between topics are considered and word correlations are ignored.

According to Egger and Yu [20], LDA is highly dependent on the initial parameters, the user must decide the number of topics in advance and “hyperparameters need to be tuned carefully”. Although the use of topic modeling with techniques such as LDA has been used widely across various domains in the past years, including for the purpose of stock market prediction [15] [16] [17], LDA has some disadvantages. LDA disregards semantic relationships among words [19] and has the following other disadvantages according to Egger and Yu [20]. First, word correlations are ignored. Furthermore, LDA is highly dependent on initial parameters and requires extensive hyperparameter tuning.

Some research on sentiment analysis of financial news for stock price prediction has been done considering only the headlines [5] [9] and not the entire content of the articles. In the case of Bakker [9], technical difficulties for executing sentiment analysis methods on the whole article text prevented the research from considering the content and therefore considered only the headlines of the articles.

Considering that headlines and the content of the articles vary in size and wording, the sentiment analysis score between headline and content possibly differs and, therefore, also the effect on stock price direction prediction may differ. Not many studies consider the impact on stock price prediction of analyzing news headlines on the one hand or the full article on the other hand. One study that considers the difference between sentiment of title and content was done in the year 2022 by Fazlija and Harder [10], in which they analyzed sentiment of financial news to predict the direction of stock prices, evaluating titles and content of the news articles. They concluded that their method achieved better results with news content than using only the title.

1.2. Objectives

Considering the previously identified problems, this thesis investigates the use of alternative approaches to sentiment analysis and topic modeling for financial news in the context of stock price prediction. In the following, the investigated approaches are briefly presented.

Jurafsky and Martin [21], refer to “language models” as models that set probabilities to upcoming words or sequences of words. Jurafsky and Martin [21] further state that different NLP tasks such as question answering, summarization, and sentiment analysis can be formulated as tasks of word prediction, meaning that pre-trained language models can later be fine-tuned for different applications.

A novel approach to language models was introduced in the year 2019 with a language model (LM) called BERT by Devlin et al. [22], which stands for Bidirectional Encoder Representations from Transformers. BERT considers the context of words in a bidirectional way (considering the words surrounding the target word to the left and to the right). In this way, the LM considers the meaning of words by their context. BERT models have had different applications and the proposal of integration of BERT models in topic modeling and sentiment analysis on financial news for stock price could yield good results.

Considering the challenges and problems of lexicons mentioned in the problem statement, Araci [14] states that NLP transfer learning methods (such as BERT) could be a good alternative solution. In the year 2019, Araci introduced FinBERT [14], which is a language model based on BERT and fine-tuned for sentiment analysis for the finance field. FinBERT was evaluated on different financial datasets, achieving state-of-the-art performance in sentiment analysis within the financial domain.

Traditional approaches to topic modeling such as LDA disregard semantic relationships among words. According to Grootendorst [19], as response to this limitation, text embedding techniques (such as BERT) have gained popularity in the NLP field as they have achieved positive results in generating word and sentence vector representations that consider the context. Taking this into consideration, Grootendorst presents BERTopic [19], which according to the author performs well compared to a variety of benchmarks and solves the problem of disregarded semantic relationships among words, generating coherent topics.

Given the advantages offered by BERT-based models for sentiment analysis and topic modeling, and considering the successful results achieved by integrating topic modeling and sentiment analysis for stock price prediction [16], it is reasonable to assume that the integration of BERT-based models in topic modeling and sentiment analysis for predicting the direction of movement of the S&P 500 index based on financial news could yield good results.

In summary, the objective of the thesis is to evaluate the performance of using language models for analysis of financial news together with classification models to predict the direction of movement of the S&P 500 index. In particular, the thesis investigates the use of FinBERT for sentiment analysis and BERTopic for topic modeling to extract sentiment scores and topics from financial news and use these data to predict the direction of movement of the S&P 500 index. Furthermore, the thesis compares the difference between using the headlines only or the content of the news articles.

1.3. Structure of the Thesis

The remainder of this thesis is structured as follows. Chapter 2 provides an overview of existing NLP approaches for stock price prediction. Chapter 3 explains the general method of using sentiment analysis and topic modeling as the input of classification models that predict the direction of movement of the S&P 500 index. Chapter 4 describes the datasets employed in the analysis. Chapter 5 then describes a BERT-based implementation of the general method. Chapter 6 describes the baseline implementation of the general method, using a lexicon-based approach for sentiment analysis and LDA for topic modeling. Chapter 7 discusses the results of using the BERT-based implementation and the baseline implementation together with different classification models. Chapter 8 concludes the thesis with a summary of insights, a discussion of limitations, and proposals for further research.

2. Literature Review

This chapter provides an overview of the relevant NLP techniques and their integration with prediction models for the purpose of stock price prediction. There have been different studies with different approaches that find good prediction models. The first section presents literature review of studies on sentiment analysis, the second one on topic modeling, and the last one the combination of sentiment analysis and topic modeling.

2.1. Sentiment Analysis

A popular use of NLP models applied to predicting stock market has been associated with sentiment analysis [5] [6] [7] [9] [10]. Following the book of Jurafsky and Martin [21], sentiment analysis is a common task in NLP that considers a text and associates a positive or negative orientation through a label or a number considering the words that the writer used.

A type of sentiment analysis technique that has been used widely is the ones related to dictionaries or lexicons [12]. Loughran and McDonald [23] use the term “word list”, which refers to a compilation of lists of words with similar sentiments (e.g., positive, negative, uncertain). Loughran and McDonald stated that with the support of these lists, the researcher can provide a measure of sentiment that is comparable by counting words that are associated with the lists.

Loughran and McDonald [23] did a compilation of the different textual analysis techniques, the specificities of the methods and the problem of how dictionaries were used at that time (to measure the tone of the text many of the previous research relied on negative word counts). In another paper, Loughran and McDonald [13] demonstrated that word lists elaborated for other disciplines do not classify correctly common words in financial text, so they developed alternative word lists that better reflect sentiment in financial text.

Another approach is the polarity lexicons which are different than a simple word counting method, as Demiroz [24] states, polarity lexicons indicate how positive or negative each term in the lexicon is. This means that the polarity lexicon weights the polarity and not only classifies the words as positive or negative. Malo *et al.* [25] highlighted that the effectiveness of polarity-lexicon-based approaches is considerably influenced by the context and domain of the subjective or opinionated statements. This means that considering only word counting methods or polarity-lexicons leaves behind the interpretation of the context of a sentence which can misrepresent the sentiment of a phrase.

Malo *et al.* [25] introduced the Linear Phrase Structure (LPS) model for semantic orientations which relies on different steps and heuristics to obtain the sentiment score. The approach showed an improvement as it includes the interaction between financial concepts and verbs (and other directional expressions) to capture the impact of the interaction within the economic and financial domain. This leads to a higher level of complexity compared to word counting methods as they divide a sentence into different segments and evaluate the type of words.

As seen, the NLP task of sentiment analysis relied on word counting methods, polarity lexicons and in analyzing the phrase structure. A revolutionary technique came to change the NLP paradigm with the publication of a language model known as BERT (Bidirectional Encoder Representations from Transformers) [22]. BERT relies on WordPiece embeddings [22] which is an approach to word representation through distributed representation (instead of complete words, the words are divided into sub-word units [26]). “Distributed word representations are called word embeddings” [27], so this approach represents words in vectors with different dimensions with numeric values. According to Turian *et al.* “Each dimension of the embedding represents a latent feature of the word” [27], with the intention of capturing useful semantic and syntactic properties. One example for understanding this concept is given in the proposal paper of *Paragraph Vector* Algorithm of Le and Mikolov [28]. Le and Mikolov [28] evaluated the quality of the resulting vector representations, expecting that words would present varying degrees of similarity (and proximity between similar words). Le and Mikolov reference the example of the previous research of Mikolov [29] using algebraic operations on word vectors: The example involved the algebraic operation of subtracting the vector for the word "Man" from the vector for the word "King" and adding the vector for the word "Woman," resulting in a vector representation closely resembling that of the word "Queen”. As a reference the *Paragraph Vector* algorithm is used in the in the commonly known genism python library Doc2Vec [30].

For understanding BERT-based models it is also important to understand how BERT is pretrained. BERT is pre trained in two tasks [22], Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In the first task, a randomly selected 15% of all WordPiece tokens in a sequence is "masked". The masked tokens are predicted with a classification layer (softmax) over the vocabulary. The second task is used to train a model that understands sentence relationships. The task consists of identifying if sentence A is followed by sentence B or not.

BERT can be finetuned to create state-of-the-art models for different tasks (question answering, sentiment analysis, language inference and others) with no need of substantial architecture modifications [22]. FinBERT, which is a BERT model trained for sentiment analysis on financial texts was developed by Araci [14]. In the results of the author, FinBERT outperforms state-of-the art models. In Table 1 results are divided in 3 sections, the first section are models that the author implemented for contrastive experiments, such as LSTM which stands for Long short-term memory models and ULMFit which stands for Universal Language Model Fine-tuning for Text Classification. The second section the author took the results from the respective papers of the models (for more detail refer to the paper [14]). In the final section are the values of the FinBERT model.

	All data			Data with 100% agreement		
Model	Loss	Accuracy	F1 Score	Loss	Accuracy	F1 Score
LSTM	0.81	0.71	0.64	0.57	0.81	0.74
LSTM with ELMo	0.72	0.75	0.70	0.50	0.84	0.77
ULMFit	0.41	0.83	0.79	0.20	0.93	0.91
LPS	-	0.71	0.71	-	0.79	0.80
HSC	-	0.71	0.76	-	0.83	0.86
FinSSLX	-	-	-	-	0.91	0.88
FinBERT	**0.37**	**0.86**	**0.84**	**0.13**	**0.97**	**0.95**

Table 1: Experimental Results on the Financial PhraseBank dataset. Adapted from [14]

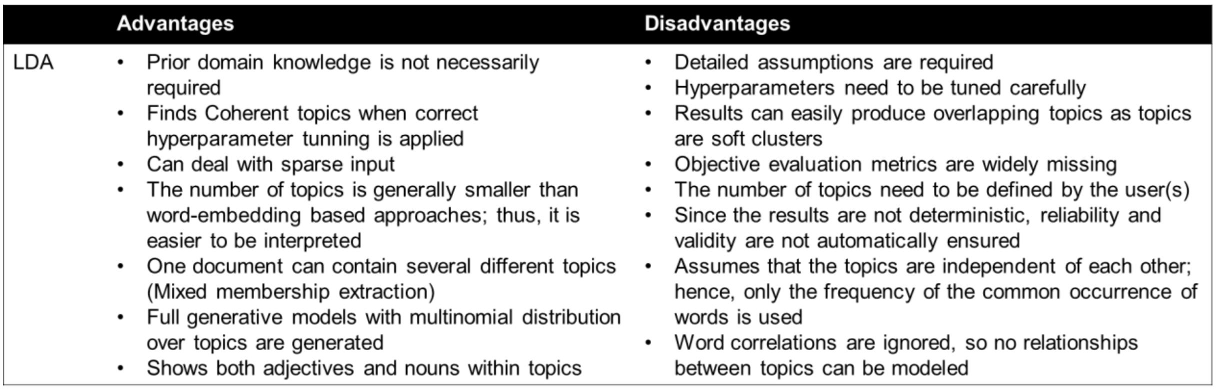
At the time of the study, in 2019, FinBERT had the best performance for financial sentiment analysis, and it has increased the state-of-the art performance in terms of accuracy in the Financial PhraseBank developed by Malo et al. [25]. As expected, FinBERT is being used for predicting stock market (direction and price) through financial news [5] [10]. Although the approach and methodologies are different, in all these papers the results are conclusive, FinBERT increases the accuracy of the predictions.

2.2. Topic Modeling

Blei states that topic modeling algorithms are statistical methods that consider the words of texts to discover the topics are present in them [31]. According to Maier et al. [18] there are many different algorithms that are used for topic modeling, and LDA (Latent Dirichlet Allocation) is a recognized as a widely used model. Referencing Yono *et al.* [16], there are numerous applications of LDA for market prediction. For example, Mahajan et al. [15] proposed a prediction system that forecasts whether the stock market has a positive or negative direction based on identifying events that affect the market using LDA on financial news.

To understand LDA, in the words of Blei *et al.* [32, p. 77] documents are presented as random combination of latent topics, with each topic being described by a distribution of words. LDA has been used in a wide variety of disciplines, in Marketing to analyze social media [20] and in Economics and Finances for visualizing financial stability [33]. LDA has allowed researchers to explore textual analysis in a new approach, being able to analyze large corpus of documents in a systematic way gathering new and useful insights. Although LDA has been considerably used among the topic modeling approaches, it has some limitations. As Grootendorst states [19, p. 1] a limitation of this type of models is that they do not consider semantic relationships among words as they rely on bag-of-words representations.

Egger and Yu did a methodological comparison in topic modeling techniques in which they assess the performance concerning their strengths and weaknesses [20]. In Table 2 can be seen the overview for LDA. In the table it is emphasized as stated By Egger and Yu [20], that LDA requires the hyperparameters to be correctly assigned and selected. This requires researchers to spend time on finding the optimal hyperparameters and this makes it difficult for researchers to get objective results since their models depend on their decision of which hyperparameters fit their needs. For example, researchers must decide the number of topics that they would like to find, which determines the results and insights of the investigation.



	Advantages	Disadvantages
LDA	• Prior domain knowledge is not necessarily required • Finds Coherent topics when correct hyperparameter tunning is applied • Can deal with sparse input • The number of topics is generally smaller than word-embedding based approaches; thus, it is easier to be interpreted • One document can contain several different topics (Mixed membership extraction) • Full generative models with multinomial distribution over topics are generated • Shows both adjectives and nouns within topics	• Detailed assumptions are required • Hyperparameters need to be tuned carefully • Results can easily produce overlapping topics as topics are soft clusters • Objective evaluation metrics are widely missing • The number of topics need to be defined by the user(s) • Since the results are not deterministic, reliability and validity are not automatically ensured • Assumes that the topics are independent of each other; hence, only the frequency of the common occurrence of words is used • Word correlations are ignored, so no relationships between topics can be modeled

Table 2: Comparison of topic models, LDA detail. Adapted from [20]

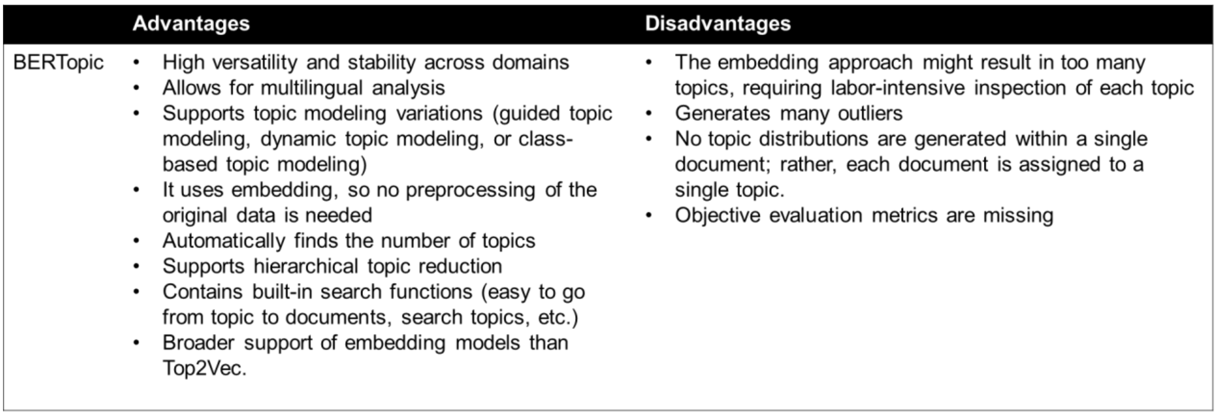
As stated in the previous section on sentiment analysis, BERT has different applications, and topic modeling is one of them. Grootendorst developed BERTopic which is a topic modeling approach with BERT [19].

BERTopic executes its topic modeling approach through the following steps [19]:

- **Document Embeddings:** The documents are converted to embedding representations using BERT (assuming semantically similar documents are associated to the same topic).
- **Dimensionality Reduction**: To optimize the clustering process the dimensionality of the resulting document embeddings is decreased.
- **Clustering:** Clusters are defined through HDBSCAN, which is an extension of DBSCAN that finds clusters of varying densities through a hierarchical clustering algorithm.
- **Topic Representation**: Topic names are associated using a variation of TF-IDF from the clusters of documents.

As stated by Grootendorst [19], BERTopic performs well compared to a variety of benchmarks, solves the problem of disregarded semantic relationships among words, and generates coherent topics.

In Table 3 can be seen the overview of Egger and Yu for BERTopic. In general, compared to LDA, BERTopic considers the context of words, it automatically finds the number of topics and it does not rely on hyper-parameters.



	Advantages	Disadvantages
BERTopic	• High versatility and stability across domains • Allows for multilingual analysis • Supports topic modeling variations (guided topic modeling, dynamic topic modeling, or class-based topic modeling) • It uses embedding, so no preprocessing of the original data is needed • Automatically finds the number of topics • Supports hierarchical topic reduction • Contains built-in search functions (easy to go from topic to documents, search topics, etc.) • Broader support of embedding models than Top2Vec.	• The embedding approach might result in too many topics, requiring labor-intensive inspection of each topic • Generates many outliers • No topic distributions are generated within a single document; rather, each document is assigned to a single topic. • Objective evaluation metrics are missing

Table 3: Comparison of topic models, BERTopic detail. Adapted from [20]

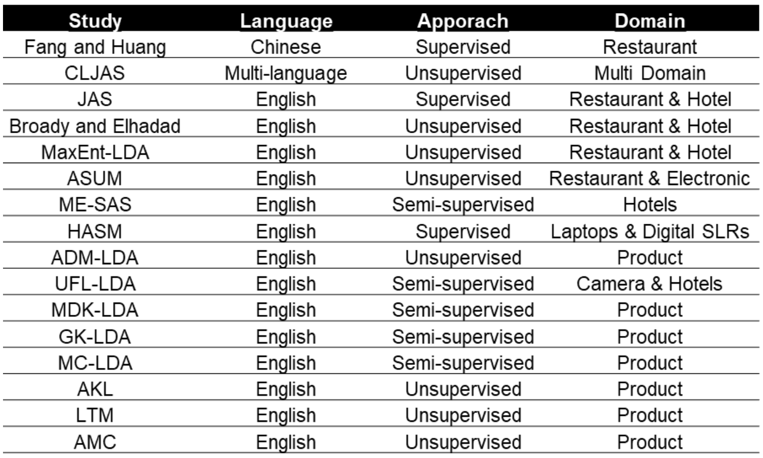
2.3. Integration of Sentiment Analysis and Topic Modeling

Considering the literature review on sentiment analysis and topic modeling is relevant to speak of the combination of these two approaches. Each approach has its own purpose and uses, but there are also applications of these approaches used together that provide a better performance in different uses.

A paper that provides a good example of the combination of these approaches was done by Lin and He [34].The paper introduces an innovative probabilistic modeling framework based on LDA, and it is named joint sentiment/topic model (JST). The approach detects simultaneously the sentiment and topic from a text. The authors analyzed a dataset of movie reviews to determine the sentiment polarity of each review. To understand the impact of considering a JST model, Lin and He [34] provided a good example from other authors, Eguchi and Lavrenko [35]: the word ‘unpredictable’ in ‘unpredictable steering’ found in an automobile review has a different connotation than in the sentence ‘unpredictable plot’ found in a movie review.

For a wide overview of the synergies of the combination of these approaches, Rana *et al.* [36] did a systematic review on sentiment analysis with topic modeling providing comparative evaluations and predominantly on LDA-based techniques. In Table 4 are presented the different studies, languages, approaches, and domains that were compared.

Rana *et al*. focused on online customer reviews of different topics (restaurants, hotels, and others). They provided an example to explain the importance of considering topic modeling with sentiment analysis: In the domain of movies, 'picture' and 'movie' are synonymous terms, however, in the context of photography, they represent different aspects [36]. The authors stated that dictionary-based approaches do not have a good performance in such cases and to avoid this problem, topic modeling is an approach that has been proven to organize cluster of topics of similar terms. With a systematic review and overview of the different approaches, Rana *et al*. [36] demonstrated that topic modeling effectively aids in aspect extraction and categorization.



Study	Language	Apporach	Domain
Fang and Huang	Chinese	Supervised	Restaurant
CLJAS	Multi-language	Unsupervised	Multi Domain
JAS	English	Supervised	Restaurant & Hotel
Broady and Elhadad	English	Unsupervised	Restaurant & Hotel
MaxEnt-LDA	English	Unsupervised	Restaurant & Hotel
ASUM	English	Unsupervised	Restaurant & Electronic
ME-SAS	English	Semi-supervised	Hotels
HASM	English	Supervised	Laptops & Digital SLRs
ADM-LDA	English	Unsupervised	Product
UFL-LDA	English	Semi-supervised	Camera & Hotels
MDK-LDA	English	Semi-supervised	Product
GK-LDA	English	Semi-supervised	Product
MC-LDA	English	Semi-supervised	Product
AKL	English	Unsupervised	Product
LTM	English	Unsupervised	Product
AMC	English	Unsupervised	Product

Table 4: Topic modeling techniques reviewed by Rana et al. Adapted from [36]

In the finance applications of the integration of sentiment and topic modeling there are two studies worth mentioning: [16] and [37].

Yono et al. proposed supervised Joint Sentiment-Topic model (sJST) [16]. Their approach consists of using text data to extract the topic and its sentiment of the market in addition with numeric data. Using the topic, its sentiment and market volatility as inputs, they applied different machine learning models to forecast the price movement of foreign exchange market. In their research they used news as the input for the adapted LDA models and the volatility of USD/JPY as the market volatility, which was used as the supervised signal of the market volatility [16]. In Table 5 results of the different models Yono et al. evaluated are presented.

Model		price only	LDA	sLDA	JST	sJST
SVM	train	52.07%	55.03%	55.01%	55.65%	55.67%
	validation	51.47%	53.64%	53.64%	53.30%	53.73%
	test	51.21%	52.57%	52.20%	52.44%	52.83%
Logistic Regression	train	52.43%	55.61%	55.83%	56.19%	56.30%
	validation	51.62%	53.49%	53.55%	52.72%	53.28%
	test	50.91%	52.19%	51.71%	52.09%	52.49%
MLP	train	51.98%	55.51%	55.70%	56.13%	56.30%
	validation	51.64%	53.18%	53.30%	52.98%	53.25%
	test	50.61%	51.84%	51.54%	51.70%	52.52%
Bi-Directional LSTM	train	52.34%	54.21%	53.95%	54.70%	54.55%
	validation	51.30%	51.92%	51.88%	53.78%	54.59%
	test	50.64%	51.13%	50.64%	52.41%	51.60%
Average	train	52.21%	55.09%	55.12%	55.67%	55.71%
	validation	51.51%	53.06%	53.09%	53.20%	53.21%
	test	**50.84%**	51.93%	51.52%	52.16%	**52.36%**

Table 5: Accuracy results of the evaluated models by Yono et al. Adapted from [16]

According to Yono et al [16], sJST is considered as the integration between supervised LDA (sLDA) and joint sentiment topic model and was evaluated with a news dataset to predict 32 currencies. The results of their investigation show that sJST achieves better results than the other evaluated models [16].

Considering that BERTopic and FinBERT are state-of-the-art techniques in topic modeling and sentiment analysis, respectively, it is reasonable to assume that a method that integrates both models could achieve good results. The combination of these models associated with the financial world is present in the research of Raju *et al.* [37]. Their research evaluates BERTopic using Consumer Financial Protection Bureau (CFPB) data. The results presented by the authors indicate that BERTopic offers more meaningful and diverse topics compared to LDA and LSA. Furthermore, Raju et al. emphasize the flexibility of BERTopic. The authors did not employ FinBERT for sentiment analysis but used it for domain-specific pre-trained embeddings for applying topic modeling, which according to Raju et al. yields better topics [37].

To the moment, there is no known research that consolidates FinBERT and BERTopic for stock price direction prediction, leaving an opportunity for investigation for this thesis to integrate these models for this purpose.

3. Method

Figure 1 gives an overview of the proposed method using BPMN notation. The method consists of integrating two different data sources: a corpus of news publications and a dataset of financial data (daily stock price). The steps of sentiment analysis and topic modeling are performed using financial news and the results are associated with the stock price to the date of the publication of the news. Using this information, prediction models are trained to classify the direction of the stock price for the next day. Once the model is trained, predictions are made, and the model is evaluated with classification metrics. The code and results for this thesis are available in the accompanying GitHub repository[1], where each step of the analysis pipeline is documented and can be replicated.

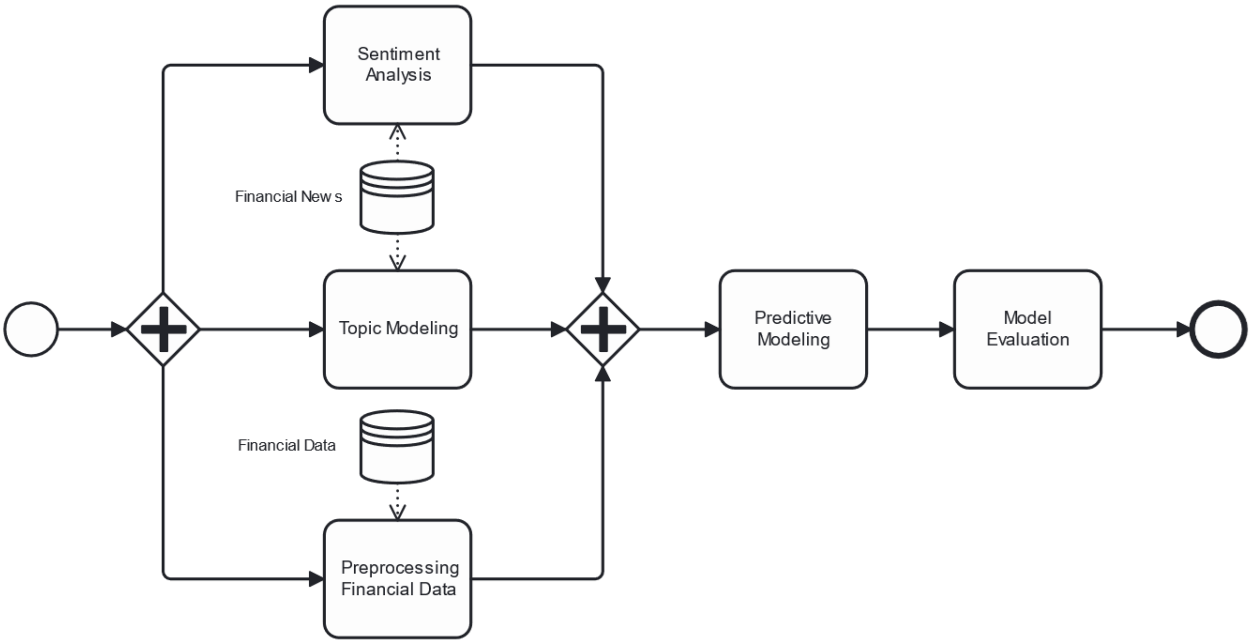


Figure 1: Overview of the proposed method

Topic modeling with sentiment analysis on news provides a novel approach to predict financial markets. Considering the sentiment of news and the topic that it is related to, provides information that can be useful to understand and interpret the financial market. After considering the literature review, the proposition of this thesis is to use the previously discussed FinBERT and BERTopic models (which are state-of-the-art language models based on BERT) on a method that considers both sentiment analysis and topic modeling to predict the direction of the financial markets (focus on the S&P500 index, more explanation on this in Chapter 4).

To provide a benchmark for the method and evaluate the integration of the BERT techniques, the same method is applied with traditional techniques for sentiment analysis and topic modeling in financial applications. LDA is used for topic modeling and a lexicon approach with the most used dictionary for sentiment analysis in financial applications. Further details on these benchmark models are presented in Chapter 5.

[1] https://github.com/miguelazue/SP500-Direction-Prediction-News-Sentiment-Topic

For better understanding of the application of this method, some details are going to be explained in this chapter. It is important to mention that some of the method details are based on the research conducted by Fazlija and Harder in [10].

3.1. Data Sources: Financial News and Financial Data

Predictive modeling consists of using explanatory variables to predict a response variable. The proposed method considers two different data sources. The financial news dataset is used to estimate the explanatory variables and the financial data which is used for estimating the response variable.

The financial news dataset is extracted from a corpus of documents that is further explained in Chapter 4. From this corpus, the text of the articles, the headlines, the publisher, and the date of publication of the news are extracted. Considering that the corpus of documents considers different types of publishers, a selection is done to consider the most relevant publisher types and this selection process is fully detailed in Chapter 4.2. The explanatory variables are then estimated through sentiment analysis and topic modeling of the financial news, the process is further explained in the following sections (3.2 and 3.3).

The financial data consists of a dataset that provides the price and the respective date of the S&P 500 index. Further explanation of this index can be found in Chapter 4.1. This data is used to estimate the response variable which is the direction of movement of the S&P 500 index. The preprocessing of this data is explained in the Section 3.4 of this chapter.

3.2. Sentiment Analysis

Sentiment analysis can be expressed in different formats, one of which involves presenting sentiment results as three numerical values: Positive, Neutral, and Negative scores. Following the approach of Fazlija and Harder [10], the numbers are consolidated in one number referred to as *sentiment score* which is calculated as the difference between the positive score and the negative score of the given text.

$$Sentiment\ Score = Positive\ Score - Negative\ Score$$

- Sentiment scores larger than 0 are considered as a positive score.
- Sentiment scores lower than 0 are considered as a negative score.

The sentiment score is estimated for the financial news calculating the difference between the positive score and the negative score and then the sentiment score is used as explanatory variable for the classification models.

3.3. Topic Modeling

As mentioned in the Chapter 2, Blei states that topic modeling algorithms are statistical methods that consider the words of texts to discover the topics which are present in them [31]. The topic modeling algorithm is used in two steps for the proposed method. First, the topics are extracted from an initial corpus of documents. Second, after the topics are defined, a different corpus of documents is associated to the already defined topics. As the use of each topic modeling algorithm has its particularities, a further explanation of each procedure can be found in the Chapter 5 for the BERT-based model and in Chapter 6 for the baseline model.

3.4. Preprocessing Financial Data

The thesis primarily revolves around a classification task aimed at predicting the direction of movement of the S&P 500 index. This task involves utilizing explanatory variables to classify whether the index's movement is positive or negative, resulting in a binary response variable for the classification. A preprocessing of the financial data is conducted to obtain the response variable as a binary variable.

Following Fazlija and Harder [10], the returns are calculated as the following formula indicates.

$$Rate\ of\ Return = \frac{P_t - P_{t-1}}{P_t}$$

where P_t is the closing price of the stock for the day t and P_{t-1} is the closing price the day t -1.

With the calculated rate of return, the direction is calculated with the following cases.

$$Direction \begin{cases} 1: & if\ rate\ of\ return \geq 0 \\ 0: & otherwise \end{cases}$$

The direction of movement of the S&P 500 index is then estimated using this procedure and used as the response variable for classification models.

3.5. Predictive Modeling

As mentioned before, predictive modeling consists of using explanatory variables to predict a response variable. In the proposed method, sentiment analysis and topic modeling are used together for calculating the explanatory variables. The explanatory variables are estimated as the average sentiment scores of the day for each topic. In essence for each day, there are 30 explanatory variables (the selection of 30 topics is explained in Chapter 5 and Chapter 6), which are the average sentiment score of the day for each topic. The response variable of the proposed method is the direction of movement of the S&P 500 index of the day that follows the explanatory variables. As a result, the combination of sentiment analysis and topic modeling of the financial news are used to predict the direction of movement of the S&P500 index for the next day.

The thesis employs six different classification model types, which are logistic regression, decision tree, random forest classifier, SVC (Support Vector Machine Classifier), gradient boosted trees for classification (GBC), and naïve Bayes. A brief explanation of each classification model is presented.

Logistic Regression. Following the book definition of Gerón [38], the most frequent use of logistic regression lies in estimating the probability of an instance being classified into a specific category.

Logistic regression is a binary classifier: If the obtained probability is above 50%, an instance is labeled as belonging to the positive category, frequently denoted as “1”. In turn, the negative category is often labeled as “0”. For an easier understanding, logistic regression estimates the coefficients (weight of the input features and a bias term) in a similar way to linear regression but estimates the result through the logistic expression (sigmoid function) for obtaining a number that spans from 0% to 100% [38].

Decision Tree. Following the book definition of Müller and Guido [39], a decision tree aims to get to the true answer by employing a sequence of if/else questions. The if/else questions are called tests in the machine learning domain. The decision tree construction process involves scanning through all potential tests to identify the most informative one regarding the target variable. For classification, the data is partitioned with each test and the process continues until each leaf (partition) in the tree exclusively contains data with a single class. When the leaf contains data points sharing the same classification, the node is termed as pure.

Following the scikit-learn documentation [40], decision tree learning has two important parameters. The first parameter is criterion, which is the selected function to measure the quality of a split (Entropy or Gini). The second parameter is maximum depth of the tree which is the maximum number of consecutive tests.

Random Forest Classifier. Following the definition of Breiman [41], Random Forest combines multiple decision trees by counting their predictions. Each tree in a Random Forest is built independently and has equal weight in making predictions. Following the scikit-learn documentation [40], the learning of a random forest classifier has the same important parameters as decision tree learning but additionally has as parameter the number of estimators, which is the number of trees that are in the forest.

Support Vector Machine Classifier. According to the book of Norris [42], the objective of SVM is to identify a boundary for segregating data into different groups. This boundary is referred to as the separating hyperplane. The ideal hyperplane is the one that maximizes the margins between support vectors. The support vectors are constructed using the data points for each classification that are closest to the separating hyperplane and the distance between these support vectors is maximized. Following the book definition of Gerón [38], there are two types of margin classifications: hard margin classification and soft margin classification. Hard margin classification consists of imposing that all the data points be correctly classified. This type of classification is sensitive to outliers and is only possible if the data is separable. In contrast, the soft margin classification addresses these by enabling margin violations. The aim of soft margin classification is to find a good balance between minimizing margin violations while maximizing the margin width. In general, margin violations are bad, however models with margin violations tend to generalize better (this is adjusted with the C hyperparameter).

As explained by Norris [42],when the data points are not linearly separable in a dimensional space, there is an alternative known as the kernel trick. The data is transformed into a higher dimensional space using a non-linear kernel function and then non-linear hyperplanes can be constructed to separate the transformed data points.

Following the documentation of scikit-learn [40], SVC have two important parameters. The C parameter and the Kernel. The options in scikit-learn for the kernel function are linear, poly, radial basis function, sigmoid or it can be precomputed.

Gradient Boosted Trees for Classification. According to Friedman [43] Gradient Boosting Trees combines multiple decision trees sequentially with a small learning rate (the trees are not independent from each other, and they have weights). This means that simple trees (trees with low depth, referred as shallow trees) are combined sequentially, each tree predicts only a portion of the data, and the iterative addition of successive trees enhances overall performance. An important characteristic is that trees added earlier are more important than trees added later [44].

Based on XGBoost [44] documentation, the “eta” parameter for the Gradient Boosted Trees is the learning rate applied during updates which is used to prevent overfitting.

Naïve Bayes Classifier. Following the conceptualization in the book of Müller and S. Guido [39], there are different kinds of naïve Bayes (NB) classifiers but the relevant for this case is the Gaussian NB. To make a prediction, the best matching class is predicted by comparing the statistics for each of the classes to a data point. In the case of Gaussian NB, it considers for each class the average value and the standard deviation of each feature. According to the authors, the NB models are efficient because they adjust the classifications by gathering per-class statistics from each feature and analyzing each feature independently. Based on scikit-learn [40] the classifier depends on the Var Smoothing parameter. The Var Smoothing is added for calculation stability as features with low variance could cause estimation of extreme values. This is done by adding a portion of the feature with the largest variance to all features’ variances.

Python libraries

Table 6 provides the libraries used in python for each classifier.

Classification Model	Library
Logistic Regression	*statsmodels.api* [45]
Decision Tree	*sklearn.tree* [40]
Random Forest Classifier	*sklearn.ensemble* [40]
Naïve Bayes	*sklearn.naive_bayes* [40]
Support Vector Machine Classifier (SVC)	*sklearn* [40]
Gradient Boosting Classifier (GBC)	*xgboost* [44]

Table 6: Python libraries used for the classification models

Each prediction model is used with the best performing parameter values obtained through a grid search procedure which is explained in Chapter 7.1.

3.6. Model Evaluation

According to Fahim [46], for evaluating diverse binary classification systems and models, accuracy has become a “gold standard”, but it is not a reliable measure to evaluate the performance of a classifier on unbalanced data. Accuracy has also been used as the metric for measuring the success of classification in different stock price direction prediction models with sentiment analysis as in [6], [5], [16]. Accuracy is calculated as the proportion of correctly identified elements over the total number of elements [38].

As outlined in the book of Muller and Guido [39], accuracy is not an adequate measure for evaluating the performance of imbalanced datasets. For example, a relatively high accuracy can be achieved by simply assigning all predictions to the category with the largest frequency. According to Norris’ book [42], this kind of classifier that assigns all predictions to the class with the largest frequency is known as the *zero-rule* algorithm and serves as a baseline to compare the performance of classification models. To illustrate, consider a scenario with 100 patients where 5 patients have cancer and the remaining 95 are cancer-free. A classifier that follows the *zero-rule* algorithm will predict that all patients are cancer-free. This classifier that follows the *zero-rule* algorithm will achieve a 95% accuracy, but the performance of the classifier is not good as it does not detect any cancer patient. This means that accuracy has limitations to differentiate a relatively good model from a simple *zero-rule* classifier. Considering the limitations associated with accuracy, Fahim [46] asserts that accuracy is not a reliable measure to evaluate the performance of a classifier on unbalanced data and states that this problem is known as the *accuracy paradox*.

Considering the *accuracy paradox*, different performance metrics are going to be considered together for evaluating the performance of the models such as accuracy, sensitivity, specificity, precision, and f1-score. These metrics depend on the classification of the predictions as follow:

- True Positives (TP): Number of correctly classified predictions as positive
- True Negative (TN): Number of correctly classified predictions as negative
- False Positives (FP): Number of incorrectly classified predictions as positive.
- False Negative (FN): Number of incorrectly classified predictions as negative.

Following the classification of the predictions, the performance evaluation metrics are explained in Table 7, as described in the book of Gerón [38]:

Metric	Formula	Interpretation
Accuracy	$\frac{TN + TP}{TN + TP + FN + FP}$	proportion of the correctly predicted classifications.
Sensitivity (Recall)	$\frac{TP}{TP + FN}$	proportion of correctly classified positive values from the total of actual positive values
Specificity	$\frac{TN}{TN + FP}$	proportion of correctly classified negative values from the total of actual negative values
Precision	$\frac{TP}{TP + FP}$	proportion of correctly classified positive values from all the predictions classified as positive
F1	$\frac{2}{\frac{1}{Precision} + \frac{1}{Sensitivity}}$	Harmonic mean of precision and recall

Table 7: Metrics used for the performance evaluation [38]

Data Variants Evaluation

To obtain valuable insights, variants of the data are evaluated and compared:

- Topic Model Training Data Size: Comparison between 250K articles and a subset of 20K articles. Comparing the totality of the articles and a sample provides insights to understand the importance of the training size of the topic model.
- Classification Test Period: Testing periods of 1 year, 2 years, and 3 years after the topic model's training (e.g., using topics discovered in 2016 to classify articles in 2019). This comparison allows to understand the impact of maintaining an updated topic model)
- Article text: Contrasting body text with headlines allows us to ascertain whether headlines alone provide sufficient information for classification or if better results are achieved using the entire article.

4. Data Sources

This chapter gives an overview of the data used in the thesis, the selection and reasoning behind it, the processing of the data, and its assembly and integration into the proposed approach. The first section explains the financial data, the S&P500 Index, data processing, and the adaptation to the method. The second section provides an overview of the sources of the news dataset, its processing, selection, and adaptation to the method.

4.1. Financial Data: Selection and Preprocessing

In this section it is important to first understand the concept of stock index and what the S&P500 means. A stock index follows the value of a hypothetical portfolio of stocks, and the Standard & Poor's 500 (S&P 500) Index is established on a portfolio of 500 stocks of different industries (financial, industrials, utilities, and transportation companies) [47]. The stocks included in the index consider about 80 percent of the total U.S. capital market and the S&P 500 is widely acknowledged as one of the most influential indicators, both reflecting and shaping movements within the U.S. stock market [48]. This index is established so that weights are proportional to market capitalization (number of shares × stock price) therefore the underlying portfolio automatically adjusts the price considering stock dividends, new equity issues, and stock splits [47]. Since the S&P 500 considers stocks from many different industries & sectors, and the objective of the thesis is to evaluate the effect of different topics on news sentiment analysis, the S&P500 index is an adequate choice for evaluating the performance of the method and the objectives of the thesis.

The stocks can have different prices throughout the day, but there is only one at the end of the trading day which is represented by the closing price. According to the Corporate Finance Institute [49] the original closing price may not offer the most accurate reflection of the stock's value. Therefore, an adjusted closing price is calculated, incorporating necessary adjustments such as those for stock splits and dividend payouts.

The US Stock Market does not operate on weekends or holidays, so the information on the stock prices is only available for around 252 days per year, which in financial terms are called “Trading Days”. Gupta and Chen [6] backfilled weekend and holiday prices in their stock price prediction research. For Saturdays the authors calculated the average between the closing price of the previous Friday and the next Monday. For Sundays they estimate the average using the already calculated price for Saturday and the value of the next Monday. Although this method is practical and useful for their study case, for the method in this thesis this is not appropriate since the idea is to see the direct effect of the news on the price. Considering this, only the trading days price is used and there is no backfilled estimation of prices.

For obtaining the closing stock prices for the S&P500 the “yahoofinancials” module was used through Python. It is a financial data module used for pulling data from Yahoo Finance developed by Connor Sanders [50]. In the Figure 2 can be visualized the price movement of the index from 2016 to the start of 2023.

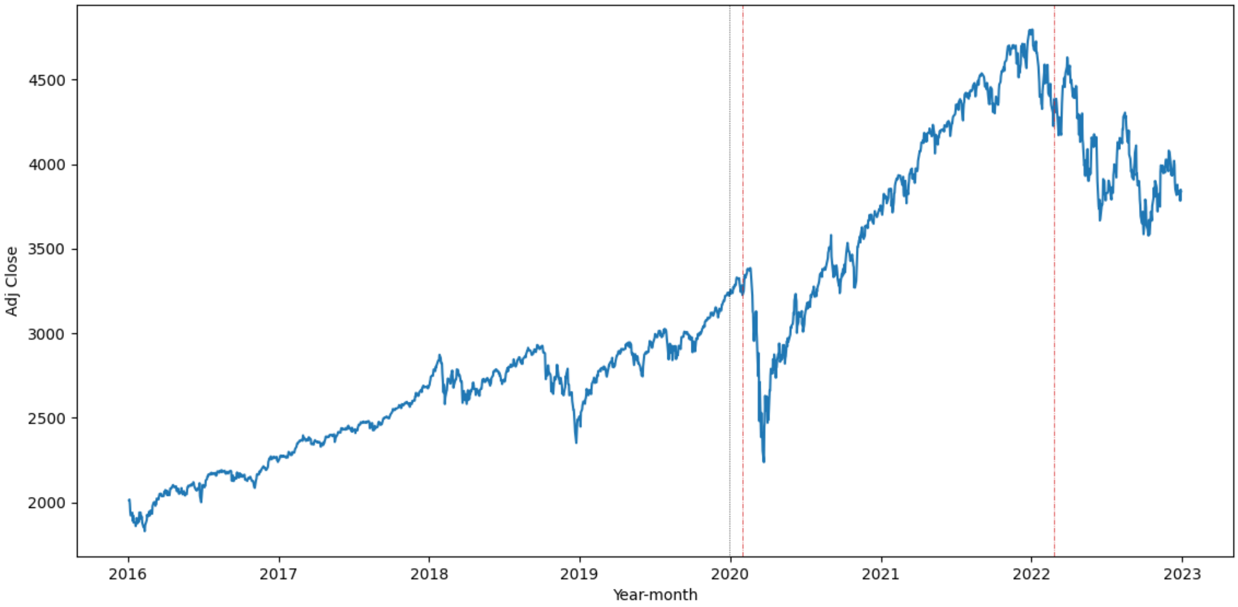


Figure 2: S&P500 adj close price from 2016 to 2023

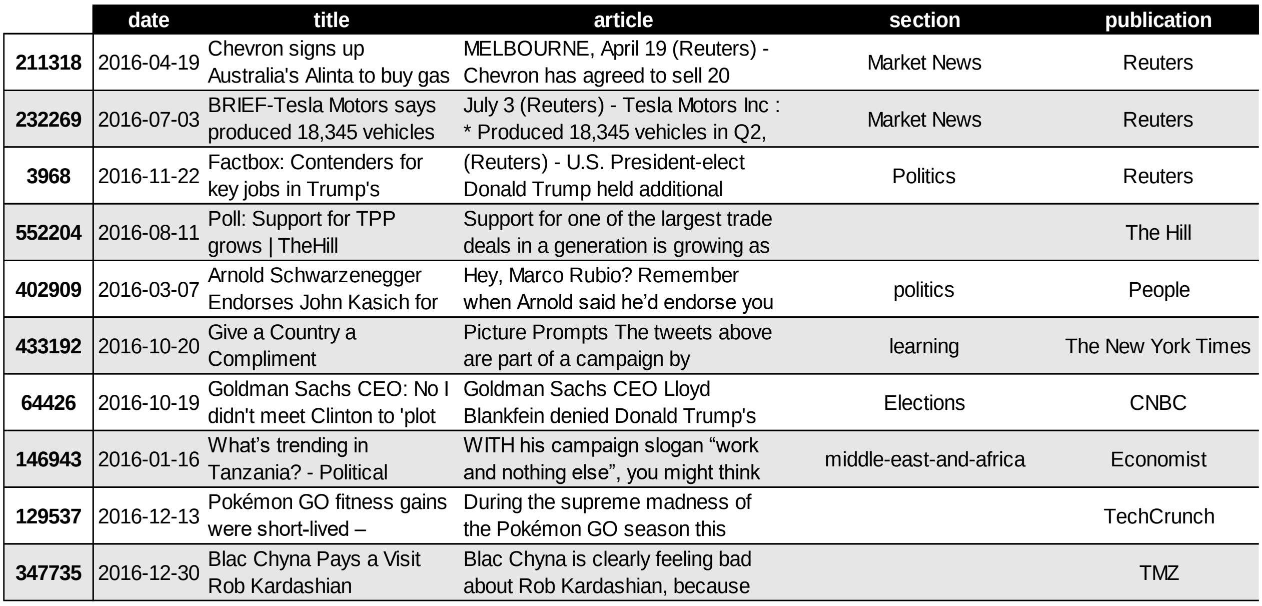
The dates 30th of January 2020, when the WHO declared the outbreak of COVID-19, and the 24th of February 2023, when Russia invaded Ukraine, are highlighted in red on the graph depicting the movement of the S&P500 index following these events. These occurrences significantly impacted the S&P500 index, as evidenced by Scott et al.'s analysis of the unprecedented stock market reaction to COVID [51] and Izzeldin et al. examination of the impact of the Russian-Ukrainian war on global financial markets [52]. Given the magnitude of these events and their profound effects on both the global economy and news coverage, it is reasonable to assume that no other major events during this period have had a comparable impact. Consequently, it is conceivable that the correlation between the S&P500 index and the sentiment of topics unrelated to the COVID pandemic and the Russian invasion of Ukraine may be overshadowed.

The thesis focuses on the period spanning from 2016 to 2019 for two primary reasons. Firstly, the selected news data source covers the timeframe from 2016 to April 2020, offering unparalleled depth and breadth crucial for the thesis. Secondly, events such as the COVID-19 pandemic and the Russian invasion of Ukraine transpired after 2019. Analyzing periods beyond 2019 would pose challenges due to the overwhelming influence of these significant events on the global economy and the S&P 500 Index price, potentially overshadowing the effects of other topics under investigation.

4.2. Financial News: Selection

The dataset used for the news articles is the one created by Andrew Thompson called "All the news 2.0" [53]. This dataset has also been used by other authors on sentiment analysis research on stock price prediction such as Bakker in its research [9].

The dataset contains more than 2.5 million news articles from 27 publishers, spanning from January 2016 to April 2020. A sample view of the data is presented in Table 8.



	date	title	article	section	publication
211318	2016-04-19	Chevron signs up Australia's Alinta to buy gas	MELBOURNE, April 19 (Reuters) - Chevron has agreed to sell 20	Market News	Reuters
232269	2016-07-03	BRIEF-Tesla Motors says produced 18,345 vehicles	July 3 (Reuters) - Tesla Motors Inc : * Produced 18,345 vehicles in Q2,	Market News	Reuters
3968	2016-11-22	Factbox: Contenders for key jobs in Trump's	(Reuters) - U.S. President-elect Donald Trump held additional	Politics	Reuters
552204	2016-08-11	Poll: Support for TPP grows | TheHill	Support for one of the largest trade deals in a generation is growing as		The Hill
402909	2016-03-07	Arnold Schwarzenegger Endorses John Kasich for	Hey, Marco Rubio? Remember when Arnold said he'd endorse you	politics	People
433192	2016-10-20	Give a Country a Compliment	Picture Prompts The tweets above are part of a campaign by	learning	The New York Times
64426	2016-10-19	Goldman Sachs CEO: No I didn't meet Clinton to 'plot	Goldman Sachs CEO Lloyd Blankfein denied Donald Trump's	Elections	CNBC
146943	2016-01-16	What's trending in Tanzania? - Political	WITH his campaign slogan "work and nothing else", you might think	middle-east-and-africa	Economist
129537	2016-12-13	Pokémon GO fitness gains were short-lived –	During the supreme madness of the Pokémon GO season this		TechCrunch
347735	2016-12-30	Blac Chyna Pays a Visit Rob Kardashian	Blac Chyna is clearly feeling bad about Rob Kardashian, because		TMZ

Table 8: Sample view of the dataset "All the News 2.0" [54]

The dataset includes the date of publication, the title of the article, the full article, the section, and the publisher (Throughout this text, the term "headlines" refers to the titles of the articles, and "Bodytext" refers to the full written content of each article). The dataset considers a wide set of different publishers, from economic-focused publishers like the Economist to entertainment-focused like TMZ. Although the variable "section" includes relevant information, it is not standardized between the different publishers (there are 7,509 different sections) and in many of the news this is an empty value (900K articles do not have a section associated).

Initially, the primary focus of this thesis was to assess the impact of financial news on stock price direction. However, recognizing that various types of news, such as Economic, Entertainment, Political, and others, could also influence the stock market, the initial scope was expanded. Consequently, a correlation analysis was conducted to examine different types of publishers and measure the correlation between the sentiment scores of various articles and the S&P500 index.

The dataset lacks a column specifying the type of publishers. To address this gap, a classification was created based on the typical content of the articles and the information provided in the "About Us" section of their respective websites.

The classification consists of 4 different categories:

- EBF (Economics, Business and Financials)
- GEN (General and Breaking News)
- ENT (Entertainment)
- POL (Politics)

The classification between the different publishers is as follows:

- Economics, Business and Financials.
  - Business Insider
  - CNBC
  - Economist
  - Reuters
  - TechCrunch

- General and Breaking News
  - Axios
  - CNN
  - Fox News
  - The New York Times
  - Vice News
  - Vox
  - Washington Post

- Entertainment
  - Buzzfeed News
  - Gizmodo
  - Hyperallergic
  - Mashable
  - People
  - Refinery 29
  - TMZ
  - The Verge
  - Vice
  - Wired

- Politics
  - New Republic
  - New Yorker
  - Politico
  - The Hill

In the Figure 3 the distribution in the dataset of articles by publisher and category can be visualized.

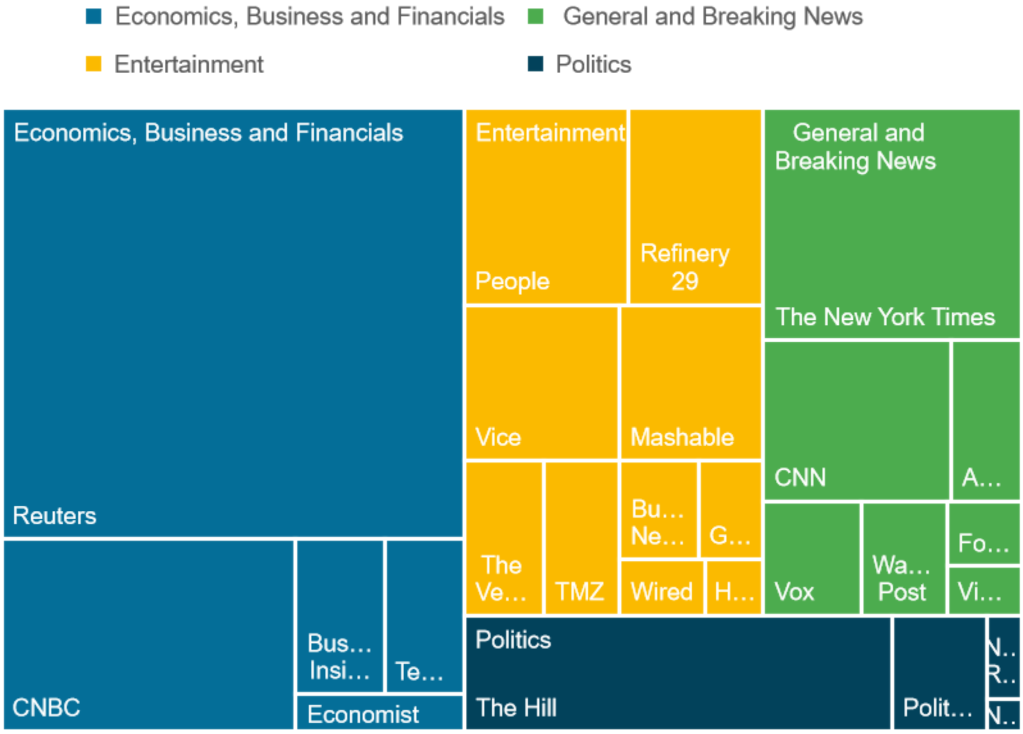


Figure 3: Publisher categorization and distribution of the articles.

An analysis was done considering the correlation between the sentiment analysis by publisher with S&P500 returns. The purpose of this analysis is to select and consider the most relevant publisher categories and therefore obtain a better performance in the prediction models of the thesis. In Figure 4 is presented a diagram that explains the correlation analysis done to select the publisher categories that are most correlated with the S&P500 Index.

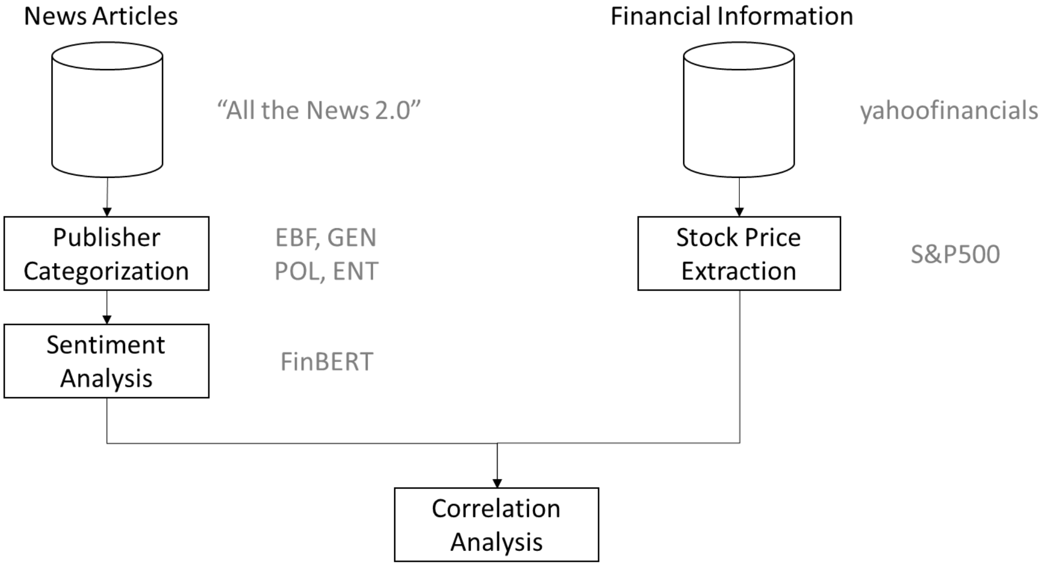


Figure 4: Diagram of publisher categories correlation analysis with the S&P500 index price.

The sentiment score was calculated for all the news using the articles headlines with FinBERT. Using the Hugging Face transformers library in python, the pretrained FinBERT model was downloaded from ProsusAI [54] and used to get the sentiment analysis scores of the article's headlines. The sentiment score was averaged and grouped for every trading day, so there is an average sentiment score for each type of publisher for every trading day. This information was then integrated with the stock price of the S&P500.

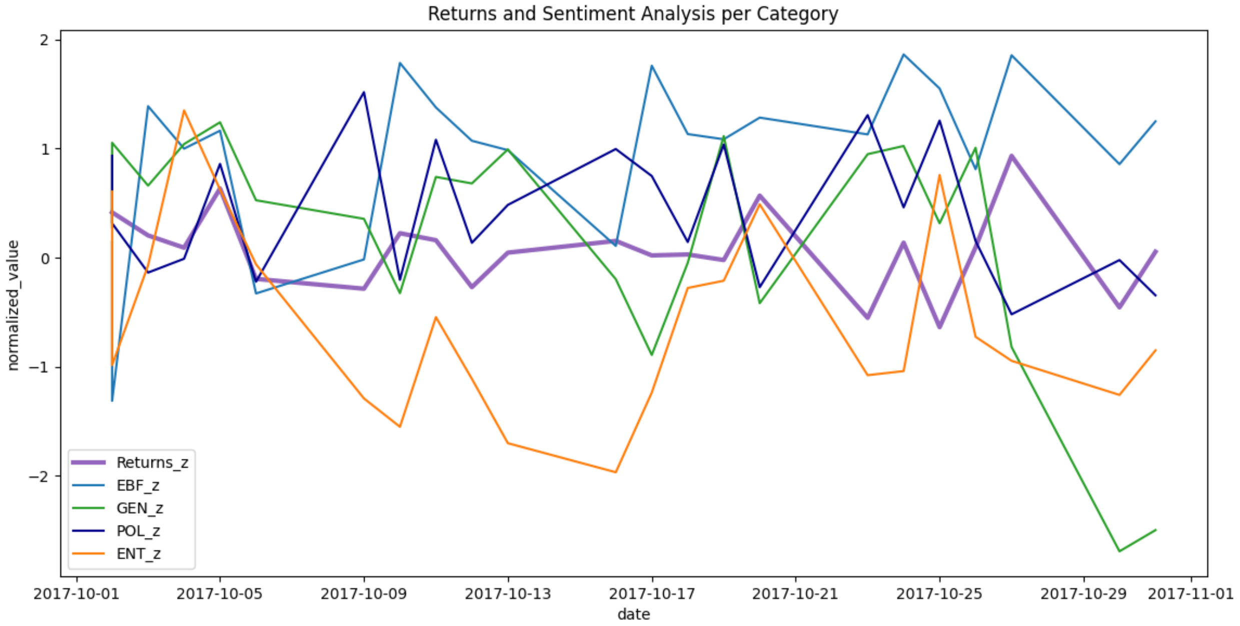


Figure 5: Correlation Analysis of publisher categories, example for October 2017

The Figure 5 visualizes the movement of the average sentiment analysis score for each publisher type with the normalized returns of the S&P500 index for a sample month. Although the graph and the normalization of the values help to visualize the movement of the S&P500 index returns and the sentiment score of the different publisher types it is still difficult to conclude any pattern or correlation.

Two correlation analyses were done: The first considering the effect of published articles sentiment score with the stock price of the same day, the second correlation analysis considering the movement of the price with the lagged sentiment score (the sentiment score of the previous day). To consider the effect of articles published on weekends and holidays, their sentiment score was associated to the previous trading day.

As an example, to understand the lag concept, in the Figure 6 for the month of January 2016 can be seen the value of the normalized returns of the S&P500, average score of EBF news, and the lagged average score of EBF news. In the visualization the light blue line is the exact same blue line, but it is moved to the right with minor differences caused by the jump from Monday to Friday (not considering the weekend days).

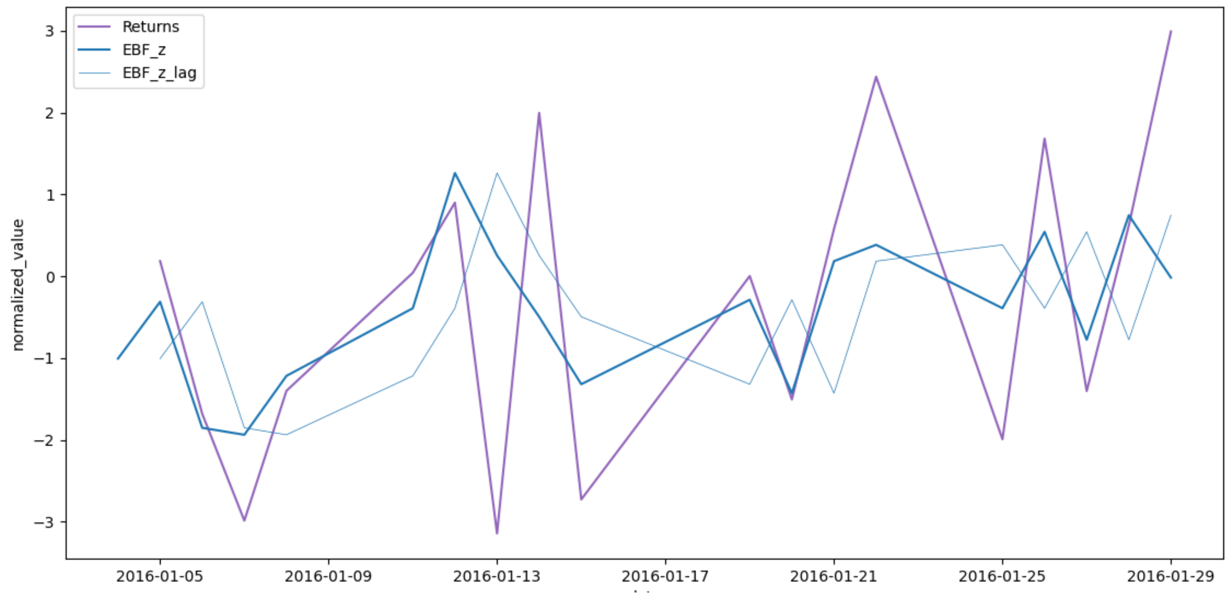


Figure 6: Correlation Analysis with lag variable of EBF publishers for January 2016.

The correlation analyses were done considering the time frame spanning from January 2016 to December 2019 with aggregated variables to the daily level. The variables considered were the price rate of returns of the S&P 500 Index, the adjusted close of the S&P500 index, time variables (such as year, month and a time trend variable starting as 1 in the first trading day of January 2016), and the average sentiment score of the day for each publisher category. The results are presented in Figure 7 which represents the correlation matrix of the analysis.

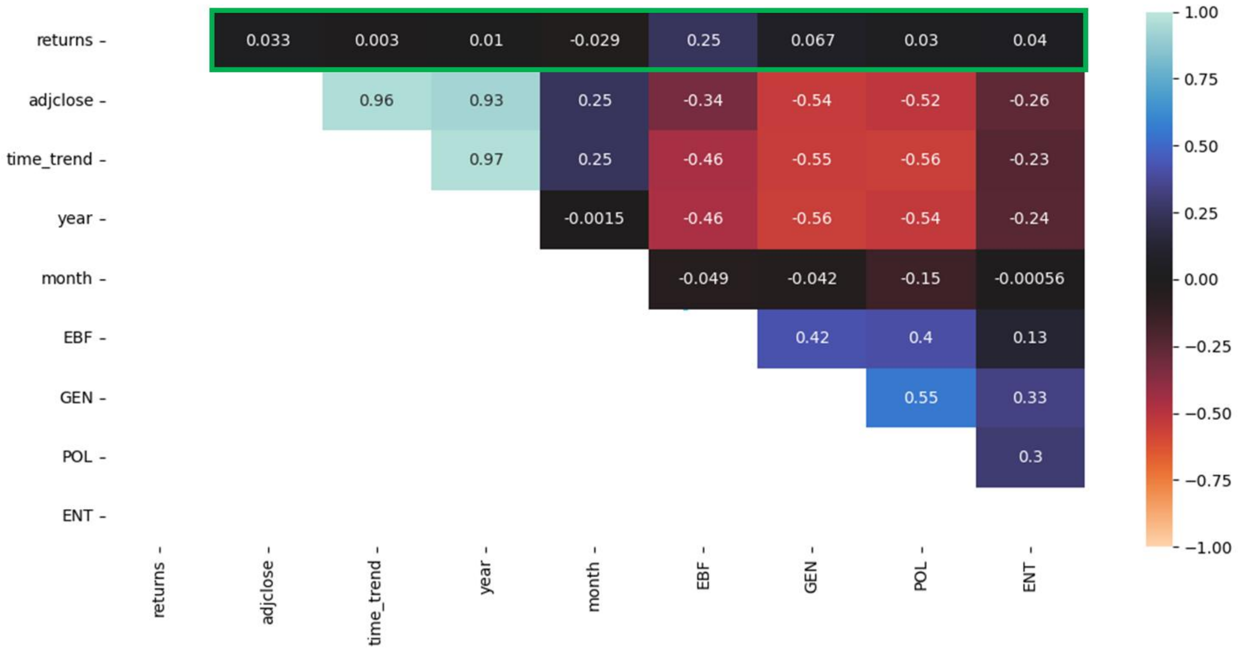


Figure 7: Correlation analysis results

For the returns (results highlighted in the green rectangle) the only correlated variable is the EBF sentiment score, it is not a high correlation, but it is the only one above 0.2 threshold. It is a positive correlation, meaning that the direction of the returns is related with the polarity of the sentiment score. In the Figure 7 some blocks of correlation can be visualized. The light-blue block in the upper left shows that the time variables are highly positively correlated with the adjusted close, which is the actual price of the S&P500, meaning that the price is highly correlated on time trend (*time_trend*) and year. This is expected as it is known that investing in the S&P500 in the long term provides profit. The blue block in the lower right corner indicates that the sentiment scores between the different publisher categories are correlated at around 0.4, except for the Entertainment category which is not as correlated with the other ones. The red block shows an interesting insight, the sentiment scores of the publisher categories are negatively correlated with time variables, indicating that from 2016 to 2019 the sentiment score on average was decreasing. The cause of this is not known, but it is interesting to highlight.

For the same period, a correlation analysis was done considering the lag variables of the sentiment scores, the price of the S&P500 (adjclose) and the returns of the S&P500. In Figure 8 the results can be visualized.

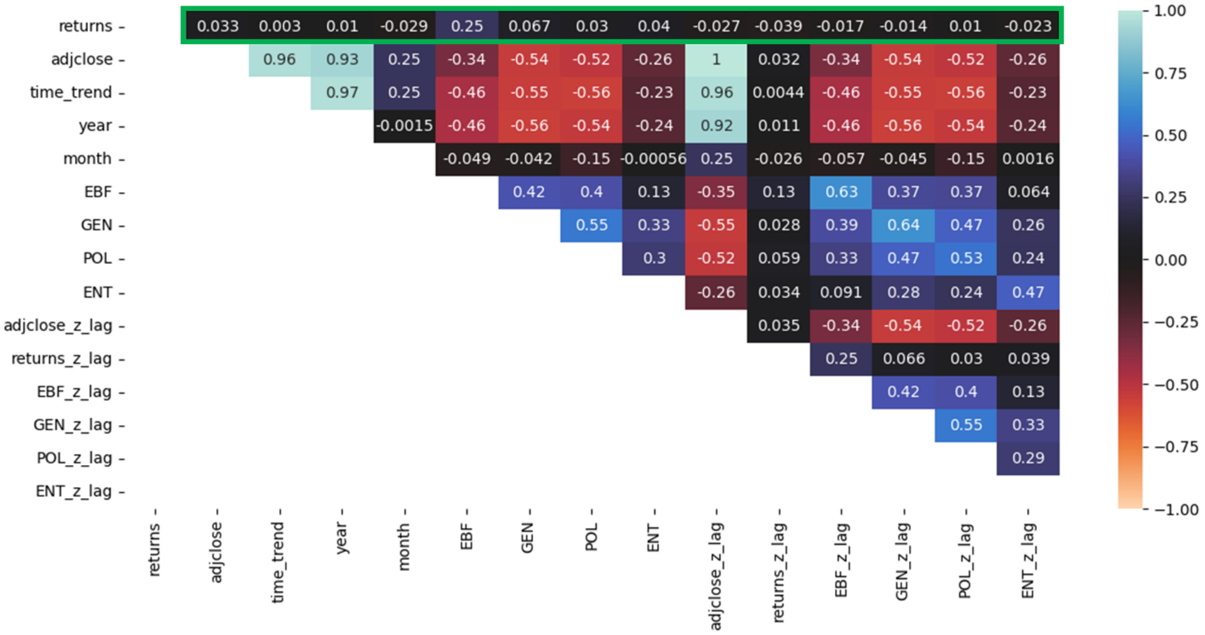
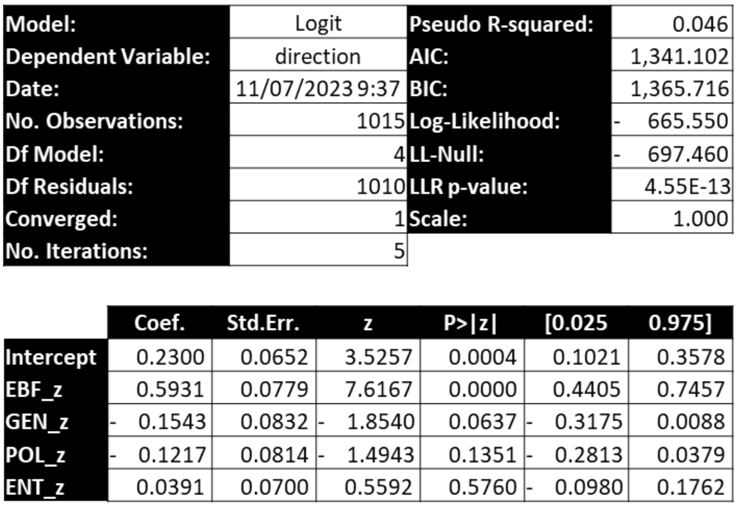


Figure 8: Correlation Analysis results Including Lag Variables

In the Figure 8, the correlation matrix shows that the returns (green rectangle) are not really correlated with any lagged sentiment score, which is not encouraging for the application of the method but considering that the topic modeling is still missing to be integrated, there is still confidence that some promising results are going to be achieved. The additional blocks of correlation with the lag variables reinforced the previously mentioned relationships between variables

Other correlation analyses were done with logistic regressions. These were done using the “statsmodels” package in python. The following are the results of one of the logistic regressions evaluated. The dependent variable is the direction of the price (1 price increased during the day, 0 otherwise). The independent variables are the daily average sentiment score of each publisher category. The results from the regression are presented in the Figure 9.



Model:	Logit	Pseudo R-squared:	0.046
Dependent Variable:	direction	AIC:	1,341.102
Date:	11/07/2023 9:37	BIC:	1,365.716
No. Observations:	1015	Log-Likelihood:	- 665.550
Df Model:	4	LL-Null:	- 697.460
Df Residuals:	1010	LLR p-value:	4.55E-13
Converged:	1	Scale:	1.000
No. Iterations:	5		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.2300	0.0652	3.5257	0.0004	0.1021	0.3578
EBF_z	0.5931	0.0779	7.6167	0.0000	0.4405	0.7457
GEN_z	- 0.1543	0.0832	- 1.8540	0.0637	- 0.3175	0.0088
POL_z	- 0.1217	0.0814	- 1.4943	0.1351	- 0.2813	0.0379
ENT_z	0.0391	0.0700	0.5592	0.5760	- 0.0980	0.1762

Figure 9: Logistic regression, stock direction and sentiment score by publisher category

The only statistically significant variable for the 5% significance level is the sentiment score of EBF news. It is a positive coefficient, meaning that a positive sentiment score in EBF articles is correlated with a positive direction in the S&P500 index price. The GEN news is statistically significant to the 10% level with a negative coefficient. “Bad News”, negative sentiment score is correlated with a positive direction in the S&P500 index price.

After these different correlation analyses some insights have been established:

- EBF articles sentiment score is positive correlated with the direction and price rate of return of the S&P500.
- GEN, POL, and ENT articles sentiment score are not really correlated with the direction or the price rate of return.
- Adjclose (Price) is highly correlated with time while the rate of return (direction) is not correlated with time.

Considering these insights, only EBF articles are going to be considered in the next steps of this research.

The pre-processing of the text data depends on the type of the implemented LM. For example, for using a lexicon-based approach for sentiment analysis, it is necessary to list the words in a text and count the words that correspond to the lists. On the other hand, with a BERT approach, the words are not separated, as BERT uses the context of the words to calculate sentiment. The pre-processing is further explained in each respective implementation.

5. BERT-Based Implementation of Sentiment Analysis and Topic Modeling

The different components of the method have already been explained separately (Data sources, Sentiment Analysis, Topic Modeling, Predictive Modeling and Model Evaluation). Following the diagram in Figure 10, this chapter explains the integration of the different components and the use of the proposed method to accomplish the objectives of this thesis.

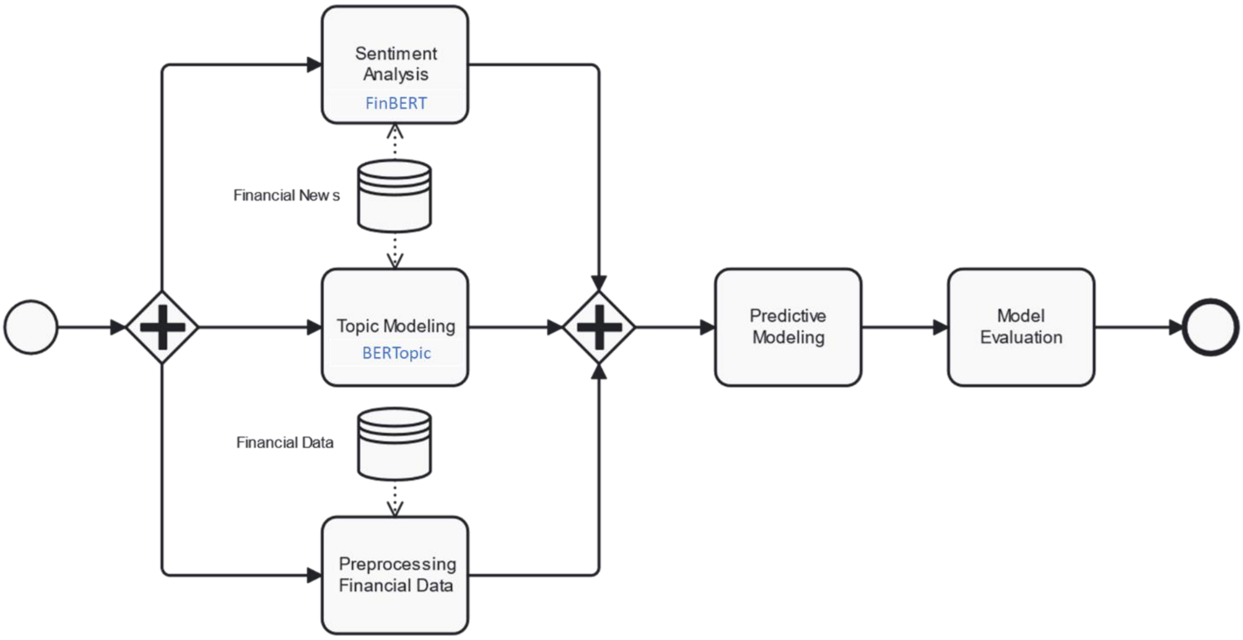


Figure 10: Diagram of the method and the selected components.

As explained in Chapter 2 FinBERT and BERTopic are state-of-the-art BERT models of sentiment analysis and topic modeling respectively. One of the objectives of this thesis is to integrate both models to consider a deeper semantic meaning of the news by using sentiment analysis together with topic modeling. The integration of the two models is as follows.

The BERTopic model is trained using the 2016 EBF news to discover common topics/themes among the articles (around 3,000 topics were discovered). A sample BERTopic topic model, considering a sample of 5K articles (2% of the 2016 EBF articles) was created to present the diagram shown in Figure 11 that allows to understand the concept of topics/themes.

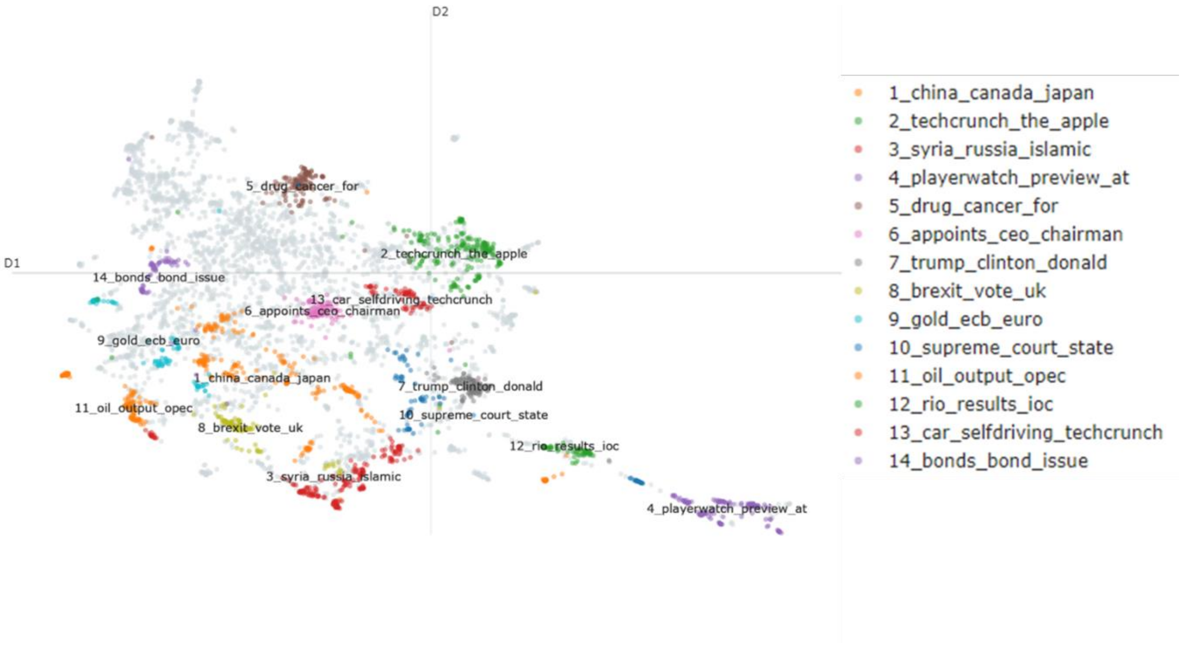


Figure 11: Topic model example using a 2016 sample of 5K news.

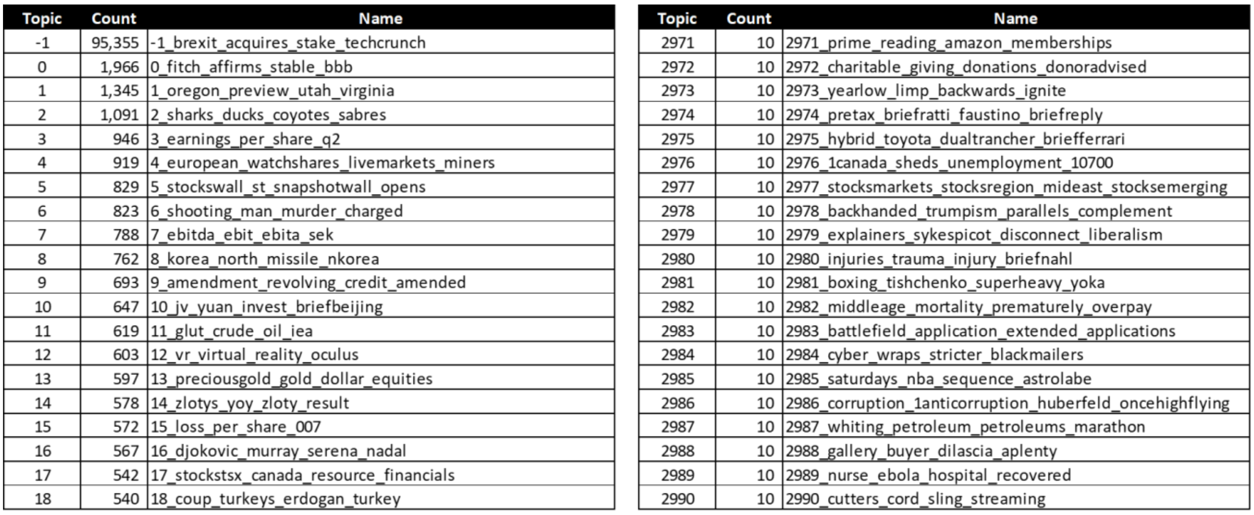
To understand Figure 11 is important to first understand some concepts such as vector semantics. Vector semantics consists of representing a word as a point in a semantic space which is derived from the distributions of its neighboring words [21]. In other terms, words that have similar meanings tend to be found in similar contexts. The idea of the link between similarity in what they mean and similarity in how words are distributed is known as the distributional hypothesis [55]. Nowadays vectors representing words are called embeddings.

In the case of the example of Figure 11, each dot represents a document and with the transformers embedding they are in a dimensional space (In the diagram the documents are in a 2-dimension space for visualization purposes). HDBSCAN algorithm is used for clustering the different documents and "discover" the topics/themes between the different documents. For giving each topic a name, a TF-IDF algorithm is used. Referencing the book of Jurafsky and Martin [21] to explain TF-IDF, lets explain each term separately. TF refers to term frequency, indicating how often a word *t* occurs in a document *d*. DF means document frequency, denoting the total count of documents in which the word *t* appears. IDF means the inverse document frequency, defined by the fraction N/df_t, where *N* is the total number of documents and, df_t signifies the number of documents containing the term t [21].

The TF-IDF algorithm is used to find words that repeat often in a document but not in others, finding in this way characteristic words for documents. For the application of the example with BERTopic, TF-IDF is used to find the most representative words and use them as the name of each topic.

In the example of Figure 11, Topic 9 refers to documents on Gold, The European Central Bank (ECB) and the Euro, Topic 11 refers to Oil production outputs and OPEC meanwhile Topic 12 refers to Rio Olympics results and the International Olympic Committee. The visualization shows that Topic 9 and Topic 11 documents are closer between each other compared to the distance with the documents associated with Topic 12 which is sports related. For more detailed explanation of the process of BERTopic a good explanation is given in reference [19].

Considering that the BERTopic Model trained with all the EBF 2016 news “discovered” around 3,000 topics and that probably not all of them are relevant, a selection process is done to keep the prediction models efficient and keep the most relevant topics. For example, it is reasonable to assume that Rio Olympics news do not have the same impact on the S&P500 Index Price as the ones related to OPEC/Oil or ECB/Euro. In the Figure 12 can be seen the distribution of the 2016 EBF articles between the topics (Only the top 20 and last 20 topics are shown, sorted by the number of documents assign to the topic).



Topic	Count	Name
-1	95,355	-1_brexit_acquires_stake_techcrunch
0	1,966	0_fitch_affirms_stable_bbb
1	1,345	1_oregon_preview_utah_virginia
2	1,091	2_sharks_ducks_coyotes_sabres
3	946	3_earnings_per_share_q2
4	919	4_european_watchshares_livemarkets_miners
5	829	5_stockswall_st_snapshotwall_opens
6	823	6_shooting_man_murder_charged
7	788	7_ebitda_ebit_ebita_sek
8	762	8_korea_north_missile_nkorea
9	693	9_amendment_revolving_credit_amended
10	647	10_jv_yuan_invest_briefbeijing
11	619	11_glut_crude_oil_iea
12	603	12_vr_virtual_reality_oculus
13	597	13_preciousgold_gold_dollar_equities
14	578	14_zlotys_yoy_zloty_result
15	572	15_loss_per_share_007
16	567	16_djokovic_murray_serena_nadal
17	542	17_stockstsx_canada_resource_financials
18	540	18_coup_turkeys_erdogan_turkey

Topic	Count	Name
2971	10	2971_prime_reading_amazon_memberships
2972	10	2972_charitable_giving_donations_donoradvised
2973	10	2973_yearlow_limp_backwards_ignite
2974	10	2974_pretax_briefratti_faustino_briefreply
2975	10	2975_hybrid_toyota_dualtrancher_briefferrari
2976	10	2976_1canada_sheds_unemployment_10700
2977	10	2977_stocksmarkets_stocksregion_mideast_stocksemerging
2978	10	2978_backhanded_trumpism_parallels_complement
2979	10	2979_explainers_sykespicot_disconnect_liberalism
2980	10	2980_injuries_trauma_injury_briefnahl
2981	10	2981_boxing_tishchenko_superheavy_yoka
2982	10	2982_middleage_mortality_prematurely_overpay
2983	10	2983_battlefield_application_extended_applications
2984	10	2984_cyber_wraps_stricter_blackmailers
2985	10	2985_saturdays_nba_sequence_astrolabe
2986	10	2986_corruption_1anticorruption_huberfeld_oncehighflying
2987	10	2987_whiting_petroleum_petroleums_marathon
2988	10	2988_gallery_buyer_dilascia_aplenty
2989	10	2989_nurse_ebola_hospital_recovered
2990	10	2990_cutters_cord_sling_streaming

Figure 12: Topics from the BERTopic EBF 2016 model.

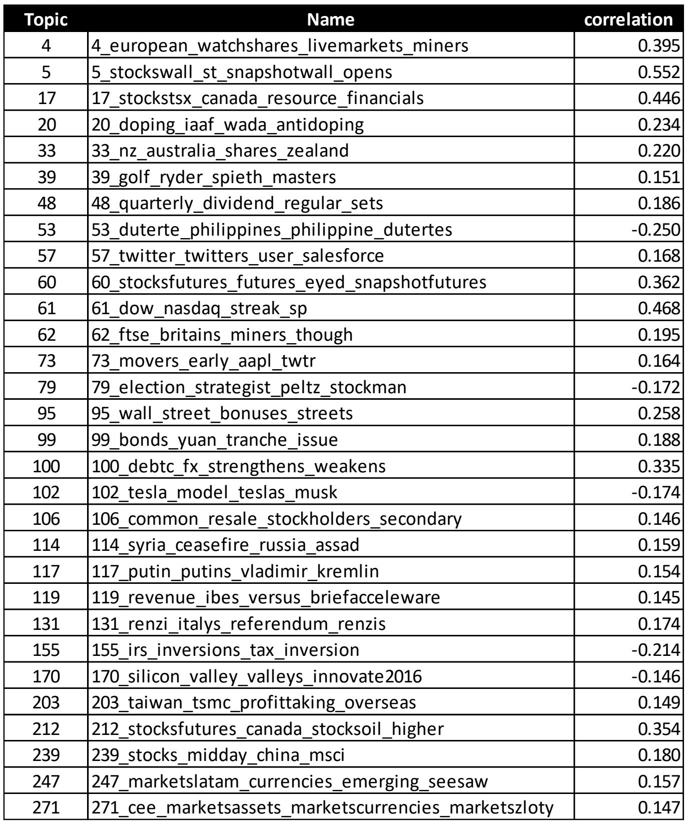
To select the most relevant topics, the selection is done with a correlation analysis process resembling the selection of publisher types explained in the Chapter 4.2. The process uses the average sentiment analysis per day for every topic using FinBERT. A correlation analysis is used to determine the most correlated topics to the S&P500 index. Two thresholds are decided to select the topics:

- A date count threshold is considered to select the topics. This means that topics that are present on at least 30% of the trading days are considered (78 days out of 260). This is done to avoid topics with a spurious correlation with the S&P500 Index, avoid topics that are very specific and/or are coincidentally correlated with the S&P500 as are only published in a small percentage of trading days.
- A correlation threshold is decided on the absolute value of 0.1 (Considering correlated topics with a correlation lower than -0.1 or higher than 0.1). This threshold was determined considering the comparison of studies done by Akoglu in [56]. The author compared “the three most commonly used interpretations of the r values” (the compared definitions are from authors from different specialties and research areas). In the comparison of these 3 interpretations, below the threshold of the absolute value of 0.1 the correlations are classified as “Zero”, “None” and “None”. Correlations coefficients with a 0.1 absolute value are classified as “Weak”, “Negligible” and “Poor”.

The top 30 topics that fulfill the date count and the correlation threshold are selected. The selected topics and their corresponding correlation with the S&P500 index can be seen in the Figure 13. An interesting observation is the diversity of domains represented by the correlated topics:

- Topics such as 17, 33, 53, 62, 203, 212, 239, 247 consider articles related with international markets such as Canada, Australia, Britain, China, Latam and others.
- Another selection of topics such as 95, 99, 100, 119, 155 are related to financial themes such as Interest rates, bonds, foreign exchange, and others.

- Topics such as 114, 117 and 131 are more related with geopolitics with Russia, Syria ceasefire and Italy constitutional referendum.
- Another group of topics 57,73, 102, 170 refer to specific corporate themes such as Twitter, Salesforce, Apple, Tesla, Elon Musk, and Silicon Valley.



Topic	Name	correlation
4	4_european_watchshares_livemarkets_miners	0.395
5	5_stockswall_st_snapshotwall_opens	0.552
17	17_stockstsx_canada_resource_financials	0.446
20	20_doping_iaaf_wada_antidoping	0.234
33	33_nz_australia_shares_zealand	0.220
39	39_golf_ryder_spieth_masters	0.151
48	48_quarterly_dividend_regular_sets	0.186
53	53_duterte_philippines_philippine_dutertes	-0.250
57	57_twitter_twitters_user_salesforce	0.168
60	60_stocksfutures_futures_eyed_snapshotfutures	0.362
61	61_dow_nasdaq_streak_sp	0.468
62	62_ftse_britains_miners_though	0.195
73	73_movers_early_aapl_twtr	0.164
79	79_election_strategist_peltz_stockman	-0.172
95	95_wall_street_bonuses_streets	0.258
99	99_bonds_yuan_tranche_issue	0.188
100	100_debtc_fx_strengthens_weakens	0.335
102	102_tesla_model_teslas_musk	-0.174
106	106_common_resale_stockholders_secondary	0.146
114	114_syria_ceasefire_russia_assad	0.159
117	117_putin_putins_vladimir_kremlin	0.154
119	119_revenue_ibes_versus_briefacceleware	0.145
131	131_renzi_italys_referendum_renzis	0.174
155	155_irs_inversions_tax_inversion	-0.214
170	170_silicon_valley_valleys_innovate2016	-0.146
203	203_taiwan_tsmc_profittaking_overseas	0.149
212	212_stocksfutures_canada_stocksoil_higher	0.354
239	239_stocks_midday_china_msci	0.180
247	247_marketslatam_currencies_emerging_seesaw	0.157
271	271_cee_marketsassets_marketscurrencies_marketszloty	0.147

Figure 13: Top 30 correlated topics with the S&P500 index

6. Baseline Implementation of Sentiment Analysis and Topic Modeling

To understand the impact of BERT models it is important to compare the results with a baseline. This baseline consists of using models that are used in the related literature and obtaining the results by following the same method. In the literature review, some previous models were discussed for sentiment analysis in Section 2.1 and for topic modeling in Section 2.2. For the sentiment analysis component, the benchmark is the Loughran and McDonald word list that is specialized for financial text [13] and in the topic modeling the benchmark used is the Latent Dirichlet Allocation model since it is general and widely used for topic modeling [18]. This means that the same method is evaluated considering the alternative for each BERT component as benchmarks, in the Figure 14 the proposed method can be visualized with the corresponding benchmark models.

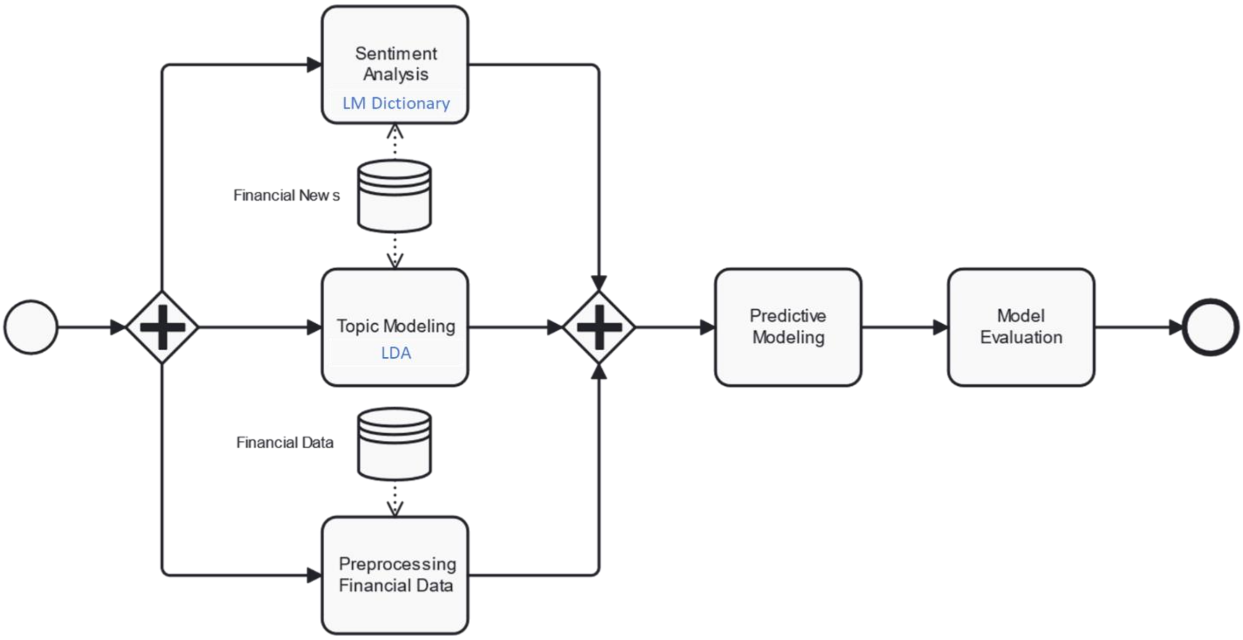


Figure 14: Diagram of the baseline implementation of the proposed method.

Sentiment Analysis - Loughran McDonald Dictionary

For the sentiment analysis benchmark, the Pysentiment2 – lm Module developed by DeRobertis [57] was used for using the dictionary of Loughran and McDonald financial sentiment dictionaries. The same procedure done with FinBERT was replicated with some adjustments. The sentiment score is calculated with the polarity as explained by DeRobertis [57].

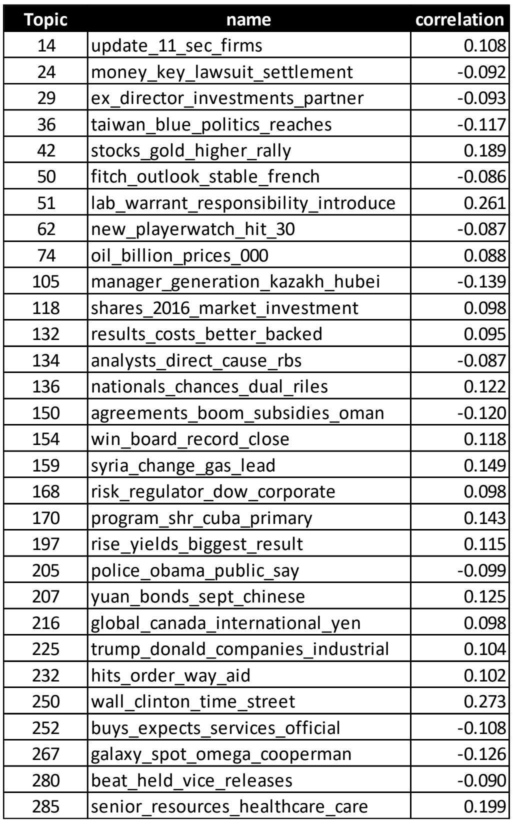
$$Sentiment\ Score = Polarity = \frac{N_{pos} - N_{neg}}{N_{pos} + N_{neg}}$$

Topic Modeling - LDA

The library *sklearn* with the module *decomposition-LatentDirichletAllocation* was used for applying LDA in the topic modeling component. The same procedure done with BERTopic was replicated but considering some adjustments. As explained in the Section 2.2. LDA differs from BERTopic in two specific features:

- BERTopic associates each document to one specific topic (Topic distribution are not generated), while in LDA a document is associated to several different topics by a percentage.
- BERTopic automatically finds the number of topics whereas in LDA the number of topics needs to be defined.

For replicating the procedure, the topic with highest percentage of association was selected as the topic of the article and the number of topics for training the LDA topic model was determined to a specific number. Considering that according to Egger and Yu [20] in topic modeling with LDA the number of topics tends to be smaller compared to word-embedding based methods. Since BERTopic found 2990 topics for the topic model for 2016 EBF news, a smaller number should be specified. Given that in the BERTopic procedure the top 30 most associated topics were selected as predictors, the number of topics in the LDA model should be larger than 30 to allow the method sufficient topics to find and select the most correlated topics with the S&P index. Considering these technicalities, the parameter chosen for the number of topics for LDA was 300 topics. After the correlation analysis the top 30 most correlated topics with the S&P Index were selected as predictors for the classification models. In Figure 15 the selected topics name (terms) can be visualized.



Topic	name	correlation
14	update_11_sec_firms	0.108
24	money_key_lawsuit_settlement	-0.092
29	ex_director_investments_partner	-0.093
36	taiwan_blue_politics_reaches	-0.117
42	stocks_gold_higher_rally	0.189
50	fitch_outlook_stable_french	-0.086
51	lab_warrant_responsibility_introduce	0.261
62	new_playerwatch_hit_30	-0.087
74	oil_billion_prices_000	0.088
105	manager_generation_kazakh_hubei	-0.139
118	shares_2016_market_investment	0.098
132	results_costs_better_backed	0.095
134	analysts_direct_cause_rbs	-0.087
136	nationals_chances_dual_riles	0.122
150	agreements_boom_subsidies_oman	-0.120
154	win_board_record_close	0.118
159	syria_change_gas_lead	0.149
168	risk_regulator_dow_corporate	0.098
170	program_shr_cuba_primary	0.143
197	rise_yields_biggest_result	0.115
205	police_obama_public_say	-0.099
207	yuan_bonds_sept_chinese	0.125
216	global_canada_international_yen	0.098
225	trump_donald_companies_industrial	0.104
232	hits_order_way_aid	0.102
250	wall_clinton_time_street	0.273
252	buys_expects_services_official	-0.108
267	galaxy_spot_omega_cooperman	-0.126
280	beat_held_vice_releases	-0.090
285	senior_resources_healthcare_care	0.199

Figure 15: Top correlated topics for LDA topic model

7. Predictive Modeling and Model Evaluation

The performance evaluation consists of measuring the classification metrics of the different classification models which differ on the selection of components and parameter values. Considering variations of the NLP models used and the type of text used as input, the testing period, and the classification model types.

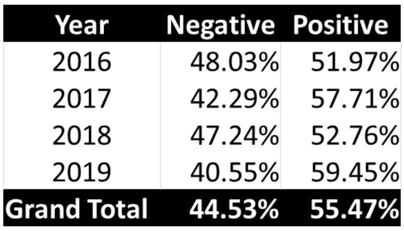
To facilitate the presentation of the results, the following notation is used in the tables:

- Classification model types
  - *log_reg:* Logistic Regression
  - *rf_model:* Random Forest
  - *svc_model*: Support Vector Machine Classifier
  - *nb_model:* Naïve Bayes Classifier.
  - *gb_model:* Gradient Boosting Classifier

- NLP models
  - *FinBERT_BERTopicModel2016Headlines250k*: Classification model with FinBERT sentiment analysis and BERTopic model trained with 2016 data with 250,000 articles headlines.
  - *FinBERT_BERTopicModel2016Headlines20k:* Classification model with FinBERT sentiment analysis and BERTopic model trained with 2016 data with 20,000 articles headlines.
  - *FinBERT_BERTopicModel2016Bodytext250k:* Classification model with FinBERT sentiment analysis and BERTopic model trained with 2016 data with 250,000 article's bodytext.
  - *FinBERT_Headlines:* Classification model considering only the FinBERT sentiment analysis of the article's headlines.
  - *FinBERT_Bodytext*: Classification model considering only the FinBERT sentiment analysis of the article's bodytext.
  - *LM_LDATopicModel2016Headlines250k:* Classification model with Loughran-McDonald dictionary sentiment analysis and LDA topic model trained with 2016 data with 250,000 articles headlines.
  - *LM_Headlines:* Classification model considering only the Loughran-McDonald dictionary sentiment analysis of the article's headlines.

- Test Period
  - *2018:* Test period 2 years after the topic model has been trained.
  - *2019:* Test period 3 years after the topic model has been trained.
  - *2018-2019*: Test period considering 2 years after the topic model has been trained after 2 years.
  - Since the year 2017 was used as the validation period for the grid search procedure, it cannot be used to measure the performance of the models.

To correctly interpret and understand the results of the direction prediction models, it is useful to know the proportion of the actual direction of the S&P500. The Table 9 presents the direction distribution per year, this means the proportion of days between positive and negative direction of the S&P500 index.



Year	Negative	Positive
2016	48.03%	51.97%
2017	42.29%	57.71%
2018	47.24%	52.76%
2019	40.55%	59.45%
Grand Total	**44.53%**	**55.47%**

Table 9: Direction distribution of the S&P500 Index

As seen in the table, the direction of the S&P500 for all the analyzed years the positive direction is the majority class, in the case of the year 2019 reaching to almost 60% and the lowest in the year 2016 to 51%. This table helps to understand the *accuracy paradox* mentioned in the Chapter 0 E.g., if a classification model always predicts that the direction of the S&P500 for the years 2016-2019 is positive, its accuracy would be 55.47%, meaning that the model is correct more than half of the time. Its sensitivity is 100% since it correctly identifies all the actual positives, but the problem is that its specificity would is 0% since does not identify any actual negative. This is the reason why it is important to consider more metrics additional to the accuracy. This example also helps understanding the use of the F1 Score metric. Since F1 Score is the harmonic mean of sensitivity and precision, the F1 score is always lower than the arithmetic mean of sensitivity and precision. This means that F1 score “punishes” extreme values, causing classification models that focus only on obtaining a high score on one of the 2 metrics to have a lower F1 Score.

In the continuation of this chapter, the results of the grid search procedure and the results of the performance evaluation are presented.

7.1. Grid Search Procedure for Parameter Selection

According to Muller and Guido [39], determining the optimal parameter values for a model to achieve the highest generalization performance is often challenging yet essential for nearly all models and datasets. Following the book of Muller and Guido [39], for choosing from a predetermined range of parameter values, one method is evaluating all possible combinations. This technique is known as grid search and is widely utilized.

Considering the importance of selecting the values of parameters, a grid search procedure was realized for each type of classification model. The grid search procedure consisted of evaluating all the possible combinations of a predefined set of values for the parameters of interest. Note that some of the classification models have many other different parameters, variations, and values but the objective of this thesis is not to do an exhaustive evaluation of the parameter’s performance.

- SVC
    - C: [0.1, 1, 10]
    - Kernel: ["linear", "poly", "rbf", "sigmoid"]

- Decision Tree
    - Criterion: ["gini", "entropy"]
    - Max Depth: [None, 3, 5, 10]

- Random Forest
    - Criterion: ["gini", "entropy"]
    - Max Depth: [None, 3, 5, 10]
    - Number of estimators: [10, 100, 500]

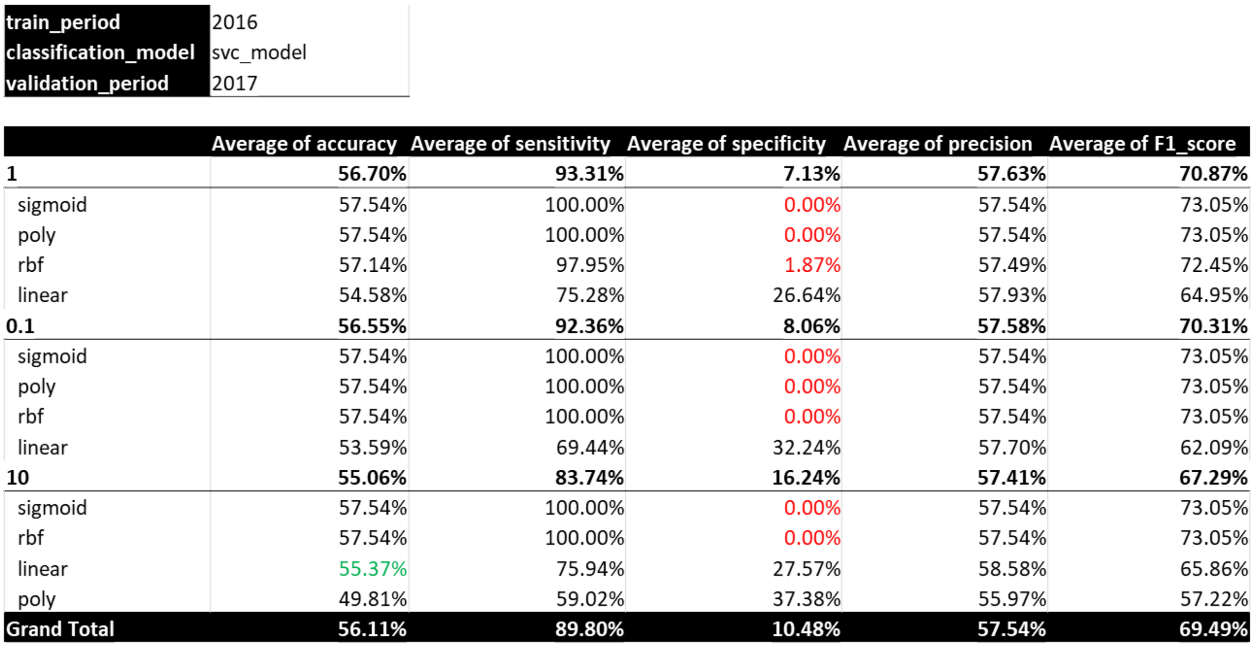
- Naïve Bayes
    - Var Smoothing: [1e-5,1e-9,1e-15]

- Gradient Boosted Trees for Classification
    - Eta: [0.1, 0.3, 0.5, 0.8]

The parameters were evaluated measuring the performance of the classification models for the year 2017, which was selected as the validation period for the selection of parameter values. The grid search evaluated two group types of models, considering BERT-based models (FinBERT and BERTopic) and the baseline models (LDA and Loughran-Macdonald Dictionary). A total of 101 different models were evaluated for the grid search procedure.

The results are organized into tables detailing the parameters under evaluation and the table results are sorted by accuracy in descending order. Parameters yielding the highest accuracy are chosen. Models with specificity or sensitivity below 5% are disregarded, as they would yield outcomes that resemble the zero-rule algorithm (explained in Chapter 0), which is not the desired outcome for employing these classification models.

Results for the grid search procedure for SVC



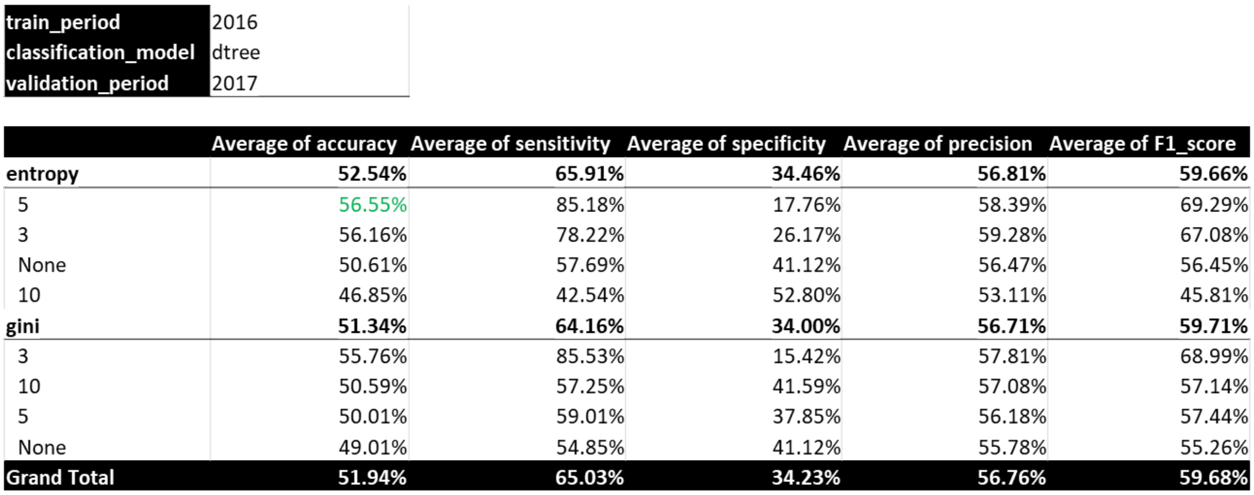
train_period	2016
classification_model	svc_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
1	**56.70%**	**93.31%**	**7.13%**	**57.63%**	**70.87%**
sigmoid	57.54%	100.00%	0.00%	57.54%	73.05%
poly	57.54%	100.00%	0.00%	57.54%	73.05%
rbf	57.14%	97.95%	1.87%	57.49%	72.45%
linear	54.58%	75.28%	26.64%	57.93%	64.95%
0.1	**56.55%**	**92.36%**	**8.06%**	**57.58%**	**70.31%**
sigmoid	57.54%	100.00%	0.00%	57.54%	73.05%
poly	57.54%	100.00%	0.00%	57.54%	73.05%
rbf	57.54%	100.00%	0.00%	57.54%	73.05%
linear	53.59%	69.44%	32.24%	57.70%	62.09%
10	**55.06%**	**83.74%**	**16.24%**	**57.41%**	**67.29%**
sigmoid	57.54%	100.00%	0.00%	57.54%	73.05%
rbf	57.54%	100.00%	0.00%	57.54%	73.05%
linear	55.37%	75.94%	27.57%	58.58%	65.86%
poly	49.81%	59.02%	37.38%	55.97%	57.22%
Grand Total	**56.11%**	**89.80%**	**10.48%**	**57.54%**	**69.49%**

Table 10: SVC grid search evaluating the C and Kernel hyperparameters

For the support vector machine classifier, the parameters with the best performance achieved an accuracy value of 57.54% but the models that yield this result present a specificity of 0%, meaning that these results are also obtained by a classifier that follows the zero-rule algorithm. As the selection criteria for the parameter values also considers that the sensitivity and specificity to be above 5%, these models are not selected. Considering the constraint and the highest accuracy, the selected parameters are a linear kernel and a C value of 10 which achieved an accuracy of 55.37%.

Results for the grid search procedure for Decision Tree



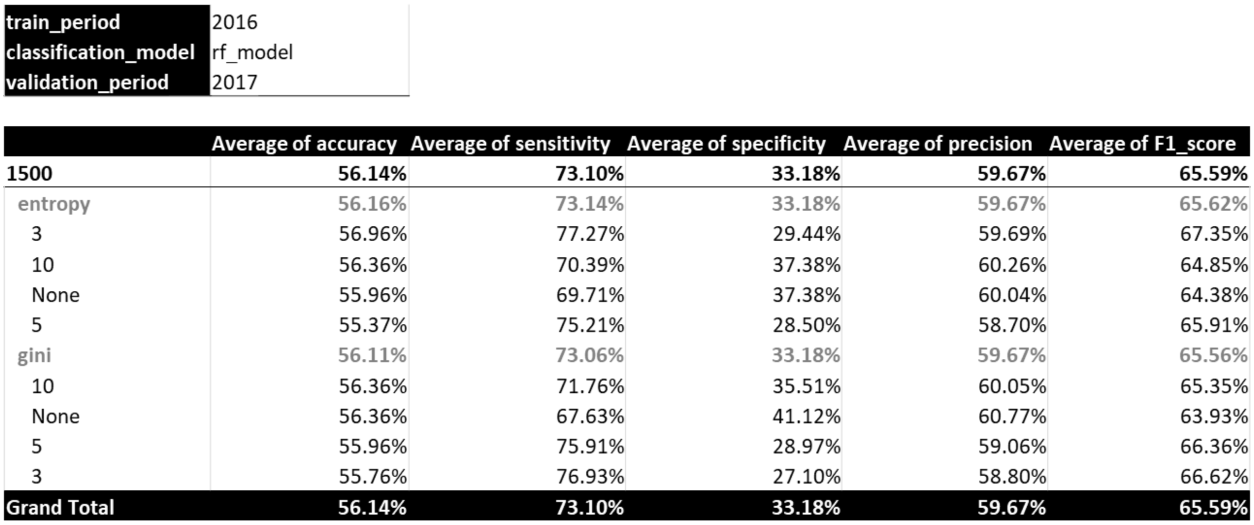
train_period	2016
classification_model	dtree
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
entropy	**52.54%**	**65.91%**	**34.46%**	**56.81%**	**59.66%**
5	56.55%	85.18%	17.76%	58.39%	69.29%
3	56.16%	78.22%	26.17%	59.28%	67.08%
None	50.61%	57.69%	41.12%	56.47%	56.45%
10	46.85%	42.54%	52.80%	53.11%	45.81%
gini	**51.34%**	**64.16%**	**34.00%**	**56.71%**	**59.71%**
3	55.76%	85.53%	15.42%	57.81%	68.99%
10	50.59%	57.25%	41.59%	57.08%	57.14%
5	50.01%	59.01%	37.85%	56.18%	57.44%
None	49.01%	54.85%	41.12%	55.78%	55.26%
Grand Total	**51.94%**	**65.03%**	**34.23%**	**56.76%**	**59.68%**

Table 11: Dec Tree grid search evaluating criterion and max depth.

For the decision tree classification models grid search, the best performance was achieved with a max depth of 5 and using the entropy criterion with an accuracy of 56.55%. The sensitivity and specificity are above 5%, therefore the parameter values selected for the next steps are the entropy as criterion and 5 as the max depth.

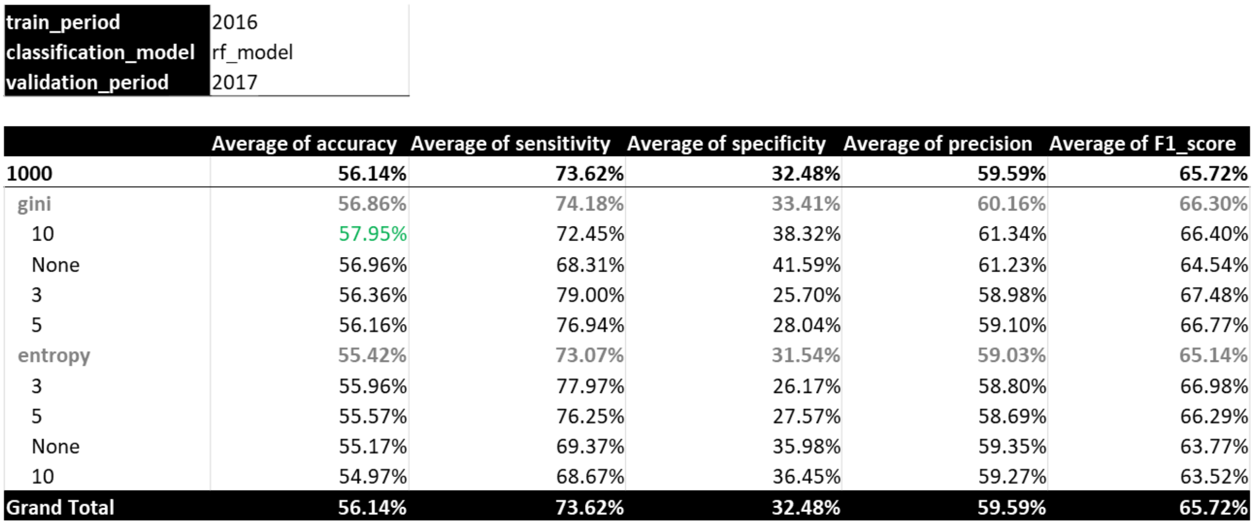
Results for the grid search procedure for Random Forest



train_period	2016
classification_model	rf_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
1500	**56.14%**	**73.10%**	**33.18%**	**59.67%**	**65.59%**
entropy	**56.16%**	**73.14%**	**33.18%**	**59.67%**	**65.62%**
3	56.96%	77.27%	29.44%	59.69%	67.35%
10	56.36%	70.39%	37.38%	60.26%	64.85%
None	55.96%	69.71%	37.38%	60.04%	64.38%
5	55.37%	75.21%	28.50%	58.70%	65.91%
gini	**56.11%**	**73.06%**	**33.18%**	**59.67%**	**65.56%**
10	56.36%	71.76%	35.51%	60.05%	65.35%
None	56.36%	67.63%	41.12%	60.77%	63.93%
5	55.96%	75.91%	28.97%	59.06%	66.36%
3	55.76%	76.93%	27.10%	58.80%	66.62%
Grand Total	**56.14%**	**73.10%**	**33.18%**	**59.67%**	**65.59%**

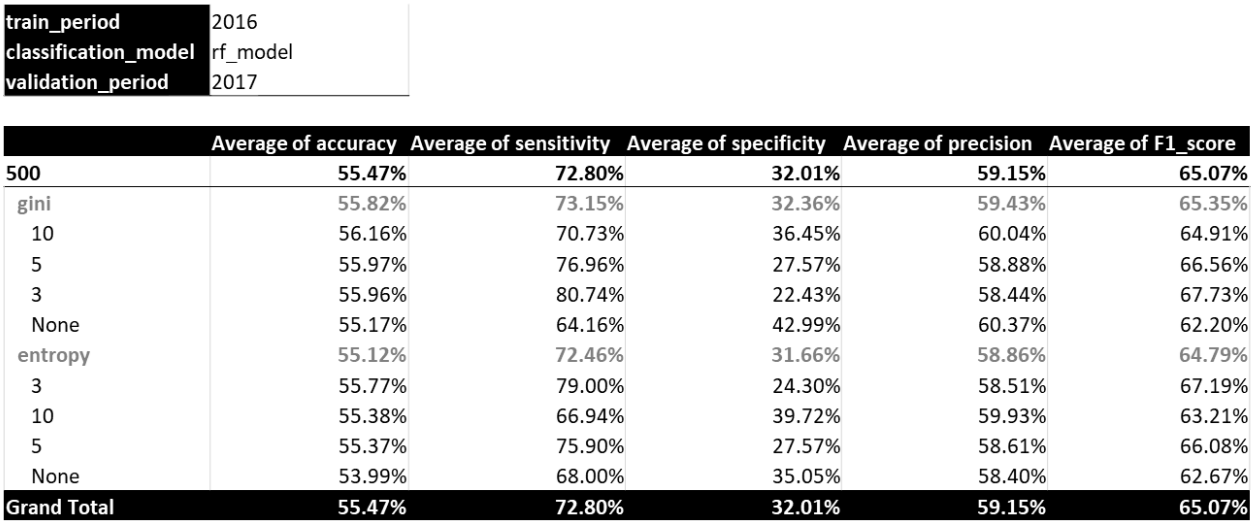
Table 12: Random Forest grid search evaluating criterion and max depth for 1500 estimators.



train_period	2016
classification_model	rf_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
1000	**56.14%**	**73.62%**	**32.48%**	**59.59%**	**65.72%**
gini	56.86%	74.18%	33.41%	60.16%	66.30%
10	57.95%	72.45%	38.32%	61.34%	66.40%
None	56.96%	68.31%	41.59%	61.23%	64.54%
3	56.36%	79.00%	25.70%	58.98%	67.48%
5	56.16%	76.94%	28.04%	59.10%	66.77%
entropy	55.42%	73.07%	31.54%	59.03%	65.14%
3	55.96%	77.97%	26.17%	58.80%	66.98%
5	55.57%	76.25%	27.57%	58.69%	66.29%
None	55.17%	69.37%	35.98%	59.35%	63.77%
10	54.97%	68.67%	36.45%	59.27%	63.52%
Grand Total	**56.14%**	**73.62%**	**32.48%**	**59.59%**	**65.72%**

Table 13: Random Forest grid search evaluating criterion and max depth for 100 estimators.



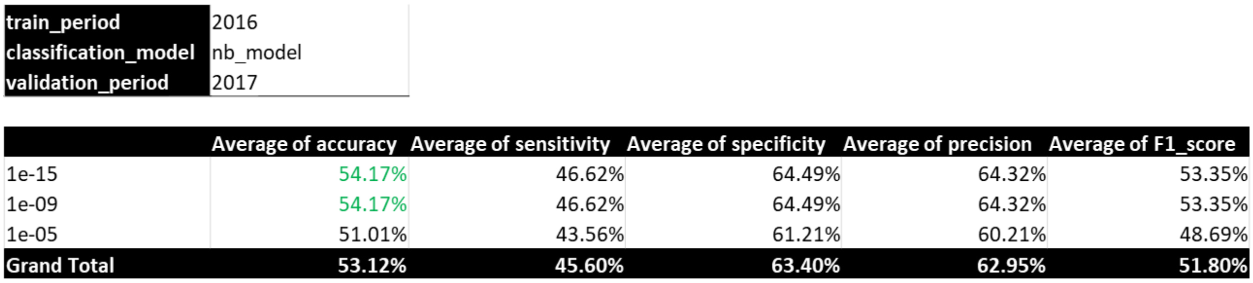
train_period	2016
classification_model	rf_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
500	**55.47%**	**72.80%**	**32.01%**	**59.15%**	**65.07%**
gini	55.82%	73.15%	32.36%	59.43%	65.35%
10	56.16%	70.73%	36.45%	60.04%	64.91%
5	55.97%	76.96%	27.57%	58.88%	66.56%
3	55.96%	80.74%	22.43%	58.44%	67.73%
None	55.17%	64.16%	42.99%	60.37%	62.20%
entropy	55.12%	72.46%	31.66%	58.86%	64.79%
3	55.77%	79.00%	24.30%	58.51%	67.19%
10	55.38%	66.94%	39.72%	59.93%	63.21%
5	55.37%	75.90%	27.57%	58.61%	66.08%
None	53.99%	68.00%	35.05%	58.40%	62.67%
Grand Total	**55.47%**	**72.80%**	**32.01%**	**59.15%**	**65.07%**

Table 14: Random Forest grid search evaluating criterion and max depth for 10 estimators.

Tables 12, 13 and 14 display the results of the grid search procedure for random forest evaluating 3 different parameters. The grid search procedure of the random forest classification models achieved a sensitivity and specificity above 5% for all the selected parameter values, therefore, the parameters with the highest accuracy were selected. The highest accuracy was 57.95% and was achieved with 1000 number of estimators, a gini criterion, and a max depth of 10 obtaining an accuracy of 57.95%. Is interesting to mention that the criterion with highest performance in random forest is different for the decision tree, which was entropy.

Results for the grid search procedure for Naïve Bayes



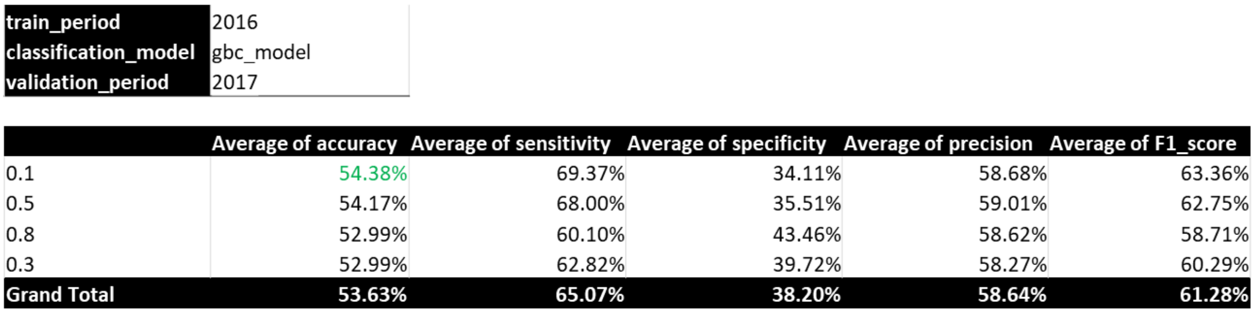
train_period	2016
classification_model	nb_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
1e-15	54.17%	46.62%	64.49%	64.32%	53.35%
1e-09	54.17%	46.62%	64.49%	64.32%	53.35%
1e-05	51.01%	43.56%	61.21%	60.21%	48.69%
Grand Total	53.12%	45.60%	63.40%	62.95%	51.80%

Table 15: Naïve Bayes grid search evaluating Var Smoothing.

The best performance for this grid search procedure achieved an accuracy of 54.17%. The parameters values of the var smoothing that achieved this result were the var smoothing of 1e-09 and 1e-15. The values of 1e-09 and 1e-15 obtained the same results, meaning that the value of 1e-09 provides sufficient stability. Considering this, the selected parameter value is 1e-09. In these models, the specificity is higher than the sensitivity meaning that these models are better in identifying the negative class than the positive class.

Results for the grid search procedure for Gradient Boosting Tree Classifier



train_period	2016
classification_model	gbc_model
validation_period	2017

	Average of accuracy	Average of sensitivity	Average of specificity	Average of precision	Average of F1_score
0.1	54.38%	69.37%	34.11%	58.68%	63.36%
0.5	54.17%	68.00%	35.51%	59.01%	62.75%
0.8	52.99%	60.10%	43.46%	58.62%	58.71%
0.3	52.99%	62.82%	39.72%	58.27%	60.29%
Grand Total	53.63%	65.07%	38.20%	58.64%	61.28%

Table 16: GBC grid search evaluating ETA

All the parameter values achieved a sensitivity and specificity above 5%, therefore, the parameters with the highest accuracy were selected. The highest accuracy was 54.38% and was achieved with an eta value of 0.1.

In summary, parameters with the highest accuracy that also meet the requirement of achieving sensitivity and specificity above 5% were chosen. The selected parameter values for each classification model type are outlined as follows:

- SVC
  - C: 10
  - Kernel: linear

- Decision Tree
  - Criterion: Entropy
  - Max Depth: 5

- Random Forest
  - Criterion: Gini
  - Max Depth: 10
  - Number of estimators: 1000

- Naïve Bayes
  - Var Smoothing: 1e-9.

- Gradient Boosted Trees for Classification
  - Eta: 0.1

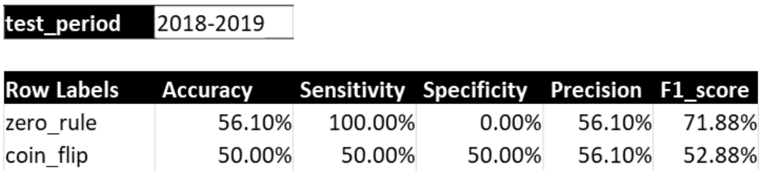
Note that no tests for statistical significance regarding the performance of the different parameters for stochastic algorithms were performed. Different executions of the algorithms may lead to different results. The goal, however, was not to find the optimal configuration but just to evaluate in general whether state-of-the-art NLP techniques for sentiment analysis and topic modeling could improve prediction accuracy with this method.

7.2. Performance Evaluation Results

The performance of various models and method variations is assessed using the parameter values selected through the grid search process (outlined in Section 7.1). The test period considers the time frame between the year 2018 and the year 2019 and the performance evaluation encompasses accuracy, sensitivity, specificity, precision, and F1 Score. The results are presented in tables that categorize each case and its model variations, with results arranged by accuracy from highest to lowest. While the results are sorted based on accuracy, the analysis also considers the other metrics to provide a comprehensive comparison.

Evaluation of reference classifiers

As stated by Norris [42], evaluating the performance of a baseline serves as a benchmark for comparing the results of the classification models. With the evaluation data, two additional models are presented as reference classifiers. The first one is a classifier that follows the zero-rule algorithm (explained in Chapter 0). The second reference classifier will be referred to as the coin flip classifier, which randomly assigns the direction with a 50% chance. To understand the results is important to consider the imbalance of the evaluated dataset, i.e., 56.10% of the days between the years 2018 and 2019 were positive. Note that expected values for the coin flip classifier are provided as they are more informative than the results of an actual experiment of a coin flip classifier.



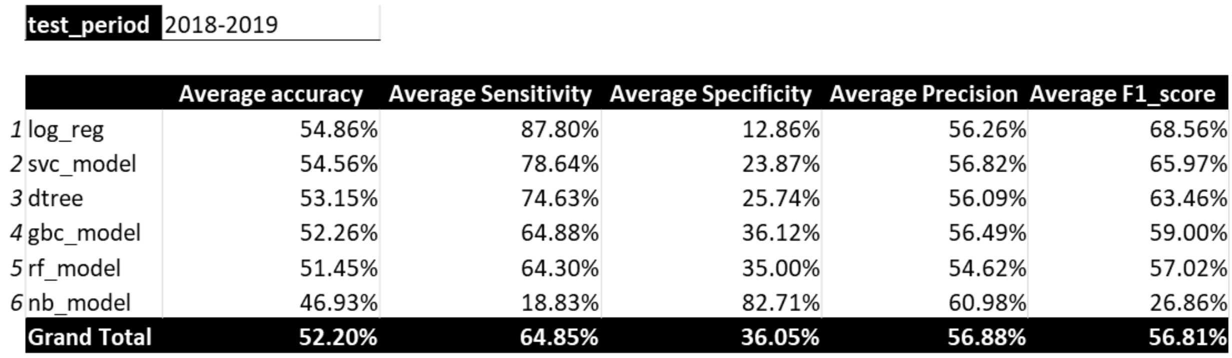
test_period	2018-2019

Row Labels	Accuracy	Sensitivity	Specificity	Precision	F1_score
zero_rule	56.10%	100.00%	0.00%	56.10%	71.88%
coin_flip	50.00%	50.00%	50.00%	56.10%	52.88%

Table 17: Results of zero-rule algorithm and coin flip classifier

Evaluation of classification model types

For a clearer interpretation of all the results, the performance between the classification model types is first presented. The Table 18 shows the average performance of *FinBERT_BERTopicModel2016Headlines250k* and *LM_LDATopicModel2016Headlines250k* for the test periods *2018-2019* for each selected type of classification model.



test_period	2018-2019

		Average accuracy	Average Sensitivity	Average Specificity	Average Precision	Average F1_score
1	log_reg	54.86%	87.80%	12.86%	56.26%	68.56%
2	svc_model	54.56%	78.64%	23.87%	56.82%	65.97%
3	dtree	53.15%	74.63%	25.74%	56.09%	63.46%
4	gbc_model	52.26%	64.88%	36.12%	56.49%	59.00%
5	rf_model	51.45%	64.30%	35.00%	54.62%	57.02%
6	nb_model	46.93%	18.83%	82.71%	60.98%	26.86%
	Grand Total	**52.20%**	**64.85%**	**36.05%**	**56.88%**	**56.81%**

Table 18: Results of the performance metrics by classification model type

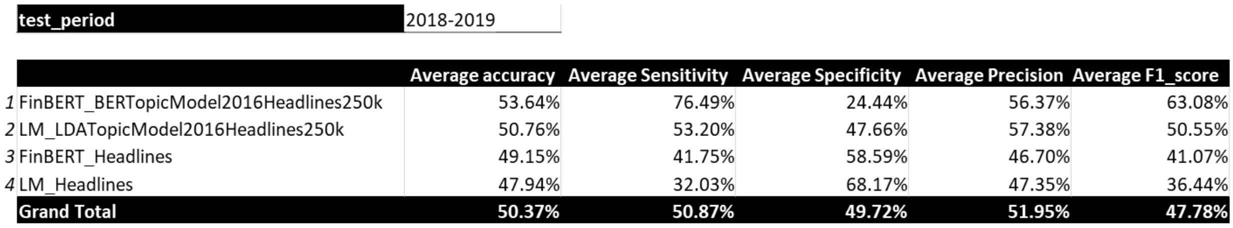
The purpose of this table is to understand the general performance of the classification model types. In other words, this table presents the average score of the different classification model types consolidating the average score of both the BERT based models and the LM-LDA based models.

The logistic regression models were the best in terms of average accuracy, but on average they heavily relied on classifying the predictions in one direction therefore being penalized with the worst specificity of 12.86%. Since the years 2018-2019 (see Table 9) the direction is more inclined to the positive direction, this benefits this type of models. The SVC stands in the 2nd position in terms of accuracy. The worst model was the Naïve Bayes model which obtained an accuracy of 46.93% although its specificity was the largest with 82.71%. The Naïve Bayes model is the best in predicting the negative classes but punishing its accuracy in predicting the classes.

In terms of accuracy on average these models performed worse than a zero-rule algorithm as the value is below 56.10% but performed better than a coin flip (except for the NB classification models).

Evaluation of method variation

The Table 19 shows the average performance of the six classification model types for the test periods *2018-2019* for each method variation. The variations are sentiment analysis with topic modeling, sentiment analysis without topic modeling and the comparison between the BERT based models and the LM-LDA models.



test_period	2018-2019				
	Average accuracy	**Average Sensitivity**	**Average Specificity**	**Average Precision**	**Average F1_score**
1 FinBERT_BERTopicModel2016Headlines250k	53.64%	76.49%	24.44%	56.37%	63.08%
2 LM_LDATopicModel2016Headlines250k	50.76%	53.20%	47.66%	57.38%	50.55%
3 FinBERT_Headlines	49.15%	41.75%	58.59%	46.70%	41.07%
4 LM_Headlines	47.94%	32.03%	68.17%	47.35%	36.44%
Grand Total	**50.37%**	**50.87%**	**49.72%**	**51.95%**	**47.78%**

Table 19: Results of the performance metrics by method variation

The 1st position for accuracy, sensitivity, precision and F1 score was achieved by the method that integrates sentiment analysis and topic modeling through FinBERT and BERTopic respectively.

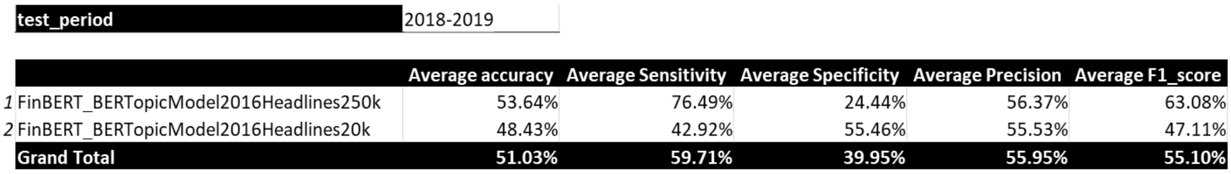
The models in the 1st and 2nd position in terms of accuracy, sensitivity, precision and F1 score were the ones integrating both sentiment analysis and topic modeling. The 3rd and 4th position are held by the models that are considering only sentiment analysis.

From the models that consider only sentiment analysis (3 and 4), a better accuracy and F1 score was achieved by the model that uses FinBERT instead of the benchmark which uses the Loughran and McDonald dictionary.

These different insights guide us in the direction that the integration of sentiment analysis and topic modeling, and the use of BERT models make a difference as they obtain a better accuracy score than the benchmark models LM-LDA.

Evaluation of topic model train size

The Table 20 shows the average performance of the selected classification model types for the test periods *2018-2019* for each train size (250.000 news vs 20.000 news) for the models that integrate topic modeling and sentiment analysis with BERT based models.



test_period	2018-2019

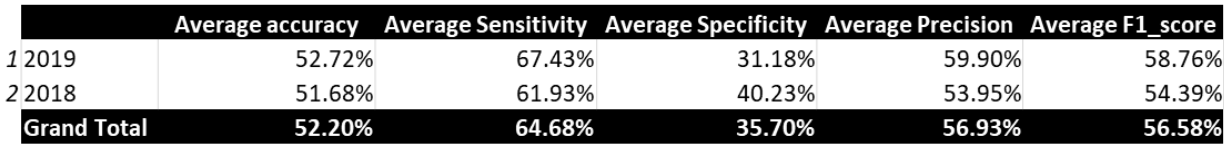
		Average accuracy	Average Sensitivity	Average Specificity	Average Precision	Average F1_score
1	FinBERT_BERTopicModel2016Headlines250k	53.64%	76.49%	24.44%	56.37%	63.08%
2	FinBERT_BERTopicModel2016Headlines20k	48.43%	42.92%	55.46%	55.53%	47.11%
	Grand Total	**51.03%**	**59.71%**	**39.95%**	**55.95%**	**55.10%**

Table 20: Results of the performance metrics by topic model train size

The accuracy of the model trained with 250K news has a better accuracy than the one trained with the sample of 20K news by 5 percentage points and the gap is larger for the F1 Score metric where the difference is around 15 percentage points. The results indicate that the size of the training of the topic modeling does matter and proves that the effort to train a larger topic model is valuable. Although it requires more resources in terms of computation and information, it proves that a larger training dataset is useful. If the data was not useful at all, the accuracy and the F1 score would not increase this magnitude.

Evaluation of test periods

The Table 21 shows the average performance of *FinBERT_BERTopicModel2016Headlines250k* of the selected classification model types for each test period.



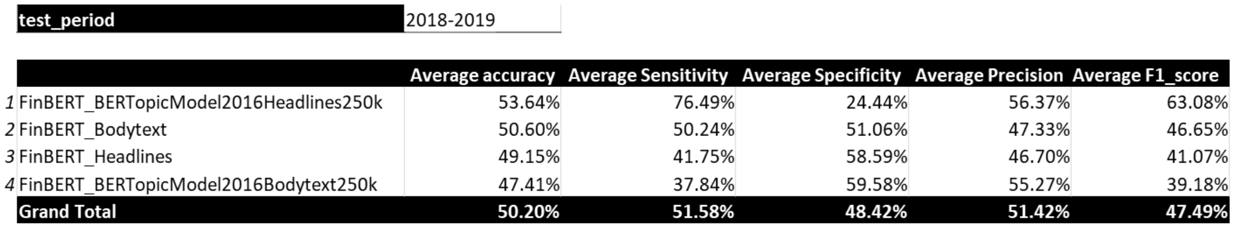
		Average accuracy	Average Sensitivity	Average Specificity	Average Precision	Average F1_score
1	2019	52.72%	67.43%	31.18%	59.90%	58.76%
2	2018	51.68%	61.93%	40.23%	53.95%	54.39%
	Grand Total	**52.20%**	**64.68%**	**35.70%**	**56.93%**	**56.58%**

Table 21: Results of the performance metrics by test period

The expectation was for predictions for the year 2018 to be more accurate than the predictions for 2019, given that topics discovered in 2016 are closer to the year 2018 than 2019. Surprisingly, the accuracy and F1 Score were higher during the 2019 test period compared to 2018. This indicates that models trained on news from 2016 performed better in predicting outcomes for 2019 than for 2018. One possible explanation is the models' bias towards predicting positive outcomes over negative ones, particularly evident in 2019, where 59% of days were positive, compared to 52% in 2018. To further investigate this, tests spanning more years would be necessary. However, the unavailability and difficulty in obtaining such data prevent us from confirming this hypothesis. Examining additional years would allow us to assess the impact on predictions when using both updated and outdated topic models.

Evaluation of type of article text

The Table 22 shows average performance of all the classification model types for the test periods *2018-2019* evaluating the bodytext and the headlines as input for the classification models.



test_period	2018-2019

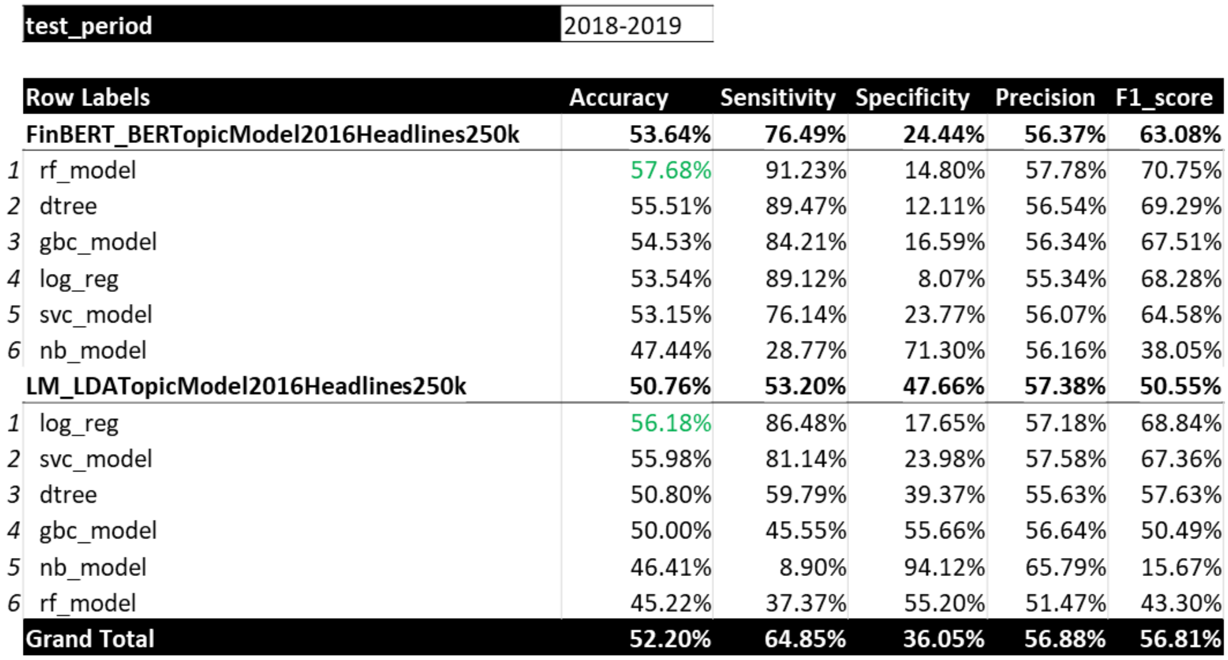
		Average accuracy	Average Sensitivity	Average Specificity	Average Precision	Average F1_score
1	FinBERT_BERTopicModel2016Headlines250k	53.64%	76.49%	24.44%	56.37%	63.08%
2	FinBERT_Bodytext	50.60%	50.24%	51.06%	47.33%	46.65%
3	FinBERT_Headlines	49.15%	41.75%	58.59%	46.70%	41.07%
4	FinBERT_BERTopicModel2016Bodytext250k	47.41%	37.84%	59.58%	55.27%	39.18%
	Grand Total	**50.20%**	**51.58%**	**48.42%**	**51.42%**	**47.49%**

Table 22: Results of the performance metrics by text type input for the method variations

The models that combine topic modeling and sentiment analysis, utilizing headlines as input, achieved the highest accuracy and F1 score. Surprisingly, the highest accuracy was achieved by the classification model that employs headlines rather than the entire article, contrary to expectations. It is important to note that the best accuracy between body text and headlines depends on the method variation. When using models that integrate topic modeling and sentiment analysis, employing headlines resulted in an accuracy that is 6 percentage points higher than when using body text. Conversely, when employing sentiment analysis without topic modeling, the best performance was achieved by utilizing body text as input. Different elements could cause this effect and more research should be done to conclude on this.

Best performing models

The previous tables showed the performance of the different models in aggregated terms, averages of the different models and variations as they summarized the insights fixing different dimensions. For comparing the actual values and not the averages, the Table 23 presents the results of each classification model for the *FinBERT_BERTopicModel2016Headlines250k* and the *LM_LDATopicModel2016Headlines250k* for the test periods *2018-2019.*



test_period	2018-2019

	Row Labels	Accuracy	Sensitivity	Specificity	Precision	F1_score
	FinBERT_BERTopicModel2016Headlines250k	**53.64%**	**76.49%**	**24.44%**	**56.37%**	**63.08%**
1	rf_model	57.68%	91.23%	14.80%	57.78%	70.75%
2	dtree	55.51%	89.47%	12.11%	56.54%	69.29%
3	gbc_model	54.53%	84.21%	16.59%	56.34%	67.51%
4	log_reg	53.54%	89.12%	8.07%	55.34%	68.28%
5	svc_model	53.15%	76.14%	23.77%	56.07%	64.58%
6	nb_model	47.44%	28.77%	71.30%	56.16%	38.05%
	LM_LDATopicModel2016Headlines250k	**50.76%**	**53.20%**	**47.66%**	**57.38%**	**50.55%**
1	log_reg	56.18%	86.48%	17.65%	57.18%	68.84%
2	svc_model	55.98%	81.14%	23.98%	57.58%	67.36%
3	dtree	50.80%	59.79%	39.37%	55.63%	57.63%
4	gbc_model	50.00%	45.55%	55.66%	56.64%	50.49%
5	nb_model	46.41%	8.90%	94.12%	65.79%	15.67%
6	rf_model	45.22%	37.37%	55.20%	51.47%	43.30%
	Grand Total	**52.20%**	**64.85%**	**36.05%**	**56.88%**	**56.81%**

Table 23: Results of the best performing models

The top-performing model within the FinBERT and BERTopic framework was the random forest with an accuracy of 57.68% and an F1 score of 70.75%. For the LM and LDA framework, the best performance model was the logistic regression with an accuracy of 56.18% and an F1 score of 68.84%. In terms of accuracy, both models performed better than the coin flip (50.0%) and the zero-rule (56.10%) classifiers. However, in terms of specificity, these models did not perform better than the coin flip classifier's 50.0%. This means that the coin flip classifier exhibited better performance in detecting the negative direction compared to the models with the highest accuracies. On average, considering aggregated metrics, the FinBERT and BERTopic models outperformed the LM-LDA models in terms of accuracy and F1 Score.

It is interesting to note that classification model types yielded diverse performances across the BERT and the LM-LDA based models, resulting in varying positions. This observation stands out considering that the models share the same types of variables. The performance disparity is highlighted by the difference between BERT-based and baseline models for random forest. Random forest was the top-performing BERT model while performing the poorest among the LM-LDA models.

8. Conclusions

In summary, this thesis evaluates the use of sentiment analysis and topic modeling over financial news for predicting the stock price direction of movement of the S&P 500 index. The proposed method integrates sentiment analysis and topic modeling through FinBERT and BERTopic, respectively, which serves as the basis for learning classification models (through Logistic Regression, SVM, Random Forest, and others) and compares them to other sentiment analysis and topic modeling approaches (lexicons and LDA). The thesis also evaluates different variations and parameters to understand the options that obtain a better performance. In particular, the thesis evaluates the impact of the size of the trained topic model, the impact of the frequency of training a topic model, and the effect of considering the headlines of the articles compared to the full article. This procedure leads to interesting insights drawn from the obtained results.

Considering that stock prices are hard to predict and that it is a constant goal of investors, increasing the accuracy, sensitivity, specificity, precision, and F1 Score of a directional prediction model of the stock market is of great value. Reference classifiers were established and evaluated for understanding the performance of the classification models trained in the research. The reference classifiers were the zero-rule algorithm, which consistently predicts the direction with the highest frequency, and a coin flip classifier, which assigns a 50% chance to each direction. While certain trained models achieved a better performance in terms of accuracy or specificity compared to the zero-rule algorithm and the coin flip classifier, none of the models concurrently outperformed both in terms of accuracy and specificity. Consequently, this finding suggests that the integration of sentiment analysis and topic modeling with economic, business, and financial articles may not consistently enhance the prediction of the S&P 500 index direction compared to approaches such as the zero-rule algorithm or flipping a coin.

Variations of the proposed method were evaluated to provide different insights and findings. BERT-based models achieved higher accuracy compared to the baseline techniques for topic modeling and sentiment analysis. The baseline used for sentiment analysis was the Loughran and McDonald dictionary, while for topic modeling, the benchmark used was the Latent Dirichlet Allocation model. Although the observed improvement may not be as substantial as expected, it is recommended for future approaches to weigh the effort of employing models that consider semantic relationships of words and the impact on the desired outcome. On average, the difference in terms of accuracy was 2.8 percentage points and the difference in terms of F1 Score was 12.5 percentage points. The best-performing model for the BERT-based models achieved an accuracy of 57.68% and a specificity of 14.80%. The best-performing model for the LM-LDA-based models achieved an accuracy of 56.18% and a specificity of 17.65%. Although these leading models achieved marginally higher accuracy rates compared to the zero-rule algorithm (56.10%) and the coin-flip classifier (50%), these models achieved a worse specificity compared to the coin-flip classifier (50%). The results indicate that these models are not better than a coin flip classifier in correctly identifying the negative direction.

On average, the integration of sentiment analysis and topic modeling performed better compared to the method of using only sentiment analysis in terms of accuracy. Specifically, the combination of BERTopic and FinBERT outperformed the sole use of FinBERT by a margin of 4.49 percentage points. Similarly, the integration of LDA and LM dictionary surpassed the performance of using sentiment analysis with the LM dictionary alone, exhibiting an improvement of 2.82 percentage points.

It was expected that a topic model trained on recent news to outperform one trained on older news. However, the results contradicted this expectation, with the 2016-trained model demonstrating better performance in 2019 than in 2018. One potential cause may lie in the bias of classification models toward predicting positive outcomes over negative ones. Notably, 2019 exhibited a higher proportion of positive days (59.45%) compared to 2018 (52.76%), which could have influenced the model's performance. Due to constraints in test periods and data availability, this could not be proven.

Another unexpected outcome is that using headlines as input for the proposed method with BERT-based models achieved a better performance in terms of accuracy and F1 score than using the body text of the articles. The difference in terms of accuracy was 6.23 percentage points. However, it is worth nothing that when it comes to using sentiment analysis without topic modeling, using the body text instead of headlines results in a better performance, the difference in terms of accuracy was 1.45 percentage points.

As explained by Egger and Yu, objective evaluation metrics are missing for evaluating topic models [20]. Due to the unsupervised nature of topic creation and assignment, objectively measuring a topic model's performance poses a challenge. An interesting approach to address this issue is the integration of topic models with sentiment analysis. This integration offers a means to objectively evaluate topic models' performance with objective metrics, particularly in tasks such as predicting the direction of movement of the S&P500 index.

While the thesis faced certain limitations and challenges, these provided ideas for future investigation. Some potential direction for future research includes:

- **Integration with Economic and Financial Metrics:** Exploring how economic and financial metrics, such as GDP, unemployment rates, interest rates, and price trends, can be integrated into the analysis to enhance predictive accuracy.

- **Performance During Exceptional Events:** Evaluating the performance of classification models when applied to news articles published during significant events, such as the COVID-19 pandemic and the Russian invasion of Ukraine.

- **Intraday News Analysis:** Investigating the intraday effects of news on stock price movements and market dynamics. Analyzing the performance of classification models with news articles published after varying time intervals from their original release (e.g., hours or minutes).

- **Specialized Fine-Tuning Procedures:** Developing and implementing specialized fine-tuning procedures to optimize model performance and achieve better results.

- **Cross-Market Analysis:** Assessing the impact of news on different financial markets across regions, including the EU, Latin America, Africa, and Asia.

References

[1] B. Malkiel, A Random Walk Down Wall Street, W. W. Norton & Company, Inc., 1973.

[2] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance,* vol. 25, pp. 383-417, 1970.

[3] G. Birz and J. Lott, "The effect of macroeconomic news on stock returns: New evidence from newspaper coverage," *Journal of Banking & Finance,* vol. 35, 2011.

[4] C. Ranganathan and C. V. Brown, "ERP Investments and the Market Value of Firms: Toward an Understanding of Influential ERP Project Variables," *Information Systems Research,* vol. 17, no. 2, pp. 145-161, 2006.

[5] T. Sidogi, R. Mbuvha and T. Marwala, "Stock Price Prediction Using Sentiment Analysis," *IEEE,* Vols. International Conference on Systems, Man, and Cybernetics (SMC), 2021.

[6] R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," *IEEE,* vol. 2020 Conference on Multimedia Information Processing and Retrieval (MIPR), 2020.

[7] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," *IEEE,* vol. Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019.

[8] S. Y. K. Mo, A. Liu and S. Y. Yang, "News sentiment to market impact and its feedback effect," *Environment Systems and Decisions,* vol. 36, 2016.

[9] M. T. Bakker, "Forecasting the Stock Market using News Sentiment Analysis," *M.S. Thesis, Department of Cognitive Science & Artificial Intelligence, Tilburg University,* 2021.

[10] B. Fazlija and P. Harder, "Using Financial News Sentiment for Stock Price Direction Prediction," *Applications and Mathematical Foundations of Machine Learning in Investments,* 2022.

[11] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal,* vol. 5, no. 4, 2014.

[12] M. Leippold, "Sentiment spin: Attacking financial sentiment with GPT-3," *Finance Research Letters,* 2023.

[13] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance,* 2011.

[14] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *University of Amsterdam,* 2019.

[15] A. Mahajan, L. Dey and S. M. Haque, "Mining Financial News for Major Events and Their Impacts on the Market," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[16] K. Yono, K. Izumi, H. Sakaji, H. Matsushima and T. Shimada, "Extraction of Focused Topic and Sentiment of Financial Market by using Supervised Topic Model for Price Movement Prediction," *IEEE,* vol. Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 2019.

[17] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li and X. Deng, "Exploiting topic based twitter sentiment for stock prediction," *51st Annual Meeting of the Association for Computational Linguistics, ACL,* vol. 2, pp. 24-29, 2013.

[18] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri and S. Adam, "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable," *Communication Methods and Measures,* vol. 12, no. 2-3 Computational Methods for Communication Science, 2018.

[19] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.

[20] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology,* vol. 7, 2022.

[21] D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice Hall, 2000.

[22] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for language understanding," 2018.

[23] T. Loughran and B. McDonald, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research,* 2016.

[24] G. Demiroz, B. Yankikoglu, D. Tapucu and Y. Saygin, "Learning Domain-Specific Polarity Lexicons," in *IEEE 12th International Conference on Data Mining Workshops*, 2012.

[25] P. Malo, S. Ankur, P. Takala, P. Korhonen and J. Wallenius, "Good Debt or Bad Debt: Detecting Semantic Orientations in economic texts," *Journal of the American Society for Information Science and Technology,* vol. 65, no. 4, 2014.

[26] Y. Wu, M. Schuster, Z. Chen, Q. V. Le and M. Norouzi, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," 2016.

[27] J. Turian, L. Ratinov and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," *Association for Computational Linguistics,* 2010.

[28] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[29] T. Mikolov, W.-t. Yih and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[30] Gensim, " Doc2vec paragraph embeddings," [Online]. Available: https://radimrehurek.com/gensim/models/doc2vec.html.

[31] D. M. Blei, "Probabilistic topic models," *Communications of the ACM,* vol. 55, no. 4, p. 77–84, 2012.

[32] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, 2003.

[33] G. Li, X. Zhu, J. Wang, D. Wu and J. Li, "Using LDA Model to Quantify and Visualize Textual Financial Stability Report," *Procedia Computer Science,* vol. 122, 2017.

[34] C. Lin and Y. He, "Joint Sentiment/Topic Model for Sentiment Analysis," *CIKM: Proceedings of the 18th ACM conference on Information and knowledge management,* 2009.

[35] K. Eguchi and V. Lavrenko, "Sentiment Retrieval using Generative Models," *Association for Computational Linguistics,* vol. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006.

[36] T. Rana, Y.-N. Cheah and S. Letchmunan, "Topic Modeling in Sentiment Analysis: A Systematic Review," *Journal of ICT Research and Applications,* vol. 10, 2016.

[37] V. Raju, B. K. Bolla, D. K. Nayak and J. Kh, "Topic Modeling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings," *IEEE,* vol. 7th International conference for Convergence in Technology (I2CT), 2022.

[38] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019.

[39] A. C. Müller and S. Guido, Introduction to Machine Learning with Python, y O'Reilly Media, 2016.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, N. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[41] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, 2021.

[42] D. J. Norris, Machine Learning with the Raspberry Pi, Apress, 2019.

[43] . J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* vol. 29, 2001.

[44] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016.

[45] S. Seabold and P. Josef, "Statsmodels: Econometric and Statistical Modeling," *Proceedings of the 9th Python in Science Conference,* 2010.

[46] M. Fahim Uddin, "Addressing Accuracy Paradox Using Enhanched Weighted Performance Metric in Machine Learning," *HCT Information Technology Trends (ITT),* 2019.

[47] J. C. Hull, Options, Futures, and Other Derivatives, Pearson College Div, 2011.

[48] L. Gaspareniene, R. Remeikiene, A. Sosidko and V. Vebraite, "Modeling of S&P 500 Index Price Based on U.S. Economic Indicators: Machine Learning Approach," *Engineering Economics,* 2021.

[49] Corporate Finance Institute, "Corporate Finance Institute - Adjusted Closing Price," [Online]. Available: https://corporatefinanceinstitute.com/resources/equities/adjusted-closing-price/. [Accessed 28 February 2024].

[50] C. Sanders, "yahoofinancials," 2018. [Online]. Available: https://github.com/JECSand/yahoofinancials. [Accessed 2023].

[51] S. R. Baker, N. Bloom, S. J. Davis, K. Kost, M. Sammon and T. Viratyosin, "The Unprecedented Stock Market Reaction to COVID-19," *The Review of Asset Pricing,* vol. 10, no. 4, 2020.

[52] M. Izzeldin, Y. G. Muradogly, V. Pappas, A. Petropoulou and S. Sivaprasad, "The impact of the Russian-Ukrainian war on global financial markets," *International Review of Financial Analysis,* vol. 87, 2023.

[53] A. Thompson, "components.one," Components, 2020. [Online]. Available: https://components.one/datasets/all-the-news-2-news-articles-dataset/. [Accessed 2023].

[54] D. Araci and Z. Genc, "Hugging Face," Prosus AI, [Online]. Available: https://huggingface.co/ProsusAI/finbert. [Accessed 30 2 2023].

[55] Z. S. Harris, "Distributional Structure," *WORD,* vol. 10, no. 2-3, 1954.

[56] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine,* vol. 18, no. 3, 2018.

[57] N. DeRobertis, "pysentiment2," 2020. [Online]. Available: https://nickderobertis.github.io/pysentiment/. [Accessed July 2023].