Analysis of Misinformation in Stock Data between 2015-2025

Gerardo Carrera, Maanas Lalwani, Garrett Power, and Ezekiel Suarez

Arizona State University, Tempe, AZ 85281, USA

May 15, 2025

1 Abstract

The increase in misinformation that coincides with our continued expansion into a digital age poses significant risk to any potential investors that use digital media to make financial decisions. This project aims to investigate how misinformation in the media may have impacted stock market behavior for the S&P500. Some real-world events are highlighted such as the significance of COVID-19's impact on an especially volatile year for the market in 2020, along with the 2021 surge in GameStop valuation and a fake White House tweet from 2023. Through this analysis, we hope to examine how misinformation during these stock market events could spur volatile events such as market disruptions, anomalies, or investor overreactions. Machine learning models and natural language processing (NLP) are leveraged with sentiment analysis. Specifically, we utilize transformer-based methods such as the BERT language model to classify content associated with stock market events (2015-2025) as misinformation or legitimate information. We web scraped as much data as we could at the time resulting in a dataset of around 2,200 records that combined both stock market events, media information, and any additional calculated financial metrics or label given by our misinformation or sentiment classification. Our findings do reinforce some insights, such as how misinformation seems to be present during many volatile stock market events. We also detail challenges to our current methodology that were presented to us during our process, such as our focus on the most volatile market events causing any analysis between volatility and the uniformly distributed misinformation classes to be moot. Challenges remain but seem to be unsurprising, future research utilizing methods such as transformer-based models will definitely be able to capture more nuanced implications and meaningful results.

2 Introduction & Background

2.1 Introduction

Misinformation can be described as the distribution of falsified information with the intent to deceive. Misinformation has played a crucial role in shaping financial markets, influencing investor sentiment, market volatility, and stock valuation. While the spread of misinformation has been utilized throughout history, seen as early as Octavian's propaganda campaign against Mark Antony around 44 BC [40], its rapid proliferation through digital media in the 21st century has heightened concerns about its impact on financial stability. Notable research indicates that misinformation does contribute to stock market volatility, especially during periods of uncertainty such as economic crises or during major global events like COVID-19 [8, 38, 15]. Misinformation has also played a role in stock market anomalies, with financial narratives fabricated purely to mislead traders, driving speculative bubbles and panic sell-offs [5, 27]. Some research has demonstrated that misinformation-driven trading has led to market deficiency and potential mispricing.[10, 47]. This digital era's increased reliance on social media and online news as primary sources for financial information has made financial markets much more susceptible to these misleading narratives, further amplifying the effects

of misinformation [28, 9, 47].

A consistent finding amongst researchers is that misinformation affects both market behavior and investor sentiment. Several analyses have suggested that misinformation influences not only short-term stock price fluctuations, but also a broader risk exposure [8, 15, 26]. Additionally, misinformation has been found to be more influential in developed markets, where investor reactions to false narratives are stronger due to the speed at which information spreads [26, 16]. Some recent studies also emphasize that misinformation's influence is not solely tied to fake news headlines, but also to a broader landscape of unverified financial information. Integrating reliable news sources can sometimes mitigate these effects of misinformation, but a key challenge remains in distinguishing credible information from the deceptive content [38, 40, 16]. This persistent challenge by misinformation highlights the need for improved detection and mitigation strategies in order to protect the populace and stock market integrity.

2.2 Background

The stock market has always been influenced by sentiment-driven movements, but even more so as investing becomes more accessible. Social media and news articles have been falsely leading these new investors into bad decisions, especially over the last decade. Fake news and social media manipulation has been shown to affect markets in greater ways than legitimate earnings reports or trends. These distort prices, momentarily leading to sharp rises and falls. For example, the SEC had a crackdown in 2017 on Lidingo Holdings, a company that was paid more than $1 million over the course of 2011-2014 to write hundreds of fake news articles. People would buy thousands of shares of a stock before paying for the release of the article and then sell the stocks the following day for a large profit [11]. Emotionally charged headlines of stock-related articles tend to be misleading, which emphasizes the need for misinformation detection to improve sentiment-based stock market prediction models [3, 11, 36].

Machine learning and sentiment analysis are not new concepts in the world of finance; models have been made to filter out misleading news before. In 2017, at University of Alberta, Canada, Dr. Golmohammadi and Dr. Zaiane built an anomaly detection framework using Twitter data to help remove 28% false positives in the detection of stock market manipulation [21]. Stocks with a presence on social media have been found to have stronger movements and are more susceptible to misinformation-driven market activity [34]. Having a model that fact-checks and recognizes linguistic patterns can improve reliability of stock prediction models.

The cryptocurrency market, which is particularly volatile and misinformation-driven, has been the focus of several studies on fraud. "Pump-and-dump" schemes have been driven by online forums such as Reddit's WallStreetBets (WSB). These are cases where a large group of people buy up a cheap stock and proceed to sell once the stock is artificially inflated. These sudden price spikes make some people rich, but usually the sharp crashes that follow affect more people. A substantial example of WSB utilizing sentiment manipulation occurred in 2021, when the group bought up a large amount of a dying GameStop (ticker: GME) stock and cost those on Wall Street who had shorted the stock billions [35, 41]. Real-time filtering of misinformation could prevent dramatic movements, therefore researchers have started using innovative neural network models to detect these pump-and-dump schemes [37].

Natural Language Processing (NLP) is a critical part of misinformation detection. Filtering out misleading Twitter sentiment and using Granger causality tests improves predictive accuracy [46]. Khedr et al. brought their model accuracy from 73% to 86% using naïve Bayes algorithm and after combining this sentiment analysis with historical stock prices, the accuracy rose to nearly 90% [31]. If done correctly and optimally, an extremely accurate stock market prediction model can be built.

However, to build an accurate stock market prediction label, we must first understand how misinformation started and its origins. Misinformation has been around since the dawn of man, but only recently started to become a widespread issue due to its ability to spread rapidly and easily in the digital age. [1, 12]

Misinformation has become a global challenge in recent years and many equate it to the same level as climate change. [24] It's also important to consider how the issue of misinformation is evolving. During 2016, misinformation mainly consisted of false stories and conspiracy theories that spread easily because algorithms distribute content to a wide audience. [2] Today the issue is far more complex, with many seeing the rise of more difficult-to-discern misinformation using tools such as deep fakes, digital forgery, or modifying the video and audio of real events to fit a certain narrative. [1]

The rise in misinformation appears to have occurred in the midst of the 2016 US election. [1, 24, 2, 22] In the months leading up to the election, between June and the day of the election, up to 500,000 articles of misinformation were published. [22] After 2016, interest in fake news increased to a level never seen before and led to misinformation becoming more popular than actual news. [22] An example of this was seen when the top 20 fake news stories generated more Facebook engagement than the top 20 election stories from major news outlets. [1] It is believed that the main reason for this proliferation of misinformation is simply the demand for it. [22] Consumers were looking for other ways to receive their news that was more engaging and compelling, and those who created misinformation did just that. Producers of misinformation often create misinformation to maximize engagement, which in turn increases advertising revenue. [22]

This increased attention to misinformation led to many asking the question of whether it affected market volatility. One of the first studies to answer this question was conducted immediately after the 2016 US election. It found that fake news had a constant negative impact on market volatility. [23] To be more specific, the results showed that on the days when disinformation was most widely shared in favor of Hillary Clinton, the market variance decreased significantly. [23] It is suggested that this was due to confirmation bias as Hillary was expected to win, therefore this fake news only further confirmed that assertion. [23] It should be noted however that the study acknowledges that its findings are a modest first step in this area and the question of how disinformation affects market volatility is not resolved yet.

Since this first study, many more studies have been conducted and they have found a lot of evidence pointing to the contrary. A study conducted in 2024 states that in instances where misinformation crosses into the realm of business and finance, it can lead to the loss of billions in seconds. [30] This was demonstrated particularly in 2013 when a fake tweet about an explosion in the White House led to a $130 billion loss in a matter of seconds.[45, 32] Investors can also superficially increase market share due to misinformation as in 2014 when Cynk Technology Corp's stock price increased 36000% over a few weeks due to fake discussions from bots.[32] It appears that misinformation does play a part in market volatility, but particularly in the short term as once the information is proved incorrect, markets stabilize. [30] However, a difference in how long it takes for stock to stabilize depends on whether misinformation was negative or positive. Negative responses faded in a week while positive responses faded in a day. [30] Despite misinformation generally only having a temporary effect on the stock market, many are still ringing the alarm bells as any misinformation can lead to conflicting opinions among investors. [30, 44] Over time, this can lead to a loss of trust in companies and increase market volatility. [30, 44] This is a large issue as once investors begin to make decisions based on information that may or may not be true [30], much of our current methods of understanding the stock market such as sentiment analysis may perform worse. We could also lose the ability to make accurate market predictions.

The stock market has always been influenced by sentiment-driven movements, but even more so as investing more accessible. The rise of social bots further complicates the spread of misinformation [17], and these distorted prices momentarily lead to sharp rises and falls. The SEC had a crackdown in 2017 on Lidingo Holdings, a company that was paid more than $1 million over the course of 2011-14 to write hundreds of fake news articles. People would buy thousands of shares of a stock before paying for the release of the article and then sell the stocks the following day for a large profit. Emotionally charged headlines of stock-related articles tend to be misleading, which emphasizes the need for misinformation detection to improve sentiment-based stock market prediction models. The ability to use and identify whether the infor-

mation is useful is a key asset for building an accurate view. Antweiler and Frank [4] explore the complexities.

Machine learning and sentiment analysis are not new concepts in the world of finance; models have been made to filter out misleading news before. It is shown that by using the right amount of information with the right signals, stock market activity and financial gains improve. This in turn impacts reliability of stock prediction models. Understanding investor psychology is key to understanding these market shifts [25].

The cryptocurrency market, which is particularly volatile and misinformation-driven, has been the focus of several studies on fraud. The behavorial aspects of financial decision-making, as described by Barberis and Thaler [7], highlight how cognitive biases and emotional factors can lead investors to make suboptimal choices. Moreover, during recessions, the effects are even more dramatic. As identified by Garcia [20] during economic recessions, this shows how negative information can affect financial markets.

To build an accurate stock market prediction label, we must first understand how misinformation started and its origins. Del Vicario et al. [13] shows that misinformation is easily spread which leads consumers needing to find new ways of filtering for good news. A large reason that these behaviors occur is due to consumers and their own expectations. There is often confirmation bias which leads them to pursue this data. Narrative economics is a major driver in this economic sector, as noted by Shiller [43].

This increased attention to misinformation led to many asking the question of whether it affected market volatility. It should be noted however that the study acknowledges that its findings are a modest first step in this area and the question of how disinformation affects market volatility is not resolved yet. The information can also be skewed to have a negative impact, which highlights why investors need to be careful, particularly due to sentiment in these markets [6]. Lazer et al. [33] shows the effects of misinformation and fake news. Moreover, behavioral traits that many have show that the spread of misinformation has a negative effect on the overall population and their investments.

Overall, the dynamics between the narratives and information drives market volatility, further reinforcing the need to study these dynamics. It is important that individuals do their own research and avoid "Thinking, Fast and Slow", instead try to be calculated with your decisions [29].

2.3 Methodologies from Literature

Machine learning models have been widely used in misinformation research and remain the most effective tools, aiming to improve accuracy and robustness for identifying misleading financial narratives. Natural Language Processing (NLP) techniques, including processes such as sentiment analysis and topic modeling, have been applied to analysis on financial news and social media discussions in order to assess their impact on stock markets [19, 14, 16]. More advanced models, such as BERT (Bidirectional Encoder Representations from Transformers) and other transformer-based architectures, have been employed to effectively categorize unstructured text and extract reliability-weighted information, thus helping quantify the severity and extent of misinformation within financial discourse [16].

One of the key challenges in misinformation detection is the difficulty of accurately quantifying and verifying misleading content. There is a lack of standardized metrics, and the asymmetric nature of misinformation complicates this evaluation on its impact within financial market dynamics [40, 16]. The continued development of structured misinformation detection frameworks has attempted to address this issue by transforming financial text into structured, comparable datasets for analysis [16]. However, research suggests that misinformation remains a persistent problem, especially during major corporate events (CE's), where deceptive financial narratives can significantly influence investor behavior [16]. Recent financial misinformation events in the U.S. underscore this necessity for more robust detection frameworks. There remains limitations in detecting nuanced misinformation, particularly when fabricated content is mixed with partial truths or biased reporting [14, 40].

During the COVID-19 pandemic, misinformation was linked to extreme market reactions, with misleading narratives affecting financial stability across multiple market sectors [26, 16]. There is strong evidence linking misinformation sentiment to both common and extreme market behavior, with studies demonstrating that increased misinformation-related sentiment corresponds to higher market volatility and lower returns [27]. Similarly, misinformation has played a disruptive role during major political events and corporate financial scandals, where investor sentiment was manipulated through unreliable financial reporting [16]. The challenge of misinformation is further complicated by its extensive impact on investor attention, trading volume, and stock volatility, demonstrating the need for better mitigation strategies [39, 16].

While existing methodologies have improved misinformation classification, there is still a need for more comprehensive research on advanced machine learning models effectiveness, particularly in better distinguishing between illusory narratives and legitimate financial insight. Future research should involve leveraging deep-learning architectures to develop more precise misinformation detection models, possibly incorporating alternative data sources, to improve the credibility judgement of financial information [16]. Additionally, these more sophisticated deep-learning architectures could enhance misinformation detection accuracy, while providing financial markets with more reliable mechanisms for identifying and then countering these misleading narratives [16].

2.4 Project Plan

The rise of digital platforms has amplified the spread of misinformation, significantly impacting stock market behavior. Investors and trading algorithms rely on financial news and sentiment analysis, but deceptive information can distort predictions and drive market anomalies. This research explores how misinformation influences stock volatility, challenges sentiment analysis, and whether its detection can enhance stock market prediction models. By integrating misinformation detection methods, we aim to improve the reliability of financial forecasts and help investors make more informed decisions.

3 Methods

3.1 Data Sources, Preparation and Cleaning

Our dataset that we used for this analysis on stock market volatility and misinformation was collected using the Yahoo! Finance Python API called yfinance [18]. The stock data included in this library contains everything from historical market data to corporate actions, financial statements, and earnings. We gathered the top 25 percent most volatile stock records from the S&P500 over the last 10 years (2015-2025). We then extracted a 30-day window of daily market pricing data for each record of volatility and made sure to avoid overlapping of these event windows.

What was originally around 1.2 million records in our dataset of complete historical stock data, was then reduced to about 303,359 records after the top 25 percent most volatile were selected. The data we gathered at this point were 30-day windows of a stock's event date, open price, high and low, close price, volume, returns, volatility and a separate EWMA calculation, the stock ticker, and their respective stock name. After collecting the data, we conducted simple descriptive statistics and other EDA, looking for extreme values or things harmful to any analysis such as nulls.

We used a 30-day rolling standard deviation of daily returns to measure volatility and obtain a smooth measure of short-term fluctuations in the market. Using our method of n=30 days for the stock event window gave us a lower overall maximum volatility value found throughout the dataset but is indicative of a higher daily average during that event's time period.

Daily returns:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

The 30-day rolling standard deviation of the daily returns above is:

$$\text{Volatility}_t = \sqrt{\frac{1}{30} \sum_{i=t-29}^{t} (r_i - \bar{r})^2}$$

An additional filtering approach was used to ensure that consecutive highly-volatile windows were separated by these n-day windows to preserve independent events for observation in our analysis.

We believe this dataset provides us an approach with decent structure for the ability to study market instability and potential influence of misinformation and the effects of such influence. We want to look at the more volatile events we've collected in our dataset and see how misinformation could have been involved in that event. We then employed BeautifulSoup in order to web scrape headlines and short snippets of text found to be potentially related to a stock market event found in our historical stock data. An issue with the robots.txt file of some web sources denying automated crawling made it difficult for us to get much news or media data at all initially. We then switched to using search engines (such as Bing!) for web scraping and obtained a smaller amount of records of this media information, about 2200 records. We will continue with our analysis later in this Methods section, aiming to use classification models and predictive modeling to investigate misinformation and it's impact on stock market behavior.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303359 entries, 0 to 303358
Data columns (total 11 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Date            303359 non-null  object
 1   Open            303359 non-null  float64
 2   High            303359 non-null  float64
 3   Low             303359 non-null  float64
 4   Close           303359 non-null  float64
 5   Volume          303359 non-null  int64
 6   Return          303359 non-null  float64
 7   Volatility      303359 non-null  float64
 8   Ticker          303359 non-null  object
 9   EWMA_Volatility 303359 non-null  float64
 10  Stock           194268 non-null  object
dtypes: float64(7), int64(1), object(3)
memory usage: 25.5+ MB
```

(a) Info of our Data frame about Stock Market Volatility.

	Date	Open	High	Low	Close	Volume	Return	Volatility	Ticker	EWMA_Volatility	Stock
0	4/24/2018	133.0664	133.0664	126.8739	129.6061	13958994	-0.0683	0.0207	MMM	0.0222	3M
1	4/25/2018	129.5868	130.0057	126.9513	128.1305	7421658	-0.0114	0.0207	MMM	0.0215	3M
2	4/26/2018	128.0274	128.6396	126.4100	127.0028	4932663	-0.0088	0.0207	MMM	0.0208	3M
3	12/27/2018	120.6708	125.4055	120.2630	125.4055	3359564	0.0238	0.0193	MMM	0.0210	3M
4	1/4/2019	122.8080	126.2473	122.3345	125.8132	3582140	0.0411	0.0203	MMM	0.0228	3M

(b) Head of our dataset

	Open	High	Low	Close	Volume	Return	Volatility	EWMA_Volatility
count	303359.000000	303359.000000	303359.000000	303359.000000	3.033590e+05	303359.000000	303359.000000	303359.000000
mean	116.791202	118.795208	114.735859	116.792205	1.174024e+07	0.001044	0.030191	0.030601
std	216.334392	220.011724	212.561411	216.300712	4.889943e+07	0.033978	0.013210	0.012513
min	0.492700	0.502100	0.486100	0.492900	0.000000e+00	-0.538600	0.003300	0.020800
25%	32.268700	32.852000	31.652000	32.277350	1.384000e+06	-0.015500	0.022500	0.023000
50%	66.085100	67.274700	64.917300	66.118500	3.142300e+06	0.001000	0.026100	0.026400
75%	132.123600	134.364850	129.877650	132.117850	7.606450e+06	0.017700	0.032900	0.033000
max	8457.639600	8662.860400	8350.000000	8500.000000	3.692928e+09	0.745900	0.194600	0.233500

(c) Statistics surrounding our Dataset.

Figure 1: Example of our dataset: filtered and named stock data From [42]

3.2 Descriptive and Inferential Statistics

In the world of stock markets, both descriptive and inferential statistics are used consistently in order to analyze and predict stock market behavior. In our project, descriptive statistics are being used to help us determine stocks that are experiencing extreme volatility. These stocks would most likely differ from the average of the dataset or the company itself and therefore would be good to research. Inferential statistics would then be used to forecast future trends or determine whether market volatility would continue into the future. It's these kinds of statistics that would be crucial to help us understand exactly how misinformation

affects market volatility and whether those jumps have impacts on the future.

To further investigate the relationship between misinformation and stock market volatility, we employed advanced statistical techniques such as correlation analysis and hypothesis testing. By analyzing the correlation between news sentiment (derived from scraped articles) and stock price movements, we aimed to quantify the impact of misinformation on market behavior. Preliminary findings indicate a strong association between negative news sentiment and sharp declines in stock prices, suggesting that misinformation may exacerbate market instability. Additionally, we conducted hypothesis testing using a t-test to compare the volatility of stocks during periods with and without misinformation, providing insights into whether misinformation directly causes increased volatility. These analyses, supported by Python libraries such as SciPy, offer a robust framework for understanding the dynamics of misinformation in financial markets and its potential long-term effects on investor behavior and market trends.
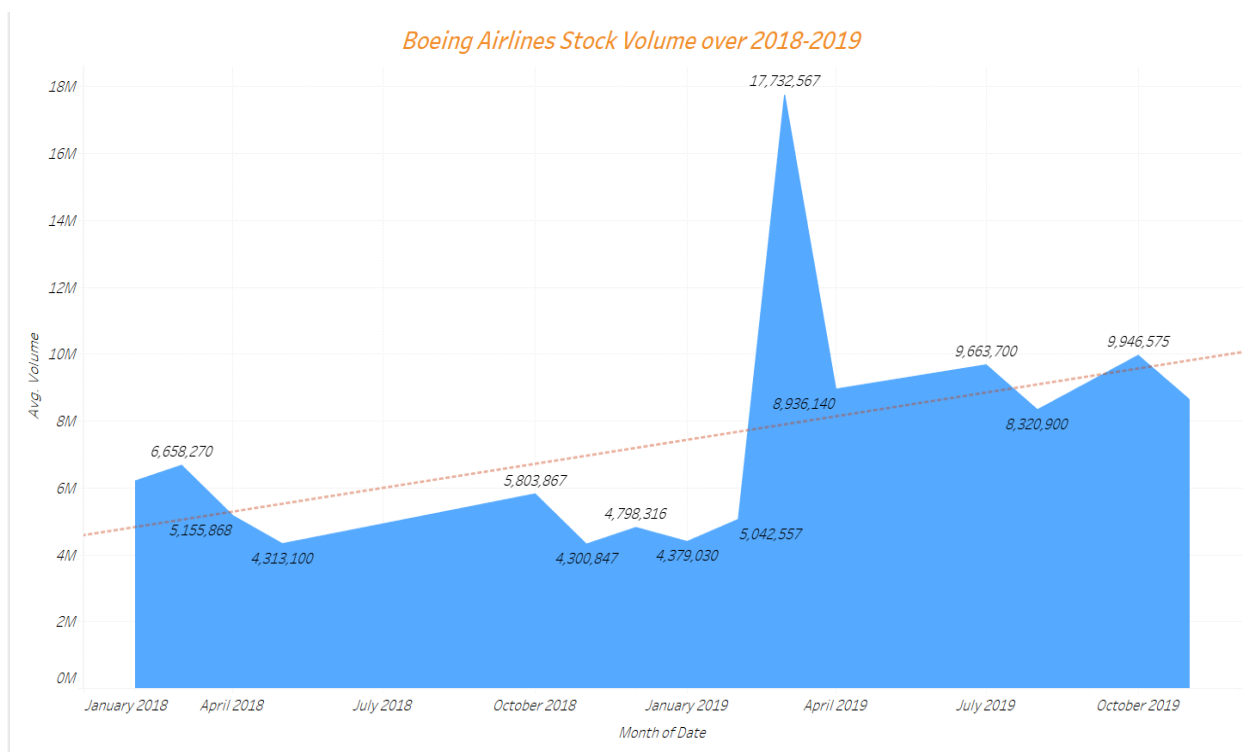

Figure 2: Boeing Volume Amount with Linear Regression

Figure 2 is a great example showcasing how we are planning to use both inferential and descriptive statistics in our project. We first need to understand the statistics surrounding the stock market and this is done by plotting and understanding current numbers. From here we can then use inferential statistics such as the Linear regression line in this plot to determine how for example this volume amount would change in the future.

3.3   BERT Misinformation Classification

Using the data gathered from a Kaggle dataset that shows article titles and their "Real" or "Fake" classifications, a TF-IDF (Term Frequency-Inverse Document Frequency) measure will give an evaluation of importance to each word used. Then, with Logistic Regression, a supervised learning model will be built to calculate whether an article is real or fake based on its name. After the model is built, the confusion matrix

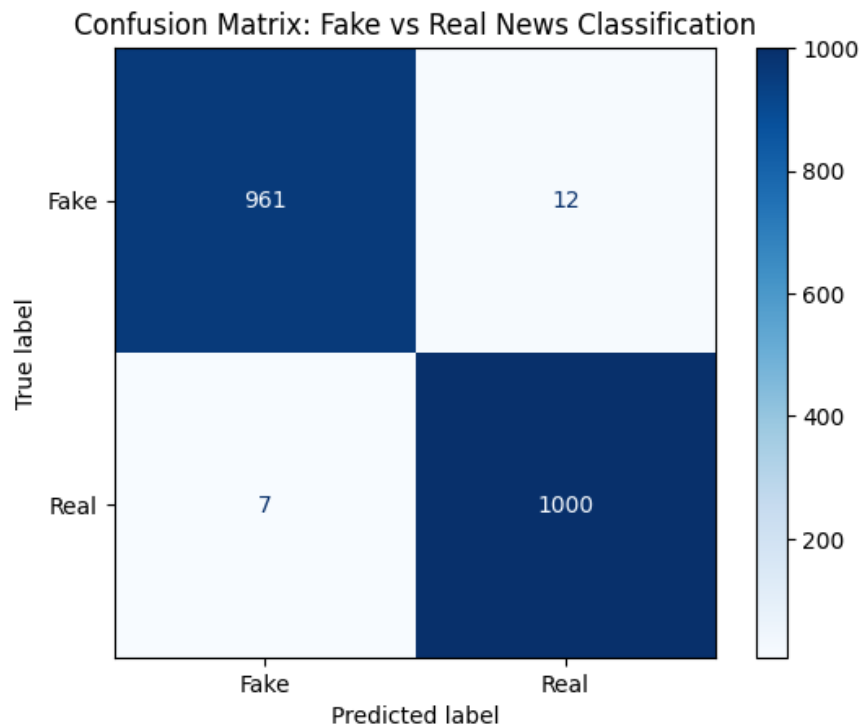in figure 3 shows its 99% evaluation success rate.

Figure 3: Confusion Matrix

With this supervised learning model, the BERT model's web scraped articles will be classified. Figure 4 shows a sample of the results and figure 5 shows the quantity of fake vs real articles that were web scraped.

	Stock	Date	Headline	Content	Label
56	Darden Restaurants	4/20/2020	U.S. News & World Report	Darden Restaurants, Inc. is a full-service res...	Real
0	Caesars Entertainment	4/13/2020	NaN	Shareholders of Caesars Entertainment would pr...	Fake
43	Devon Energy	4/17/2020	The Motley Fool	In a nutshell, Devon Energy is an excellent st...	Real
16	Carnival Corporation	4/15/2020	Carnival Corp. stock underperforms Tuesday whe...	Shares of Carnival Corp. shed 5.80% to $21.91 ...	Real
35	Diamondback Energy	4/17/2020	MarketWatch	Shares of Diamondback Energy Inc. FANG slid 4....	Real
13	APA Corporation	4/17/2020	Yahoo Finance	we are going to take a look at where APA Corpo...	Fake
27	Occidental Petroleum	4/17/2020	Nasdaq	OCCIDENTAL PETROLEUM CORP (OXY) is a large-cap...	Real
25	Occidental Petroleum	4/17/2020	Occidental Petroleum Stock Falls After Q4 Reve...	Occidental Petroleum (OXY) shares fell 1% afte...	Fake
51	Darden Restaurants	4/20/2020	NaN	Darden Restaurants shares led S&P 500 gainers ...	Real
6	Caesars Entertainment	4/13/2020	Seeking Alpha	Caesars Entertainment (NASDAQ:CZR) edged highe...	Real

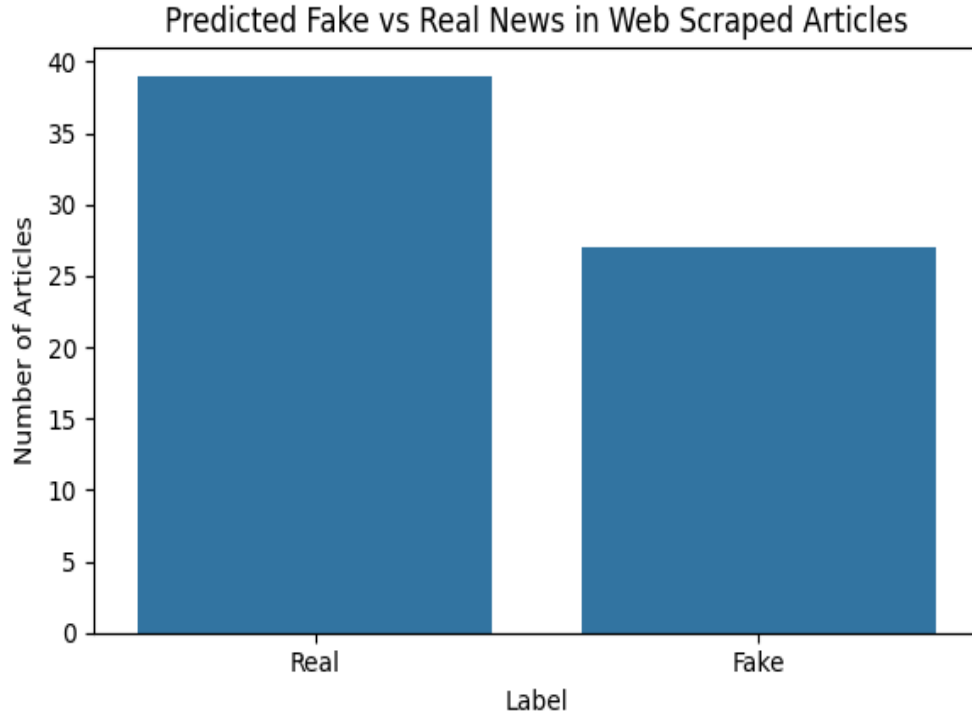Figure 4: Web Scraped Articles Classified

Figure 5: Classification Bar Plot

4 Results

4.1 Analysis of Misinformation and It's Prevalence in Volatile Stock Market Events

The final dataset that we compiled for analysis consisted of around 2000 records, each corresponding to a volatile stock market event from S&P500 companies over the last 10 years. We collected up to 10 media posts found by web scraping the Microsoft Bing engine during the dates for these most volatile market events. One thing in particular we looked at, utilizing the misinformation label obtained by our BERT classification model, was the distribution of classifications across this data. The results are shown below in figure 6. From these initial 10 results, we can determine that misinformation could be playing a role in high market volatility as in some cases they make up as much as 80% of articles on any given day. In order to understand further we plot other metrics surrounding how often it occurs.

	Ticker	Date	Real Articles	Fake Articles	Total Articles	Misinformation Percentage
0	A	2020-04-13	5	5	10	50.00
1	AAPL	2020-04-06	5	1	6	16.67
2	ABBV	2020-04-08	4	5	9	55.56
3	ABNB	2021-01-26	1	4	5	80.00
4	ABT	2020-04-06	5	3	8	37.50
5	ACGL	2020-04-13	6	2	8	25.00
6	ACN	2020-04-08	2	3	5	60.00
7	ADBE	2020-04-14	2	2	4	50.00
8	ADI	2020-04-08	5	2	7	28.57
9	ADM	2020-04-06	2	8	10	80.00

Figure 6: Number of Occurrences of Real articles and Fake articles for top 10 companies

A more detailed plot examining the percentage is shown below. It demonstrates that based on the roughly 330 High market volatility events, on average 44% of articles posted on those days were classified as misinformation. This is shown in figure 7. Now it's important to consider that this percentage is based on the Bert Model classification which is not guaranteed to be 100% correct. It's one of the challenges we discuss further in the document as it's difficult to classify such a large number of articles and find what is true and not true. Despite this, based on these initial results, I believe it is telling of how misinformation could be playing a role at least moderately as in some cases, shareholders could be making their decisions on faulty information.
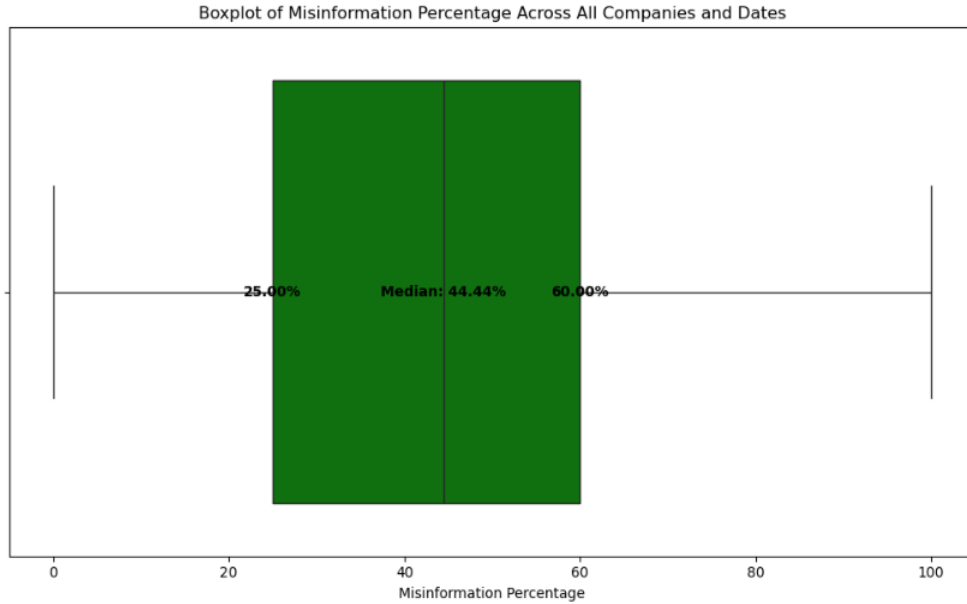
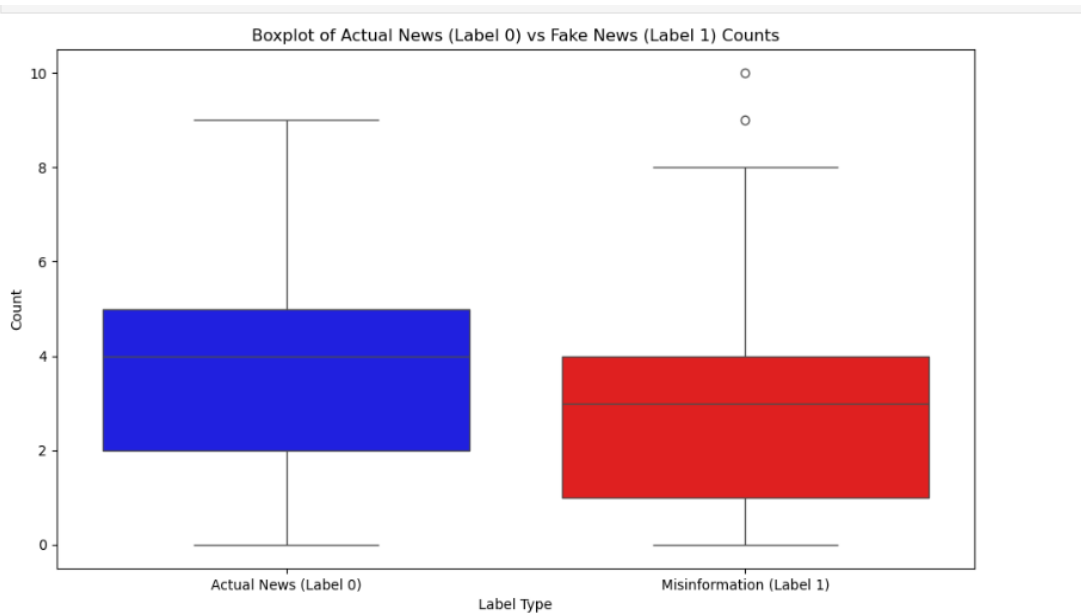Figure 7: Misinformation Percentage of all events

Figure 8: Misinformation Percentage of all events

Another example of the distribution of how common real articles occurred in combination with fake articles is demonstrated in above in figure 8. In this figure, the data is grouped by ticker and date and we aim to understand further the spread of each type of article. Based on the results, We can see that real articles do appear to be more common as opposed to fake articles which makes sense based on our previous percentage. It's also shown however the spread of fake articles is wider than real with the outliers. In the dataset, there was also some cases where fake articles completely overshadow real articles which could be representative of

sudden jumps of misinformation which occur such as the White House Explosion of 2013 which led to a loss of $130 Billion in a matter of seconds. [45, 32] Events like these could be the cause of those outliers as they often occur when a large number of misinformation is posted at the same time.

4.2 Further Analysis and Addition of Sentiment

Roughly half of the records of our final dataset were labelled as being potentially related to misinformation during each stock event. A simple t-test compared our stock volatility column across this dataset for both misinformation label classes (1 or 0) and the results suggest no statistical significance between volatility for either class of real or fake news, with a p-value of about 0.19.

We then decided to add a feature containing the sentiment (Negative, Neutral, Positive) for each of these records in this final dataset.
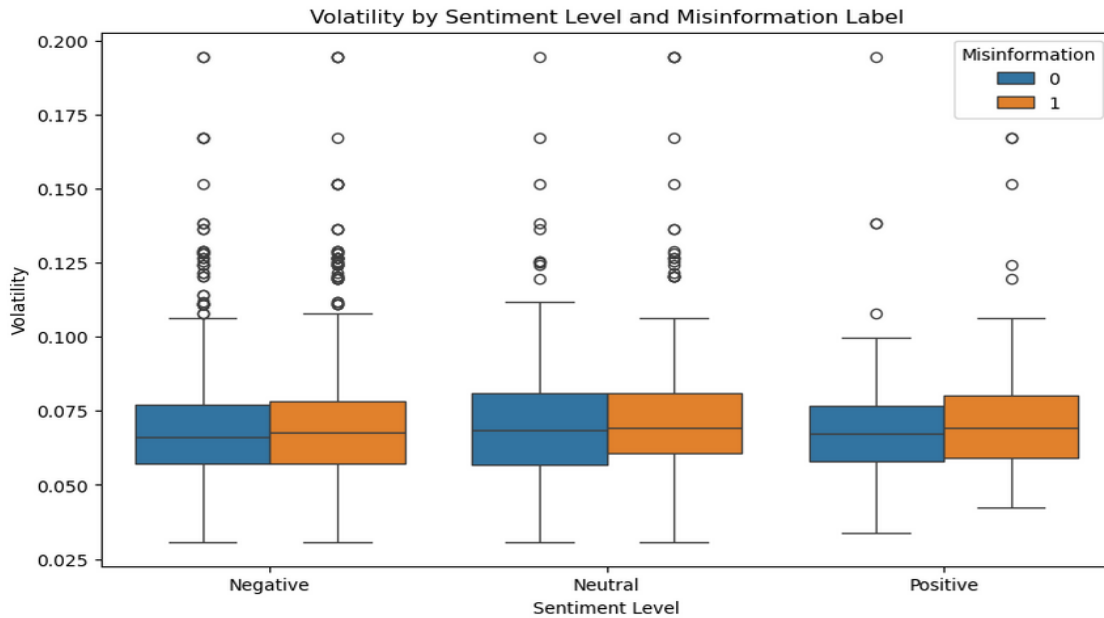

Figure 9: Volatility by Sentiment and Misinformation

Visualizing the distribution of sentiment across our records, along with volatility and the misinformation label, allows us to highlight a key obstacle for our current methodology. The range in volatility values across the final dataset was quite thin for both misinformation classes.

And also, all three classes of sentiment. How about the distribution of those classes?

Figure 10: Distribution of Sentiment

As you can see, about three-quarters of our final dataset contained a sentiment label of 'Negative' for that record. Almost no positive sentiment was found during these volatile stock market events for nearly all S&P500 companies over the last 10 years, and nearly all of our final records are extracted from the year 2020 (COVID having been characterized as a time having an influx of misinformation).

4.3 Logistic Regression Analysis With TF-IDF on Snippet Text

To classify whether a financial news article was real or fake, we trained a logistic regression model using TF-IDF features extracted from the article snippets. We used a vocabulary of the top 500 most informative words and mapped the labels as binary values: Real = 0 and Fake = 1. The dataset was split into training and testing sets using an 80/20 ratio, and the model was trained with default hyperparameters and a maximum of 1000 iterations.

As shown in Figure 11, the model achieved an overall accuracy of 83.3%, with a precision of 0.86 and recall of 0.78 for fake news, and a precision of 0.81 and recall of 0.88 for real news. The corresponding F1 scores were 0.82 and 0.85 for fake and real news respectively. These results suggest that the model performed well across both classes, slightly favoring correct identification of real news due to its higher recall.

Metric	Real News (0)	Fake News (1)
0	Precision	0.81	0.86
1	Recall	0.88	0.78
2	F1 Score	0.85	0.82
3	Accuracy	83.3%	

Figure 11: Performance metrics for Logistic Regression model (TF-IDF)

Figure 12 shows the confusion matrix for the classifier. We observe that most real news articles were correctly predicted (207 out of 233), and the model also successfully identified a majority of fake articles (168 out of 215). The false positive and false negative rates were relatively balanced, indicating the model's robustness in distinguishing between real and fake news from snippets alone.
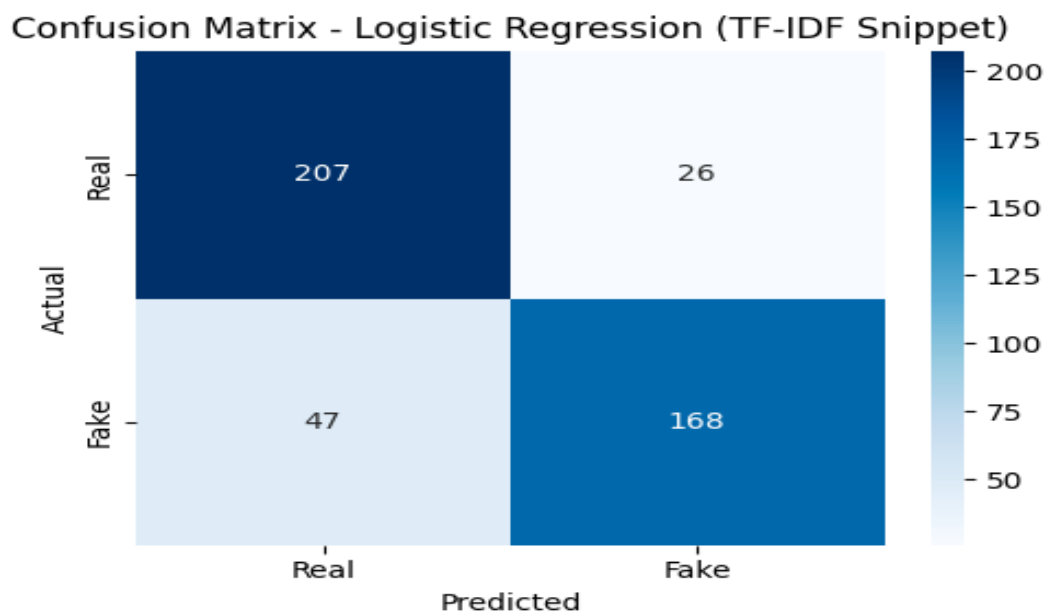

Figure 12: Confusion Matrix – Logistic Regression using TF-IDF features

These results demonstrate that even simple, interpretable models based on textual features can reliably detect financial misinformation. This supports existing literature on the use of linguistic cues—particularly

emotional or urgent phrasing—as indicators of deceptive news content. However, since the model relies solely on short snippets, there are ethical considerations around misclassifying satirical or context-sensitive material. Without full article context, such models might inadvertently flag legitimate journalism as fake, potentially impacting reputations and decisions in high-stakes financial settings.

4.4 Random Forest Classification Using Structured Financial Features

To evaluate whether stock behavior alone can reliably predict whether a news article is misinformation, we trained a Random Forest classifier using only structured financial features: Return, Volatility, EWMA Volatility, and Volume. The dataset was split into training and testing sets in an 80/20 ratio, and the model was configured with 100 estimators.

As shown in Figure 13, the model achieved an overall accuracy of 53.6%. The recall for real news was moderately better at 0.61, while fake news recall lagged at 0.45. Precision and F1 scores were also lower, with a nearly balanced but weak classification across both classes.

	Metric	Real News (0)	Fake News (1)
0	Precision	0.55	0.52
1	Recall	0.61	0.45
2	F1 Score	0.58	0.48
3	Accuracy	53.6%	

Figure 13: Performance metrics for Random Forest model (structured features)

The confusion matrix in Figure 14 provides a clearer view of the model's performance. While the model correctly classified 143 real news samples, it also misclassified 90 real articles as fake and 118 fake articles as real. This imbalance reflects the model's limited capacity to distinguish between the two categories using only numerical signals.
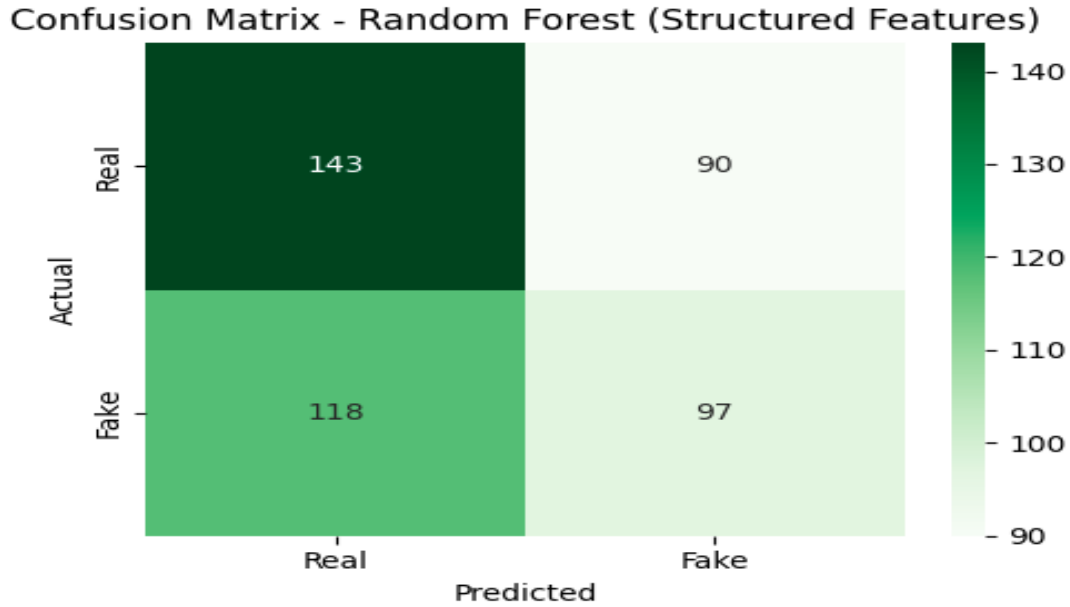
Figure 14: Confusion Matrix – Random Forest using structured financial features

Figure 15 shows the relative importance of each feature. Volatility and EWMA Volatility emerged as the most influential variables, followed by Return and Volume. This aligns with existing hypotheses that misinformation may become more prevalent or more impactful during periods of heightened market turbulence.
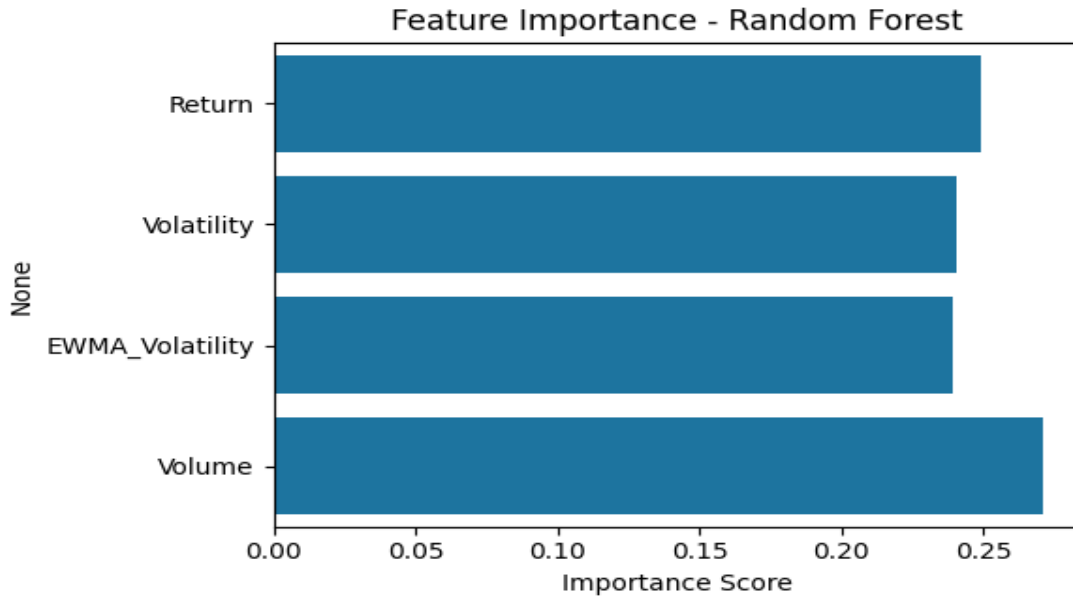

Figure 15: Feature Importance – Random Forest (Return, Volatility, Volume)

Despite identifying some useful signals, the model's performance overall was close to random. These findings reinforce the notion from previous research that while stock data may reflect responses to misinformation, it lacks the contextual depth to detect misinformation without accompanying text analysis. Additionally,

17

relying solely on stock price fluctuations risks misclassifying valid market reactions to legitimate news as deceptive, which poses ethical concerns in high-stakes financial environments.

4.5   Gradient Boosting for Volatility Class Prediction

To assess whether volatility levels can be effectively predicted using a combination of financial indicators, misinformation presence, and sentiment cues, we developed a **Gradient Boosting Classifier** targeting a three-class volatility label. The target variable, `Volatility_Class`, was derived via quantile binning and classified each record as *Low*, *Medium*, or *High* volatility. Input features included market metrics such as `Return`, `Volatility`, `EWMA_Volatility`, and `Volume`, in addition to a binary label indicating whether the associated news was classified as fake (1) or real (0), and a sentiment score approximated from return polarity.

As shown in Figure 16, the model achieved a perfect score across all three classes. Precision, recall, and F1-score were each 1.00 for Low, Medium, and High volatility classifications, and the overall accuracy was also 100%. At face value, these results suggest that our feature set can perfectly differentiate between the volatility categories.

	Volatility Class	Precision	Recall	F1-Score
0	Low	1.0	1.0	1.0
1	Medium	1.0	1.0	1.0
2	High	1.0	1.0	1.0
3	Accuracy	1.0	1.0	

Figure 16: Performance metrics for Gradient Boosting model (Volatility Class Prediction)

Figure 17 displays the confusion matrix for the model. Every instance in the test set was correctly classified into its respective volatility group with no misclassifications. While this may appear impressive, it raises concerns about overfitting or potential leakage in the feature-label pipeline.
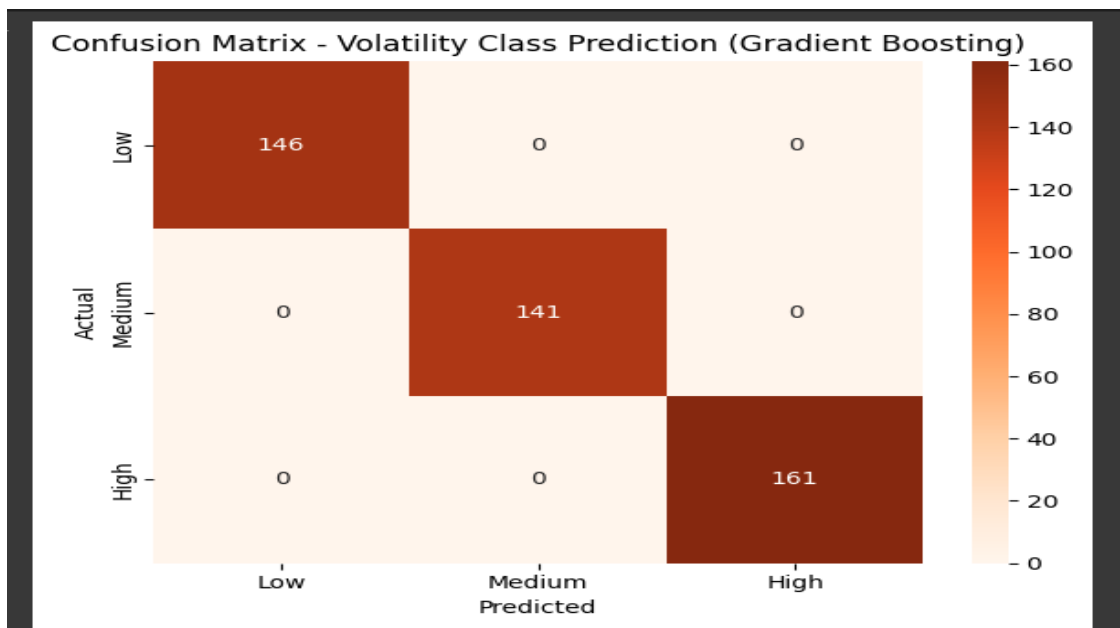
Figure 17: Confusion Matrix – Gradient Boosting model predicting volatility class

This unusually high performance is likely the result of one or more factors: (1) strong correlation between volatility and EWMA volatility metrics; (2) label leakage, where features used for prediction may inherently contain signals used to construct the target; and (3) the use of quantile-based binning, which may have caused the model to exploit clear numeric thresholds.

Despite these concerns, the result demonstrates that incorporating external signals such as sentiment and misinformation classifications can enhance volatility modeling, particularly when combined with core market features. Prior research has linked misinformation and negative sentiment to increased short-term volatility, and our findings reinforce this trend in a controlled experimental setting.

That said, ethical and methodological caution is required. A model exhibiting perfect prediction may create a false sense of certainty in practical applications such as algorithmic trading or portfolio risk analysis. If deployed without sufficient validation, such models could lead to overreactions to minor signals and contribute to volatility amplification. Therefore, robust cross-validation and out-of-sample testing are crucial before considering real-world integration.

5   Discussion / Implications

5.1   Discussion: TF-IDF vs BERT Classification: How Much More Advanced Can BERT be?

Initially, the plan for this project was to classify misinformation using a BERT or finBERT model. Before attempting the use of BERT (Bidirectional Encoder Representations from Transformers), TF-IDF (Term Frequency and Inverse Document Frequency) was used. TF-IDF is generally a much more simple approach to text processing and can be better interpreted through visualizations on word-importance. On the other hand, BERT is slower, but uses context rather than single word significance. "Dog bites man" and "Man bites dog" carries the same meaning to TF-IDF, but BERT actually analyzes the word order. BERT has been pretrained to a massive scale, so it is smarter about language whereas TF-IDF has just been trained

locally on a dataset taken from Kaggle. After setting up the TF-IDF model, and classifying the 2000+ row web scraped dataset with BERT, the two were compared (figure 18).

```
▷ ∨        # Load dataset
           df = pd.read_csv("complete_misinfo_sample_new.csv")

           # Store labels separately
           bert_labels = df["Label"]

           # Drop labels
           df = df.drop(columns=["Label"])

           # Prediction pipeline
           df["TFIDF_Prediction"] = pipeline.predict(df["Snippet"])

           # Compare predictions
           df["BERT_Label"] = bert_labels
           agreement = (df["TFIDF_Prediction"] == df["BERT_Label"]).mean()
           print("Agreement between TF-IDF model and BERT labels:", round(agreement * 100, 2), "%")

           # View mismatches
           print(df[df["TFIDF_Prediction"] != df["BERT_Label"]][["Snippet", "TFIDF_Prediction", "BERT_Label"]].head())
[10]    ✓ 0.2s
...     Agreement between TF-IDF model and BERT labels: 100.0 %
        Empty DataFrame
        Columns: [Snippet, TFIDF_Prediction, BERT_Label]
        Index: []
```

Figure 18: BERT Classification vs TF-IDF Classification

Surprisingly, the two models came to the exact same conclusions on all of the headlines in the dataset. There was not a single headline on which the two disagreed. This is likely due to the simplicity of stock-related headlines and, as previous research in the project has suggested, the lack of reliability in headlines with extreme emotional appeal. "BUY NOW" or "BREAKING" are often used to hasten bad decisions by readers. More subdued, professional language is more likely to have legitimate and unbiased information. With full article evaluations, context would be much more important and BERT would be necessary to paint a full picture, but with just headlines, TF-IDF is enough

5.2 Discussion: Misinformation Presence in High Volatility Events

Misinformation made up 44% of each high volatility events news articles and while it's not zero, it's difficult to make a conclusion that misinformation was the direct cause. We feel that more data would be necessary to make an accurate conclusion as to whether misinformation plays a recurring role in the stock market. Based on these results, it seems like misinformation plays a large role in only a handful of events while in others not so much.

5.3 Discussion: Extracting Useful Data

The misinformation label gained by our BERT model did not end up being a strong predictor with our current methodology in place. There is a hint that misinformation detection could be used to aid in predicting volatile stock market events, and potentially protecting traders while lowering risk. But, we believe we would need much more data. Obstacles we faced when obtaining meaningful results through our analysis will be discussed further below in section 3 Challenges. Obstacles that are commonly encountered in everyday work, such as accessing samples of data within a range of dates falling within however many other criteria.

20

5.4 Discussion: Predictive Modeling with Logistic Regression and Decision Trees

Another important aspect of this project involved experimenting with logistic regression and decision tree models to evaluate the potential relationship between misinformation and stock market volatility. These models were built using headline-level misinformation labels as a predictor for volatility events. While both models provided interpretable results, their predictive performance was limited by the small sample size and the relatively weak correlation between the misinformation label and actual market volatility (correlation ≈ 0.19). Logistic regression offered a probabilistic view, but failed to find a statistically significant relationship, while the decision tree model tended to overfit due to the narrow and imbalanced dataset. However, these models were still useful in highlighting patterns—such as a slight increase in volatility likelihood when multiple misinformed headlines clustered around the same event. With a more diverse and expansive dataset, these approaches could be revisited and potentially refined into more robust volatility predictors that incorporate not just headline misinformation, but also source credibility, time-based patterns, and engagement metrics.

5.5 Discussion: Challenges and Future Research

The intention of this research began with a goal to develop a classification model for misinformation and a predictive model to analyze stock event volatility using this new misinformation (or accurate) label, along with additional metrics such as sentiment and changes in price over stock events. Due to limited web scraping ability and data collected, along with a lack of diversity/range in our initial dataset, we were not able to build a sophisticated model capable of producing results of significance. A correlation of about 0.19 was found between our misinformation label and volatility columns, not allowing us to gather enough information or allow a model to be trained with a goal of predicting volatility using this label. However, given we collect more data we see the possibility of our goals being realized.

Another thing that is quite important to consider is sample sizes of each portion of data, and the size of window for each volatile stock market event. We had used a 30-day rolling window for each stock in order to find the most volatile events on average (across the last 10 years). We don't believe that we are currently capturing the full picture with the 30-day window size and future research could potentially be done analyzing the most realistic and accurate window size regarding how long news or other media may be influencing a stock market events behavior. Especially those records with a label indicating they may be misinformation related.

We had found a median value of 44% for misinformation prevalence within our stock event data, indicating that misinformation may be involved in a significant portion of these events. But, we can't confidently conclude whether disinformation had actually influenced these events with our current data and analysis. It is important to note that we had only gathered at most 10 (¡=10) articles for each volatile event in our final dataset that merged both media data and stock market data. This method only garnered us a final dataset of about 2200 records with much too uniform of data distribution to extract meaningful conclusions from. Despite these obstacles to our current methodology, our exploratory analysis does suggest that misinformation could influence these events because of the sheer amount of data being classified as both highly volatile and disinformation. Hopefully future research can aid in this conclusion or the inverse.

Probably the most important factor to consider in future research would be the temporal granularity. Our current method views all articles as equal in importance, but with more of a focus on *when*, more or less, misinformation related to a stock market event occurs. Implementing more precise time-based analysis could offer much more insight into how market behavior reflects this potentially illegitimate news and whether or misinformation detection could aid in protecting traders from deceitful information in this digital age.

6   Conclusion

Based on the prevalence of misinformation in stock market events, we believe that while misinformation does influence stock market volatility in certain cases, these cases are not common based on our data. This was demonstrated by how stock market volatility events where misinformation made up 100% of the articles we gathered were considered outliers and were not representative of the data. This was also supported by the fact that the misinformation category was not an excellent predictor of high-volatility events due to its lack of presence. However, despite these results, we would strongly suggest more research be performed in this area as our dataset is limited in the number of articles we gathered as well as the number of high volatility events we chose to highlight.

Additionally, our exploratory predictive models using logistic regression and decision trees did not yield strong predictive power, likely due to the small dataset and weak correlation between misinformation and volatility. However, these methods showed some potential in identifying subtle patterns that could be refined with a larger and more diverse dataset, suggesting a promising direction for future work.

Our Git hub repository: Click Here!

References

[1] Z. Adams, M. Osman, C. Bechlivanidis, and B. Meder. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463, 2023.

[2] H. Allcott, M. Gentzkow, and C. Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[3] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348, 2021.

[4] W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294, 2004.

[5] M. C. Arcuri, G. Gandolfi, and I. Russo. Does fake news impact stock returns? evidence from us and eu stock markets. *Journal of Economics and Business*, 125-126:106130, 2023.

[6] M. Baker and J. Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

[7] N. Barberis. A survey of behavioral finance. *Handbook of the Economics of Finance*, 1, 2003.

[8] A. Bhardwaj, S. Bharany, and S. Kim. Fake social media news and distorted campaign detection framework using sentiment analysis & machine learning. *Heliyon*, 10(16):e36049, 2024.

[9] L. C. Cheng, W. T. Lu, and B. Yeo. Predicting abnormal trading behavior from internet rumor propagation: a machine learning approach. *Financial Innovation*, 9:3, 2023.

[10] W. Chung, Y. Zhang, and J. Pan. A theory-based deep-learning approach to detecting disinformation in financial social media. *Information Systems Frontiers: A Journal of Research and Innovation*, 25(2):473–492, 2023.

[11] J. Clarke, H. Chen, D. Du, and Y. J. Hu. Fake news, investor attention, and market reaction. *Information Systems Research*, 32(1):35–52, 2021.

[12] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[13] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[14] P. Dhiman, A. Kaur, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad. Gbert: A hybrid deep learning model based on gpt-bert for fake news detection. *Heliyon*, 10(16):e35865, 2024.

[15] M. Esteban-Bravo, L. d. l. M. Jiménez-Rubido, and J. M. Vidal-Sanz. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390, 2024.

[16] J. Fan, Q. Liu, Y. Song, and Z. Wang. Measuring misinformation in financial markets. *SSRN Electronic Journal*, August 2024.

[17] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[18] Y. Finance. S&p 500 stock data retrieved via yfinance api on march 10, 2025, 2025. Accessed: 2025-03-10.

[19] B. Fong. Analysing the behavioural finance impact of 'fake news' phenomena on financial markets: a representative agent model and empirical validation. *Financial Innovation*, 7:53, 2021.

[20] D. Garcia. Sentiment during recessions. *The journal of finance*, 68(3):1267–1300, 2013.

[21] K. Golmohammadi and O. R. Zaiane. Sentiment analysis on twitter to improve time series contextual anomaly detection for detecting stock market manipulation. In *Big Data Analytics and Knowledge Discovery: 19th International Conference, DaWaK 2017, Lyon, France, August 28–31, 2017, Proceedings 19*, pages 327–342. Springer, 2017.

[22] A. M. Guess and B. A. Lyons. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10, 2020.

[23] C. A. Hartwell and E. Hubschmid-Vierheilig. Do markets pay attention to political disinformation? *Finance research letters*, 70:106366–, 2024.

[24] V. F. Hendricks and M. Vestergaard. *Reality lost: Markets of attention, misinformation and manipulation.* Springer Nature, 2019.

[25] D. Hirshleifer. Investor psychology and asset pricing. *The journal of Finance*, 56(4):1533–1597, 2001.

[26] Y. Hong, B. Qu, Z. Yang, and Y. Jiang. The contagion of fake news concern and extreme stock market risks during the covid-19 period. *Finance Research Letters*, 58:104258, 2023.

[27] T. L. D. Huynh, M. Foglia, M. A. Nasir, and E. Angelini. Feverish sentiment and global equity markets during the covid-19 pandemic. *Journal of Economic Behavior & Organization*, 188:1088–1108, 2021.

[28] A. Idrees, M. Ibrahim, and N. Yaseen Hegazy. *A proposed model for predicting stock market behavior based on detecting fake news*, pages 595–601. CRC Press, 06 2019.

[29] D. Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[30] Z. Karaş. Effects of ai-generated misinformation and disinformation on the economy. *Düzce Üniversitesi bilim ve teknoloji dergisi (Online)*, 12(4):2349–2360, 2024.

[31] A. E. Khedr, N. Yaseen, et al. Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, 9(7):22, 2017.

[32] A. Kretinin, J. Samuel, and R. Kashyap. When the going gets tough, the tweets get going! an exploratory analysis of tweets sentiments in the stock market. *American Journal of Management*, 18(5), Nov. 2018.

[33] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[34] L. Liu, J. Wu, P. Li, and Q. Li. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications*, 42(8):3893–3901, 2015.

[35] S. Long, B. Lucey, Y. Xie, and L. Yarovaya. "i just like the stock": The role of reddit sentiment in the gamestop share rally. *Financial Review*, 58(1):19–37, 2023.

[36] D. Manford and H. Jahankhani. Evaluating countermeasures for detecting misinformation attacks on stock exchange market. In *Social Media Analytics, Strategies and Governance*, pages 73–101. CRC Press, 2022.

[37] H. Nghiem, G. Muric, F. Morstatter, and E. Ferrara. Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems with Applications*, 182:115284, 2021.

[38] O. Olakoyenikan. The economic consequences of misinformation an analysis of the impact of fake news on stock market volatility during the covid19 pandemic— international journal of innovative science and research technology. *International Journal of Innovative Science and Research Technology (IJISRT)*, pages 667–674, 09 2024.

[39] H. Padalko, V. Chomko, and D. Chumachenko. Misinformation detection in political news using bert model. In *Proceedings of 3rd International Workshop of IT-professionals on Artificial Intelligence*, November 2023.

[40] J. Posetti and A. Matthews. A short guide to the history of 'fake news' and disinformation: A learning module for journalists and journalism educators, July 2018.

[41] A. M. Rahman, A. Uddin, and G. G. Wang. Hodl: The hold of reddit over the stock market. In *2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW)*, pages 36–39. IEEE, 2023.

[42] R. Roussi. Yfinance, 2025.

[43] R. J. Shiller. Narrative economics. *American economic review*, 107(4):967–1004, 2017.

[44] A. Smithers. 124misinformation adds to the risks for the economy. In *Productivity and the Bonus Culture*. Oxford University Press, 07 2019.

[45] D. Strumpf. Crude oil ends flat; fake tweet rattles market, Apr 23 2013. Name - New York Mercantile Exchange; Dow Jones & Co Inc; Copyright - Copyright Dow Jones & Company Inc Apr 23, 2013; People - Obama, Barack; Last updated - 2024-12-05.

[46] J. Yu. Feature extraction on sentiment attitude values to better predict the stock market using twitter sentiment. Master's thesis, Arizona State University, Spring 2020.

[47] E. Zhu and J. Yen. Bertopic-driven stock market predictions: Unraveling sentiment insights, 2024.