Feedback draft

bbaasan

2023-11-7

Contents

1 Abstract

Natural Language Processing (NLP), a field of artificial intelligence, shows great potential in predicting stock market trends. However, applying NLP in this domain can be challenging, as it may capture irrelevant or misleading information and encounter shifting sentiments. These factors, combined with the challenge of making inferences from a relatively smaller dataset, which is common in existing literature, could potentially reduce the reliability of predictions. To address these issues, our research introduces the Hybrid Market Movement (HMM) predictor model.

This model blends traditional stock market analysis (technical analysis) with modern NLP techniques. It analyzes the content and sentiment of earnings calls, which are key discussions companies have with investors about their financial results. By combining these two approaches with a significantly larger dataset, the HMM model has demonstrated impressive results. It accurately predicted stock market trends more than 80 percent of the time, both on training and test data, as well as when applied to new datasets extending beyond the original time frame.

Our study investigates the question: 'How effective are machine learning classifiers in categorizing documents for predicting stock market movements?' To this end, we employ sophisticated machine learning classifiers and compare the results. Additionally, we examine the common limitations of these classifiers in the realm of stock price prediction, especially when analyzing a larger dataset with complex features including earnings calls, to better understand the constraints and potential biases inherent in NLP methodologies within financial contexts. Furthermore, we delve deeper into identifying the most crucial aspects of our model for making accurate predictions. We use the Random Forest classifier to assist us in pinpointing the most vital features in the model. We also conduct permutation testing, a method to validate the significance of these features, and feature importance testing to enhance our understanding of how various elements interact within it, thereby bolstering the reliability of our predictions.

2 Introduction

In the intricate realm of financial markets, the pursuit of informed decisions is paramount. This research embarks on an all-encompassing journey, blending corporate communication, sentiment analysis, topic exploration, and technical indicators to elucidate the connections between information, market sentiment, and price shifts, similar to the studies by [1], [2], [3], [4], [5], [6], [7], and [8]. Central to our investigation are earnings calls, which are widely recognized for their capacity to unveil valuable insights into corporate performance and strategic directions. Originating from corporate boardrooms, these texts furnish a unique vantage point into a company's emotional state, strategies, and market significance. Our model benefits from a significantly larger dataset compared to existing literature, enabling a more comprehensive exploration of these calls and yielding findings that are both more robust and statistically significant. By doing so, we aim to address previous limitations and offer a deeper understanding of the insights that earnings calls can provide, ultimately contributing to a more comprehensive body of knowledge in this field.

Our foremost objective is discerning the correlations between earnings calls, market sentiment, and ensuing price alterations. Earnings calls are recorded conference calls wherein public companies' managements declare and deliberate on quarterly or annual financial results [9]. These calls serve multiple purposes: guiding price recommendations for investors [10], influencing stock price shifts [7], interpreting signals relayed during earnings announcements, and more. In instances, investors assign higher value to novel information in analyst reports when managers might withhold critical data [4]. By meticulously analyzing data, we hoped to strive if positive or negative sentiments in these earning calls, coupled with particular themes, associates with stock price shifts. We are confident that exposing these ties can bestow invaluable predictive instruments upon market players.

Fundamentally, earnings calls can sway stock prices through sentiment. While financial metrics proffer a historical perspective, the qualitative nuances voiced during the call grant context and indicators of potential future performance. For instance, despite a company announcing favorable outcomes, a somber tone or bleak future projections can precipitate a decline in stock price [7]. Infusing sentiment derived from earnings calls into our model introduces an extra dimension of qualitative analysis that is similar to [6]. Earnings calls are

momentous events for publicly traded firms, serving as platforms where executives discuss their financial outcomes, operational advancements, and prospective forecasts. Employing natural language processing (NLP) in text mining aids stakeholders and investors in gleaning information from unstructured data, like earnings calls, to anticipate stock behavior. Generally, scholars leverage publicly accessible data in numerical formats using methods like TF-IDF, Word2Vec, Doc2Vec, and feed them into algorithms such as Random Forest, Support Vector Machine, Logistic Regression, and Recursive Neural Network (RNN).

NLP facilitates tasks like sentiment analysis, crucial in predicting post-call modifications in price targets [10]. Sentiment analysis discerns the emotional undertone of a text fragment, gauging the author's stance toward a topic. Some academic works suggest that the analytically prowess of transcripts exceeds human interpretation of earnings indicators [7]. Furthermore, NLP encompasses features for sentiment analysis vital for anticipating post-call price target shifts. Sentiment analysis, or opinion mining, employs automated tools to detect subjectivity—opinions, attitudes, or feelings—in text [11]. In certain scenarios, sentiments in earnings calls can assist analysts in predicting stock price trajectories with accuracies ranging from 50+% [12] to a staggering 73% [6].

While we recognize the importance of backtesting when evaluating predictive models, it's essential to clarify that our main focus isn't on crafting trading strategies. We understand that earnings call transcripts represent a form of unstructured data, making them susceptible to noise and overfitting. Our goal is to deepen the existing knowledge base, highlighting the intricate ties between corporate communication, sentiment, and market trends. To address these challenges, we are introducing the Hybrid Market Movement (HMM) predictor model. This model employs topic modeling to address issues arising from high dimensionality, overfitting, and noise — characteristics that are often more predictive in the training set but not as much in new data. It integrates this approach with technical indicators, such as daily closing prices, daily moving averages, and market sentiments. Consequently, we delve deeper into the model by comparing its features, employing methods like feature importance and feature permutation to assess the quality of each feature and its contribution to the model. We aimed to answer a question: 'How effective are machine learning classifiers in categorizing documents for predicting stock market movements? and what are the limitations of natural language processing as an analytical instrument.

Subsequent sections will delve into our literature review, research methodologies, data reservoirs, empirical discoveries and discussion and conclusion.

3 Literature Review

A significant body of work has centered on predicting stock market trends using natural language processing (NLP). This interest might be attributed to the observation that stock price movements around earnings announcement events exhibit certain predictable characteristics [7]. These characteristics can bolster an analyst's accuracy in forecasting future events [13], assist in trend analysis [8], and extend the economic implications beyond the predictive capacities of quantitative firm data. Furthermore, employing NLP in conjunction with machine learning techniques offer a promising avenue, as this approach outperforms many traditional methods found in literature reviews [13]. Notably, it can accurately categorize forward-looking statements without necessitating user interaction or extensive tuning.

[8] analyzed transcripts from five years' worth of quarterly earnings calls from 10 NASDAQ-listed public companies. Their research contrasted five different text-based predictive analysis methodologies across two forecasting horizons. Stock performance was gauged by comparing each company's stock price trajectory to the NASDAQ index over an identical period. The research also highlighted a novel sentiment feature extraction technique using the Word2Vec's most_similar function, aiming to derive a list of words potent enough to provide industry-specific insights.

In their study, [10] delved into analysts' decision-making processes regarding the linguistic content of earnings calls. They identified 20 pragmatic attributes within analysts' inquiries and drew correlations with analysts' pre-call investor recommendations. The intention was to ascertain the extent to which semantic and pragmatic elements from an earnings call complement market data in forecasting any post-call modifications to analysts' price targets. Their findings suggested that while earnings calls had a moderate impact on analysts' decision-

making, these decisions were concurrently influenced by other factors, such as confidential discussions with company executives and the overall market environment.

Lastly, [6] found that the sentiment derived from earnings calls could predict stock price movements with a commendable accuracy of 73%. The forecasting strength of earnings call sentiment paralleled that of earnings-per-share (EPS) and revenue surprise metrics, both of which are utilized as stand-ins for earnings call content. By employing a range of methodologies, including OLS regression, the research confirmed the formidable predictive prowess of earnings call sentiment, particularly in gauging stock price fluctuations following an earnings announcement.

3.1 Topics Modeling

In the context of our model, an important aspect is the incorporation of topic modeling, which is an unsupervised machine learning technique, for the purpose of dimensionality reduction. Topic modeling is a type of statistical modeling aimed at uncovering hidden topics or themes within a collection of documents or text data. Functioning as a dynamic belief model, it represents text document collections through approaches such as a corpus or a bag-of-words [14].

As [15] notes, a topic model envisions each document in a collection as a multinomial distribution over topics, with each topic being a multinomial distribution of words. Commonly used in NLP applications, these models (LDA and Latent Semantic Analysis or LSA) assist in word sense disambiguation, text classification, and information retrieval. They offer a method to systematically gauge word similarity, benefiting tasks like document classification or clustering based on word similarity. This form of statistical modeling classifies text within a document to specific topics. The utility of topic modeling as a review technique helps identify and compare machine learning (ML) research trends across business organization verticals [16]. This unsupervised ML method scrutinizes a series of documents to discern patterns of words and phrases, thereby automatically clustering these word groups.

Historically, topic modeling has been a mainstay in computational linguistics, employed to determine sets of words (or "topics") that encapsulate the hidden semantics of a document or a collection of them [15]. It's also been used as a review technique to recognize and contrast ML research trends [5], analyze risk factors in the 10-K financial statement's Section 1A [3], quantify the coherence of document sets [17], gauge semantic meaning in inferred topics [18], refine or partition models [19], and for longitudinal analyses [20].

[4] leveraged topic modeling, a methodology rooted in computational linguistics, to contrast the thematic content of a vast sample of analyst reports with the content of corresponding earnings conference calls. Their findings underscored that analysts frequently broach unique topics not covered in the conference calls and offer interpretations on call content. They also revealed that investors place heightened importance on novel information in analyst reports when corporate managers are more inclined to withhold significant information. Particularly, analyst interpretations become invaluable when the costs associated with processing conference call information rise. Their research underlines the pivotal role of analysts as information intermediaries who unearth insights beyond standard corporate disclosures while also clarifying and authenticating them.

In his research, [19] segmented documents into distinct semantic topic units. His enhanced Latent Dirichlet Allocation (LDA) topic model, based on partitioning (LDAP), is optimized for medium to lengthy texts. The LDAP not only retains the advantages of the original LDA but also improves the granularity from the document to the semantic topic level. Comprehensive tests on the Fudan University and Sougou Lab corpora revealed that LDAP outperformed other topic models like LDA, HDP, LSA, and doc2vec.

The work by [3]' is centered on linguistic changes in financial reports that might predict firms' future returns and operations. By contrasting language in the Question-and-Answer and Presentation sessions of earnings conference calls, he identified tendencies for "lazy prices" — prices that react slowly to information changes. This research bifurcated the dataset to compare language use in different sections, asserting that methods like topic overlap and comparison language channels could diminish the predictive power of textual changes for future returns.

Introduced by [21], the Wu-Palmer Similarity (WPS) is a metric for computing the semantic similarity between words by examining the depth of their closest shared ancestor within a taxonomy [21]. Unlike other

metrics that may emphasize semantic relatedness or the shortest path between concepts [22], [23], WPS gives precedence to word pairs that are closely connected within the hierarchical structure. Complementing this, the Resnik Information Content (IC) approach, proposed by Philip Resnik in 1995, offers another angle by measuring semantic similarity. Resnik's method quantifies the likeness of two words based on the information content shared by their most informative common ancestor (MICA) within a taxonomy or ontology [24]. These methodologies provide nuanced insights into the semantic connections between terms, each from a distinct perspective that highlights different aspects of word similarity.

The work by [11] pioneered a distinct measure of semantic similarity based on information entropy. This method computes the similarity between two words by using the harmonic mean of their Information Content (IC) values within a taxonomy or ontology, such as WordNet. It assigns a higher similarity score to word pairs with higher-IC MICA and a lower score to those with lower-IC MICA. Through this approach, [25] expanded the information theoretic framework in NLP, resulting in a more nuanced measurement of similarity between words.

[26] introduced Jiang-Conrath Similarity (JCN), introduced in 1997. The model estimates the similarity between two words considering the difference in their IC values and the IC of their Least Common Subsumer (LCS) within a taxonomy or ontology. This similarity is defined as 1 / (1 + d), with 'd' being the IC difference between the words and their LCS. This measure attributes a higher similarity score to word pairs with smaller IC differences and a lower score to those with larger differences. [27] did similar work where it revealed that the accuracy of similarity measures with varying back-off strategies ranges from 45 to 60 percent across three corpora. However, within the Semeval dataset, accuracy lingers between 35 and 37 percent, possibly due to the inclusion of only nouns and verbs and the restriction that sentences contain a maximum of 16 words.

Introduced by Michael Lesk in 1986, the Lesk algorithm is a word sense disambiguation (WSD) methodology [28]. It postulates that the most probable sense of a word in a context is the one with the most overlap with surrounding words. The Lesk algorithm operates by juxtaposing the definition of each sense of a word with the words in its context, selecting the sense with the maximum overlap as the likeliest.

[29] proposed a distributional semantics model in 1998. This NLP branch delves into word meanings in context. The foundational idea is that words in similar contexts often bear similar meanings. In his model, words are represented as high-dimensional vectors, each dimension corresponding to a context feature, like neighboring words. The seminal contribution from [29] lies in this model's ability to encapsulate word meanings based on their distribution in vast text corpora.

3.2 Technical Indicators

We recognize the nuanced interplay between textual data and market dynamics. To piece this puzzle together, we immerse ourselves in technical analysis, emphasizing historical prices, trading volumes, and technical indicators consistent with [30], [31], [32], [8], [16], [6], [11],[20]. Fusing quantitative metrics with qualitative insights, we forge an exhaustive system for forecasting price fluctuations.

Technical analysis revolves around the study of historical price patterns and trading volumes in financial markets to forecast future price movements. Unlike fundamental analysis, which examines the intrinsic value of securities, technical analysis is solely anchored on historical trading data [33]. The primary objective of technical analysis is to recognize patterns or consistencies in stock price data. These patterns could manifest as specific formations (e.g., "head and shoulders" or "double bottom"), trends (whether upward, downward, or lateral), or other recurrent dynamics [34]. Discerning these patterns equips traders and analysts with insights for predicting potential future price fluctuations [33].

It is important to note that historically, the use of technical analysis and indicators has yielded mixed results. Specifically, [35] pointed out that returns vary not only from market to market but also are influenced by different time periods. This variation in returns is observable not only in domestic markets but also in foreign exchange markets. The patterns of variation demonstrate that the impact of technical analysis and indicators is not confined to a single type of financial market. Instead, it spans across different markets globally, indicating a widespread, albeit inconsistent, influence of these analytical tools.

Stock price movements often exhibit nonlinear trajectories that can elude conventional linear models. However, nonparametric techniques, like kernel regression, can adept at pinpointing these patterns. They weigh observations based on closeness, eschewing strict data assumptions, and are capable of elucidating intricate nonlinear relationships even in data laden with noise [33]. Although this doesn't imply that technical analysis can consistently yield above-average trading profits, it does suggest the potential value it could bring to the investment process [33].

Among various indicators, the Daily Moving Average (DMA) stands out. It's often seen as a low-risk gauge for transactions since it mirrors the average price traders have paid over a specific duration. For instance, a 50-day moving average encapsulates the average price traders paid for an asset over the last 10 trading weeks, making it a widely acknowledged support level [36]. On the other hand, the 200-day moving average reflects the average price across the previous 40 weeks, hinting at a relatively affordable price when juxtaposed with the range prevalent for the majority of the past year. If prices dip below this average, it could serve as resistance since those who've ventured into the market might contemplate exiting to prevent substantial losses [36]. Additionally, moving averages help filter out price chart noise. The orientation of the moving average provides a rudimentary sense of price direction—angled upward signifies a rise, downward indicates a decline, and sideways suggests a range [37].

Additionally, moving averages can also double up as support or resistance levels. In a bullish phase, metrics like the 50-day or 200-day moving averages can offer support, analogous to a floor from which prices can rebound. Conversely, during a bearish phase, these averages might act as resistance, akin to a ceiling which prices touch before receding [36], [37], [33], [35].

It is our view that once technical indicators are widely used, the information they reveal is likely to diminish in value like it is demonstrated in [35]. This is because the price would already reflect the information, as described by [38]. In essence, the widespread adoption of these indicators might lead to a scenario where their predictive power is reduced, aligning with the Efficient Market Hypothesis, which suggests that current asset prices reflect all available information. It is also important to note that as technology advances, the use of technical indicators continues to reveal more information. This is due to new technologies being introduced into the market, which enhance the capabilities of these indicators to analyze and interpret market data more effectively.

4 Methodology

The aim of our research is to answer the question: 'How effective are machine learning classifiers in categorizing documents for predicting stock market movements?'. This innovative approach to predicting stock price movement is an extension inspired by the works of [12] and [6]. By crafting an ensemble model that deftly combines LDA topic modeling [39], sentiment analysis [40], [41], and technical indicators [42], we are taking a page from their playbook and pushing the boundaries in financial analytics. Their foundational principles serve as a bedrock upon which this comprehensive method stands, allowing for a richer and more nuanced understanding of stock market dynamics. Such a fusion not only pays homage to their contributions but also charts a new path forward in the realm of stock prediction.

4.1 Data Source

Earnings Call
Text data can be sourced to number of ways: Harvard Business School[1] suggests Bloomberg, Factiva, Standard and Poor's Capital IQ. It is also available at Motley[2]. The paper used Seeking Alpha, a crowd-sourced content service that publishes news on financial sectors. Earning Calls transcripts for this research are scraped from the website, using python libraries including BeautifulShop4, requests, and Pandas. Codes can be available in Github repository except those codes used for web scrapping.

Stock Market Data

[1] https://asklib.library.hbs.edu/faq/47473
[2] https://www.fool.com/earnings-call-transcripts/

Close and Volume data: Historical closing price and volume data can be downloaded using yfinance a python library, which offers a reliable method of downloading historical market data from Yahoo! Finance API. It is maintained by [43]. It is open source and free and provides high granularity of data with intervals of "1m, 2m, 5m, 15m, 30m, 60m, 90m, 1h, 1d, 5d, 1wk, 1mo, 3mo".

Ticker information
The data encompasses 6,126 tickers[3], which consist of those listed on National Association of Securities Dealers Automated Quotations Stock Market (NASDAQ) and New York Stock Exchange (NYSE) other symbols. Information regarding NASDAQ listings can be found on their website[4]. Data for NYSE and other listings can be accessed here[5].

4.2 Data Size

The data separated into two sets where the main (see Table 1) is used for modeling (136522 documents out of 199000+) which gathered on 2023-06-08. And other (see Table 2) used as a unseen data set for the model evaluation which consists of (3942 documents out of 4000) gathered on 2023-09-19.

4.3 Data Period

Time period of the earning calls in the main data set spans between 2011-03-31 and 2023-06-08, while unseen data set covers period of 2023-07-28 and 2023-09-19.

	tick	rcount	info	from_date	to_date
0	T	139	['NYSE', 'AT&T Inc.']	2012-10-24	2023-05-22
1	F	127	['NYSE', 'Ford Motor Company Common Stock']	2012-10-25	2023-05-31
2	V	126	['NYSE', 'Visa Inc.']	2012-10-31	2023-05-31
3	CSCO	119	['NASDAQ', 'Cisco Systems, Inc. - Common Stock']	2012-11-14	2023-06-08
4	MA	115	['NYSE', 'Mastercard Incorporated Common Stock']	2012-10-31	2023-06-06
5	MU	113	['NASDAQ', 'Micron Technology, Inc. - Common Stock']	2012-12-20	2023-05-31
6	GILD	112	['NASDAQ', 'Gilead Sciences, Inc. - Common Stock']	2012-10-23	2023-06-08
7	BMY	112	['NYSE', 'Bristol']	2012-10-24	2023-06-06
8	MSFT	109	['NASDAQ', 'Microsoft Corporation - Common Stock']	2012-10-18	2023-06-08
9	REGN	103	['NASDAQ', 'Regeneron Pharmaceuticals, Inc. - Common Stock']	2012-10-24	2023-06-08

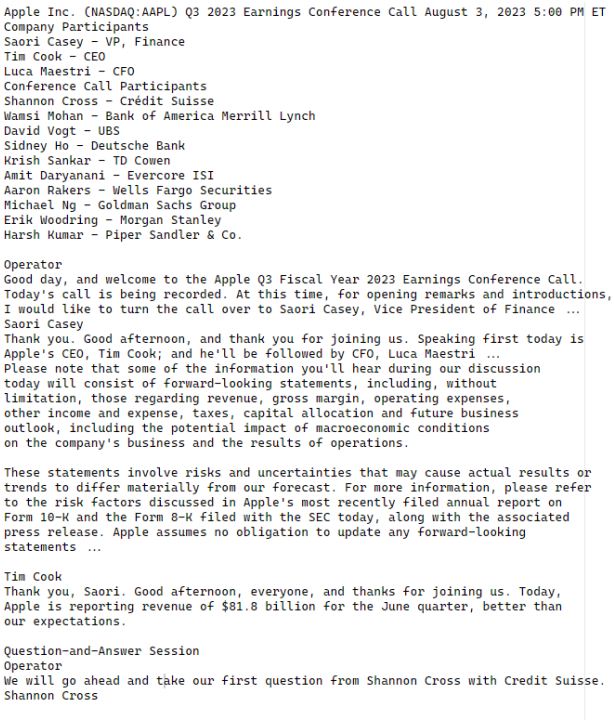
Figure 1: Main dataset: Top 10 Companies with the most Earnings calls.

	tick	rcount	info	from_date	to_date
0	ANET	7	['NASDAQ', 'Arista Networks Inc. Common Stock']	2023-07-31	2023-09-12
1	PFE	5	['NASDAQ', 'Pfizer Inc. Common Stock']	2023-08-01	2023-09-18
2	NVDA	5	['NASDAQ', 'NVIDIA Corporation Common Stock']	2023-08-23	2023-09-11
3	ON	4	['NASDAQ', 'ON Semiconductor Corporation Common Stock']	2023-07-31	2023-09-07
4	WDC	4	['NASDAQ', 'Western Digital Corporation Common Stock']	2023-07-31	2023-09-07

Figure 2: Unseen dataset: Top 5 companies with the most Earning calls

[3]Top 2200+ companies and their information are available here
[4]https://www.nasdaq.com/market-activity/stocks/screener
[5]https://datahub.io/core/nyse-other-listings#resource-nyse-other-listings_zip

- A Screenshot of the Earning Call



```
Apple Inc. (NASDAQ:AAPL) Q3 2023 Earnings Conference Call August 3, 2023 5:00 PM ET
Company Participants
Saori Casey - VP, Finance
Tim Cook - CEO
Luca Maestri - CFO
Conference Call Participants
Shannon Cross - Crédit Suisse
Wamsi Mohan - Bank of America Merrill Lynch
David Vogt - UBS
Sidney Ho - Deutsche Bank
Krish Sankar - TD Cowen
Amit Daryanani - Evercore ISI
Aaron Rakers - Wells Fargo Securities
Michael Ng - Goldman Sachs Group
Erik Woodring - Morgan Stanley
Harsh Kumar - Piper Sandler & Co.

Operator
Good day, and welcome to the Apple Q3 Fiscal Year 2023 Earnings Conference Call.
Today's call is being recorded. At this time, for opening remarks and introductions,
I would like to turn the call over to Saori Casey, Vice President of Finance ...
Saori Casey
Thank you. Good afternoon, and thank you for joining us. Speaking first today is
Apple's CEO, Tim Cook; and he'll be followed by CFO, Luca Maestri ...
Please note that some of the information you'll hear during our discussion
today will consist of forward-looking statements, including, without
limitation, those regarding revenue, gross margin, operating expenses,
other income and expense, taxes, capital allocation and future business
outlook, including the potential impact of macroeconomic conditions
on the company's business and the results of operations.

These statements involve risks and uncertainties that may cause actual results or
trends to differ materially from our forecast. For more information, please refer
to the risk factors discussed in Apple's most recently filed annual report on
Form 10-K and the Form 8-K filed with the SEC today, along with the associated
press release. Apple assumes no obligation to update any forward-looking
statements ...

Tim Cook
Thank you, Saori. Good afternoon, everyone, and thanks for joining us. Today,
Apple is reporting revenue of $81.8 billion for the June quarter, better than
our expectations.

Question-and-Answer Session
Operator
We will go ahead and take our first question from Shannon Cross with Credit Suisse.
Shannon Cross
```

Figure 3: An Example of Earnings Call: First line includes company name, stock exchange where listed, company ticker, date and time when the earning call is recorded. Starting the third line company representatives with name and title, followed by participants in the call. Earning Call usually comprises from sections Presentation and Questions and Answer, where company representatives discuss summary of the financial performance for the period and answers questions. In some cases, it contains only Presentation part.

4.4 Features and their type

Our model is designed with a diverse array of features, each chosen for its potential impact on predicting stock market movements. The core features include the closing price, which serves as a fundamental indicator of market sentiment at the end of each trading day. Additionally, we incorporate two types of Simple Moving Averages (SMAs) – the 50-day SMA and the 200-day SMA – that help in identifying medium and long-term trends and market momentum.

To delve deeper into the qualitative aspects, we analyze ten different topics (labeled Topic 1 to Topic 10) identified through Gensim LDA topic modeling [39]. This approach allows us to capture the thematic structure of financial documents and earnings calls. Alongside, we incorporate sentiment analysis [41], gauging the emotional tone behind the text using a sentiment score. We also employ the VADER (Valence Aware Dictionary and sEntiment Reasoner) score, a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media [40].

The final feature is the label, which categorizes each dataset entry based on our classification criteria. This ensemble of features – from technical indicators to sophisticated NLP tools – equips our model with a comprehensive view, enabling it to effectively analyze and predict stock market trends.

1. Closing Price (float): The closing price of the stock on a given day. It refers to the last price at which a stock trades during a regular trading session.

2. Volume (integer) is the amount of an asset or security that changes hands over some period of time, often over the course of trading day.

3. 50-Day Moving Average of Price (float): A moving average of the stock's closing price over the last 50 days.

4. 200-Day Moving Average of Price (float): A moving average of the stock's closing price over the last 200 days.

5. 50-Day Moving Average in Volume (float): A moving average of the stock's volume over the last 50 days.

6. 200-Day Moving Average in Volume (float): A moving average of the stock's volume over the last 200 days.

7. 7-16 Topics 1-10 (float): These are derived features from earnings call transcripts using the Gensim topic modeling algorithm [39] available in the Python library. It is one a few popular tools for topic modeling and is widely used for implementing techniques like LDA. Topics represent themes or subjects found within the text data, and an adequate selection of model parameters is crucial [44]. The proper number of topics in a topic model can be evaluated based on coherence and perplexity scores. By examining both metrics, one can more effectively determine the appropriate number of topics for the model (see Appendix E: coherence and perplexity scores).

8. 17-19 Loughran-McDonald Sentiment (integer): Sentiment scores calculated using the Loughran-McDonald financial sentiment word lists developed by Tim Loughran and Bill McDonald of Notre Dame in 2011 [41]. It is a dictionary which tailored for financial text analysis. It is one of the most popular financial lexicon available [45]. Lexicon, often called the "bag of words" approach to NLP use a dictionary of words or phrases that are labelled with sentiment. The motivation for the LM dictionary was that in testing the popular and widely used Harvard Dictionary, negative sentiment was regularly mislabeled for words when applied in the financial context. The LM dictionary was built from a large sample of 10 Q's and 10 K's from the years 1994 to 2008. The sentiment was trained on approximately 5,000 words from these documents and over 80,000 words from the Harvard Dictionary [45].

9. 20-22. VADER Sentiment (float): Sentiment scores calculated using the VADER sentiment analysis tool, which measures sentiment (positive, negative, neutral) in text data. The work is credited to [40] and its python library[6] is available at the footnote. It a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media[40].

10. Label: it is binary label where value takes either 1 or 0, which is consistent with [42], [1]. The label is based on the "Golden Cross" and "Death Cross" conditions in stock trading [42], specifically the crossover of the 50-day and 200-day moving averages. In this context:

    - 1 indicates a "Golden Cross" where the short-term moving average (DMA50) crosses **above** the long-term moving average (DMA200). This is often seen as a bullish signal.

    - 0 indicates a "Death Cross" where the short-term moving average (DMA50) drops **below** the long-term moving average (DMA200). This is often seen as a bearish signal.

[6] https://pypi.org/project/vaderSentiment/

4.5 Process Flow

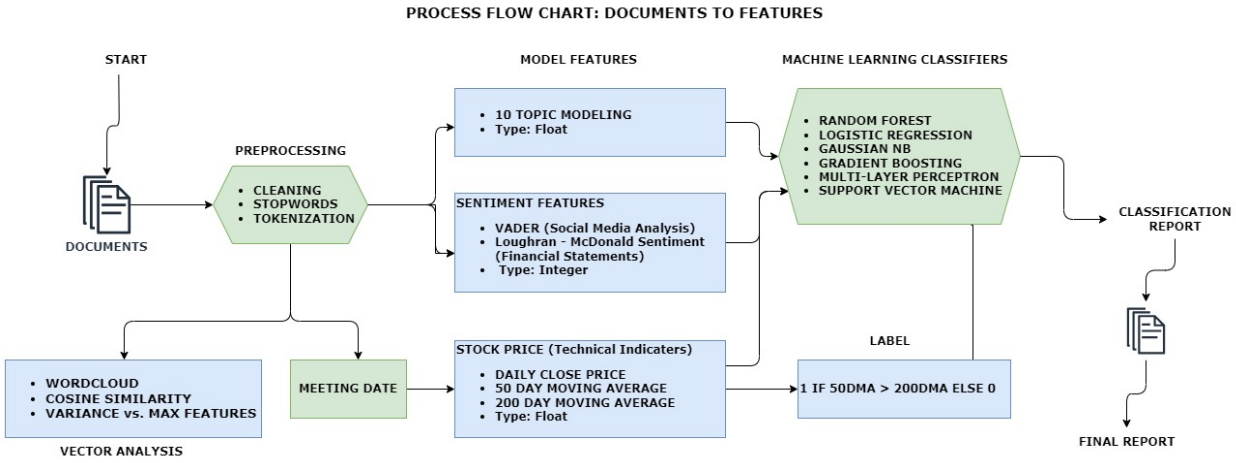


Figure 4: Overview Flow Chart

In the structure, the paper presents an ambitious endeavor to forecast stock price movements. This forecasting hinges on an amalgamation of several factors: technical indicators closing price and moving averages, the themes broached during earnings calls, and market sentiment as deciphered from the earning calls. On the other hand, lower left corner of the flowchart includes features like word clouds, cosine similarity, and variance versus max features. These elements are not directly involved in the model; however, they play a crucial role in understanding the data and selecting feature reduction methods. For example, a word cloud demonstrates word frequency, while cosine similarity, a widely used metric in information retrieval[46], measures the similarity between two documents term vectors[47], [48]. For technical review [46] is highly recommended while earlier work review [47]. With respect to Explained variance and the maximum number of features help determine the number of features that can be selected (Review Appendix 9.2 for detailed explanation). It is important to note that these elements are part of the initial stages of the project data analysis and are not directly part of the actual modeling process.

To bring this vision to fruition, the paper employs a plethora of sophisticated machine learning techniques. This diverse range includes the Random Forest[49], Logistic Regression, Gaussian Naive Bayes, and Gradient Boosting. Additionally, it explores the potential of the Multi-Layer Perceptron, the Support Vector Machine (SVM), and Neural Network architectures. By deploying such a vast array of algorithms, the research seeks to construct a robust classification model that can effectively gauge the nuances of stock market movements.

5 Results

5.1 Topics Modeling

Using the topic modeling tool from Gensim, we've been able to sort out the main themes from a collection of earnings call transcripts. We looked at the data from 2011 to 2023, including some new data from July to September 2023, to see which words pop up most in Topic 1. The dominant appearance of the term 'market' in our data indicates its significance in corporate discussions. It could reflect how companies strategize, position themselves, and adapt to market dynamics. Topics related to market positioning, competitive analysis, market conditions as well as the corporate communications are prevalent, underscoring their relevance to investors and analysts. However, to gain a deeper understanding of these discussions and their impact on corporate decision-making, further in-depth and comprehensive analysis is warranted. It is because firms with different market capitalization employ different strategy [50]. This result could serve as the basis for a separate research endeavor or paper focusing on the complexities of market-related discourse in earnings calls.

Training and Testing Data[7]

Topic 1 Words:

0.033*market + 0.014*slide + 0.012*performance + 0.011*costs + 0.011*prices +

0.009*ebitda + 0.008*indiscernible + 0.008*increased + 0.007*negative + 0.007*volumes

On the other hand, when we examine Topic 1 within the unseen data, spanning from July to September 2023, it encompasses a distinctly different set of terms. Words like "cancer," "program," "disease," "studies," "therapy," and "vaccine" stand out, indicating a significant shift in the focus of the discussions. This variation suggests that the conversations in the more recent earnings calls could have pivoted towards healthcare-related themes. Understanding this shift is crucial, as it may reflect changes in company priorities, industry trends, or global health events that have become more pressing or relevant in that period.

Unseen Data[8]

Topic 1 Words:

0.015*cancer + 0.010*program + 0.010*disease + 0.009*studies + 0.008*therapy +

0.007*programs +0.007*vaccine + 0.006*financial + 0.006*efficacy + 0.006*research

Thus, the analysis of our datasets reveals the intricate tapestry of industry-specific dialogues captured through earnings call transcripts. The diversity of terms associated with Topic 1 uniquely underscores thematic elements inherent to different sectors. Notably, the emergence of specialized terms in the unseen data highlights evolving industry dialogues, particularly within the healthcare sector. These insights demonstrate the proficiency of our model in navigating the complex lexicon of corporate communications and underscore the significance of contextual nuances in understanding the multifaceted nature of earnings calls. Such sector-specific linguistic patterns offer a window into the evolving priorities and strategic focuses across the corporate spectrum.

5.2 Classifier performances

The following figure displays the performance of various classification models that were used to predict stock price movements using features derived from earnings call data. It outlines the precision, recall, and F1 score for each model, which are crucial indicators of accuracy under default settings of the classifiers, without any hyperparameter tuning. The F1 score serves as a balanced measure of model accuracy, as it is the harmonic mean of precision and recall, with its most favorable score at 1 and least at 0, thereby reflecting a balance between precision and recall in the assessment.

5.3 Feature Analysis

The idea of feature importance[9] (column 1, Figure 6.) comes from the notion of how useful or valuable each feature contributes to the construction of the decision tree within the model, in this case, Random Forest[51]. We expect higher the importance, the more influential the feature making the prediction. In Random Forests, for example, it is often computed from the average depth at which the feature splits data across trees. Given that the sentiment's importance score is between 3% to 4%, it suggests that while earnings call sentiment plays a role in price prediction, it's supplemental to the more dominant predictors like the daily close, 50DMA, and 200DMA. However, a 3-5% influence in a domain as complex as stock prediction is non-trivial. It might be the edge that sets this model apart, especially if the model is used in conjunction with other analyses. On the other hand, daily moving average of 200 business day is the highest among the features, suggesting that substantial significance in predicting the target. It's high importance indicates the 200 days average plays a pivotal role in the decision-making process of the model.

[7](see Appendix B for full top 10 topics Or visit here

[8](see Appendix C for full Top 10 topics. Or visit here }

[9]https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

	Algorithm	Movement	Precision	Recall	F1-Score
0	Random Forest	0	0.9	0.72	0.8
1	Random Forest	1	0.82	0.94	0.88
2	Logistics Regression	0	0.5	0.16	0.24
3	Logistics Regression	1	0.6	0.88	0.71
4	Gaussian Naïve Bayes	0	0.68	0.99	0.8
5	Gaussian Naïve Bayes	1	0.26	0.01	0.01
6	Gradient Boosting	0	0.96	0.79	0.87
7	Gradient Boosting	1	0.68	0.93	0.79
8	Multi-Layer Perceptron	0	0.74	0.27	0.4
9	Multi-Layer Perceptron	1	0.61	0.81	0.7
10	Support Vector Machine	0	0.71	0.47	0.57
11	Support Vector Machine	1	0.35	0.6	0.44
12	Neural Network	0	0.76	0.22	0.34
13	Neural Network	1	0.34	0.86	0.49

Figure 5: Gradient Boosting and Random Forest are two top performing classifiers on the main data set and able to predict 80+ in F1-Score. They followed by Multi-Layer Perceptron.

	Algorihm	Movement	Precision	Recall	F1-Score
0	Random Forest	0	0.85	0.76	0.81
1	Random Forest	1	0.8	0.87	0.83
2	Gradient Boosting	0	0.96	0.8	0.88
3	Gradient Boosting	1	0.69	0.94	0.8

Figure 6: Random Forest and Graident Boosting classifiers performance on the Unseen data. We notice that F1-score indicates that both classifiers dropped their scores in predicting price movements but less than 5 point scores.

	FeatureName	FeatureImportance	FeaturePermutationMean	FeaturePermutationSD
0	close	0.105152	0.273356	0.00280756
1	volume	0.0308552	-0.00976122	0.000895997
2	sma50_close	0.131209	0.307675	0.00308581
3	sma200_close	0.197274	0.240238	0.00209998
4	sma50_vol	0.0348396	-0.0105767	0.00125179
5	sma200_vol	0.0326016	-0.0133551	0.000846967
6	Topic_1	0.0319111	-0.00460797	0.00111746
7	Topic_2	0.028903	-0.00378599	0.000736541
8	Topic_3	0.0315252	-0.0058857	0.00067175
9	Topic_4	0.028879	-0.00284845	0.000647483
10	Topic_5	0.0281301	-0.00341162	0.000797671
11	Topic_6	0.0303102	-0.00422058	0.000813062
12	Topic_7	0.0302171	-0.00384784	0.000843266
13	Topic_8	0.0289777	-0.00240083	0.000796809
14	Topic_9	0.0318232	-0.00264173	0.000757775
15	Topic_10	0.0303106	-0.00298354	0.000633336
16	LM_Positive	0.0260341	-0.002575	0.000718205
17	LM_Negative	0.0293785	0.00129889	0.000850231
18	LM_Uncertainty	0.0254051	-0.000763384	0.000768723
19	neg	0.0314476	-0.00107915	0.000982885
20	neu	0.0263183	-0.00437359	0.000776869
21	pos	0.0284979	-0.00573596	0.000916465

Figure 7: Feature Analysis: Importance, Permutations Mean and Standard Deviation.

Feature permutation mean and standard deviation show that when a feature randomized while holding others constant, model's performance drop in average and consistency of the drop performance across different permutations, in this case, 30 different runs. For example, permutation mean and standard deviation of daily moving average of 50 days are 0.31 and 0.0031. It can be interpreted as when 50 days moving average is permuted or randomized or shuffled, the model performance drops by an average of approximately 0.31, signifying that the model is heavily relied on the feature to make accurate predictions. Therefore, standard deviation value represents that the decline in performance is consistent and not to due to random chance.

One of the interesting observation is positive sign of mean value of the Loughran - McDonald sentiment Negative score in (Feature Analysis[LM_Negative, FeaturePermutaionMean] in Figure 5). As the sign indicates that the permutation of the feature is the only feature other than technical indicators follows the direction the model expected during the feature permutation run. On the other hand, the negative sign is indication of model performance improvement when other features are holding constant. However, the way to see is that is not so much improving, specially when we are comparing the performance null, which is un-randomized or have done no changes to the features. It is because during the permutation, we are expecting the performance to drop, since we are altering the original information and the model is no longer has access to the correct information from that feature. When the performance improves, we raise some suspicion that the feature might be noisy and misleading. It might mean that in its original form the feature was adding some noise, making the model over-fitting to some noise or patterns in that particular feature. When it was randomized, model is effectively avoiding the over-fitting leading to better performance. However, in this case, range between 0.000 and 0.005 may not be as significant. Thus, some considerations are understanding the complexity of the text and domain specific knowledge.

It is noteworthy that the model's performance tends to decline when both the features: volume and market sentiment, move in tandem. This trend could potentially be attributed to the linkage between market sentiment and trading volume, as indicated by studies like [52]. The trading volume in the stock market, driven by specific company news, economic declarations, or overarching market updates, often swells as investors adjust to fresh insights. The magnitude of resultant price fluctuations, however, hinges on the overall effect of this news and the extent to which it was previously factored into prices. Contrarily, volume simply quantifies the shares exchanged within a given period, and the price depicts the worth at which these dealings transpire. Hence, consistent trading at a stable price can bolster volume without necessarily altering price volatility.

6 Discussion

The findings outlined above align with the conclusions of other studies, such as those by [6], [3], [2]. Furthermore, there are significant discussions that future research should consider, which include the following points.

6.1 Fixed Vocabulary and Topic Consistency

In the domain of topic modeling, ensuring consistent topic extraction across multiple documents is paramount. One approach that promises greater consistency is the utilization of a fixed vocabulary. Ideally, this vocabulary is derived from a combination of training datasets or, in some cases, curated from external sources[41]. The benefit of such an approach is that it ensures a consistent set of terms for the topic modeling process, irrespective of the specific data run. However, while this strategy stabilizes the term set, it introduces another layer of complexity: the determination of an optimal number of topics between data sets. Determining the right number is not straightforward and requires careful consideration to ensure that the underlying themes within the data are adequately captured(See Appendix B: Explained Variance vs. Max Features). A method attempted for this purpose involved examining the explained variance as a function of the number of features. This approach, although promising, requires further scrutiny and validation to ascertain its efficiency in capturing the overarching themes of the data.

Contrarily, relying on static, predetermined topics can sometimes mean overlooking the evolving and dynamic nature of data. Especially in rapidly changing fields such as company news and market sentiment on a specific

company, where the ebb and flow of events, opinions, and revelations continually reshape the discourse. In these arenas, what was relevant yesterday might not hold the same weight today. For instance, a breakthrough product announcement or an unexpected financial setback can drastically alter the narrative around a company. If we remain tethered to a fixed set of topics, we risk missing out on capturing these pivotal moments, thus rendering our analyses less accurate or even obsolete. Moreover, as the volume of documents increases over time(see Figure 8, Appendix 7.0.1: Earning Call aggregated monthly), the limitations of fixed topics become even more pronounced (see Figure 9, and 10, Appendix B. Feature Selection). They may fail to capture the new nuances introduced by the sheer influx of data, and the evolving themes that emerge with time. A fixed topic model, though consistent, may struggle to stay relevant, demanding constant re-evaluation to ensure its alignment with the present state of affairs.

6.2 Look-ahead Bias

In time series analysis and related domains, particularly finance, the concept of "look-ahead bias" is paramount to the integrity of modeling and predictions. Look-ahead bias is introduced when a model inadvertently incorporates information that originates from future data points, thereby skewing its predictive capabilities. The implications of this bias are multifaceted. Firstly, it compromises the analysis's validity, as using future data in retrospective predictions artificially inflates a model's perceived effectiveness. This often manifests as overfitting, where a model seemingly performs exceptionally on training data, given its access to future insights, but displays poor generalization to new, unseen data. Furthermore, the results derived from a model tainted with look-ahead bias can be deceptively precise, leading to unwarranted confidence in its forecasts. Such misleading precision can catalyze strategic and financial miscalculations when applied to real-world scenarios. To mitigate this bias, analysts must exercise rigorous data management, ensuring models are isolated from future data during training and validation. Techniques like rolling forecasting and time-series cross-validation can be instrumental in preserving the temporal sanctity of data, thereby safeguarding models from the perils of look-ahead bias. In essence, ensuring the absence of look-ahead bias is not just a methodological obligation, but a cornerstone of ethical and accurate predictive modeling.

6.3 Feature Engineering Bias

One of the challenges emphasized is inherently tied to the specific characteristics of the data. We turned to topic modeling techniques, specifically to extract discernible topics from the vast textual data. However, a distinct complication can arise once this data undergoes processing: merging the X_train and X_val datasets becomes an impractical endeavor. The core of this issue lies in the dynamics of topic modeling. The contribution of individual terms to the broader topics is not merely a matter of their presence but is intrinsically linked to the overall volume of the text being analyzed. When contemplating entirely separate, unseen data, the true challenge emerges: ensuring that topics identified from the training phase resonate consistently. It's crucial to understand that this unseen data has been processed independently, with its topic modeling outcomes being entirely isolated from the main dataset. This scenario accentuates the importance of a topic modeling approach that's both robust and broadly applicable. It is not enough for the model to be finely tuned to the nuances of the training data; it should ideally be designed to generalize effectively across different textual landscapes. The consistency in the model's performance on new unseen data, particularly the balance between recall and precision, aligns with the trends we identified during its tuning on the training and validation sets. We believe this consistency is encouraging. It hints at the model's ability to generalize effectively, even in the face of the unique time series intricacies in the training and validation datasets. More optimistically, the minimal impact of these time series quirks on the model's performance with the new data underscores its resilience. This suggests that the model is not merely mimicking specific patterns from the training set but is genuinely discerning the foundational trends crucial for accurate predictions. In the realm of time series data, a model's capability to accurately predict future observations is paramount. The consistent performance of the model across familiar and unfamiliar data indicates its adeptness at identifying the core characteristics of the data, sidestepping potential pitfalls associated with time-bound peculiarities.

7 Limitations

This study encounters several significant limitations, high dimensionality, tedious features selection process, the application of topic modeling to longitudinal analyses. These limitations become particularly pronounced as the size of the documents in our dataset grows. When dealing with lengthy documents, such as comprehensive earnings call transcripts spanning multiple quarters or years, the challenges related to high-frequency words and semantic overlap become exacerbated. As the volume of text data increases, high-frequency terms can dominate topics, skewing the representation of underlying themes. Moreover, longer documents often involve more diverse vocabulary and complex semantic structures, leading to topics with greater potential for overlap.

Additionally, the application of topic modeling introduces questions and considerations that pertain to both static and cross sectional analyses. Specifically, while applying topic modeling to a specific time period for cross-sectional analysis provides valuable insights into the content within that time frame, it raises questions when considered alongside time series analysis. The cross-sectional approach offers a snapshot of topics and themes at a single point in time, but when dealing with a time series, the dynamic nature of topics over multiple time points comes into play. Questions may arise regarding the continuity or evolution of topics across different time slices, as well as how external events or trends influence these dynamics. Balancing the cross-sectional perspective with a time series analysis requires careful consideration of both static and temporal aspects, aiming to provide a comprehensive understanding of topic evolution and trends over time.

One of the main challenges in our analysis lies in determining which machine learning algorithms are most suitable for our modeling purposes. We have explored a range of classifiers, each designed to address specific problems and scenarios. For instance, the support vector machine (SVM) classifier, as employed by [13], is a commonly used approach for text classification, particularly in cases with smaller sample sizes, such as analyzing three documents from three distinct periods. However, it's important to note that the computational demands of the SVM classifier can become increasingly burdensome as the scale of the data expands.

The logistic regression classifier, a method akin to the approach used by [6], is applicable in conjunction with the Loughran-McDonald sentiment dictionary. However, as we scaled our analysis to a larger dataset and introduced additional data types, such as continuous values, we encountered a notable drop in classifier performance.

Finally, we explored the application of a neural networks classifier, which is often considered an obvious choice for classification tasks due to its capacity to model complex relationships within data. However, during our analysis, we observed a drop in performance when employing neural networks. This drop was primarily attributed to the neural network's inherent demand for a significantly larger dataset and a high degree of normalization compared to most other classifiers. Neural networks, with their deep and intricate architectures, thrive when provided with extensive data and carefully normalized inputs. In our case, despite their potential, the size of our dataset and the diversity of data types introduced challenges. The demand for a more extensive dataset and high degree of normalization to train neural networks effectively became evident.

8 Conclusion

In conclusion, while natural language processing (NLP) shows potential for predicting stock market trends for various companies, it can also unintentionally introduce noise that compromises the model's ability to generalize. Our study contributes to the field by introducing the Hybrid Market Movement (HMM) predictor model, a novel approach that integrates technical analysis with insights gleaned from topic modeling and sentiment analysis of earnings call transcripts. The model demonstrates strong performance, with F1-scores surpassing 80 percent across training and testing datasets, and maintaining this level of accuracy with new, unseen data. Additionally, our research delves into the relationship and influence of different features on the model's predictions, using Random Forest to assess feature importance and permutation testing to validate these findings.

Looking ahead, the potential enhancements to our approach are promising, particularly with the integration of rolling window techniques in topic modeling. By implementing a rolling window analysis, future iterations of the Hybrid Market Movement (HMM) model could capture temporal shifts in market sentiment and topic relevance more accurately, reflecting the evolving nature of financial discourse over time. This would allow the model to dynamically adjust to recent trends, providing a more nuanced and timely prediction of stock market movements. The incorporation of such temporal methodologies stands to further refine the predictive accuracy of NLP applications in financial analysis, marking a significant stride forward for real-time market analytics.

9 Appendix

9.1 Appendix A: Earnings Call Time Period, 2011 and 2023 (main data)

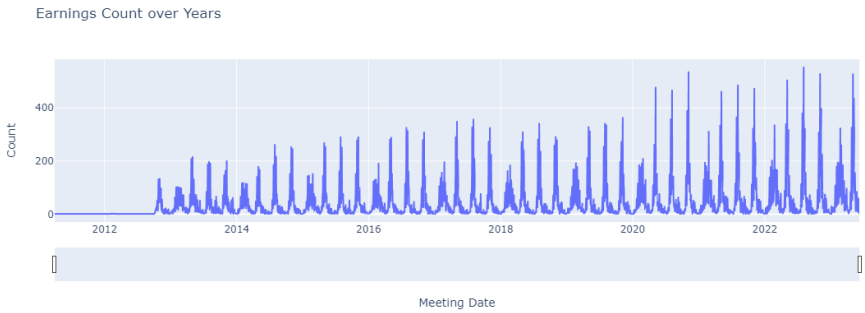


Figure 8: Earning Call aggregated monthly

9.2 Appendix B: Feature Selection - Explained Variances vs. Max Features

Using explained variance function as a criterion for selecting the appropriate number of max features is helpful for classifiers to work effectively. In general,explained variance measures the proportion of target variable variance that is 'explained' by the features used in the model. The max features is a parameter controls number of features to consider when looking for the best split at each tree node. More documentation are available at scikit-learn.
The graphs show that our features and explain variability in our target model. The inverse relations of following two graphs tell us as increase of the max features is not significantly improving the model. Therefore, as it converges to the mean of the dependent variable, the features become less capable of capturing the variations. Thus, additional features contribute little to explained variance and therefore the model might be struggling to use given features for accurate predictions.

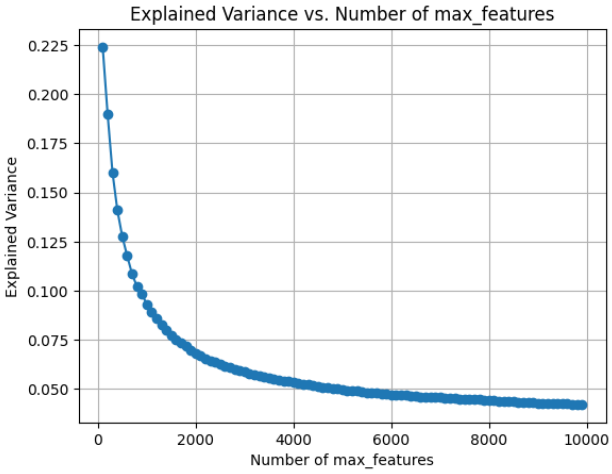


Figure 9: Explained Varience vs. Max Number of Features, Preliminiary data set 1: During the initial period doing an analysis on smaller datasetit was helpful. This dataset comprises of first 5000 rows of documents of the the main dataset. The inverse relation between variables shows first 2000 features explains most variations in the documents and quickly fade away after, showing it would be reasonable to choose max features somewhere between 1700 and 2000.

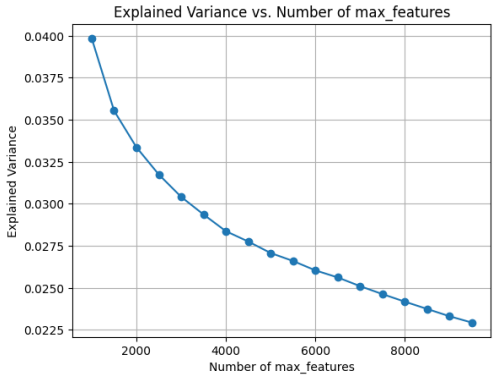


Figure 10: Explained Varience vs. Max Number of Features, Main data: As we were adding more features to the bigger dataset from earnings calls, we noticed a proportional rise in explained variance. However, the new features seemed to add little substantive information, indicating that the addition could be noise. When features are derived from data that is noisy or irrelevant, they might create the illusion of enhancing the model by artificially inflating the explained variance. This leads to a situation where each new feature appears crucial, but in reality, may not be. With an increasing number of features, the model risks adjusting to this noise, focusing on peculiarities in the training data, which can result in overfitting. This becomes particularly concerning when the increase in explained variance fails to correspond to improved predictive accuracy on new, unseen data. The objective is to develop a model that captures the true variance in the current dataset and maintains accuracy and consistency when applied to new data.

9.3 Appendix C: Feature Selection: Topics Modeling - Coherence and Perplexity Scores

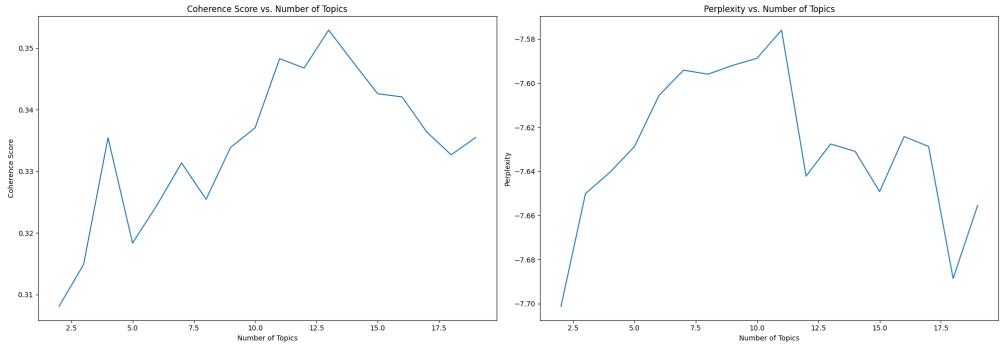


Figure 11: Coherence and Perplexity vs. Number of Topics

- Coherence
  It measures the degree of semantic similarity between high scoring words within each topic, which is intended to reflect how interpretable and meaningful topics are to humans. In general corresponds to topics that are more understandable and coherent to human readers. It evaluates the quality of the topics generated by considering the pairwise similarity between words within a topic. Though can vary based on the metric used (e.g., UMass, C_v), but generally, higher values indicate more coherent and interpretable topics.
- Perplexity
  It is a statistical measure of how well a probability model predicts a sample, and in the context of topic models, it evaluates how well the model describes the distribution of words across documents.Lower perplexity indicates that the probability distribution of the model is closer to the true distribution of words in the documents. It's often used as an indicator of how well the model will generalize to unseen

data. Lower values suggest better generalization. While it's a common metric, perplexity doesn't always correlate with human judgment of topic quality and interpretability. It also tends to favor models with more topics.

9.4 Appendix D: Cosine Similarity

The cosine_similarity function in scikit-learn package computes the similarity between two vectors(or documents), and returns a NumPy array of floating-point values. It is a measure of the cosine of the angle between two vectors in a high dimensional space. When the angle between two vectors is small (close to 0 degrees), their cosine similarity score will be close to 1, indicating a high degree of similarity and vice verso.

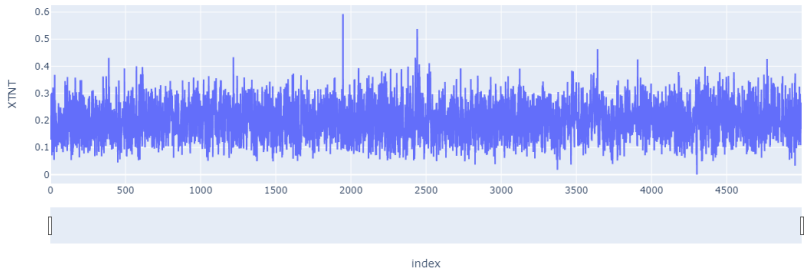


Figure 12: Cosine Similarity Analysis, First 5000 Documents

9.5 Appendix E: Topics Modeling Main dataset

Topic: 1 Words: 0.033*"market" + 0.014*"slide" + 0.012*"performance" + 0.011*"costs" + 0.011*"prices" + 0.009*"ebitda" + 0.008*"indiscernible" + 0.008*"increased" + 0.007*"negative" + 0.007*"volumes"

Topic: 2 Words: 0.021*"revenue" + 0.018*"market" + 0.018*"product" + 0.016*"health" + 0.014*"products" + 0.010*"medical" + 0.009*"financial" + 0.006*"operating" + 0.006*"approximately" + 0.006*"system"

Topic: 3 Words: 0.023*"market" + 0.017*"credit" + 0.015*"investment" + 0.013*"rates" + 0.012*"financial" + 0.011*"assets" + 0.011*"management" + 0.009*"increased" + 0.009*"expense" + 0.008*"expenses"

Topic: 4 Words: 0.037*"revenue" + 0.012*"revenues" + 0.011*"financial" + 0.010*"ebitda" + 0.010*"marketing" + 0.009*"advertising" + 0.009*"increased" + 0.009*"market" + 0.008*"operating" + 0.008*"adjusted"

Topic: 5 Words: 0.012*"program" + 0.010*"disease" + 0.010*"cancer" + 0.009*"studies" + 0.009*"product" + 0.008*"therapy" + 0.007*"financial" + 0.007*"programs" + 0.007*"market" + 0.006*"approval"

Topic: 6 Words: 0.015*"projects" + 0.014*"project" + 0.014*"market" + 0.011*"assets" + 0.009*"financial" + 0.008*"operating" + 0.008*"approximately" + 0.008*"opportunities" + 0.007*"costs" + 0.007*"slide"

Topic: 7 Words: 0.020*"store" + 0.013*"consumer" + 0.013*"product" + 0.012*"customers" + 0.011*"market" + 0.010*"customer" + 0.009*"performance" + 0.009*"operating" + 0.009*"increased" + 0.008*"marketing"

Topic: 8 Words: 0.028*"market" + 0.025*"customers" + 0.024*"revenue" + 0.021*"product" + 0.017*"products" + 0.015*"technology" + 0.014*"customer" + 0.009*"operating" + 0.009*"industry" + 0.008*"financial"

Topic: 9 Words: 0.018*"market" + 0.017*"operating" + 0.017*"adjusted" + 0.014*"research" + 0.014*"margins" + 0.012*"increased" + 0.012*"revenue" + 0.011*"ebitda" + 0.011*"performance" + 0.010*"products"

Topic: 10 Words: 0.041*"revenue" + 0.033*"customers" + 0.021*"services" + 0.020*"customer" + 0.015*"market" + 0.013*"service" + 0.011*"clients" + 0.011*"solutions" + 0.010*"financial" + 0.010*"product"

9.6 Appendix F: Topics Modeling Unseen dataset

Topic: 1 Words: 0.015*"cancer" + 0.010*"program" + 0.010*"disease" + 0.009*"studies" + 0.008*"therapy" + 0.007*"programs" + 0.007*"vaccine" + 0.006*"financial" + 0.006*"efficacy" + 0.006*"research"

Topic: 2 Words: 0.041*"customers" + 0.022*"customer" + 0.020*"market" + 0.015*"revenue" + 0.014*"product" + 0.012*"enterprise" + 0.010*"opportunity" + 0.010*"network" + 0.010*"technology" + 0.010*"products"

Topic: 3 Words: 0.025*"market" + 0.013*"rates" + 0.012*"financial" + 0.012*"credit" + 0.012*"investment" + 0.010*"assets" + 0.009*"increased" + 0.009*"management" + 0.009*"performance" + 0.007*"environment"

Topic: 4 Words: 0.017*"market" + 0.014*"projects" + 0.012*"project" + 0.010*"prices" + 0.009*"financial" + 0.009*"ebitda" + 0.008*"costs" + 0.008*"customers" + 0.008*"opportunities" + 0.008*"assets"

Topic: 5 Words: 0.017*"revenue" + 0.013*"market" + 0.012*"product" + 0.011*"consumer" + 0.011*"marketing" + 0.011*"ebitda" + 0.010*"adjusted" + 0.010*"performance" + 0.009*"operating" + 0.009*"customers"

Topic: 6 Words: 0.040*"revenue" + 0.020*"customers" + 0.014*"market" + 0.012*"financial" + 0.012*"adjusted" + 0.012*"customer" + 0.011*"operating" + 0.010*"ebitda" + 0.010*"product" + 0.009*"solutions"

Topic: 7 Words: 0.033*"health" + 0.012*"market" + 0.012*"revenue" + 0.009*"medical" + 0.009*"members" + 0.008*"medicare" + 0.008*"ebitda" + 0.007*"adjusted" + 0.007*"financial" + 0.006*"costs"

Topic: 8 Words: 0.017*"slide" + 0.013*"project" + 0.012*"costs" + 0.010*"ebitda" + 0.010*"market" + 0.010*"financial" + 0.010*"performance" + 0.009*"prices" + 0.009*"projects" + 0.008*"operations"

Topic: 9 Words: 0.017*"product" + 0.015*"market" + 0.011*"revenue" + 0.009*"financial" + 0.009*"products" + 0.008*"approval" + 0.007*"medical" + 0.007*"therapy" + 0.007*"program" + 0.007*"expenses"

Topic: 10 Words: 0.037*"market" + 0.016*"products" + 0.011*"product" + 0.011*"margins" + 0.011*"operating" + 0.011*"increased" + 0.010*"slide" + 0.009*"backlog" + 0.009*"rates" + 0.009*"performance"

9.7 Appendix G: Stopwords from the Earning Call

inc, officer, conference, this, call, company, participants, with, office, presentation, operator, what, about, everyone, could, although, instructions, question, com, name, video, please, know, actual, looking, afternoon, help, available, between, two, passed, actually, another, five, six, eight, ten, three, twenty, second, third, together, already, including, very, stood, includes, important, partly, months, hereof, website, ladies, gentleman, considers, ended, joined, some, introduce, publ, couple, mentioned, things, altogether, always, thing, most, roughly, wanted, months, certainly, therefore, roughly, certainly, something, explain, longer, address, included, wondering, really, rapidly, month, throughout, everything, some, versus, participation, update, continue, absolutely, expect, different, ongoing, approzimately, continuing, talked, relates, related, inside, continue, appreciation, depends, ultimately, watching, additionaly, least, getting, benfited, compared, want, helps, help, paraphrase, accompany, terms, nothing, started, least, obviously, absolutely, conclude, conclusion, conlcudes, terms, term, helping, issue, activity, students, continuously, result, results, resulting, currently, partially, unable, solid, objectives, companies, conlcude, differnt, difference, group, want, forefront, happen, recieved, behavior, almost, reference, references, compared, continue, earlier, early, started, happening, eats, character, gradually, documents, document, brief, followed, follow, contains, executive, remarks, remark, placed, annually, annual, several, whatever, thinking, think, usually, usual, relatively, seventeen, pleased, brought, bring, showed, squeeze, relative, impressive, oversee, clarify, clarifying, several, considered, consider, considering, unidentified, identify, understood, understand, continually, previously, percent, space, visit, experince, largely, betteer, well, several, refer, necessarily, necessary, existing, explain, explains, explained, except, behalf, mentioning, figures, figure, closest, close, substantially, substantial, examples, represent, represents, primary, primarily, begining, begin, begins, fifty, thirty, fourty, analyze, analyzing, continued, none, example, examples, consideration, amount, separate, worked, order, adding, division, efforts, effort, naturally, natural, anything, total, addition, hundred, additions, provide, provides, thoughts, sized, size, sports, players, order, thereafter, therfore, course, opened, preliminary, starting, looks, look, beginning, dramatically, dramatic, repeat, located, locate, urgency, urgent, seventy, eighty, ninty, concludes, chair, holdings, intend, described, billion, million, relations, relation, locally, local, reminder, following, afterwards, report, publication, records, recording, record, analylist, mixed, required, require, appears, appeared, paragraph, remind, virtually, biggest, big, largest, building, without, requires, whereas, allows, allow, source, regard, statement, introduction, section, executives, news, herein, beliefs, www, succesfully, succesful, success, information, inform, direct, directly, comparable, compare, webcast, expressed, implied, imply, contain, contains, contained, events, event, periods, period, somewhere, addition, additionally, publicly, amazingly, amaze, amazing, aattributed, attribute, discussed, readily, ready, position, initial, initially, eleven, completed, complete, single, repaid, effective, business, eventually, executing, exclude, excluding, allow, allowed, insist, closely, specific, affect, affecting, ranges, range, serve, single, fewer, few, fighter, eloborate, eloborrated, eighteen, midterm, employer, employers, somebody, thirdly, particularly, particular, thirteen, stopped, stop

References

[1] S. Medya, M. Rasoolinejad, Y. Yang, and B. Uzzi, "An exploratory study of stock price movements from earnings calls," in *Companion proceedings of the web conference 2022*, 2022, pp. 20–31.

[2] P. Grafe, "Topic modeling in financial documents," *Department of Computer Science Stanford University*, 2011.

[3] C. Zhang, "Earnings conference calls and lazy prices." 2021.

[4] A. H. Huang, R. Lehavy, A. Y. Zang, and R. Zheng, "Analyst information discovery and interpretation roles: A topic modeling approach," *Management science*, vol. 64, no. 6, pp. 2833–2855, 2018.

[5] P. Pramanik and R. K. Jana, "Identifying research trends of machine learning in business: A topic modeling approach," *Measuring business excellence*, 2022.

[6] J. Jayaraman and A. Dennis, "Can earnings call sentiment predict stock price movement?" *Proceedings of the Northeast Business & Economics Association*, 2020.

[7] N. Liu, "Predicting stock returns using natural language processing of earnings calls," 2019.

[8] M. Brennan, "Predicting stock performance using quarterly analyst s call transcripts and natural language processing (NLP): An exploration." ProQuest Dissertations Publishing, 2021.

[9] CFITeam, "Earnings call." https://corporatefinanceinstitute.com/resources/valuation/earnings-call/, 2022.

[10] K. A. Keith and A. Stent, "Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls," *arXiv preprint arXiv:1906.02868*, 2019.

[11] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on information and knowledge management*, 2009, pp. 375–384.

[12] D. Kennett, "Earnings call NLP analysis." Github. Available: %7Bhttps://github.com/daniellkennet t/Earnings_Call_NLP_Analysis%7D

[13] L. Noce, A. Zamberletti, I. Gallo, G. Piccoli, and J. A. Rodriguez, "Automatic prediction of future business conditions," in *ADVANCES IN NATURAL LANGUAGE PROCESSING*, vol. 8686, in Lecture notes in computer science, vol. 8686., Cham: Springer International Publishing, 2014, pp. 371–383.

[14] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *2017 international conference on intelligent computing and control systems (ICICCS)*, IEEE, 2017, pp. 745–750.

[15] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, 2010, pp. 100–108.

[16] N. Dumont, "Sentiment analysis natural language: Processing techniques for capital markets disclosure," *The Corporate Governance Advisor*, vol. 25, no. 6, pp. 16–23, 2017.

[17] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on web search and data mining*, 2015, pp. 399–408.

[18] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Advances in neural information processing systems*, vol. 22, 2009.

[19] C. Guo, M. Lu, and W. Wei, "An improved LDA topic modeling method based on partition for medium and long texts," *Annals of data science*, vol. 8, no. 2, pp. 331–344, 2021.

[20] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink, "Machine learning in finance: A topic modeling approach," *European financial management: the journal of the European Financial Management Association*, vol. 28, no. 3, pp. 744–770, 2022.

[21] Z. Wu and M. Palmer, "Verb semantics and lexical selection," *arXiv preprint cmp-lg/9406033*, 1994.

[22] C. Leacock, M. Chodorow, and G. A. Miller, "Using corpus statistics and WordNet relations for sense identification," *Computational Linguistics*, vol. 24, no. 1, pp. 147–165, 1998.

[23] G. Hirst, D. St-Onge, *et al.*, "Lexical chains as representations of context for the detection and correction of malapropisms," *WordNet: An electronic lexical database*, vol. 305, pp. 305–332, 1998.

[24] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[25] D. Lin, "Automatic retrieval and clustering of similar words," in *36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, volume 2*, 1998, pp. 768–774.

[26] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[27] S. Torres and A. Gelbukh, "Comparing similarity measures for original WSD lesk algorithm," *Research in Computing Science*, vol. 43, pp. 155–166, 2009.

[28] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th annual international conference on systems documentation*, 1986, pp. 24–26.

[29] H. Schütze, "Automatic word sense discrimination," *Computational linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[30] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *PloS one*, vol. 15, no. 9, pp. e0239441–e0239441, 2020.

[31] D. Roozen and F. Lelli, "Stock values and earnings call transcripts: A dataset suitable for sentiment analysis," 2021.

[32] D. Roozen and F. Lelli, "Stock values and earnings call transcripts: A sentiment analysis dataset." DataverseNL.

[33] A. W. Lo, H. Mamaysky, and J. Wang, "Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation," *The journal of finance*, vol. 55, no. 4, pp. 1705–1765, 2000.

[34] R. D. Edwards, J. Magee, and W. C. Bassetti, *Technical analysis of stock trends*. CRC press, 2018.

[35] C.-H. Park and S. H. Irwin, "The profitability of technical analysis: A review," 2004.

[36] C. Murphy, "How do 50-day, 100-day, and 200-day simple moving averages differ?" https://www.investopedia.com/ask/answers/06/differencebetweenmas.asp, 2021-12-31.

[37] C. Mitchell, "How to use a moving average to buy stocks"." https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp, 2022-04-08.

[38] E. F. Fama and K. R. French, "Multifactor explanations of asset pricing anomalies," *The journal of finance*, vol. 51, no. 1, pp. 55–84, 1996.

[39] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[40] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014, pp. 216–225.

[41] T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-ks," *The Journal of finance*, vol. 66, no. 1, pp. 35–65, 2011.

[42] Y.-S. Tsai, C.-P. Chang, and S.-W. Tzang, "The impact of golden cross and death cross frequency on stock returns in pre-and post-financial crisis," in *Innovative mobile and internet services in ubiquitous computing: Proceedings of the 11th international conference on innovative mobile and internet services in ubiquitous computing (IMIS-2017)*, Springer, 2018, pp. 708–713.

[43] R. Anaroussi, "yfinance." Dec. 2023. doi: 10.5281/zenodo.1234.

[44] D. Maier *et al.*, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," *Communication methods and measures*, vol. 12, no. 2–3, pp. 93–118, 2018.

[45] C. Kantos, D. Joldzic, G. Mitra, and K. Thi Hoang, "Comparative analysis of NLP approaches for earnings calls," *Available at SSRN 4210529*, 2022.

[46] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th international student conference on advanced science and technology ICAST*, 2012, p. 1.

[47] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Cornell University, 1987.

[48] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Intelligent data engineering and automated learning–IDEAL 2013: 14th international conference, IDEAL 2013, hefei, china, october 20-23, 2013. Proceedings 14*, Springer, 2013, pp. 611–618.

[49] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[50] C. Blankson, K. Cowan, J. Crawford, S. Kalafatis, J. Singh, and S. Coffie, "A review of the relationships and impact of market orientation and market positioning on organisational performance," *Journal of Strategic Marketing*, vol. 21, no. 6, pp. 499–512, 2013.

[51] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[52] S. M. So and V. U. Lei, "On the relationship between investor sentiment, VIX and trading volume," *Risk Governance & Control: Financial Markets & Institutions*, vol. 5, no. 4, pp. 114–122, 2015.