Words that Move Markets: Leveraging TF-IDF and Machine Learning to Forecast Stock Market Trends from Financial News

Minwoo Sohn
Vanderbilt University
minwoo.sohn@vanderbilt.edu

ABSTRACT

This study investigates the predictive capability of linguistic patterns in financial news on stock market trends, focusing on the "Magnificent Seven" (M7) companies using Natural Language Processing (NLP) with the term frequency-inverse document frequency (TF-IDF) metric. These firms significantly influence the U.S. economy, accounting for 32% of the economic profit from all listed firms as of 2022 and are thus critical to understanding broader economic trends and market behavior. Analyzing a dataset from HuffPost (January 2012 - September 2022) and weekly stock data from Yahoo Finance, this research correlates linguistic patterns in news articles with weekly market movements classified into 'bull' and 'bear' markets. The study aims to enhance the understanding of financial terminology and market behavior for individual investors by leveraging NLP to explore how media influences market performance. The primary research question investigates whether TF-IDF metrics can distinguish between bull and bear markets for the M7 companies, with secondary questions focusing on the association of specific terms with stock price shifts and market conditions.

Keywords

NLP, TF-IDF, financial news, machine learning, stock market prediction.

1. INTRODUCTION

In the intricate financial markets, individual investors often find themselves at a disadvantage, lacking the depth of understanding and expertise possessed by professional analysts. Financial news articles serve as the most accessible information source for individuals; however, the challenge lies in extracting actionable insights that correlate with market performance. To bridge this gap, Natural Language Processing (NLP) was leveraged with an ambitious goal: to extract valuable investment insights from the vast ocean of financial news. A rich dataset from HuffPost, encompassing news articles from January 2012 to September 2022 with approximately 210,000 unique entries, serves as the foundation for this analytical endeavor.

Central to my analysis are the stocks nicknamed as "Magnificent Seven"(M7)—Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla—have a profound impact on the U.S. economy and stock market, accounting for approximately 32% of the economic profit of all listed firms as of 2022 [1]. The movements in their stock prices can have ripple effects across the entire market, making them critical indicators of broader economic health and investor sentiment. These high-performing companies are cornerstones of the U.S. but also key drivers of the global economy through their technological innovation. The R&D spending of the M7 companies has increased dramatically, growing from 12% of the total R&D expenditure by publicly listed U.S. firms in 2012 to 38% in 2022. This investment in innovation not only enhances their market dominance but also enables them to outpace competitors, significantly influencing industry standards and economic trends. Consequently, analyzing these companies offers crucial insights into market leadership and the mechanics of innovation-driven economic success [1].

Given their significant impact, these firms area also ideal subjects for exploring how linguistic patterns in financial news can provide predictive insights into stock movements.

The analysis employs the weekly percentage change in stock index price as the performance indicator. Weeks with a positive percentage change from the previous week are labeled 'bull' markets, whereas those with a negative change are deemed 'bear' markets for simplicity. The study will aggregate data for the seven companies from Yahoo Finance into a comprehensive index, with details on this aggregation methodology provided in the data section.

This project seeks to demystify the complex language of finance through NLP. By pinpointing language patterns that align with stock performance and market trends, this research not only aims to make financial analysis more accessible to individual investors but also enhances our collective understanding of the influence of information dissemination on market behavior.

2. PURPOSE STATEMENT AND RESEARCH QUESTION

The purpose of this study is to investigate the predictive power of linguistic patterns in financial news articles on stock market movements. By employing natural language processing (NLP) techniques, specifically the term frequency-inverse document frequency (TF-IDF) metric, this research aims to identify and quantify the impact of specific terms and phrases found in news coverage of the "Magnificent Seven" (M7) companies on the dynamics of the stock market.

News articles and stock prices are both aggregated weekly. This synchronization enables a direct comparison of linguistic patterns with stock market movements, enhancing the accuracy of the analysis. The weekly aggregation of articles involves concatenating all

articles published within the same week, while stock data is processed to calculate the percentage change week-over-week. Weeks are classified as 'bull' if there is a positive percentage change and 'bear' if there is a negative change. This alignment allows for the application of TF-IDF to text data and subsequent machine learning analysis to assess correlations between news language and market behavior.

The following research questions guide this study:

Primary Research Question: Can the linguistic patterns in financial news articles, as quantified by TF-IDF metrics, predict stock market conditions—specifically, distinguishing between bull and bear markets—for the M7 companies? This question focuses on determining if textual analysis of financial news can be effectively used to predict market trends.

Secondary Research Questions:

- What specific terms or phrases from the financial news articles are most strongly associated with shifts in the stock prices of the M7 companies?
- How do variations in the frequency and significance of these terms correlate with the performance indicators of bull and bear markets?

These questions aim to drill down into more detailed aspects of the primary question, providing a pathway to understanding not only if TF-IDF metrics can predict market conditions but also how specific linguistic elements contribute to or correlate with these conditions.

3. LITERATURE REVIEW

Influence of the news on stock markets

In analyzing the influence of news on stock markets, research from Qing Li et al. [2] has taken quantitative approach to evaluate how media impacts stock performance. The study revealed that market sensitivity is influenced by nature of the news and the specific attributes of companies involved, with a pronounced reaction to news about corporate restructuring and earnings reports. It underscores that, in the age of social media, stock markets are increasingly sensitive to the flow of public information. Additionally, the research highlighted that investor sentiment is shaped by the emotional content within news articles and the public's reactions on social media platforms, which in turn influences their decision-making processes. This observation is particularly pertinent in the Web 2.0 era, where various social media platforms serve as arenas for public discourse. The implications of this research support the use of NLP to detect patterns in financial news, offering a predictive lens on stock market trends.

Broad Applications of TF-IDF

The Term Frequency-Inverse Document Frequency (TF-IDF) metric is highly effective in enhancing query retrieval by identifying key terms that are uniquely significant to individual documents. This method simplifies the process of connecting search queries with the most relevant documents. For example, the higher TF-IDF score suggests that the words in a document have a strong relevance to the search query [3]. Furthermore, TF-IDF is not only robust in evaluating keyword relevance within document corpora but also demonstrates a straightforward implementation process, which supports its widespread use [4].

The TF-IDF technique has been applied in various research contexts, such as conducting sentiment analysis on English tweets from Saudi Arabian telecommunication companies [5]. In another study, Twitter datasets were used to examine the sentiment polarity of specific Indian government initiatives, classifying sentiment into positive, negative, and neutral categories **[6]**. Further, TF-IDF has served as a feature set for different machine learning models. Ravinder et al. suggest that for text classification, TF-IDF is a more effective feature choice compared to N-Gram [7]. Additionally, Babacar Gaye proposed a sentiment analysis approach for general Twitter text datasets using a combination of long short-term memory (LSTM) approaches, TF-IDF, and other supervised machine learning algorithms [8]. Features were extracted using the TF-IDF method and classified using LSTM to achieve optimal results.

Past research has utilized texts from Twitter on various topics, but such approaches have not specifically addressed financial or stock analysis. One study explored models that extract features using TF-IDF for different classification algorithms using tweets, classifying these tweets into positive, negative, and neutral categories [9]. However, this study has not extended to include actual financial news.

Analyzing financial news offers several advantages over tweets. Firstly, news articles generally provide deeper and more comprehensive coverage of events than the brief snippets found on Twitter, allowing for a more thorough analysis of financial events and trends. Secondly, traditional news outlets adhere to editorial standards and are often more reliable than Twitter, where content can be posted by anyone without accuracy checks. Relying on news articles reduces the risk of basing analyses on unverified or false information. Lastly, news articles provide context, expert opinions, and background, which help readers understand the significance of events. This is particularly important in financial analysis where the background, regulatory environment, and market history play crucial roles, and oftentimes those are omitted from the limited character space of tweets.

Thus, this study aims to apply TF-IDF to financial news to focus on more articulated articles. Rather than merely identifying sentiment, this study will align stock price performance with the labeling of market increases and decreases, expected to enhance model training and provide better insights.

4. METHOD

4.1 Data

The news data was sourced from publicly available datasets on Kaggle, while the stock price information was retrieved through calls to the Yahoo Finance API.

4.1.1 HuffPost News Data

The dataset comprises approximately 210,000 news headlines collected from HuffPost between 2012 and 2022, making it one of the largest archives of its kind and a valuable benchmark for computational linguistics tasks. The collection spans a decade of news coverage, with a significant portion (about 200,000 headlines) from 2012 to May 2018. Following changes to HuffPost's archival policies, the dataset includes a smaller subset (around 10,000 headlines) from May 2018 to 2022.

Each record within the dataset consists of several attributes:

- ‘category’: The news category under which the article was published, encompassing a wide range of subjects including politics, wellness, entertainment, travel, style & beauty, parenting, healthy living, among others.

- 'headline': The article's title, capturing the essence of the news item.
- 'authors': Names of the contributors to the article.
- 'link': A URL directing to the full article on the HuffPost website.
- 'short_description': An abstract summarizing the article's content.
- 'date': The publication date, documenting the article's timeliness.

The data was obtained from Kaggle and the more information can be found at:

https://www.kaggle.com/datasets/rmisra/news-category-dataset?resource=download

In the actual analysis, the 'short_description' column was primarily used for text analysis and TF-IDF computation.

4.1.2 Yahoo Finance API

The primary stock data will be sourced from the Yahoo Finance API, a comprehensive suite of libraries, APIs, and methods designed to provide historical and real-time data across various financial markets. The focus is on extracting data for seven key companies, corresponding to the same period covered by the news dataset from 2012 to 2022.

The variable of interest are:

- 'Date': the date of the stock exchange.
- 'ticker': An abbreviated symbols representing company names.
- 'adjusted close price': The value of a stock after adjusting for any corporate actions, ensuring that our calculations are precise and meaningful.
- 'equal_weight_index': An aggregated stock index number that incorporates the 7 stocks with equal weight for simplicity.

An index with equal weighting for these companies - Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla will be constructed. This aggregated index approach simplifies the complex fluctuations of the stock market into a single metric, facilitating an easier comparison and analysis of overall market performance.

The analysis will center around the weekly change as a performance indicator, distinguishing between bull and bear markets. This method will not only highlight the general market direction but also provide insights into the volatility and potential trends within these periods.

4.2 Data Preprocessing

4.2.1 News Data

The dataset comprises news articles from various sectors. The data preprocess involves refining this dataset by filtering out articles relevant only to the M7 companies, using their names and stock tickers as keywords. This targeted filtering reduced the initial count from 209,527 articles to 2,634.

Subsequent steps included ensuring there were no duplicate articles. After removing duplicates, the dataset was further reduced to 2,347 articles. Upon reviewing the content, it became apparent that many entries were promotional rather than informative. Examples of such content include:

- Want more? Be sure to check out HuffPost Style on Twitter, Facebook, Tumblr, Pinterest, and Instagram at @HuffPostStyle. Between
- Keep in touch! Check out HuffPost Weddings on Facebook, Twitter, and Pinterest. Check out more hilarious last-name fails in…
- Have something to say?
- Follow Tia Ghose on Twitter and Google+. Follow us @livescience, Facebook & Google+. Original article on LiveScience.
- See Camilla's style evolution! Check 'em out. Want more? Be sure to check out HuffPost Style on Twitter, Facebook, Tumblr

These entries were identified as not significantly relevant to deriving financial insights on any day. Moreover, many articles merely mentioned the names of M7 companies like Google and Facebook without providing substantial insights, adding to the noise. Finally, any text starting with '@' and '#' was removed to streamline the dataset for analysis, ensuring that the focus remained on relevant financial information. After this process, the data was decreased to 1390 rows.

The dataset was sorted chronologically, and news articles were grouped by publication week. The text from these articles was then concatenated into a single string for each week.

4.2.2 Stock Data

Initially, stock data was retrieved from the Yahoo Finance API. To access the data, the ticker symbols for the companies—'GOOGL', 'AMZN', 'AAPL', 'META', 'MSFT', 'NVDA', 'TSLA'—were specified in a list. Additionally, the start and end dates were set, along with specifying 'Adj Close' to obtain the adjusted close prices.

For the analysis, a simple average of the stock prices was computed daily using an equal weighting method. This involved assigning an equal weight of 1/7 to each of the seven key companies and then multiplying this weight by the daily average price across all stocks. This approach allows for a standardized comparison of collective stock performance over time.

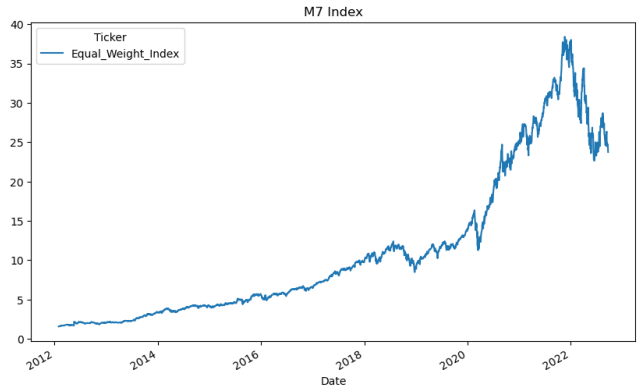

Figure 1. The Equal Weight Index of M7 Stocks

The stock data is aggregated weekly to align with the news data's timeframe. Weeks exhibiting a positive percentage change compared to the previous week in stock prices are labeled as 'Bull' markets, whereas those with a negative change are classified as 'Bear' markets.

Table 1. Weekly Market Performance

Date	Weekly_Percentage_Change	Market_Condition
2/12/12	2.023782	Bull
2/19/12	1.564829	Bull
2/26/12	0.945401	Bull
3/4/12	1.95066	Bull
3/11/12	-0.591682	Bear
...	...	...
8/28/22	-5.22923	Bear
9/4/22	-5.530199	Bear
9/11/22	5.302813	Bull
9/18/22	-5.86719	Bear
9/25/22	-2.933344	Bear

4.3 Linguistic Analysis

The term frequency-inverse document frequency (tf-idf) technique is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. It is achieved by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. The detailed equation is shown below[10] :

$$\mathrm{TF} = \frac{\text{Number of times keyword in document}}{\text{Number of words in document}} \quad (1)$$

$$IDF = log\left(\frac{\text{Number of documents}}{\text{Number of documents containing the keyword}}\right) \quad (2)$$

$$\text{TF-IDF} = \text{TF} * \text{IDF} \quad (3)$$

This results in words that are common in a specific document but rare in the overall document set being assigned a high score, highlighting their importance in the context of that document.

In the context of this project, TF-IDF is employed to sift through thousands of financial news articles to identify and weigh the significance of words that are most relevant to stock market movements. By analyzing the frequency and importance of specific terms in news articles related to the "Magnificent Seven" companies, the technique aids in pinpointing linguistic patterns that may have predictive power over stock market trends. It allows us to quantify the impact of certain keywords or phrases on market conditions, distinguishing between common language and terms that could signal shifts in the financial landscape.

4.4 Statistical Analysis (Machine Learning)

After extracting features from the TF-IDF, it was used for different classification machine learning models. This study explored three different algorithms. All the algorithms are to classify which particular words would have the strong association with the 'increase' or 'decrease' in the stock market. The section briefly discusses these algorithms and the evaluation metrics for the good model.

Following the extraction of features using the TF-IDF technique, this study employed these features in various machine learning classification models. The objective of these models is to identify which specific words strongly associate with an 'Bull' or 'Bear' market. This section will outline the three different algorithms used and discuss the evaluation metrics applied to determine the effectiveness of each model.

4.3.1 Logistic Regression

Logistic regression models analyze the impact of dependent variables on categorical outcomes, typically binary. The model is referred to as simple logistic regression when it includes one predictor, and as multivariable logistic regression when it involves multiple predictors [11]. In the context of our project, Logistic Regression is used to predict the market condition (bull or bear) based on the TF-IDF scores of terms found in financial news articles. By modeling the relationship between the terms' TF-IDF scores and the market's direction, we can assess the likelihood of market movements based on current news trends, making it a powerful tool for forecasting.

4.3.2 Multinomial Naive Bayes

Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem under the assumption that the features within a dataset are independent of each other. This "naive" assumption simplifies computation, particularly benefiting high-dimensional data environments like text classification, where it efficiently handles large volumes of input variables [12].

Unlike the standard Naive Bayes, which typically models binary features (often used in spam detection), MNB is well-suited for frequency data, such as the word counts or TF-IDF transformations commonly found in text classification tasks. This capability makes MNB particularly effective for categorizing text where the frequency and presence of particular words are more predictive of the class label.

In this study, Multinomial Naive Bayes is employed to classify news articles as indicators of either bull or bear markets, using the frequency of text features transformed through TF-IDF.

4.3.3 Gradient Boosting Machines

Gradient Boosting Machines (GBMs) employ a sophisticated ensemble approach, constructing a predictive model through the sequential integration of multiple weak models, commonly decision trees. This technique enhances the model by systematically minimizing errors using a gradient descent-based approach, tailored to various differentiable loss functions [13].

Within this project, Gradient Boosting is leveraged to predict market trends based on the aggregated TF-IDF features of financial news articles. This method is particularly useful for reducing bias and variance in our predictions, offering a more nuanced understanding of how specific terms or combinations of terms in the news correlate with market movements.

4.3.4 Evaluation Metrics

Four most common evaluation metrics for classification tasks are used in this study on predicting the market condition. The evaluation metrics are accuracy, precision, recall and f1[14].

Accuracy: This metric represents the proportion of correctly classified instances out of the total number of instances. It is calculated using the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

Precision: Defined as the ratio of true positives to the sum of true positives and false positives

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

Recall: Measures the ability of a model to identify all relevant instances. It is defined as the ratio of true positives to the sum of true positives and false negatives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

F1-Score: A harmonic mean of precision and recall, this metric is useful for comparing two classifiers, especially when the classes are imbalanced. It is calculated as follows:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

5. RESULTS

This section discusses the outcomes of applying machine learning models to predict stock market trends. Three classification models—Logistic Regression, Multinomial Naive Bayes, and Gradient Boosting—were utilized, leveraging TF-IDF extracted linguistic features to forecast whether the stock market would experience an increase or decrease.

Initial results from the baseline models indicated suboptimal performance, particularly in accurately predicting bear market trends. This prompted an investigation into different sampling methods to address the imbalance between the majority class (bull market) and the minority class (bear market). To improve model performance, strategies such as oversampling the minority class and downsampling the majority class were explored and tested.

5.1 TF-IDF

TF-IDF values were calculated using the 'TfidfVectorizer' from the Scikit-learn library, which allows the specification of a tokenizer as an argument. The tokenizer employed in this study converted all words to lowercase. This facilitated the transformation of the weekly aggregated news articles into a matrix representing the TF-IDF values for each vocabulary term per week. The resulting matrix dimensions were 384 by 5172, indicating the data spanned 384 weeks and included 5172 unique vocabulary terms, each with its corresponding TF-IDF score.

A quick analysis was conducted on the TF-IDF distribution by summing all the TF-IDF scores presented in the table and outlining the distribution. Instead of averaging the values, the sum was chosen to highlight differences on a larger scale, given that most TF-IDF scores are below 0.01, which does not sufficiently illustrate significance on its own. The term frequency distribution graph highlights the broad and varied vocabulary within the dataset. It shows that most words are used infrequently, with a sum of TF-IDF scores below 3. The graph also illustrates a sharp decline in the count of words as their frequency increases, indicating a steep drop-off in the number of words that occur more frequently.

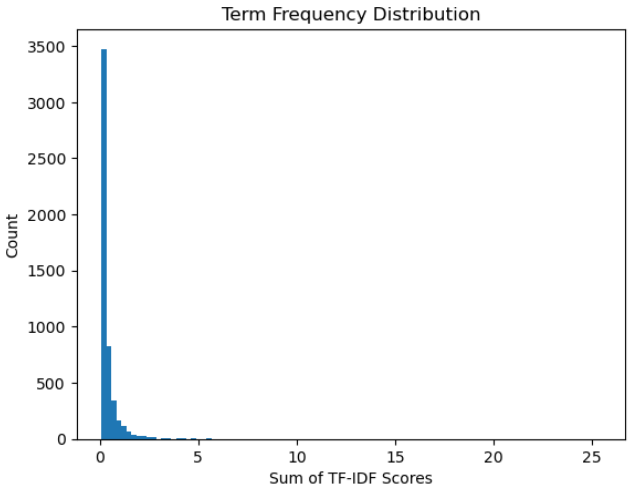

Figure 2. Term Frequency Distribution

5.2 Logistic Regression

In this analysis, three logistic regression models were evaluated: a baseline model, an oversampled model, and an undersampled model. Each model was assessed based on precision, recall, f1-score, and overall accuracy across two classes, with class 0 potentially representing the bear market and class 1 the bull market.

Table 2. Evaluation Metrics for Baseline Logistic Regression

Class	Precision	Recall	F1
Class 0	**0.00**	**0.00**	**0.00**
Class 1	**0.60**	**1.00**	**0.75**
Accuracy			**0.60**

Table 3. Evaluation Metrics for Oversampled Logistic Regression

Class	Precision	Recall	F1
Class 0	**0.84**	**0.74**	**0.79**
Class 1	**0.80**	**0.88**	**0.83**
Accuracy			**0.82**

Table 4. Evaluation Metrics for Undersampled Logistic Regression

Class	Precision	Recall	F1
Class 0	0.44	0.53	0.48
Class 1	0.46	0.38	0.41
Accuracy			0.45

The baseline model shows low performance with a precision of 0.60 and a recall of 1.00 for class 1, resulting in an f1-score of 0.75. For class 0, all metrics—precision, recall, and f1-score—register at 0.00, indicating that the model fails to correctly identify any instances of this class. The overall accuracy stands at 60%, but this figure is misleading given the model's complete inability to detect class 0.

The oversampled model exhibits a substantial improvement across all metrics. For class 0, the precision is 0.84 with a recall of 0.74, yielding an f1-score of 0.79. Class 1 shows a precision of 0.80 and a recall of 0.88, resulting in an f1-score of 0.83. The overall accuracy significantly increases to 82%, demonstrating a more robust model performance compared to the baseline.

The undersampled model shows a more balanced performance between the two classes but with lower scores than the oversampled model. It achieves a precision of 0.44 and a recall of 0.53 for class 0, and a precision of 0.46 with a recall of 0.38 for class 1, with respective f1-scores of 0.48 and 0.41. The overall accuracy is considerably lower at 45%, reflecting a generally weaker model capability compared to the oversampled approach.

Next, the analysis focused on identifying which words from the logistic regression model were most influential in predicting stock market movements, either upward or downward. Positive coefficients indicate words closely associated with an increasing market, while negative coefficients suggest a stronger link to a declining market. For this analysis, an oversampled logistic regression approach was employed.

Table 5. Positive and Negative Coefficients for the Best Logistic Regression Model

Coeff_1	Vocab_1	Coeff _2	Vocab_2
-0.994	instagram	0.675	apple
-0.681	new	0.598	ceo
-0.631	search	0.509	write
-0.570	book	0.437	’
-0.568	review	0.424	support
-0.559	facebook	0.413	issue
-0.557	compare	0.400	life
-0.548	man	0.394	word
-0.536	educational	0.392	article
-0.522	election	0.390	join
-0.520	coronavirus	0.377	giant
-0.517	team	0.360	government
-0.508	reduce	0.358	service
-0.506	social	0.357	tweet
-0.503	disclose	0.350	accuse
-0.480	regard	0.349	movie
-0.473	2016	0.348	child
-0.471	lawyer	0.339	spread
-0.468	cite	0.335	morning
-0.462	moment	0.330	john

Words such as "apple," "ceo," "write," "support," "issue," "life," and "government" exhibit positive coefficients, signaling their association with bull market trends. For example, the strong positive coefficient for "apple" suggests that discussions about Apple are likely indicators of positive market performance. Similarly, terms linked to business and governance, like "ceo" and "government," correlate with upward market movements, reflecting optimism when these terms are prevalent.

Conversely, terms like "instagram," "new," "search," "book," "review," and "facebook" have negative coefficients, indicating an association with bear market conditions. For instance, "instagram" shows the most negative coefficient, suggesting its mention predicts market downturns. One hypothesis could be that these social media platforms are frequently mentioned in contexts that relate to privacy issues, advertising controversies, and various legal challenges. These topics could contribute to negative sentiment surrounding the companies, potentially impacting their public perception and, subsequently, their market performance. Other terms such as "election," "coronavirus," "social," and "lawyer" also have negative coefficients, implying that discussions around these topics might relate to negative market sentiments or downturns.

Words with negative associations often reflect market uncertainties, typically triggered by political, social, or health-related issues. In contrast, words with positive coefficients usually relate to corporate and governmental activities, which generally lead to positive market reactions.

5.3 Multinomial Naïve Bayes

A similar method was applied using three variants of the Multinomial Naive Bayes model: baseline, oversampled, and undersampled. Each version was evaluated for its effectiveness using metrics such as precision, recall, f1-score, and overall accuracy. These models were tested across two categories, where class 0 likely corresponds to bear market conditions and class 1 to bull market conditions, to determine their accuracy in predicting market trends.

Table 6. Evaluation Metrics for Baseline Multinomial Naïve Bayes

Class	Precision	Recall	F1
Class 0	0.00	0.00	0.00
Class 1	0.60	1.00	0.75
Accuracy			0.60

Table 7. Evaluation Metrics for Oversampled Multinomial Naïve Bayes

Class	Precision	Recall	F1
Class 0	0.59	0.84	0.69
Class 1	0.77	0.49	0.60
Accuracy			0.65

Table 8. Evaluation Metrics for Undersampled Multinomial Naïve Bayes

Class	Precision	Recall	F1
Class 0	0.47	0.57	0.52
Class 1	0.50	0.41	0.45
Accuracy			0.48

The baseline model shows a strong bias towards Class 1 (bull market), with a precision of 0.60 and recall of 1.00, resulting in an F1-score of 0.75. However, it completely fails to identify Class 0 (bear market), with all metrics at 0.00. The overall accuracy of the model is 0.60, indicating that while it predicts bull markets effectively, it is unable to recognize bear markets at all.

The oversampled model demonstrates improved balance in recognizing both market conditions. Class 0 shows a significant improvement with a precision of 0.59, recall of 0.84, and F1-score of 0.69. Class 1's metrics, however, decrease compared to the baseline, with a precision of 0.77, recall of 0.49, and F1-score of 0.60. Overall accuracy slightly improves to 0.65, reflecting a better balanced but still imperfect model.

The undersampled approach results in the lowest overall accuracy of 0.48. It offers more balanced performance between classes than the baseline model but still shows moderate effectiveness. Class 0 has a precision of 0.47, recall of 0.57, and F1-score of 0.52. Class 1 scores are closely aligned with those of Class 0, with a precision of 0.50, recall of 0.41, and F1-score of 0.45.

The results from the Naive Bayes model, as displayed, reveal the log probabilities assigned to words in predicting market movements categorized as "Decrease" or "Increase". Log probabilities nearer to zero signify a higher original probability and a stronger influence on the outcome. Comparing these values across categories helps identify which words have a greater impact on specific market movements. Additionally, the presence of NaN values indicates that certain words were absent from the training data for specific categories.

Table 9. Log probabilities for the Multinomial Naïve Bayes Model

Terms	Decrease	Increase
facebook	-6.051	-6.205
amazon	-6.711	-6.819
google	-6.843	-6.788
apple	-6.856	-6.560
new	-6.901	-7.346
post	-7.024	-7.069
instagram	-7.096	NaN
twitter	-7.174	-7.344
like	-7.232	-7.406
social	-7.298	NaN
people	-7.303	-7.539
want	-7.348	NaN
page	-7.359	-7.433
president	-7.391	NaN
search	-7.397	NaN
photo	-7.407	NaN
million	-7.411	NaN
use	-7.433	NaN
medium	-7.456	NaN
week	-7.463	-7.440
ceo	NaN	-7.319
day	NaN	-7.363
year	NaN	-7.377
time	NaN	-7.427
write	NaN	-7.438
change	NaN	-7.446
tesla	NaN	-7.521
company	NaN	-7.535
right	NaN	-7.542

Words like "facebook", "amazon", "google", and "apple" have negative log probabilities in both categories, but the values are closer for "Decrease", suggesting a slightly stronger association with negative market trends. The words exclusively mentioned in the "Increase" category, such as "CEO," "company," "Tesla," "day," "year," "time," "write," "change," and "right," are predominantly linked to business, corporate actions, and temporal events. They reflect optimism associated with corporate progress, strategic changes, and significant temporal markers and view as indicators of positive developments in the market.

Certain terms such as "instagram", "social", "want", "president", "search", "photo", "million", "use", "medium", and a few others appear only under the "Decrease" category, or have a NaN (not a number) under "Increase", indicating these words are strongly associated with market declines. These words are commonly relate to social media, public communication, and personal interactions. These terms suggest a focus on general social engagement and media discourse, which might be associated with periods of negative sentiment when frequently mentioned in financial news. Additionally, the presence of words like "president" and "million" could

indicate discussions surrounding political events or high-stake financial news.

5.4 Gradient Boosting

The baseline Gradient Boosting Model's efficacy was assessed through metrics including precision, recall, f1-score, and overall accuracy. These metrics were analyzed across two designated categories: Class 0, likely representing bear market conditions, and Class 1, indicative of bull market trends. This evaluation was aimed at gauging the model's ability to accurately predict directional market movements.

Table 10. Evaluation Metrics for Gradient Boosting Model

Class	Precision	Recall	F1
Class 0	**0.27**	**0.13**	**0.17**
Class 1	**0.56**	**0.76**	**0.65**
Accuracy			**0.51**

For Class 0, representing a decreasing market, the model shows a precision of 0.27, recall of 0.13, and an F1 score of 0.17, suggesting a lower predictive accuracy for this category. In contrast, for Class 1, corresponding to an increasing market, the model achieves better precision at 0.56, recall at 0.76, and an F1 score of 0.65, indicating a relatively higher accuracy in identifying bullish trends. Overall, the model attains an accuracy of 0.51.

The Gradient Boosting model yielded the top 20 features with the highest importance scores. It's important to note that this model does not differentiate the features by their impact on bull or bear market conditions. Instead, the model assigns a weight of importance to each feature, which it uses in its decision-making process to determine the market direction.

Table 11. Ranking of Words with High Importance

Rank	Words	Importance
1	**netflix**	**0.0670**
2	**moment**	**0.0421**
3	**co**	**0.0361**
4	**inc.**	**0.0339**
5	**lady**	**0.0248**
6	**instagram**	**0.0198**
7	**medium**	**0.0187**
8	**challenge**	**0.0180**
9	**live**	**0.0173**
10	**celebrate**	**0.0169**
11	**post**	**0.0158**
12	**day**	**0.0157**
13	**york**	**0.0150**
14	**reveal**	**0.0146**
15	**search**	**0.0143**
16	**panel**	**0.0139**
17	**enforcement**	**0.0138**
18	**effort**	**0.0135**
19	**beauty**	**0.0131**
20	**war**	**0.0129**

The feature importance from the Gradient Boosting model reveals key influences on market trends, with high-ranking words like "Netflix," "Instagram," and corporate identifiers like "inc." and "co." suggesting the impact of media, entertainment, and company-specific news. Social dynamics captured by terms such as "challenge" and "celebrate," along with references to events ("moment," "live") and locations ("York"), indicate a connection between public sentiment and market behavior. Additionally, "enforcement" and "war" reflect the market's sensitivity to regulatory actions and geopolitical events, underscoring the complex interplay between diverse societal factors and financial market reactions.

6. CONCLUSION AND FUTURE WORK

The analysis of machine learning models including Logistic Regression, Multinomial Naive Bayes, and Gradient Boosting has provided deep insights into stock market trends. Logistic Regression showed clear patterns linked to market sentiment, while Multinomial Naive Bayes highlighted key terms driving market movements. The Gradient Boosting Model indicated that media influence, social dynamics, corporate activities, and broader socio-economic factors significantly impact market behavior.

Among these, the Oversampled Logistic Regression model excelled, showing superior performance with an overall accuracy of 0.82. It achieved a precision of 0.84 and recall of 0.74 for bear market predictions (class 0), and a precision of 0.80 and recall of 0.88 for bull markets (class 1), outperforming other models including Gradient Boosting.

This study underscores the utility of TF-IDF extracted linguistic features and demonstrates the potential of NLP in financial market analysis. The results confirm that even with varying levels of model success, valuable market insights can be gleaned from the careful application of machine learning techniques to financial news text data.

For future work, enriching the dataset with content from specialized financial platforms such as The Economist and Financial Times could deepen the insights. Inclusion of comprehensive news coverage from leading outlets like The New York Times would also broaden the scope of financial news analysis.

The choice of weekly data aggregation was made to streamline computations, yet it limited the dataset's size. Future research should consider a daily analysis to capture more granular market fluctuations. Moreover, redefining bear and bull markets with a more refined threshold than simply the zero percent change could significantly improve the precision of the analysis.

GitHub link to the project: https://github.com/minu803/nlp-final-project

7. REFERENCES

[1] Erencin, C., Evenett, S. J., Ledan, P., and Reitz, F. 2023.The 'Magnificent Seven' (M7): Assessing their contribution to the performance of US publicly listed firms. Crux of Capitalism. Insight #10

[2] Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. 2014. The effect of news and public mood on stock movements. Information Sciences, 278, 826-840. https://doi.org/10.1016/j.ins.2014.03.096

[3] Ramos, J.E. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c

[4] Qaiser, S., & Ali, R. 2018. Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181. https://doi.org/10.5120/ijca2018917395

[5] Qamar, A. M., Alsuhibany, S. A., & Ahmed, S. S. 2017. Sentiment classification of Twitter data belonging to Saudi Arabian telecommunication companies. International Journal of Advanced Computer Science and Applications (IJACSA), 8(1). http://dx.doi.org/10.14569/IJACSA.2017.080150

[6] Rekha, V., Raksha, R., Patil, P., Swaras, N., & Rajat, G. L. 2019. Sentiment analysis on Indian government schemes using Twitter data. International Conference on Data Science and Communication (IconDSC) (pp. 1-5). Bangalore, India. https://doi.org/10.1109/IconDSC.2019.8817036

[7] Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. 2019. The impact of features extraction on the sentiment analysis. Procedia Computer Science, 152, 341-348. https://doi.org/10.1016/j.procs.2019.05.008

[8] Gaye, B., Zhang, D., & Wulamu, A. 2021. A tweet sentiment classification approach using a hybrid stacked ensemble technique. Information, 12(9), Article 374. https://doi.org/10.3390/info12090374

[9] Popoola, G., Abdullah, K.-K., Fuhnwi, G. S., & Agbaje, J. 2024. Sentiment analysis of financial news data using TF-IDF and machine learning algorithms. In 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). Houston, TX, USA. https://doi.org/10.1109/ICAIC60265.2024.10433843

[10] J.C. C. S. 2023. TF-IDF in python (scikit-learn example) - term frequency-inverse document frequency. JC Chouinard. https://www.jcchouinard.com/tf-idf/

[11] Nick, T.G., Campbell, K.M. (2007). Logistic Regression. In: Ambrosius, W.T. (eds) Topics in Biostatistics. Methods in Molecular Biology™, vol 404. Humana Press. https://doi.org/10.1007/978-1-59745-530-5_14

[12] Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. In: Webb, G.I., Yu, X. (eds) AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science, vol 3339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30549-1_43

[13] Natekin, A., Knoll, A. 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, Volume 7.

[14] Hossin M. A Review on Evaluation Metrics for Data Classification Evaluations 2015. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, March 2015