

# CDS-492: Capstone Science Article Fall 2025

Eric Wu



**Abstract**—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.

## CONTENTS

- 1 Introduction
- 2 Prior Research
- 3 Data
- 4 Theory
- 5 Results
- 6 Conclusion

## References

## 1 INTRODUCTION

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?

## 2 PRIOR RESEARCH

As the landscape of stock market movement has evolved, the use of sentiment betas has become increasingly important [1] [2]. Sentiment scores and the derived sentiment categories have been oft used by financial analysts and researchers in stock market movement prediction models sentiment [3] [4] [5] [6]. The use of techniques have advanced as technology and text mediums have evolved. Social media, online news sources, and online forums have become valuable sources of data to derive sentiment [3] [4] [5] [6]. Studies have used dictionary and lexicon based models to derive sentiment from corpora and been shown to be performant on news headlines and article and on individual's social media posts [3] [5]. The use of topic modeling and joint sentiment-topic models have become popular as a predictor in stock price movement models [7].

## 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 IXC, the NASDAQ 500. Yahoo finance partners with trading exchanges and providers of financial market data to collect and provide historical stock quote data [8] [9].

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 [10].

The Yahoo Finance finance datasets did not require extensive cleansing. I derived the following features for the index datasets: daily volatility, daily change, and daily return.

The effect of the news cycle, or in this case the "Trump post cycle", required a mapping were if a post was posted before or during the open hours of the stock market, the post would be attributed to the same day. Otherwise, if the post was posted after the closing time of the market, then the post would be attributed to the following day.

The truth social dataset was dirty and required filtering for blank content, removal of miss-encoded symbols, and removal of posts including links or reposts of other posts. The textual content of the truth social data set was then

conveted to a corpus to clean the text in the remaining posts. Following this cleansing, a sentiment analyzer using the VADER lexicon, a LDA topic modeler, and seededLDA topic modeler where applied to the corpus to assign sentiment and topic catagories to Trump's posts.

#### 4 THEORY

The original hypothesis 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:

$$H_0 : \mu_{positive} = \mu_{neutral} = \mu_{negative} \quad (1)$$

My second hypothesis was that the topic of Donald Trump's posts would effect the stock market. I used index funds to abstract the stock market, in this case, I only used the S&P500 index, because the performance of both indexes was similar over the timeframe of Trump's second presidency.

$$H_0 : \mu_{topic_i} = \mu_{topic_{i+1}} = \mu_{topic_n} \quad (2)$$

#### 5 RESULTS

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. The following ANOVA test serves to illustrate the lack of significance sentiment classes had on daily\_volatility at a 5 percent significance level:

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
```

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.

```
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
```

The TukeyHSD found a significant 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.

	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 1  
Post-Hoc Tukey LDA

1

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. 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. 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
```

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:

	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 2  
Post-Hoc Tukey SLDA

2

From the TukeyHSD, we find that Trump's posts about trade, trade wars, and tarriifs had significant differences in the mean daily volatility of the S&P500.

#### 6 CONCLUSION

The tests I conducted on the use of sentiment to predict market movement were unexpectedly unfruitful. There where

1. The topics are descrived as: Topic 2: Negative posts; Topic 3: Trade War; Topic 7: Positive posts; Topic 8: War; Topic 9: Border control/immigration

2. The topics are descrived as: Trade: Posts about tariffs and trade wars; Other1: Negative Posts; Other2: Positive Nationalism; Other3: Digs at Democrats; War: Posts about actual war

previous studies that found correlations between sentiment categories 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 reliance 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.

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 findings on the influence of the topic of DJT's social media posts, the substantive content of the posts, on the market are measurable. Unfortunately I was unable to chain together the influence of the sentiment and the topic. If there 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.

## REFERENCES

- [1] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *Journal of Economic Perspectives*, vol. 21, no. 2, pp. 129–152, Spring 2007. [Online]. Available: <https://ideas.repec.org/a/aea/jecper/v21y2007i2p129-152.html> 1
- [2] M. Nofer, "The value of social media for predicting stock returns : Preconditions, instruments and performance analysis," Wiesbaden, 2015. 1
- [3] D. L. John, B. Stantic, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki, and N. T. Nguyen, "Machine learning or lexicon based sentiment analysis techniques on social media posts," in *Intelligent Information and Database Systems*, ser. Lecture Notes in Computer Science. Switzerland: Springer, 2022, vol. 13758, pp. 3–12. 1
- [4] D. Shah, H. Isah, and F. Zulkernine, "Predicting the effects of news sentiments on the stock market," pp. 4705–4708, 2018. 1
- [5] N. Das, B. Sadhukhan, T. Chatterjee, and S. Chakrabarti, "Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 92, 2022. [Online]. Available: <https://doi.org/10.1007/s13278-022-00919-3> 1
- [6] C. Hutto and E. 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*, vol. 8, no. 1, pp. 216–225, May 2014. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550> 1
- [7] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert systems with applications*, vol. 42, no. 24, pp. 9603–9611, 2015. 1
- [8] "Historical stock quote for IXC," *Commodity Systems, Inc.*, 2025. [Online]. Available: <https://finance.yahoo.com/quote/IXIC/history/?period1=1704067200&period2=1761190431> 1
- [9] "Historical stock quote for GSPC," *Commodity Systems, Inc.*, 2025. [Online]. Available: <https://finance.yahoo.com/quote/GSPC/history/?period1=1704067200&period2=1761190302> 1
- [10] M. Stiles, "Trump-truth-social-archive," <https://github.com/stiles/trump-truth-social-archive>, 2025. 1