THE EFFECT OF DONALD J. TRUMP'S SOCIAL MEDIA POSTS ON STOCK MARKET MOVEMENT

by

Eric Wu

A Capstone Project Proposal
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Bachelor of Science
Computational and Data Science

Committee:

_____ Dr. Kent Miller, Capstone Project Proposal Director

_____ Dr. First Last, Committee Member

_____ Dr. First Last, Committee Member

_____ Dr. First Last, Department Head

Date: _____ Spring Semester 2025
George Mason University
Fairfax, VA

The Effect of Donald J. Trump's Social Media Posts on Stock Market Movement

A capstone project proposal submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science at George Mason University

By

Eric Wu
Bachelor of Science
My Other Former School, Year of first degree

Director: Dr. Kent Miller, Professor
Department of Computational and Data Science Department

Spring Semester 2025
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to ...

Acknowledgments

I would like to thank the following people who made this possible ...

Table of Contents

Page

List of Tables vi
List of Figures vii
Abstract viii
1 Introduction 1
2 Prior Research 2
2.1 Machine Learning or Lexicon Based Sentiment Analysis Techniques on Social Media Posts by David L. John and Bela Stantic (2022) 2
2.2 Predicting the Effects of News Sentiments on the Stock Market by Dev Shah, Haruna Isah, and Farhana Zulkernine (2018) 4
2.3 Sentiment Analysis on Social Media for Stock Movement Prediction by Thien H. Nguyen, Kiyoaki Shirai, and Julien Velcin (2015) 5
2.4 Effect of Public Sentiment on Stock Market Movement Prediction During the COVID-19 Outbreak by Das, Sahukhan, Chatterjee, and Charkrabarti (2022) 9
2.5 Using Twitter to Predict the Stock Market: Where is the mood effect? (2015) 11
2.6 Investor Sentiment in the Stock Market by Malcom Baker and Jeffery Wurgler (2007) 14
3 Data 16
4 Theory 21
5 Results 23
6 Conclusion 35
7 Software 38
Bibliography 57

List of Tables

Table		Page
5.1	Top Terms per Topic	28
5.2	Post-Hoc Tukey LDA	30
5.3	Seeded Topic Dictionary	31
5.4	Post-Hoc Tukey SLDA	34
6.1	LDA Topic Definitions	36
6.2	Seeded LDA Topic Definitions	37

List of Figures

Figure		Page
3.1	Box plots of S&P500 Numerical Features	17
3.2	Box plots of NASDAQ Composite Numerical Features	17
3.3	S&P 500 vs. NASDAQ Normalized Performance (Closing Price)	19
3.4	Counts of Various Data Quality Issues In Trump Truth Social Archive	20
5.1	Sentiment Distribution of Social Media Posts	23
5.2	Index Closing Prices by Sentiment Class	24
5.3	Index Daily Change in Price by Sentiment Class	24
5.4	Top 30 Most Frequent Terms Found in Donald Trump's Truth Social Posts	27
5.5	Spread of Daily Volatility in the S&P500 by Topic	28
5.6	Topic Distribution of the Seeded LDA Topic Model	32
5.7	Spread of Daily Volatility in the S&P500 by Topic	33

Abstract

THE EFFECT OF DONALD J. TRUMP'S SOCIAL MEDIA POSTS ON STOCK MARKET MOVEMENT

Eric Wu, BS

George Mason University, 2025

Capstone Project Proposal Director: Dr. Kent Miller

The effect of sentiment and news on the stock market has been used by institutional and retail investors to predict the potential for arbitrage. The aim of this paper is to identify and establish a link between the social media ramblings from Donald Trump during his second term on the stock market. The outsized influence of an individual on the stock market could effect the worldwide economy and have far reaching consequences. This research builds off of the behavioral finance framework of the stock market, the documented effects of predicting market movement using textual data derived from news sources and social media content, and the natural language processing tools developed by open source developers. The historical stock quotes used in this project were obtained from Yahoo Finance and Donald Trump's social media posts were obtained from a webscrape of Truth Social. In this paper we explore the effect of derived sentiment classes on the daily volatility of the stock market, and the effect of the topics Donald Trump posts about on the stock market. Although we failed to establish a link between the sentiment classes of Trump's tweets, we were able to identify a relationship between the topics of his tweets on the daily volatility of the S&P500.

Chapter 1: Introduction

How powerful can a man's words be?

During his second term, the 47th president of the United States, Donald Trump, has inflicted chaos and pandemonium on the markets via his enlightened posts on social media. This unease has caused retail and institutional investors to second-guess their risk tolerance, long-term holdings, and confidence in the continued financial growth of the United States.

Do Donald Trump's colorful social media posts actually effect market performance in a measurable and meaningful manner the short and long-term?

The stock market affects all of us indirectly and directly for those with exposure in said markets. Some of us depend on continued growth in the stock market for our savings or retirement, obtaining stable employment from publicly traded companies who raise capital from healthy stock markets, or rely on the growing stock market as an economic signal that the economy is thriving, bolstering consumer confidence, which in turn boosts consumer spending.

Chapter 2: Prior Research

2.1 Machine Learning or Lexicon Based Sentiment Analysis Techniques on Social Media Posts by David L. John and Bela Stantic (2022)

This text is a chapter from the textbook, "Intelligent Information and Database Systems". This chapter explains how social media platforms provide an accessible and plentiful source of data because individuals post their thoughts and opinions across a vast sea of domains. The chapter states, "social media posts are a plentiful data source that offers a great opportunity for analysis for decision-making purposes". [1]

Sentiment analysis is also known as opinion mining. Sentiment analysis is a natural language processing technique that aims to quantify qualitative textual data. This means, the goal of sentiment analysis is to analyze individual's sentiments, opinions, attitudes, emotions, and appraisals. [1]

Classifying the sentiment of text relies largely on a predetermined lexicon. This pre-existing lexicon typically takes the form of a dictionary of pre-labeled words that have a known sentiment or opinion. An example would be "bad", which typically has a negative sentiment. For example, an individual who says "I had a bad day" would usually feel and have a negative emotion, opinion and sentiment. The converse is also true. An individual saying, "I had a good day", would generally be associated with a positive sentiment.

However, a distinction of opinion and comparative-type words in sentiment analysis needs to be made. Opinion-type words express an absolute opinion, for example "good" and "bad". Comparative-type words, like "better", "best", and "worse", cannot be used in sentiment analysis as they do not express an absolute sentiment; rather these words simply

compare A and B. Even though it is theoretically possible to assign a positive sentiment to the word "better", a comparison does not express an absolute opinion, instead it only compares a relationship like already established.

John and Stantic (2022) [1] state, as of the date of the textbook's publishing, the majority of sentiment analysis had been on written text. This may limit the applicability of predicting sentiment as where the text's sentiment does not follow the norm of an opinion word's sentiment.

The cases where sentiment analysis on written text struggles are orientation and context, sentiment-less opinion words, sarcasm or irony, and statements lacking opinion words. [1]

Orientation and context in the context of sentiment analysis refers to positive or negative opinion words having their actual sentiment reversed in a sentence because of context or domain. Opinion words may lack sentiment when observed in sentences involving questions or conditionals. Sarcasm and irony are an obvious case where sentiment can be hard to analyze as the word's individual sentiment maybe positive, but the intended sentient of the writer was negative. Again, in this edge-case, the word's typical sentiments did not match the individual's expressed intent. Statements lacking opinion words may have an actual sentiment, however sentiment analysis may not be able to identify the actual sentiment due to lacking opinion words.

Sentiment analysis can be performed using Bayesian network classifiers, support vector machines (SVMs) and random forests. John and Stantic (2022) [1] compared several of the most common sentiment analysis techniques, including the ANEW, LIWC, SentiWordNet, and machine learning derived techniques like Naive Bayes classifiers, SVMs and decision tree-based classifiers. They found the open-sourced VADER lexicon, was the most accurate sentiment analysis tool in accurately predicting the sentiment of tweets. An advantage of the VADER lexicon is in its pre-complied nature; this fact enables the lexicon to be more computationally efficient. They discovered neural network-based methods may be less sensitive to changes in sentiment as the standard deviation between the maximum and minimum sentiment values for their dataset. [1]

These findings are particularly relevant to the stated goal of this capstone project. The finding that VADER was particularly efficient and effective at assessing the sentiment of short text strings scraped from twitter should mean using VADER to assign sentiment scores to Donald Trump's truth social posts as Truth Social is essentially an alternative analogue to twitter posts.

2.2 Predicting the Effects of News Sentiments on the Stock Market by Dev Shah, Haruna Isah, and Farhana Zulkernine (2018)

This article states the large amount online text data has an observable influence on the stock market by way of influence through public sentiment. The paper focused public sentiment derived from the public domain via Wikipedia, social media posts, and news outlets. In particular, they derived sentiment from news outlets to achieve a 70.59% accuracy in predicting the direction of stock pricing from stocks in the pharmaceutical industry. [2]

Public sentiment can be measured or proxied through individuals expressing sentiments like moods and emotions through social media posts. However, an air of skepticism must be acknowledged. Social media in its nature of acting as a public forum, enables malicious users, bots, and unintentional misinformation to propagate, causing users to value or devalue products, values, and ideologies. Quality measures and controls need to be created and assessed to mitigate the influence of these users.

Public sentiment through societal mood can contribute to an equities' perceived value or lack thereof value. They also state other research has shown the value in incorporating public sentiment measures can improve the predictive accuracy of analytical stock models. [2]

This paper used a dictionary-based approach to compute public sentiment. They also note lexicon-based approaches have an inherent advantage versus other rule-based approaches because the generally have good cross domain performance due to being trained and designed on a variety of data sources. [2]

The authors extracted news articles and preprocessed and transformed them using traditional NLP techniques. Then they assigned the news articles sentiment scores from a pre-compiled dictionary. After, they cross-validated the buy and hold positions based on the assigned sentiment values to understand the effects of sentiment on the decision-making process of stock traders.

The pre-compiled dictionary was manually labeled with text that pertained to pharmaceutical companies. They stated a word weight measure may enhance the accuracy of their model as some words are likely to be more significant in the pricing of stock. [2] In addition to that statement, the introduction of a hybrid model may also enhance model performance.

The findings regarding a manually created lexicon on a specific use-case may mean an altered version of the VADER lexicon could be needed for my capstone project. Donald Trump does have a particular way of speaking and posting on social media. He uses rather short sentences with an emphasis on talking about nouns or people, in particular.

2.3 Sentiment Analysis on Social Media for Stock Movement Prediction by Thien H. Nguyen, Kiyoaki Shirai, and Julien Velcin (2015)

In this article, the authors extracted mood information from social media data using sentiment analysis. Then, they integrated this mined data to predict stock movement. They note, social media post often contains slang, shorthand, misspellings, and uncommon grammar constructions.[3] This diversity in language can make sentiment analysis difficult.

This article proposes a "topic-sentiment" feature to improve a stock movement model's predictive ability; this feature is a derivation of the extracted topics from the texts in tandem with the associated sentiments. The paper uses two such topic models to test their theory. They use a latent model called the joint sentiment/topic model (JST), which is an existing model where the extracted topics and sentiments are latent, meaning they are hidden. Whereas their second model is of their own design and topic and sentiment are not hidden.

Predictive stock movement models ought to operate under the efficient market hypothesis (EMH). The efficient market hypothesis states that the stock market should reflect all of the publicly available information. This should mean as publicly available information iterates, in this case from the news, the stock market should reflect the changes in information in real time. When considering the news cycle, as it updates, the stock market should reflect the changes from the news cycle.

Because the news operates in an unknowable and random pattern, it can be modeled with a random walk; a random walk should only be about 50% predictable. Therefore, stock movements should similarly follow a random walk. This would mean the accuracy of a stock movement model would be no more than 50%.

However, there have been researchers who have disproven or disregarded this line of reasoning. The researchers who state that stock market prices do not follow a random walk, and are somewhat predictable, use the evidence of a 56% directional accuracy of predictive models; this ought to negate the first assertion that the stock market's pricing operates under a random walk. [3]

Other than the efficient market hypothesis and the random walk theories, the trading philosophies of investors need to be considered. There are two distinct schools of trading philosophies: fundamental analysis and technical analysis. [3]

Fundamental analysis refers to the use of the company's fundamentals to predict a stock price. [3] Fundamental analysis would consider the company's financial conditions, operations, and macroeconomic indicators.

The other philosophy, technical analysis, focuses on using historic information and time-series data to predict a stock's price.[3] Technical analysis operates under the belief prices follow historical trends. Since history tends to repeat itself, some researchers have tried to use historical prices to predict a stock's future performance. To discover these historical patterns, researchers have used Bayesian networks, auto-regressive models, moving average models, auto-regressive average models, etc.

While all of these models are valid approaches to predicting stock movement, they mostly fail to account for the psychology of investors in pricing the stock market. [3] As sentiment has been long known to be a measure of an individual's mood, opinion, and emotion, it may improve these models' predictive ability. A simple example of which would be the usage of a linear regression model for modeling historical pricing data in combination with sentiment scores derived from news articles.

The joint sentiment/topic model previously mentioned to asses text content's combined topic and sentiment was created with the intent of extracting topic and sentiment from movie reviews. Because the model simultaneously extracts topic and sentiment, it is an applicable analog to compare with the proposed topic-sentiment model by the author.

The authors, Nguyen, Shirai, and Velcin, extracted historical stock quotes from Yahoo finance for 18 large cap stocks. The data contained the following features: date, open, high, low, close, and adjusted close to account for splits and dividends. Since adjusted close prices are more representative in the long term, most researchers chose the adjusted close as the predicted feature.

The textual features were extracted from the Yahoo Finance Message Board, an online forum where investors post about company news, investing opinions, and comments. The extracted messages span a period of one year from July 23, 2012, to July 19, 2013. A feature of these comments was the annotations of buy, sell, or hold that 15 percent of messages had been annotated by forum users.

The text extracted from these forum messages contained errors similar to those in the previous study. [3] The text often contained uncommon grammar constitutions, misspellings, and was rather short. The text also faces an issue of validity, which is similar to the findings of the use of Twitter as a source for sentiment analysis to predict stock movement.

The authors used a support vector machine for classification of the stock's predicted movement. SVMs are known to efficiently handle high dimensional data and perform well at classification tasks. Therefore, the authors selected a linear kernelled SVM model to predict stock movement.

They used $price_{t-1}$, $price_{t-2}$, $Hsent_{i,t}$, $Hsent_{i,t-1}$ as the features for training the SVM. The $Hsent_{i,t}$ and $Hsent_{i,t-1}$ are the percentages of the messages extracted from the forums at time, t, that belong to the sentiment classes of strong buy, buy, hold, sell, and strong sell. As only about 15% of messages were user annotated, the researchers first trained a model to train a topic model to extract classes from the remaining 84.4% of messages. [3]

To extract those classes, they removed stopwords, then lemmatized the messages, and weighed the features using a TF-IDF, and finally chose a SVM with a linear kernel to classify the remaining message's sentiment classes.

After the models where trained, the aspect-based sentiment model, they proposed was the most performant, achieving an average accuracy of 54.41%. [3] This average was calculated from the average of the performance of the model on each individual stock. They found that the model was more performant for some stocks like 71.05% for Amazon and 64.47% for Dell. [3]

The aspect-based sentiment model outperformed the sentiment classification model by 2.54%, the LDA-based method by 2.14%, and the JST-based model by 2.87%. [3] They do state there are several limitations associated with their approach. For some of the individual stocks, the price-based time series model more accurately predicted the stock's prices. Another potential explanation to low performance could be due to the sentiments

of the individual's on the Yahoo Finance Messaging board not being clean enough or the individual's assessments of the market overall being faulty.

To build upon their research, the authors adding features to the model and fine-tuning the model. They presented a novel approach for capturing topics and sentiments together from investor's messages an investment related forum. This approach is similar to the approach I am going to use in my capstone project. I plan on following in their research by combining a sentiment metric with a topic class to improve the predictive accuracy of a stock pricing model. However, my approach diverges from this research article by focusing on an individual's social media posts instead of a mass of individual's messages to gauge public sentiment. Instead of being reactionary, my approach assumes the importance and gravity of the subject's opinions, emotions, and ideology.

2.4 Effect of Public Sentiment on Stock Market Movement Prediction During the COVID-19 Outbreak by Das, Sahukhan, Chatterjee, and Charkrabarti (2022)

Das, Sahukhan, Chatterjee, and Charkrabarti (2022) present a stock movement prediction model that aggregates sentiment data from four different online sources and integrates to sentiment data with historical stock data to predict a stock's future movement. The online text sources to gauge public sentiment are stock related news articles, user posts from twitter, financial news from the "Economic Times", and comments from Facebook. They then extracted sentiment using a process similar to a consensus model to ascertain an average sentiment score from seven tools. The tools are as follows: VADER, logistic regression, Loughran-McDonald, Henry, TextBlob, Linear SVC, and Stanford.

The authors reason that given the existing mass of literature on the applicability and efficacy of sentiment analysis tools, in this case opinion mining to identify the polarity of text sources' messaging relating to individual's opinions of the stock market.

Opinion mining was used by the authors to identify the polarity of the sentiments associated with the text sources regarding stocks. They used lexicon-based and machine learning-based tools to discern the public sentiment associated with the stock market.

The previous works researched in the prior research section seem to support the authors' assertions that the text sources of the media types of social media and news outlets are valid data sources for opinion mining regarding the public sentiment of the stock market. [2] [3] The authors' selection of dictionary and lexicon-based and ML-based sentiment analysis tools is also supported by the previous articles in the prior research section of this paper. [2] [3]

Regarding the sentiment analysis tools used in this model, compared to the researchers, the authors used lexicon-based and ML-based tools, rather than simply using pre-compiled and pre-computed lexicon-based tools. The dictionary-based tools were: Loughran-McDonald and Henry, The Lexicon-based tool was textblob. The ml-based tools used in their model were support vector machines (SVMs), Stanford, a large classifier neural network sentiment analysis tool, VADER, a rule-based classifier, probabilistic classifiers, and logistic regression.

Their model sourced online textual data from the before mentioned sources and applied the standard natural language processing pre-processing pipeline to clean the data. This included extracting the data from the web and applying the appropriate preprocessing steps per model, for example, lemmatizing the text to then apply VADER's .polarity_scores() [1] function. Then, the model applied each of the seven sentiment analysis tools to each entry and obtained an average sentiment score across the seven sentiment analysis tools. Finally, the average sentiment score in addition to the relevant stock market data were fed into a long short-term memory (LSTM) model to assess the performance and impact of public sentiment on stock movement.

[1] nltk.sentiment.vader.SentiText.polatity_scores() is a python function that returns the sentiment score of an input.

An LSTM model is applicable because stock data is inherently time-series data. [4] LTSMs are and are an advanced branch of the RNNs, that are efficient at processing time-series data. The model's performance was then measured using regression and classification metrics including r-squared, MSE, MAE, accuracy, recall, and F1.

The combination of features that resulted in the most performant model used sentiment scores applied to Facebook comments regarding stock information in addition to stock data. The linear SVC, VADER, and Loghran-McDonald sentiment analysis tools were found to be the most performant. They found that the linear SVC sentiment scores had a significant influence on the model's stock movement prediction accuracy. [4] The linear SVC achieved a 98.32% accuracy score in predicting stock market movement prediction. The highest prediction accuracy was achieved using average sentiment scores from Facebook comments from a linear SVC paired with stock data with an accuracy of 98.11%. [4]

Overall, it appears the combination of the sentiment mining of news headlines and historical stock quotes had a meaningful improvement to stock forecasting accuracy.

2.5 Using Twitter to Predict the Stock Market: Where is the mood effect? (2015)

The author notes, "the stock market can be driven by emotions of stock market participants". [5] This behavioral finance notion builds upon the notion that individual's sentiments, being their emotions, ideologies, and attitudes cause a meaningful effect on the stock market's prices and performance. This chapter analyzes the effect of mood states, mood contagion, and mood states scrapped from twitter and their effects on the German stock market in a behavioral finance perspective.

They used twitter as a proxy for approximating mood states. They point to evidence supporting the notion that "lead-users" exert influence on their followers, which in turn causes a reaction impacting their followers mood states, known as mood contagion. [5]

In behavioral finance, investors are humans who may be driven to make errors due to emotions. [5] The author points to evidence in market abnormalities that at least contradict how the market ought to behave under the efficient market hypothesis. For example, a prediction of stock market prices should not be possible, if each piece of information is continuously priced into the market.

Examples of abnormalities that may contradict the efficient market hypothesis include: calendar abnormalities and technical abnormalities. [5] Calendar abnormalities are time-based, like seasonal, trends that move stock prices. The authors point to the January effect as one such example. The January effect is a trend that states the average return in the month of January tends to be higher than other months of the year, however, a hypothesis that may explain this phenomenon may be tax-loss harvesting for the previous tax year. Technical abnormalities involve the use of historical trends to predict or explain future performance. An example cited is the momentum effect, where previous winners will continue to win, with the converse being true. This phenomenon directly counters the EMH as it would cause an inefficiency in the market.

Now, integrating investor's psychology into an explanation for stock market movements requires understanding two groups. The two groups are defined by the authors as rational arbitragers and noise traders. Rational arbitragers are investors who are well informed and are not prone to sentiment related trading errors.[5] Noise traders are the investors who seemingly irrationally give weight to non-fundamental related information; these traders are prone to rely on sentiment to make trades.[5]

Under the EMH, price increases are driven by the selling and buying of stocks between the rational arbitragers and the noise traders. An exogenous shock would drive noise traders to buy or sell and then rational arbitragers will then drive the prices back up to fundamental value.[5] The author points to an example where rational arbitragers may predict the noise, instead of following the fundamentals, which in turn contradicts the EMH. [5] The author believes, using the behavioral financial lens, better explains this phenomenon.

The author proposes a framework that explains the attribution of mood, its effect on risk attitude, and the effect of risk attitude on investment behavior. They argue a fluctuation in mood can cause an increase in risk attitude, which in turn increases an individual's willingness to invest in a riskier asset.

The next step is to understand the predictive value of social media. The author points to a study where overall sentiment measured from twitter had predictive value over the Dows Jones Industrial Average (DIJA). [5]

Emotional/mood contagion are contagious; even in environments where direct social interaction is absent, like social media interactions, emotional contagion may still occur. [5] The crux of this phenomenon can be observed in an example illustrated by the author where individuals on Facebook produced fewer positive status updates as the were subjected to an increasingly large volume of negatively charged posts on social media feeds. [5] They hypothesize a relationship between increased user mood states from social media and greater stock market returns.

They derived a weighted social mood index from:

$$WMSI = \frac{positive\ mood * followers}{(positive\ mood * followers) + (negative\ mood * followers)} \tag{2.1}$$

They found an increased WMSI compared to the previous day, tended to cause an increase in stock price and a decreased WMSI compared to the previous day tended to cause a decrease in stock price. This was confirmed as their trading strategy was to buy the DAX index on days where the WMSI increased and shorted the DAX on days where the WMSI decreased, which led to their portfolio outperforming the average weighted return over 6 months by at least 6%. [5] They conclude that the network structure of twitter should be considered when analyzing the relationship between mood levels and share returns.

2.6 Investor Sentiment in the Stock Market by Malcom Baker and Jeffery Wurgler (2007)

Baker and Wurdler (2007) build upon a theoretical framework that combines the notions that investors as individuals that are affected by sentiment and the macroeconomic investor sentiment approach proposed by them.

The investor sentiment approach they proposed is a top-down model, unlike the bottom-up approach used by previous researchers used by previous researchers to explain the effects of sentiment related mood states of individual investor's under-reaction or overreactions of past returns or fundamentals. [6] The top-down approach they presented is distinctly macroeconomic. They reason in behavioral finance, the bottom-up approach developed by other researchers, is unable to fully capture the psychology of individual investors. [6] For example, this approach fails to account for the variance of selected biases and trading frictions of individuals. Instead, the top-down approach aims to capture the aggregate sentiment of investors and traces the aggregate sentiment of investors as a whole to explain the overall market and individual stocks.

The top-down model is compatible with the irrefutable assumptions of behavioral finance; those assumptions being sentiment and the limits of arbitrage to explain why specific stocks are more affected by sentiment instead of a blanket notion that the overall market is affected by sentiment. [6]

There are two key ideas that come with this approach. The first is the possibility demand shocks may vary across firms due to sentiment, i.e. when sentiment increases, the speculative stocks are likely to have higher than average returns. The second is the stocks that are most difficult to price are the ones that are the most difficult to arbitrage, i.e. stocks that are most prone to sentiment-based demand shocks.

They propose a sentiment beta. This measure relates to the effect sentiment plays on the valuation of a stock. [6] For example, a stock with a sentiment beta greater than 1 will

indicate that a stock has a relatively higher market risk, with the converse being true. This means market returns are negatively related to changes in sentiment.

At the time of this paper's writing, sentiment analysis and opinion mining from textual sources to gauge public sentiment to ascertain a sentiment score were not as developed. Therefore, this paper makes do with sentiment proxies that derive a sentiment measure from data sources including momentum, retail investor trades, investor surveys, exogenous mood-related factors, and fundamental data. [6] The fundamental data included: dividend premiums to measure market confidence, closed-end fund discounts, IPO first day returns, short-term IPO volume, and the volume of equity issued over time.

The effect of this approach was measured as sentiment betas increase, stocks become more difficult for arbitrage. The found a significant correlation of 0.43 between the market indexes' returns and changes in their sentiment index. [6] Therefore, they assert the cause of investor sentiment is exogenous and has a quantifiable effect on the market.

Chapter 3: Data

In this project, 3 datasets were used. Two datasets come from Yahoo Finance, a leading media company that provides finance related news, including stock quotes. These two data sets contain stock movement data for the following index funds, the SP500, Standards and Poors 500, and IXIC, the NASDAQ 500. Yahoo fiance partners with trading exchanges and providers of financial market data to collect and provide historical stock quote data [7, 8]. The index funds' data sets contain the following features:

Column	*Comment*
Date	Date
Open	Opening price of the equity
High	Maximum price of the equity for the given day
Low	Minimum price of the equity for the given day
Close	The closing price of the equity

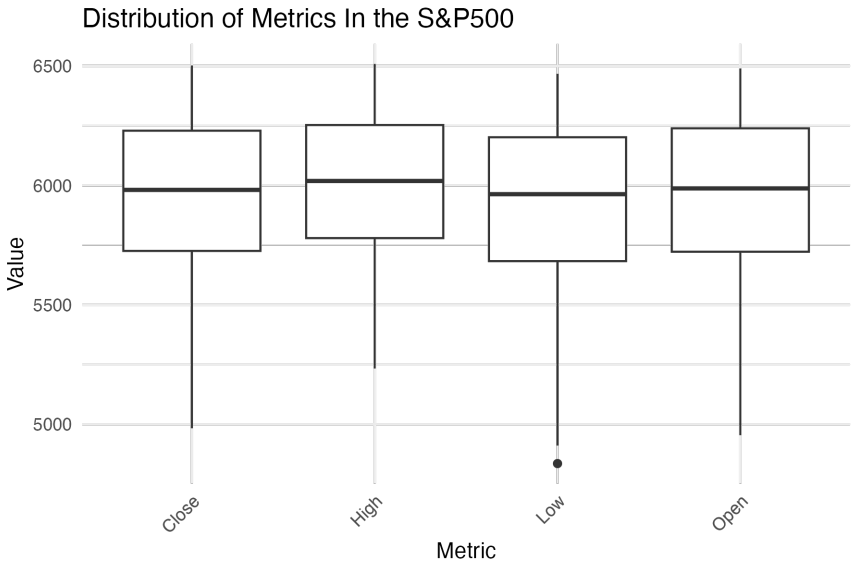

Figure 3.1: Box plots of S&P500 Numerical Features

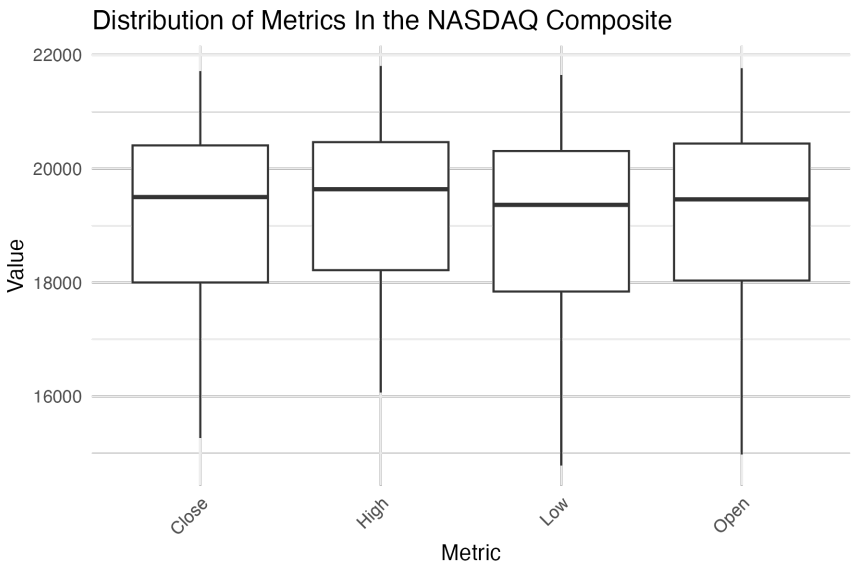

Figure 3.2: Box plots of NASDAQ Composite Numerical Features

The third data set used for this project is a web scrape of all of Donald J. Trump's posts on his social media platform, Truth social. This data includes his direct posts, re-posts, and his interactions with various truth social users. The web scrape was provided using an open sourced tool from the github user, stiles, who provides a scraped mirror of Donald Trump's posts for research purposes [9]. This dataset contains the following features:

Column	*Comment*
id	The unique identifier for the post
created at	Timestamp when the post was made
content	The text content of the post
url	Direct link to the post on Truth Social
media	An array of image and video URLs if the post contains media
replies count	Number of replies to Trump post
reblogs count	Number of re-posts, or re-truths, to Trump post
favourites count	Number of favorites to Trump post

As index funds typically aim to capture a diverse subset of the overall market, therefore increasing diversification, leading to reducing risk, while not reducing reward. While, the NASDAQ composite index fund had historically been seen and regarded as the more performant and safer fund, I found in this case the normalized closing price of the index funds performed very similarly in the time-frame of Trump's second presidency.

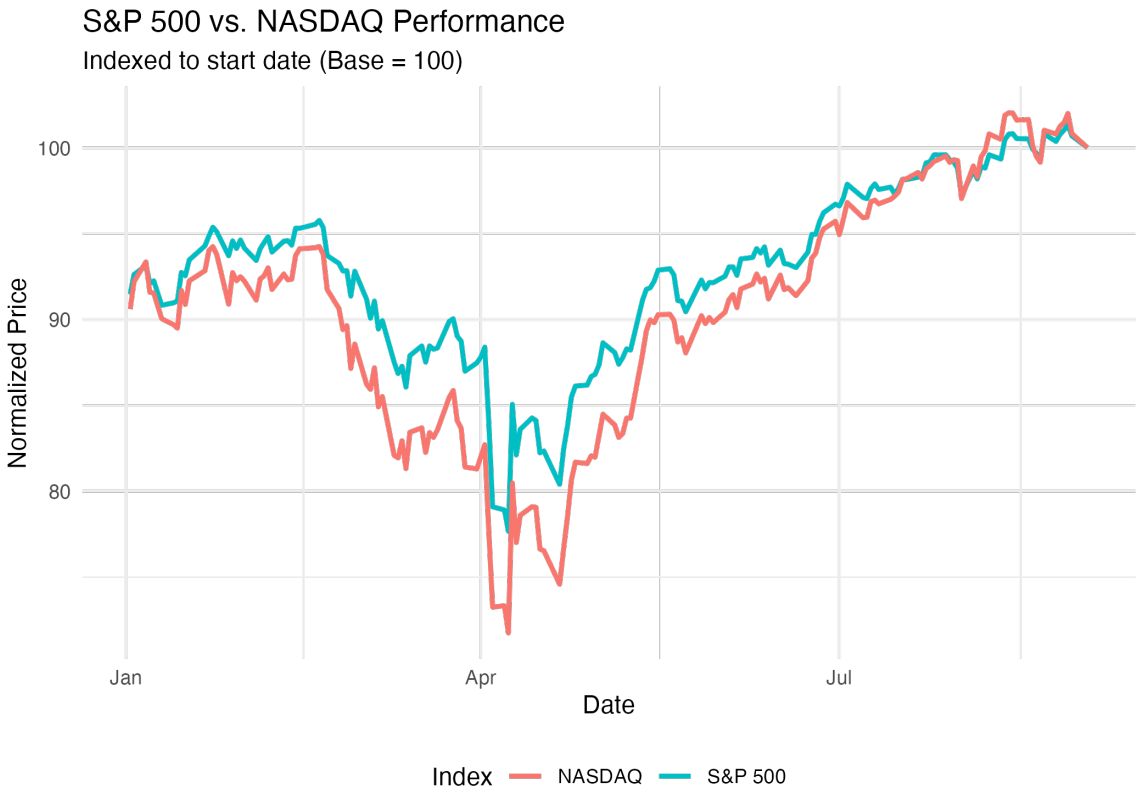

Figure 3.3: S&P 500 vs. NASDAQ Normalized Performance (Closing Price)

Another point in the similarity or the lack of meaningful differentiation in performance, between the two index funds from 01/01/2025 until 09/20/2025, to abstract the overall performance of the United States stock market is the relative similarity in performance between both indexes in both sets of hypothesis tests, as shown in the theory and results chapters.

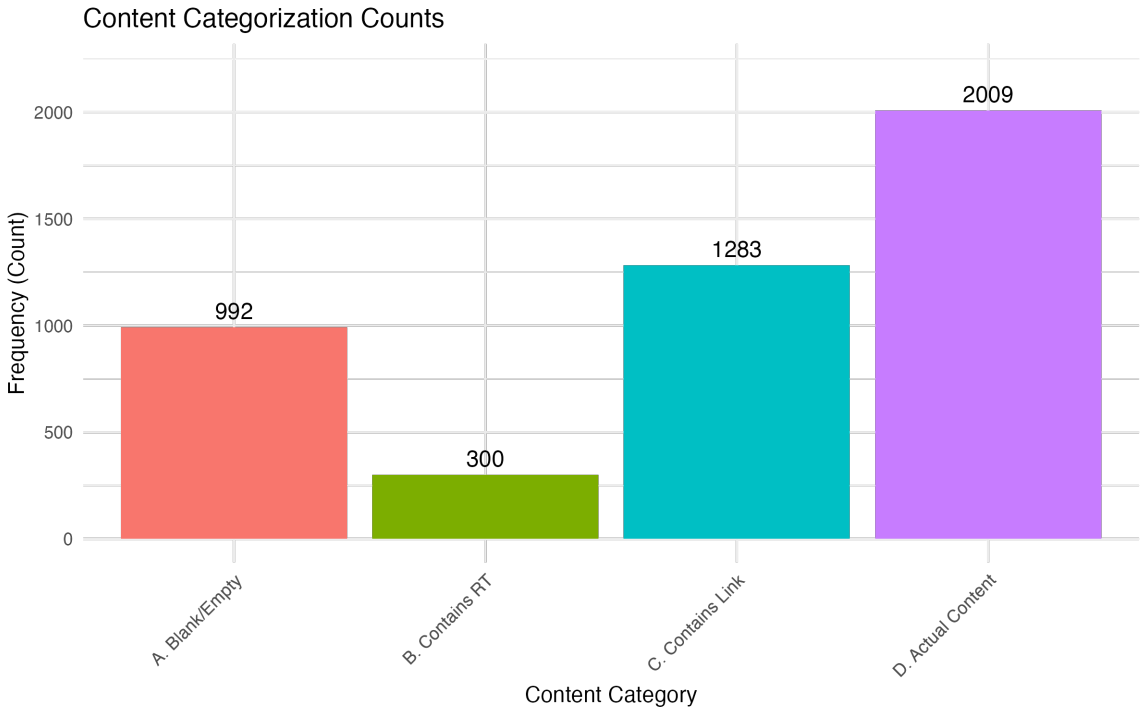

Figure 3.4: Counts of Various Data Quality Issues In Trump Truth Social Archive

While the index fund data was clean, the same is not the case for the Trump Truth Social Archive. The archive had empty rows without text content, retweets, and responses to other users in addition to Trump's direct posts during the time frame of his second presidency. This finding regarding the unprocessed nature of the Trump Truth Social Archive will require extensive cleaning, preprocessing, and filtering to yield usable and clean data for modeling.

The primary goal of this project is to explain the phenomena of the April stock market crash, which seems to be universally understood or partially attributed to Donald J. Trump's social media posts on tariffs and trade wars. To explain this, this project aims to either prove or disprove an empirical link between Donald J. Trump's posts on Truth Social and stock market movement.

Chapter 4: Theory

My original theory was that the sentiment of Donald Trump's posts would effect the stock market. I used index funds to abstract the stock market, in this case the S&P500 and NASDAQ Composite.

My null hypothesis for this initial theory was:

$$H_0 : \mu_{positive} = \mu_{neutral} = \mu_{negative} \tag{4.1}$$

In the initial stages of exploratory data analysis after cleaning and preprocessing Trump's posts dataset to a usable state, I applied a sentiment model to the textual data stored in content. Using the VADER lexicon to assign each post a sentiment score, we can then derive sentiment categories of positive, neutral, and negative with a threshold of any compound sentiment of 0.05 is positive, less than minus 0.05 is negative, and those falling between the threshold are considered neutral. An interesting finding here was the distinct lack of any difference in means for either index's performance, when measuring the sentiment category versus the closing price in a faceted box plot.

The VADER lexicon for sentiment analysis was created in by Hutto and Gilbert in 2014 [10]. The equation for deriving sentiment from text corpus is as follows:

$$Compound\_Score_i = \frac{x_i}{\sqrt{x_i^2 + a}} \tag{4.2}$$

[1] [10]

Following the failure to reject that null hypothesis, I pivoted to using topic modeling to extract the latent topics in the content of Donald Trump's posts.

In this new theory, I theorized that there may be a difference in the mean daily change, daily volatility, or closing price given a topic category.

My null hypothesis for this theory was that the sample means for each topic are the same:

$$H_0 : \mu_0 = ... = \mu_{i-1} = \mu_i \tag{4.3}$$

The Latent Dirichlet Allocation topic analysis model applied to machine learning was proposed by Blei, David M., and Ng, Andrew Y (2003). [11] The LDA topic model is a three-layered hierarchal Bayesian model, which is a generative probabilistic model that collects discrete terms to provide an explicit representation for each document. [11]

There is no singular equation for LDA topic modeling, instead Gibb's sampling or other algorithmic methods are used to identify latent structures. [11]

[1] x = the sum of all the sentiment valence scores as calculated by the VADER lexicon's heuristic rules. a = the normalization parameter; usually 15, which should match the maximum expected x. [10]

Chapter 5: Results

The results of my first hypothesis regarding the impact of sentiment categories derived from Donald J Trump's posts from Truth Social using the Vader Lexicon sentiment analyzer model on Stock Market Movement are as follows:

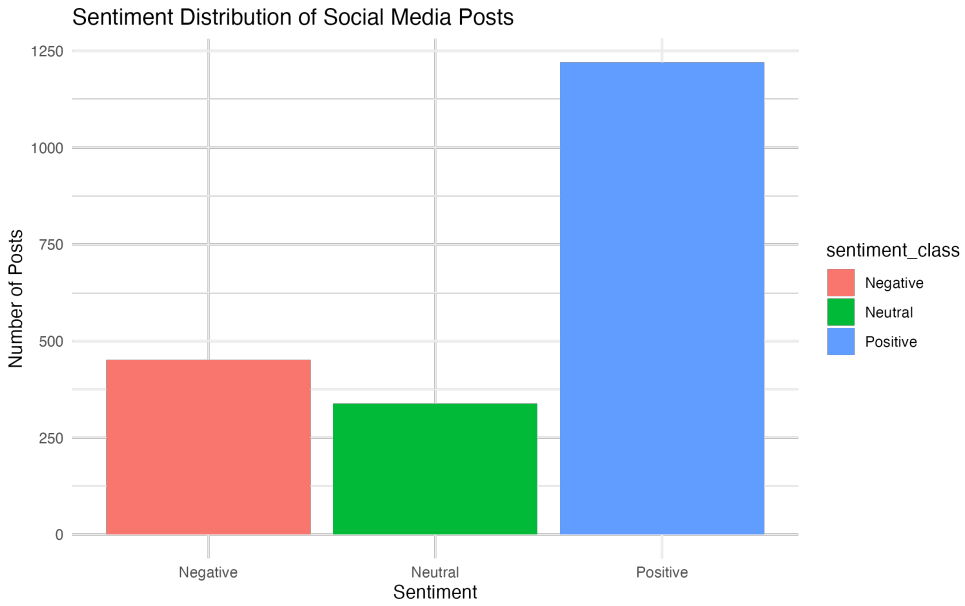

Figure 5.1: Sentiment Distribution of Social Media Posts

The distribution of sentiment by social media posts seems suspect. From manually sampling Donald Trump's posts, a majority of the posts would either be interpreted as negatively emotionally charged, meaning they seem to be written with a strong emotional or opinionated undertone, which does not seem to line up with this distribution of mostly positively labeled sentiments per post.

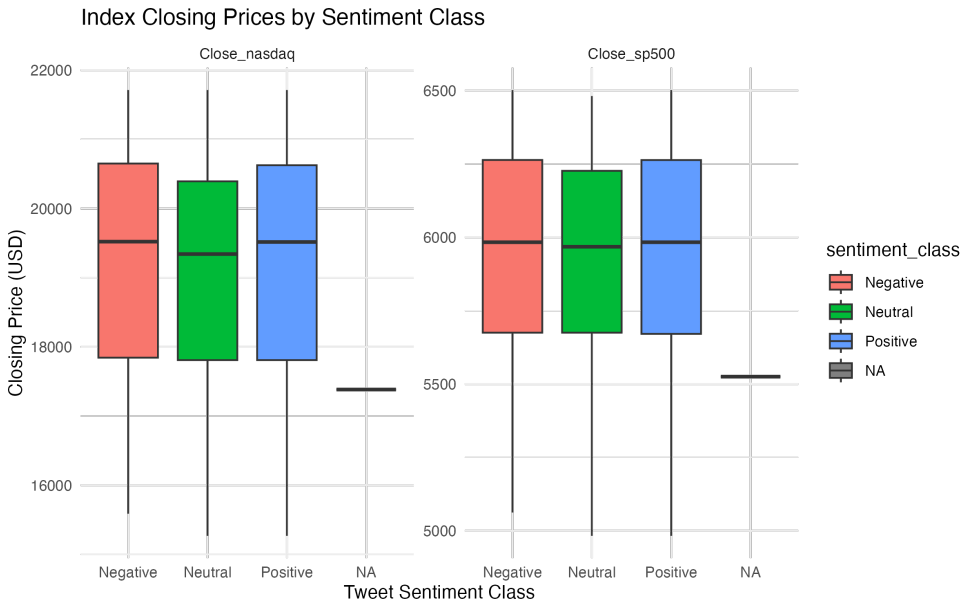

Figure 5.2: Index Closing Prices by Sentiment Class

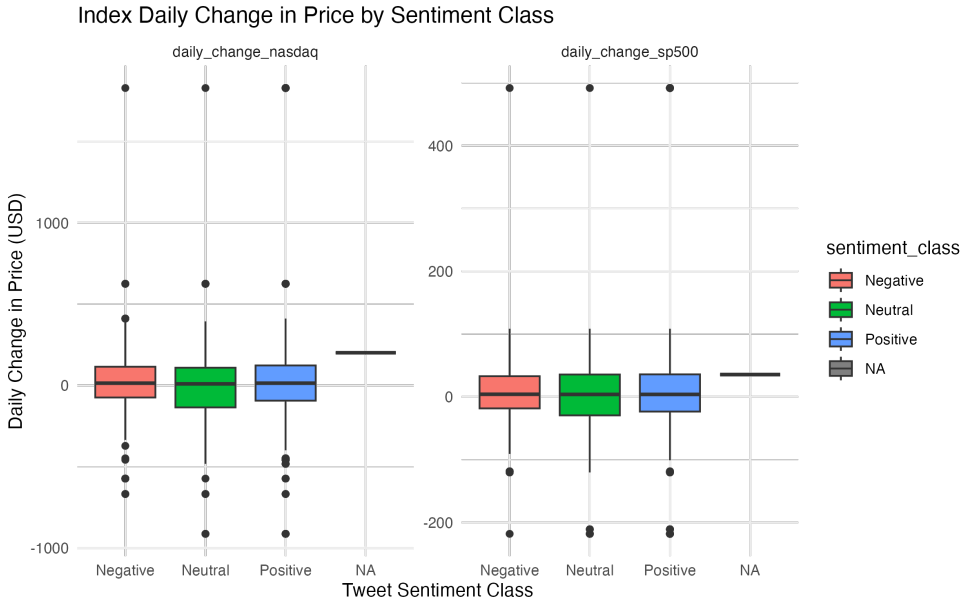

Figure 5.3: Index Daily Change in Price by Sentiment Class

Given that there was not a significant difference in means between the sentiment classes when measuring the sample means from closing_price and daily_change, we do not fit the necessary conditions to perform an ANOVA test.

Statistical Significance for the ANOVA test on Closing price of S&P500 Sentiment Category:

```
    Analysis of Variance Table


Response: Close_sp500
                  Df    Sum Sq Mean Sq F value Pr(>F)
sentiment_class    2    158986   79493  0.6511 0.5216
Residuals       1613 196938196  122094
```

Statistical Significance for the ANOVA test on the daily volatility of the price of S&P500 Sentiment Category

```
    Analysis of Variance Table


Response: daily_volatility_sp500
                  Df  Sum Sq Mean Sq F value Pr(>F)
sentiment_class    2    7886  3942.8  0.8504 0.4274
Residuals       1613 7478507  4636.4
```

Statistical Significance for the ANOVA test on Closing price of NASDAQ Composite Sentiment Category:

```
    Analysis of Variance Table


Response: Close_nasdaq
                  Df     Sum Sq Mean Sq F value Pr(>F)
sentiment_class    2    3233757 1616878  0.6429 0.5259
Residuals       1613 4056464149 2514857
```

Statistical Significance for the ANOVA test on the daily volatility of the price of NASDAQ Composite Sentiment Category

```
    Analysis of Variance Table


Response: daily_volatility_nasdaq
                  Df   Sum Sq Mean Sq F value Pr(>F)
sentiment_class    2   105530   52765  0.8919 0.4101
Residuals       1613 95423509   59159
```

As the criteria necessary to perform a single factor ANOVA were not met, meaning a significant difference in means was not fulfilled, the the p_values for any of these 4 ANOVA tests are not statistically significant. The p_value from none of these tests were less than the significance level of $\alpha = 0.05$, therefore there is not sufficient evidence to reject the null hypothesis.

The results of my second hypothesis on the impact of the topic categories derived from Donald J Trump's posts from Truth Social using the Latent Dirichlet Allocation Topic Model on Stock Market Movement are as follows:

In order to perform topic modeling using Latent Dirichlet Allocation, a document term matrix needs to be constructed from a text corpora. The most frequent terms in the Trump Truth Social Archive text corpus are:

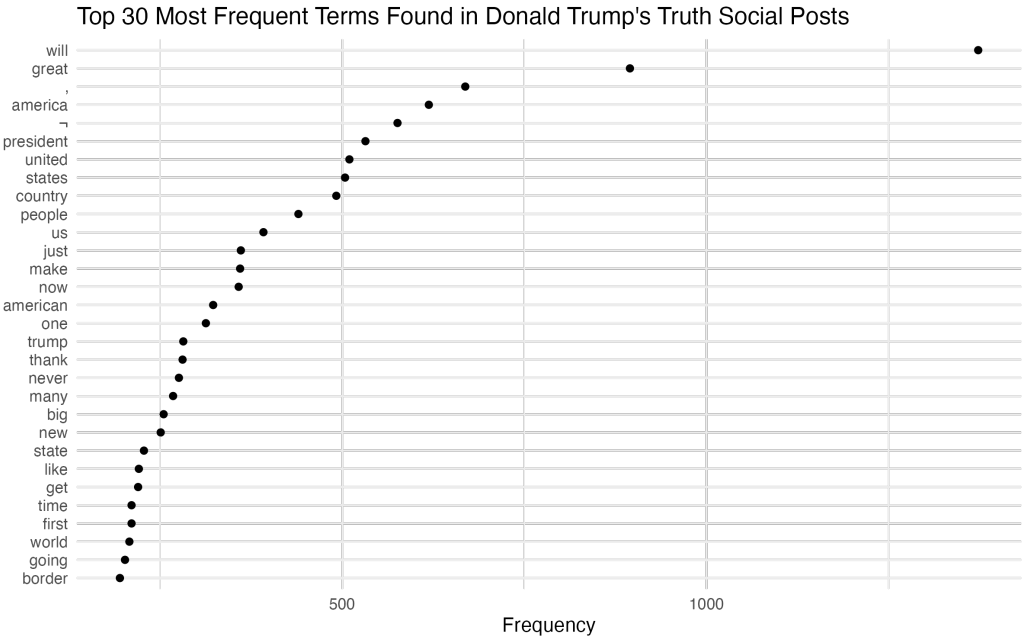

Figure 5.4: Top 30 Most Frequent Terms Found in Donald Trump's Truth Social Posts

Due to a double encoding error, there are some non ASCII characters that show up. This is likely due to the text in the Trump Truth Social Archive corpus being encoded in a non-standard uft encoding then re-encoded into uft-8. However, this is not a prevalent issue, nor will it cause issues, since these symbols are removed in a preprocessing step before applying the topic model.

The LDA topic model was applied using the following parameters $a = 15, k = 10, seed = 123$. The 10 topics found where:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
will	fake	will	great	president	great	will	will	secure	country
great	new	america	will	trump	will	states	president	will	biden
country	news	united	america	will	one	united	war	great	will
crime	york	tariffs	make	administration	big	announce	russia	complete	just
people	great	states	republican	including	beautiful	pleased	just	total	people
now	people	great	house	law	america	great	ukraine	endorsement	joe
america	bad	country	republicans	pro	bill	congratulations	iran	border	democrats
thank	case	countries	senate	donald	tax	secretary	peace	always	america
just	judge	trade	bill	justice	people	serve	united	let	states
states	even	china	election	bono	ever	district	great	second	election

Table 5.1: Top Terms per Topic

After the topics had been extracted using the unseeded LDA model, I compared the topics against the daily volatility of the S&P500. Below is the boxplot shoeing the spreads of daily volatility in the S&P500 across the discovered topics.

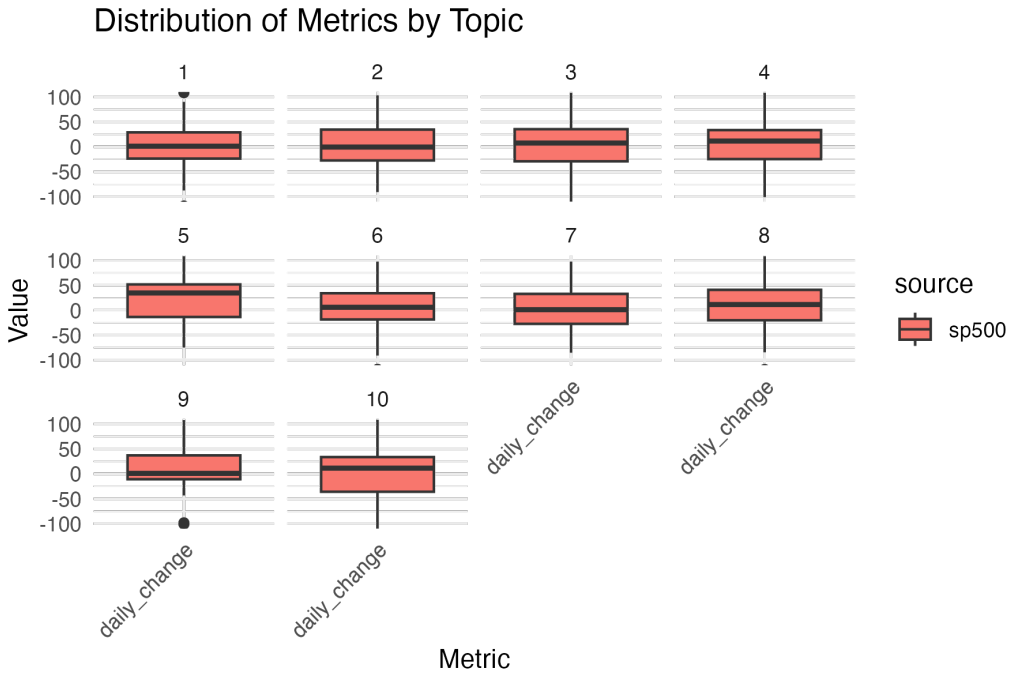

Figure 5.5: Spread of Daily Volatility in the S&P500 by Topic

Although the means did not appear to be distinct, i.e. the boxplots did not appear to have considerable differences in means, I still elected to concuct an ANOVA test. This ANOVA test is of the explanatory variable, topic, and the response variable, the daily volatility of the S&P500.

```
  Terms:
                  topic Residuals
Sum of Squares   151055   6063361
Deg. of Freedom       9      1265

Residual standard error: 69.23273
Estimated effects may be unbalanced
              Df  Sum Sq Mean Sq F value   Pr(>F)
topic          9  151055   16784   3.502 0.000272 ***
Residuals   1265 6063361    4793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] ""
```

The ANOVA test found a p value of 0.000272, which is very significant at a 5 percent significance level, meaning I can reject the null hypothesis of no difference in means between groups. I then conducted a post-hoc test using TukeyHSD. Below is the Tukey test:

My null hypothesis for the post-hoc test is:

$$H_0 : \mu_i = \mu_j \tag{5.1}$$

	diff	lwr	upr	pval
3-2	28.20	0.82	55.59	0.04
7-3	-33.48	-58.98	-7.98	0.00
8-3	-25.02	-49.28	-0.75	0.04
9-3	-33.83	-62.33	-5.33	0.01

Table 5.2: Post-Hoc Tukey LDA

The TukeyHSD found a signifcant difference between topic 3 and topics 2, 7, 8, and 9. From this result, having significant p values, I can reject the null hypothesis. I believe there is evidence to support the possibility that topic 3 has a tangible effect on the daily volatility of the S&P500.

I then tried a seeded LDA model where I defined 5 topics, immigration, rates, trade, war, and social in a dictionary. These topics where meant to model topics that Donald Trump frequently posted about that might move the market. The topics where seeded with as follows:

	IMIGRATION	RATES	TRADE	WAR	SOCIAL
1	aliens	bill	canada	defense	elon
2	border	budget	china	hostages	musk
3	borders	capital	india	ceasefire	cost
4	enforcement	inflation	negotiations	iran	cut
5	homeland	job	steel	israel	costs
6	illegal	jobs	tariff	military	cuts
7	illegally	jerome	tariffs	putin	cutting
8	illegals	powell	trade	russia	security
9	migrant	price	war	russian	social
10	mexico	prices	commerce	ukraine	veterans
11	patrol	rate		vladimir	
12		rates		zelenskyy	
13		tax			
14		treasury			
15		economy			
16		economic			
17		taxes			
18		fed			

Table 5.3: Seeded Topic Dictionary

I also allowed the seededLDA to discover 3 nondefined topics, leading to a total number of 8 topics for the seedLDA. The other parameters used in the seededLDA were: $batch\_size = 0.01, auto\_iter = TRUE, verbose = TRUE, residual = 3$. The distribution of topics is as follows:

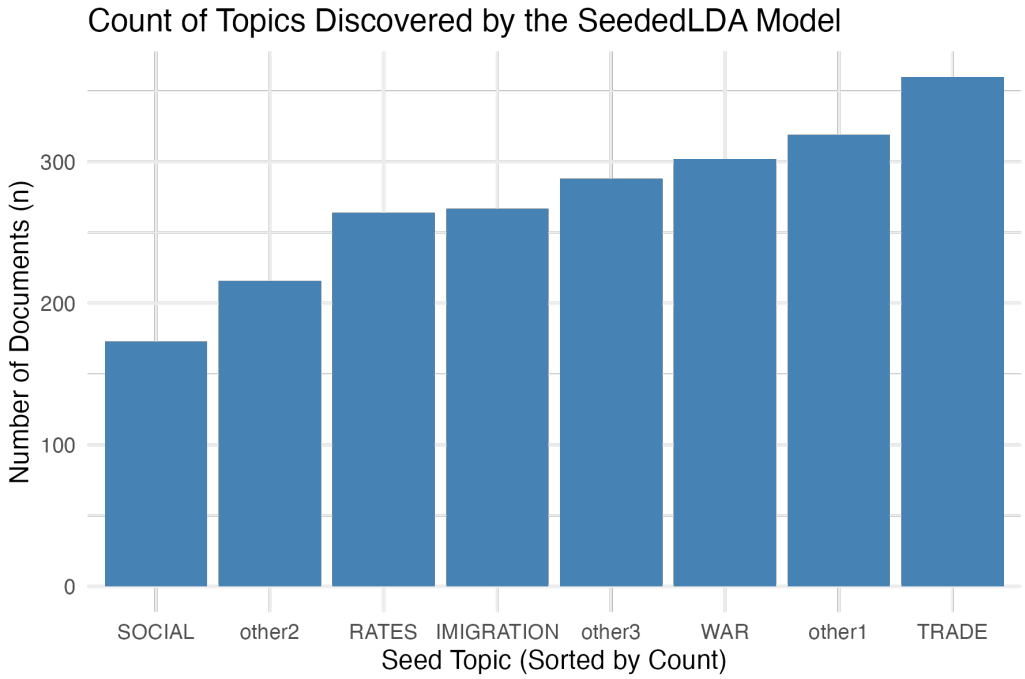

Figure 5.6: Topic Distribution of the Seeded LDA Topic Model

The next test visual was for testing the differences in means requirement for the ANOVA test. Once again, it is not easy to see the differences in means of the topics, but I still thought it pertainent to conduct an ANOVA test, given some of the means seemed like they could be different enough.

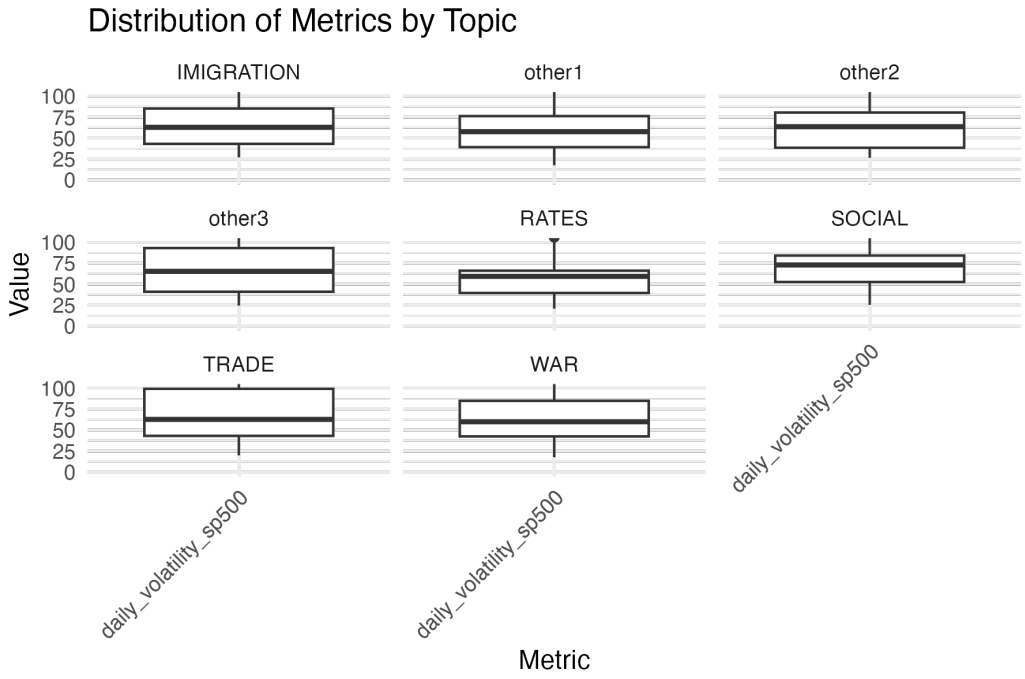

Figure 5.7: Spread of Daily Volatility in the S&P500 by Topic

I then conducted an ANOVA test on the explanatory variable, topic, and the response variable, the daily volatility of the S&P500, using the topics extracted from the seededLDA model.

```
  Terms:
                slda_topic Residuals
Sum of Squares       80064   4291314
Deg. of Freedom          7      1087

Residual standard error: 62.83193
Estimated effects may be unbalanced
              Df  Sum Sq Mean Sq F value  Pr(>F)
slda_topic     7   80064   11438   2.897 0.00528 **
Residuals   1087 4291314    3948
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] ""
[1] "Post-hoc test"
```

The ANOVA test found a p value of 0.00528, which is very significant at a 5 percent significance level, meaning I can reject the null hypothesis of no difference in means between groups. Although, this p value for the topics extracted from the SLDA topics was less signifcant than that of the LDA topics. I then conducted a post-hoc test using TukeyHSD. Below is the Tukey test:

My null hypothesis for the post-hoc test is:

$$H_0 : \mu_i = \mu_j \tag{5.2}$$

	diff	lwr	upr	pval
TRADE-other1	24.99	4.34	45.64	0.01
TRADE-other2	23.11	1.22	45.01	0.03
TRADE-other3	21.42	0.09	42.76	0.05
WAR-TRADE	-25.48	-46.77	-4.18	0.01

Table 5.4: Post-Hoc Tukey SLDA

From the TukeyHSD, we can reject the null hypothesis. We can see that Trump's posts about trade, trade wars, and tarrifs had significant differences in the mean daily volatility of the S&P500.

Chapter 6: Conclusion

Through this capstone project, I have delved deep into the financial and behavioral theoretical frameworks of the stock market. Namely, the effects of human psychology in the field of behavioral finance. The tests I conducted on the use of sentiment to predict market movement were unexpectedly unfruitful. There where previous studies that found corelations between sentiment catagories extracted from forms of textual media, like articles and social media posts about the markets, however, due to my singular focus on an individual's posts on an arguably smaller and potentially more niche social media platform and due to the relience on pretrained and precomputed lexicon based sentiment analyzers, I was unable to reproduce the previous studies' results.

This inability to reproduce the previous studies's results could also be down to a small dataset of about 3000 useable texts, of which were around 200 words or less. Another potential error may be due to the VADER lexicon having specifically been trained on written text compared to the tendency of social media users to type out what is more similar to what they may say in conversation.

The lack of differences in the means of the daily volatility of both stock indexes, given the sentiment class, could be due to the VADER lexicon not performing well on my textual dataset, but it could also be due to the markets being more interested in the substance of specific DJT social media posts rather than the sentiment.

Despite these issues in the sentiment analysis component of my research, I was able to come up with some interesting findings using topic modeling on the textual dataset.

In this stage of my research, I elected to only test against the S&P500, instead of both indexes. This is because, I was able to establish a similar normalized performance between the two over the tested time frame. The ANOVA Test of the 10 topics discovered from the LDA Topic model versus the daily volatility of the S&P500 found enough evidence to

reject the null hypothesis, that there was no significant difference in means. This gave me reason to conduct a post-hoc test using TukeyHSD. The post-hoc test found a significant difference between topic 3 when compared to topics 2, 7, 8, and 9. These topics are defined as follows:

Topic Number	*Topic Description*
Topic 2	Negative posts; keywords include: fake news, New York
Topic 3	Trade War; keywords include: China, tarrif, and trade
Topic 7	Positive posts; keywords include: United states, pleased, congratulations
Topic 8	War; keywords include: Russia, Ukraine, war, peace
Topic 9	Border control/imigration; keywords include: border, secure, second

Table 6.1: LDA Topic Definitions

Then I tried another LDA topic model, this time using a partially preseeded model. The preseeded topics, where topics of DJT posts that I thought had the potential to influence the market. This time the ANOVA test between the topics discovered by the seededLDA model versus the daily volatility of the S&P500 found enough evidence to reject the null hypothesis. From the post-hoc test, a significant difference between topic on trade wars when compared to topics other1, other2, other3, and actual war. These topics are defined as follows:

Topic/Number	Topic Description
Trade	Posts about tariffs and trade wars; keywords include: negotiations, tariff, tariffs, and China
Other1	Negative Posts; fake news, bad, never, don't
Other2	Positive Nationalism; US, administration, America, announce
Other3	Digs at Democrats; keywords include: dems, Joe Biden, left
War	Posts about actual war; keywords include: Russia, military, defense, and Ukraine.

Table 6.2: Seeded LDA Topic Definitions

These findings on the influence of the topic of DJT's social media posts, the substantive content of the posts, on the market are measurable. Unfortnately I was unable to chain together the influence of the sentiment and the topic. If thete were to be more research on the subject of this research, I would recommend compiling a custom lexicon based sentiment analyzer based on DJT and other politicians or influential public figure's social media posts and their stated sentiments. I would also recommend exploring other topic models like top2vec or text embending extraction.

Chapter 7: Software

The IDE used in this project was RStudio and the programming language used was GNU R.

The code for my analysis is as follows:

```
library(conflicted) #fix dplyr and stats conflicts
library(tidyverse, tidytext, reshape2)
#prefer dplyr library for filter() and lag()
conflicts_prefer(dplyr::filter(),dplyr::lag(),.quiet=TRUE)
library(readr) #for reading in .csv files with automatic encoding detection
library(lubridate) #used for date + time objects
library(TTR) #used for SMA
library(quanteda, quanteda.textstats) #used for NLP text preprocessing
library(udpipe, stringr)
library(vader) #used for sentiment analysis
library(tm, topicmodels, seededlda) #topic modeling
library(xtable, knitr, forcats)


sp500 <- read.csv('SPX.csv') %>%
  mutate(
    Date = mdy(Date), # mdy() is for Month-Day-Year format
    across(c(Open, High, Low, Close), parse_number)
  )
nasdaq <- read.csv('IXIC.csv') %>%
  mutate(
```

```
    Date = mdy(Date), # mdy() is for Month-Day-Year format
    across(c(Open, High, Low, Close), parse_number)
  )
ts <- read.csv('truth_archive.csv') %>%
  mutate(
    created_at = ymd_hms(created_at)
  )

#Filter 'truth social' for posts within the time frame of 01/01/2025 - 09/26/2026
ts$created_at <- ymd_hms(ts$created_at)
#define start and end dates of the time frame
start_date <- ymd("2025-01-01")
end_date <- ymd("2025-09-26")

#filter df
ts1 <- ts %>% filter(created_at >= start_date, created_at <= end_date)

### Barchart of TS1

#define the link and RT regex patterns
link_pattern <- "https?://\\S+|www\\.\\S+"
rt_pattern <- "^\\s*RT\\b"

df_counts <- ts1 %>%
  mutate(
    # First, trim whitespace to correctly identify blank/whitespace strings
    content_clean = trimws(content),
    # Categorize the content
```

```
    category = case_when(
      #filter following cases: blank or no content
      is.na(content_clean) | content_clean == "" ~ "A. Blank/Empty",
      str_detect(content_clean, regex(rt_pattern, ignore_case = TRUE)) ~ "B. Contains RT
      str_detect(content_clean, regex(link_pattern, ignore_case = TRUE)) ~ "C. Contains
      TRUE ~ "D. Actual Content"
    )
  ) %>%
  #count occurrences
  count(category, name = "count")

#filter df for blank content
ts1 <- ts1 %>% filter(trimws(content) != '')

#filter df for retweets and links due to retweeting
ts1 <- ts1 %>% filter(!str_detect(content, pattern = c("https|RT: |RT @")))
#S&P500
sp500 <- sp500 %>%
  mutate(
    #finance features
    daily_change = Close - Open, #add a change in price per day
    daily_volatility = High - Low,
    daily_return = (Close - lag(Close)) / lag(Close), #pct change day by day
    sma20 = SMA(Close, n = 20), #20 day simple moving avg

    #time based features
    day_of_week = wday(Date, label = TRUE, abbr = FALSE), #add a day of week label
    month = month(Date, label = TRUE, abbr = FALSE) #add a month label
  )
```

```
#add a movement feature
sp500 <- sp500 %>%
  mutate(
    Result = case_when(
      daily_change > 0 ~ "gain",
      daily_change < 0 ~ "loss",
      daily_change == 0 ~ "none",
      TRUE ~ NA_character_ #fallback for NA values
    )
  )

#NASDAQ
nasdaq <- nasdaq %>%
  mutate(
    #finance features
    daily_change = Close - Open, #add a change in price per day
    daily_volatility = High - Low,
    daily_return = (Close - lag(Close)) / lag(Close), #pct change day by day
    sma20 = SMA(Close, n = 20), #20 day simple moving avg

    #time based features
    day_of_week = wday(Date, label = TRUE, abbr = FALSE), #add a day of week label
    month = month(Date, label = TRUE, abbr = FALSE) #add a month label
  )

  #add a movement feature
nasdaq <- nasdaq %>%
```

```
  mutate(
    Result = case_when(
      daily_change > 0 ~ "gain",
      daily_change < 0 ~ "loss",
      daily_change == 0 ~ "none",
      TRUE ~ NA_character_ #fallback for NA values
    )
  )
ts1_vader_sentiment <- vader_df(ts1$content) # apply vader lexicon based sentiment analy
ts1 <- cbind(ts1, ts1_vader_sentiment) #join sentiment results back to original df

threshold_sent = 0.4

ts1 <- ts1 %>%
  mutate(sentiment_class = case_when(
    compound >= threshold_sent ~ "Positive",
    compound > -threshold_sent & compound < threshold_sent ~ "Neutral",
    compound <= -threshold_sent ~ "Negative",
    TRUE ~ NA_character_ # Handle any unexpected cases
  ))
vader_sentiment_counts <- ts1 %>%
  count(sentiment_class)

  #TS
ts1 <- ts1 %>% mutate(
  post_date = as_date(created_at),
  post_time = hms::as_hms(created_at)
)
```

```
ts1 <- ts1 %>%
  mutate(
    sentiment_date = if_else(
      post_time < hms::as_hms("16:30:00"),
      post_date,
      #if after 4:30 PM, calculate the next business day
      case_when(
        wday(post_date) == 6 ~ post_date + 3, # Friday -> Monday (+3)
        wday(post_date) == 7 ~ post_date + 2, # Saturday -> Monday (+2)
        TRUE ~ post_date + 1 # All other days -> next day (+1)
      )
    )
  )

indices <- sp500 %>%
  inner_join(nasdaq,
             by = 'Date',
             suffix = c('_sp500','_nasdaq')
  )
#Join indices with sentiment data
indices_sent <- indices %>%
  left_join(ts1, by=c('Date' =  'sentiment_date'))

indices_union <- bind_rows(
  "sp500" = sp500,
  "nasdaq" = nasdaq,
  .id = "source"
)
```

```
indices_long <- indices_union %>%
  pivot_longer(
    cols = where(is.numeric),
    names_to = "metric",
    values_to = "value"
)

#Normalize indices\$Close_i for i in {\_sp500, nasdaq}
indices_normalized <- indices %>%
  mutate(
    Close_sp500_normalized = (Close_sp500 / first(Close_sp500)) * 100,
    Close_nasdaq_normalized = (Close_nasdaq / first(Close_nasdaq)) * 100
  )

indices_sent_long <- indices_sent %>%
  pivot_longer(
    cols = c(Close_sp500, Close_nasdaq),
    names_to = "metric",
    values_to = "values"
  )

indices_sent_long <- indices_sent %>%
  pivot_longer(
    cols = c(daily_change_sp500, daily_change_nasdaq),
    names_to = "metric",
    values_to = "values"
  )
```

```
#ANOVA Test on sentiment vs closing price S&P500
sp500_sent_anova = lm(Close_sp500 ~ sentiment_class, data=indices_sent)
sp500_sent_anova <- anova(sp500_sent_anova)
sp500_sent_anova

#ANOVA Test on sentiment vs closing price NASDAQ Composite
nasdaq_sent_anova = lm(daily_volatility_nasdaq ~ sentiment_class, data=indices_sent)
nasdaq_sent_anova <- anova(nasdaq_sent_anova)
nasdaq_sent_anova

#Assign a 'doc_id' to ts1 to add back the topics after they have been discovered by topi
ts1$doc_id <- as.character(1:nrow(ts1))

#create vcorpus
trump_vcorpus <- VCorpus(VectorSource((ts1$content)))
#Cleaning
trump_vcorpus <- tm_map(trump_vcorpus, content_transformer(tolower))
trump_vcorpus <- tm_map(trump_vcorpus, removePunctuation) #remove punctuation
#trump_vcorpus <- tm_map(trump_vcorpus, removeNumbers)     #remove numbers
trump_vcorpus <- tm_map(trump_vcorpus, removeWords, tm::stopwords('english'))
#trump_vcorpus <- tm_map(trump_vcorpus, stripWhitespace)   #remove dupe white space
trump_qeda_corpus <- corpus(trump_vcorpus)
trump_dfm <- dfm(tokens(trump_qeda_corpus))

word_frequencies <- textstat_frequency(trump_dfm)
print(word_frequencies)

tokens_trump <- tokens(trump_qeda_corpus, remove_symbols = TRUE)
```

```
tstat_col_trump <- tokens_select(
  tokens_trump,
  pattern = "^[0-9A-Z]",
  valuetype = "regex",
  case_insensitive = TRUE,
  padding = TRUE
) %>% textstat_collocations(
  min_count = 10,
  size = 4
)
head(tstat_col_trump, 20)

trump_dtm = DocumentTermMatrix(trump_vcorpus)
inspect(trump_dtm)

set.seed(123) #set seed for reproducability
n_terms = 10 #n_terms
trump_lda <- LDA(
  trump_dtm,
  k = n_terms,
  control = list(seed = 123)
)

#get top n words for each topic
trump_topic_terms <- topicmodels::terms(trump_lda, n_terms)
print(trump_topic_terms)

#terms >>>> matrix >>>> data.frame (manually assign col names)
```

```
trump_topics_matrix <- as.matrix(trump_topic_terms)
trump_topics_df <- as.data.frame(trump_topics_matrix)
latex_table <- xtable(trump_topics_df, caption = "Top Terms per Topic", label = "tab:lda

topic_terms_tidy_manual <- trump_topics_matrix %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Rank") %>%
  tidyr::pivot_longer(
    cols = starts_with("Topic"),
    names_to = "Topic",
    values_to = "Term"
  )

topic_dists_trump <- tidy(
  trump_lda,
  matrix = "gamma"
)
trump_max_topics <- topic_dists_trump %>%
  group_by(document) %>%
  summarise(max_gamma = max(gamma)) %>%
  ungroup()

#get topic distribution for documents
trump_topic_dominant <- topic_dists_trump %>%
  group_by(document) %>%
  slice_max(gamma, n = 1) %>%
  ungroup()
```

```
#check if the left join has already occured
if (!("topic" %in% names(ts1))) {
  # If the column does not exist, run the left_join
  ts1 <- left_join(ts1, trump_topic_dominant, by = c('doc_id' = 'document'))
}

ts1_topic_indices2 <- indices_union %>% left_join(ts1,by=c("Date" = "sentiment_date"),re

ts1_topic_indices_long2 <- ts1_topic_indices2 %>% pivot_longer(
  cols = c("daily_change"),
  values_to = "values",
  names_to = "metric"
)

print("Anova test")
test2 <- aov(daily_volatility_sp500 ~ topic, data=ts1_topic_indices %>% filter(gamma > 0
print(test2)
print(summary(test2))
print("Post-hoc test")
posthoc_tukey <- as.data.frame(TukeyHSD(test2)$topic)
colnames(posthoc_tukey) <- c("diff", "lwr", "upr", "pval")
subset(posthoc_tukey, pval < 0.05)

seed_topic_dict <- dictionary(
  list(
    IMIGRATION = c('aliens', 'border', 'borders', 'enforcement', 'homeland', 'illegal',
    RATES = c('bill', 'budget', 'capital', 'inflation', 'job', 'jobs', 'jerome', 'powell
    TRADE = c('canada', 'china', 'india', 'negotiations', 'steel', 'tariff', 'tariffs',
    WAR = c('defense', 'hostages', 'ceasefire', 'iran', 'israel', 'military', 'putin', '
```

```
    SOCIAL = c('elon', 'musk', 'cost', 'cut', 'costs', 'cuts', 'cutting', 'security', 's
  )
)
set.seed(6969)
trump_seeded_lda <- textmodel_seededlda(trump_dfm,
                                        seed_topic_dict,
                                        batch_size = 0.01,
                                        auto_iter = TRUE,
                                        verbose = TRUE,
                                        residual = 3)

slda_theta <- data.frame(trump_seeded_lda$theta)
slda_theta$doc_id <- substring(rownames(slda_theta),5)
slda_topic_max <- slda_theta %>%
                      pivot_longer(
                        cols = where(is.numeric),
                        names_to = "slda_topic",
                        values_to = "slda_gamma"
                      ) %>%
                        group_by(doc_id) %>%
                        slice_max(slda_gamma, n = 1) %>%
                        ungroup()

if (!("slda_topic" %in% names(ts1_topic_indices))){
  ts1_topic_indices2 <- left_join(ts1_topic_indices, slda_topic_max, by = c("doc_id" = "
}

print("Anova test")
```

```
test_slda <- aov(daily_volatility_sp500 ~ slda_topic, data=ts1_topic_indices2 %>% filter
print(test_slda)
print(summary(test_slda))

print("")
print("Post-hoc test")
posthoc_tukey <- as.data.frame(TukeyHSD(test_slda)$slda_topic)
colnames(posthoc_tukey) <- c("diff", "lwr", "upr", "pval")
subset(posthoc_tukey, pval < 0.05)

p0 <- bar_plot <- ggplot(df_counts, aes(x = category, y = count, fill = category)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = count), vjust = -0.5, size = 4) +
  labs(
    title = "Content Categorization Counts",
    x = "Content Category",
    y = "Frequency (Count)"
  ) +
  theme_minimal() +
  scale_y_continuous(limits = c(0, max(df_counts$count) * 1.1)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

p1 <- ggplot(indices_long, aes(x = metric, y = value, fill = source)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Metrics by Index",
    x = "Metric",
    y = "Value",
    fill = "Source Index"
```

```
  ) +
  # Rotate axis labels for better readability
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#sp500
sp500_long <- sp500 %>%
  pivot_longer(
    cols = c('Open', 'High', 'Low', 'Close'),
    names_to = "metric",
    values_to = "value"
)

p2_1 <- ggplot(sp500_long, aes(x = metric, y = value)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Metrics In the S&P500",
    x = "Metric",
    y = "Value"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
plot(p2_1)

#nasdaq
nasdaq_long <- nasdaq %>%
  pivot_longer(
    cols = c('Open', 'High', 'Low', 'Close'),
```

```
    names_to = "metric",
    values_to = "value"
)

p3_1 <- ggplot(nasdaq_long, aes(x = metric, y = value)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Metrics In the NASDAQ Composite",
    x = "Metric",
    y = "Value"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
plot(p3_1)

p4 <- ggplot(data = indices_normalized, aes(x = Date)) +
  geom_line(aes(y = Close_sp500_normalized, color = "S&P 500"), linewidth = 1) +
  geom_line(aes(y = Close_nasdaq_normalized, color = "NASDAQ"), linewidth = 1) +
  labs(
    title = "S&P 500 vs. NASDAQ Performance",
    subtitle = "Indexed to start date (Base = 100)",
    y = "Normalized Price",
    x = "Date",
    color = "Index"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
plot(p4)
```

```
p8 <- ggplot(vader_sentiment_counts, aes(x = sentiment_class, y = n, fill = sentiment_cl
  geom_bar(stat = "identity") +
  labs(
    title = "Sentiment Distribution of Social Media Posts",
    x = "Sentiment",
    y = "Number of Posts"
  ) +
  theme_minimal()
plot(p8)

p10 <- ggplot(indices_sent_long, aes(x = sentiment_class, y = values, fill = sentiment_c
  geom_boxplot() +
  labs(
    title = "Index Closing Prices by Sentiment Class",
    x = "Tweet Sentiment Class",
    y = "Closing Price (USD)"
  ) +
  facet_wrap(~ metric, scales = "free_y") +
  theme_minimal()
plot(p10)

p11 <- ggplot(indices_sent_long, aes(x = sentiment_class, y = values, fill = sentiment_c
  geom_boxplot() +
  labs(

    title = "Index Daily Change in Price by Sentiment Class",
    x = "Tweet Sentiment Class",
```

```
    y = "Daily Change in Price (USD)"
  ) +
  facet_wrap(~ metric, scales = "free_y") +
  theme_minimal()
plot(p11)

p13 <- trump_dfm %>%
  textstat_frequency(n=30) %>%
  ggplot(aes(x=reorder(feature,frequency),y=frequency)) +
  geom_point() +
  coord_flip() +
  labs(
    title = "Top 30 Most Frequent Terms Found in Donald Trump's Truth Social Posts",
    x = NULL,
    y = "Frequency"
  ) +
  theme_minimal()
plot(p13)

col_trump_20 <- as.data.frame(head(tstat_col_trump, 20))

col_trump_20 %>% ggplot(aes(x=fct_reorder(collocation,count,.desc=TRUE), y=count)) +
  geom_col() +
  labs(title = "Barchart of Collocations from the Trump Truth Social Archive",
       x = "Collocation",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
ts1_topic_indices_long2_filtered <- ts1_topic_indices_long2 %>% filter(source == "sp500"

topic_sp500_boxplot <- ts1_topic_indices_long2_filtered %>% ggplot(aes(metric, values,fi
  geom_boxplot()+
  facet_wrap("topic")+
  ggplot2::coord_cartesian(ylim = c(-100, 100))+
  labs(
    title = "Distribution of Metrics by Topic",
    x = "Metric",
    y = "Value",
    fill = "source"
  ) +
  #rotate axis labels for better readability
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

slda_topic_dist <- slda_topic_max %>%
  ggplot(aes(x = reorder(slda_topic, slda_topic, FUN = length))) +
  geom_bar(fill = "steelblue") +
  labs(
    title = "Count of Topics Discovered by the SeededLDA Model",
    y = "Number of Documents (n)",
    x = "Seed Topic (Sorted by Count)"
  ) +
  theme_minimal()

ts1_topic_indices_long3 <- ts1_topic_indices2 %>% pivot_longer(
```

```
  #cols = c("daily_change_sp500", "daily_volatility_sp500","daily_change_nasdaq", "daily
  cols = c("daily_volatility_sp500"),
  values_to = "values",
  names_to = "metric"
)

slda_topic_sp500_boxplot <- ts1_topic_indices_long3 %>%
  filter(slda_gamma > 0.5) %>%
  ggplot(aes(metric, values))+
  geom_boxplot()+
  facet_wrap("slda_topic")+
  ggplot2::coord_cartesian(ylim = c(0, 100))+
  labs(
    title = "Distribution of Metrics by Topic",
    x = "Metric",
    y = "Value"
  ) +
  #rotate axis labels for better readability
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Bibliography

[1] David L. John, Bela Stantic, Tien Khoa Tran, Ualsher Tukayev, Tzung-Pei Hong, Bogdan Trawiński, Edward Szczerbicki, and Ngoc Thanh Nguyen. Machine learning or lexicon based sentiment analysis techniques on social media posts. In *Intelligent Information and Database Systems*, volume 13758 of *Lecture Notes in Computer Science*, pages 3–12. Springer, Switzerland, 2022.

[2] Dev Shah, Haruna Isah, and Farhana Zulkernine. Predicting the effects of news sentiments on the stock market, 2018.

[3] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert systems with applications*, 42(24):9603–9611, 2015.

[4] Nabanita Das, Bikash Sadhukhan, Tanusree Chatterjee, and Satyajit Chakrabarti. Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak. *Social Network Analysis and Mining*, 12(1):92, 2022.

[5] Michael. Nofer. The value of social media for predicting stock returns : Preconditions, instruments and performance analysis, 2015.

[6] Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152, Spring 2007.

[7] Historical stock quote for IXIC, 2025.

[8] Historical stock quote for GSPC, 2025.

[9] Matt Stiles. Trump-truth-social-archive. https://github.com/stiles/trump-truth-social-archive, 2025.

[10] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

Biography

Include your *biography* here detailing your background, education, and professional experience.