WallESense

Tanvi Bhaskarwar
bhaskarw@usc.edu

Venkata Meghana Achanta
vachanta@usc.edu

Vaibhav Rungta
vrungta@usc.edu

Abstract

This project proposes the development of a predictive model that leverages Natural Language Processing techniques to forecast the VIX(CBOE) index. For this purpose the model will analyze news articles to provide insights into potential market movements.

1 Introduction

The predictive approach of this model integrates NLP-derived sentiment analysis to forecast stock market volatility. It analyzes the text derived from news to predict future volatility in the stock market. Through the analysis of implied volatility with advanced NLP techniques, this model seeks to offer a comprehensive and quantitative perspective on market volatility. Therefore, we use the techniques to predict the VIX(CBOE) that models the volatility for majority of the US stock market.

The novelty of this project lies in its approach to sentiment analysis within the context of predicting volatility, rather than the directionality in market trends. Traditional sentiment analysis in financial markets has primarily focused on classifying sentiments as either positive or negative, aiming to correlate these sentiments directly with upward or downward movements in stock prices. However, volatility represents a measure of market uncertainty and fluctuation, encompassing both positive and negative movements. On the other hand, high volatility can result from both overwhelmingly positive news (such as groundbreaking product launches) and significantly negative events (such as financial scandals).

This project explores whether advanced language models like BERT and FinBERT, which have been primarily utilized for sentiment analysis as shown in studies like FinBERTAraci [1] and StonkBERT[7], can surpass traditional models like LDA and Naive Bayes in predicting market volatility.

2 Differentiating Volatility and Sentiment Analysis

While both volatility and sentiment analysis are instrumental in financial markets, they serve distinct purposes and operate on different dimensions of market analysis. Volatility focuses on the magnitude of price movements, regardless of the direction, and is a statistical measure of the dispersion of returns for a given security or market index. It inherently deals with the variability and risk inherent in the price movements of securities, providing a gauge of market turbulence and investor sentiment.

In contrast, sentiment analysis evaluates the qualitative aspects of market data, often derived from textual information in news articles and social media, to gauge the collective attitude of investors toward a particular security or the market as a whole. This form of analysis involves interpreting the emotional tone behind words used in market-related communications, categorizing them into positive, negative, or neutral sentiments. Sentiment analysis aims to predict how these collective attitudes might influence future market behavior, focusing primarily on directional cues rather than the intensity of price fluctuations.

3 Measuring Volatility

Volatility in finance is quantified by the standard deviation of logarithmic returns. It represents the risk associated with the asset's price movements and is a crucial measure in assessing market sentiment and risk management strategies.

The Implied Volatility Index, often symbolized as the IV Index, is a metric that reflects the market's expectation of volatility in the price of an underlying asset over a specific period, derived from the prices of options on that asset.

It is calculated by inputting the market price of an option into an option pricing model (such as the Black-Scholes model) and solving for the volatility value that sets the model price equal to the market price. Mathematically, for a call option, the Black-Scholes model is represented as:

$$C(S,t) = SN(d_1) - Ke^{-rt}N(d_2)$$

where:

- $C(S,t)$ is the price of the call option,
- S is the current price of the underlying asset,
- K is the strike price of the option,
- r is the risk-free interest rate,
- t is the time to expiration,
- $N(\cdot)$ is the cumulative distribution function of the standard normal distribution,
- $d_1 = \frac{1}{\sigma\sqrt{t}}\left(\ln\frac{S}{K} + \left(r + \frac{\sigma^2}{2}\right)t\right)$,

- $d_2 = d_1 - \sigma\sqrt{t}$,
- σ is the volatility of the underlying asset, which in the context of implied volatility, is the variable solved for.

The CBOE Volatility Index (VIX) is a real-world embodiment of implied volatility, predicated on the prices of near-term S&P 500 options traded on the Chicago Board Options Exchange (CBOE). Distinct from other volatility measures, the VIX is forward-looking and reflects the market's consensus on expected volatility over the coming 30 days. It calculates implied volatility by aggregating the weighted prices of a wide array of S&P 500 index options

4 Previous Work

Historical studies have significantly demonstrated the direct correlation between public news information and trading activity in financial markets. One seminal work in this area is "The Impact of Public Information on the Stock Market" by Mark L. Mitchell and J. Harold Mulherin (1994) Mitchell and Mulherin [5], which highlights the substantial influence that news has on trading volumes. Their study reports a correlation coefficient of 0.367 between the number of daily Dow Jones news announcements and trading volume, emphasizing the statistical significance of this relationship (p-value < 0.0001). Furthermore, they found that a 100% increase in the volume of news stories is associated with a 38% increase in trading volume. This relationship persists even after controlling for day-of-the-week effects, illustrating how news can act as a primary driver of market behavior. These findings underscore the importance of news as a critical factor in the dynamics of market activity, affirming the foundational role of public information in influencing investor behavior and market trends.

Furthermore, in the study "Financial news predicts stock market volatility better than close price" by Atkins, A., Niranjan, M., Gerding, E. (2018) Atkins et al. [2], the value of text data in predicting market volatility is distinctly highlighted. This research reports that the predictive accuracy for market volatility using news-derived information reaches 56%, underlining the complex patterns within text data that lend themselves more effectively to forecasting volatility than to predicting directional price movements, which only achieve a 49% accuracy—essentially equivalent to random guessing. These findings underscore the nuanced capabilities of textual analysis over traditional sentiment analysis, which often targets the direction of price movements. Utilizing machine learning techniques such as Latent Dirichlet Allocation (LDA) and Naive Bayes classifiers, the study illustrates how sophisticated text-based models can more accurately capture the essence of market volatility, offering insights that are crucial for managing risk and making informed investment decisions.

FinBERT Araci [1] is a model specifically developed for financial sentiment analysis. While it has shown effectiveness in interpreting the sentiment from various textual sources within the financial domain, its application has not yet been extended to predict market volatility. This omission is notable given that recent studies, such as those involving StonkBERTPasch and Ehnes [7], have demonstrated the potential of textual data not just for sentiment analysis but also for predicting market phenomena like stock price performance using transformer-based classifiers. Such classifiers have been used to analyze diverse types of company-related text data including news articles, blogs, and annual reports. Traditional statistical models have often been the benchmark in these studies, compared against more modern approaches like RNNs and LSTMsPuh and Babac [8]. However, the literature still lacks a thorough investigation into the use of BERT or any language models for directly estimating the Volatility Index through regression. This gap highlights a significant opportunity: leveraging the established capability of text data in forecasting market volatility, as shown by its superior performance in predicting volatility over directional price movements, to develop language model-based approaches for volatility prediction.

5 Data Preprocessing

The selection of news articles for this project is informed by an extensive analysis of GDELT[6]'s Global Knowledge Graph (GKG) files, which serve as a comprehensive repository of global news events and their underlying contexts. Specifically, the GKG files are mined to identify news articles pertinent to financial markets, with a particular focus on those tagged with relevant financial metadata.

For the development of our prototype model we are using the articles in GDELT for the 4 year period from beginning of 2019 to the end of 2022. This gives us 1010 day that CBOE was open for trading, which translates to 1010 data points for our model. For each day we have collected close to 6500 financial articles from GDELT.

The preprocessing of textual data from news articles for this project involves a systematic procedure to prepare the content for NLP analysis. Initially, the textual content is extracted from the given news sources, leveraging dynamic content loading mechanisms to ensure comprehensive retrieval.

The subsequent cleaning phase addresses several key aspects to refine the data: HTML tags, which are irrelevant to text analysis, are removed to isolate pure textual content; non-alphanumeric characters that could introduce noise into the analysis are eliminated; excessive white space, which might skew text parsing algorithms, is condensed; and all text is converted to lowercase to standardize the dataset and facilitate uniform analysis.

Furthermore, we log-normalize the vix values obtained for over the same time period. Log-normalization of VIX closing prices, as depicted in the histograms, serves several crucial statistical purposes. Primarily, this transformation is applied to handle the skewness towards higher values observed in raw VIX data. By converting the data into a logarithmic scale, we reduce the variance and stabilize the spread of data points, making the distribution more sym-

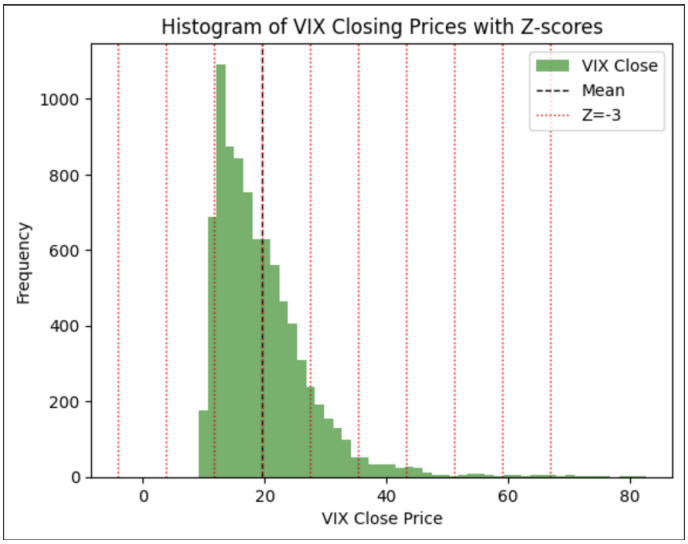

Figure 1: Histogram of VIX Closing Prices

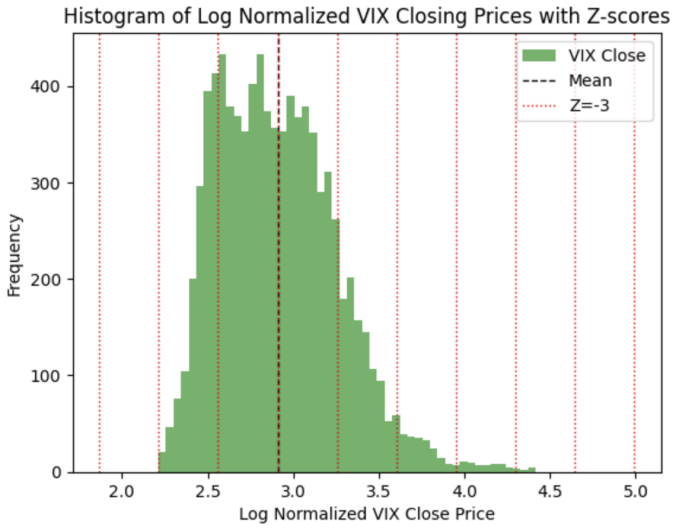

Figure 2: Histogram of Log Normalized VIX Closing Prices

metric and closer to a normal distribution. Additionally, log-normalized data dampens the impact of extreme outliers, which are common in financial time series, thereby providing a more accurate and robust analysis of central tendencies and variability. This normalization facilitates a clearer understanding of the underlying patterns in the data, enhancing the reliability of volatility forecasts generated by predictive models.

6 Regression-Based Forecasting Approach

Our main approach to the problem statement is to implement a BERT model with a regression layer such that we can train the model based on the VIX values of the previous 4 years. The model would be tested to estimate the VIX values for the past 2-3 years. This would especially be beneficial for studying the market trends closely and ensuring that profit can be extracted from volatility.

For example, given the text articles for the day, 'Wall street ends dismal volatile year on a bright note wall street closed out a dismal turbulent year for stocks on a bright note on monday' results in a VIX score of 23.989. (Moderate Volatility)

'Stocks plunge 508 points, a drop of 22.68%; 604 Million volume nearly doubles record' results in a VIX score of 70.844 (High Volatility)

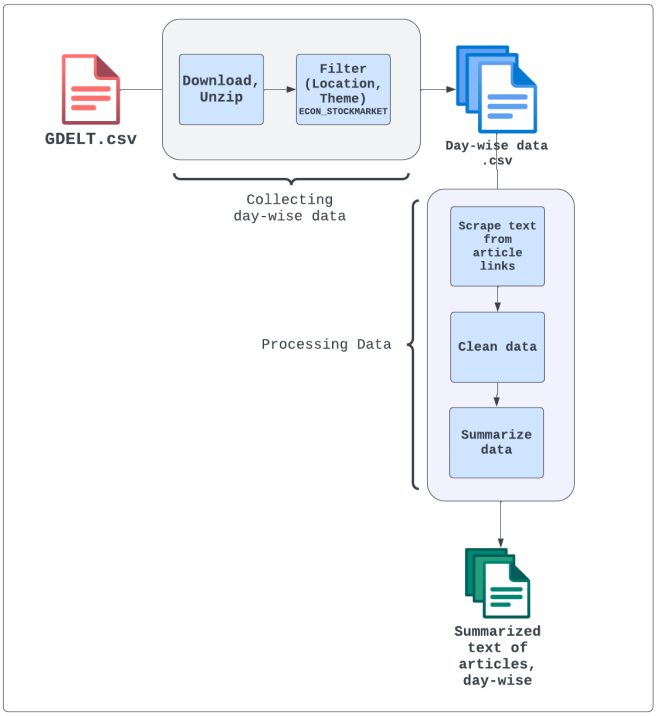

Figure 3: Data pre-processing pipeline

7 BERT: Baseline model

Devlin et al. [3] BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model developed by Google in 2018, notable for its transformer architecture and bidirectional context understanding.

7.1 Key Features of BERT:

- **Transformer Architecture:** BERT's transformer architecture allows it to process and understand text data efficiently by capturing complex dependencies and relationships among words. This capability is essential for analyzing the intricate language used in financial news articles.

- **Bidirectional Context Understanding:** BERT considers bidirectional context, meaning it comprehensively grasps the relationship between words by leveraging information from both preceding and succeeding words. This feature is crucial for interpreting the nuanced language often found in financial reports and news articles.

- **Attention Mechanism:** BERT utilizes attention mechanisms to focus on relevant words and their connections within a sentence, enabling a nuanced understanding of language semantics. This capability is pivotal for capturing sentiment and market dynamics embedded in textual data.

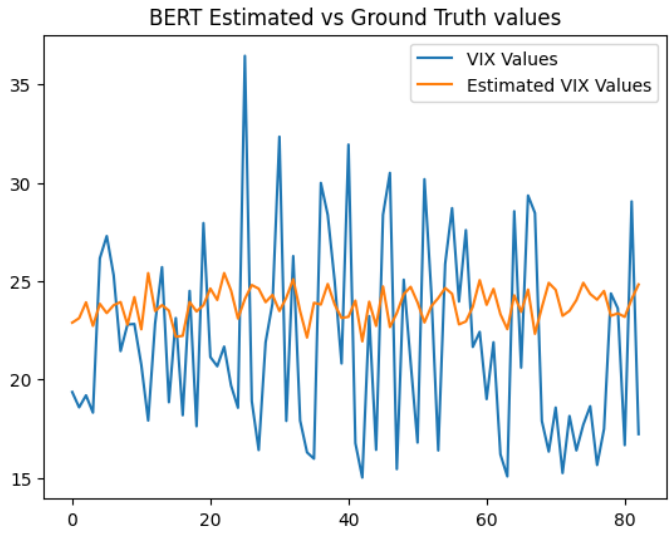

Figure 4: BERT output

7.2 Importance for Predicting VIX Values:

In predicting VIX values based on news articles and sentiment analysis, BERT's transformer architecture and bidirectional context understanding are instrumental. By leveraging BERT, the model can effectively capture the complex language used in financial contexts, interpret sentiment dynamics, and correlate these insights with market volatility, ultimately enhancing the accuracy of VIX predictions.

7.3 Model Training and Fine-Tuning:

The dataset was split into training and validation sets using a test size of 20% and a random seed of 42 for reproducibility. This allowed for assessing the model's performance on unseen data during validation.

The BERT-based predictive model underwent fine-tuning using a batch size of 16 during training, which encompassed 5 complete epochs. The learning rate for this training process was set to 1e-5. This meticulous fine-tuning process was essential to optimize the model's performance, ensuring its ability to effectively predict the VIX index based on inputs from news articles.

8 RoBERTa

Liu et al. [4] RoBERTa, short for "A Robustly Optimized BERT Approach," is an advanced transformer-based language model developed by Facebook AI, in 2019. It represents a significant evolution of BERT, incorporating optimizations that enhance model performance and robustness. RoBERTa achieves state-of-the-art results in various NLP tasks, offering improvements in language understanding and representation learning.

8.1 Key Features of RoBERTa Compared to BERT for VIX Prediction:

- **Enhanced Pre-training Methodology:** RoBERTa's pre-training process involves longer training times and larger batch sizes compared to BERT. This extended pre-training results in more thorough language understanding, allowing RoBERTa to capture intricate linguistic nuances within news articles more effectively.
- **Dynamic Masking during Pre-training:** Unlike BERT, RoBERTa utilizes a more sophisticated masking strategy during pre-training. RoBERTa employs dynamic masking patterns that change at every training epoch, encouraging the model to learn more robust representations of the text and improving its ability to generalize to diverse inputs.
- **Removal of Next Sentence Prediction (NSP) Task:** RoBERTa excludes the NSP task during pre-training, focusing solely on the masked language modeling (MLM) objective. This modification helps RoBERTa develop a deeper understanding of context within sentences, leading to more accurate predictions based on sequential information in news articles.

8.2 Performance Advantages Over BERT:

RoBERTa's enhancements enable superior prediction of VIX movements from news articles. Its extended pre-training and dynamic masking capture subtle language patterns and dependencies critical for accurate market forecasts, while omitting the NSP task refines its contextual understanding, outperforming standard BERT models in this specialized domain.

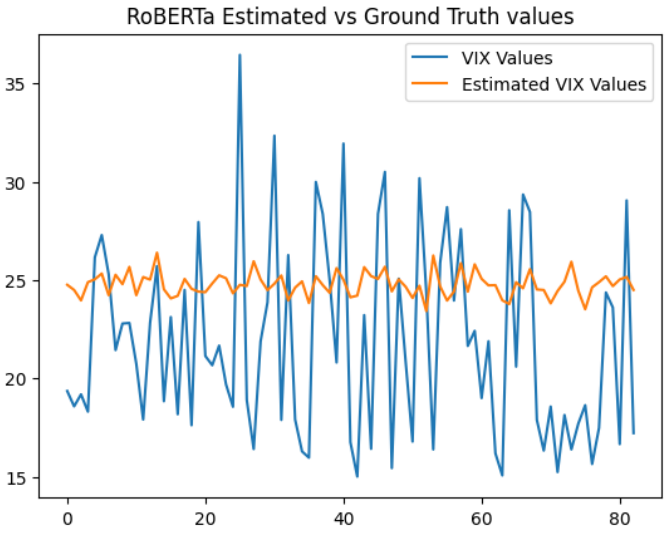

Figure 5: RoBERTa output

8.3 Model Training and Fine-Tuning:

During model development, the dataset was split into training and validation sets using a test size of 20% and a random state of 42 to ensure consistency. The model was fine-tuned with a training batch size of 10 and a validation batch size of 8, trained over 5 epochs with a learning rate set to 1e-5.

9 FinBERT

FinBERT [1] is a specialized version of BERT that has been developed for financial analysis. FinBERT is exclusively

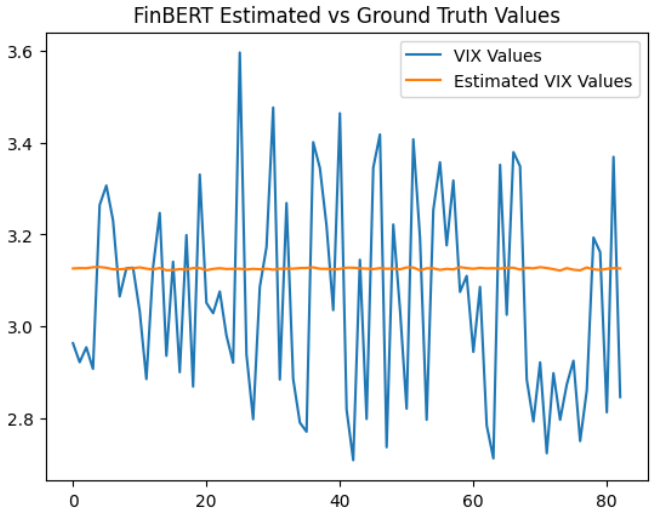

Figure 6: FinBERT Output

trained on financial articles, finance news blogs and reports. It is used for sentiment analysis to gauge the sentiment of financial news articles, reports and social media posts.

9.1 Key Features of FinBERT Compared to BERT and RoBERTa for VIX Prediction:

- **Domain-Specific Pretraining:** FinBERT is pretrained on financial domain-specific text, including news articles, financial reports, and other financial documents. This domain-specific pretraining can help FinBERT better understand financial jargon, terminology, and context, which may be particularly relevant for estimating VIX values.

- **Familiarity with Financial Vocabulary:** FinBERT may have learned embeddings for financial terms and concepts during pretraining, making it more familiar with the vocabulary used in finance. This familiarity can help FinBERT better capture the nuances of financial text and make more accurate predictions for tasks related to financial markets, including VIX estimation.

- **Robustness to Financial Noises:**Financial texts often contain noise, such as typographical errors, abbreviations, and slang terms. FinBERT's training on a diverse range of financial documents enables it to learn robust representations resilient to such noise. By capturing the underlying semantics and context of financial language, FinBERT can effectively filter out irrelevant information and focus on extracting relevant signals for financial analysis tasks like VIX estimation.

- **Interpretability of Results:**Due to its domain-specific training and familiarity with financial language, the outputs generated by FinBERT are often more interpretable in the context of financial analysis. For example, when FinBERT predicts sentiment scores for financial news articles or social media posts, users can better understand the rationale behind the model's predictions, as they are grounded in financial vocabulary and concepts. This interpretability enhances trust in the model's outputs and facilitates decision-making in financial markets.

9.2 Model Training and Fine-Tuning:

During model development, the dataset was split into training and validation sets using a test size of 20% and a random state of 42 to ensure consistency. The model was fine-tuned with a training batch size of 8 and a validation batch size of 8, trained over 5 epochs with a learning rate set to 1e-4.

10 Evaluation and Results

We trained and fine-tuned BERT, RoBERTa and FinBERT over 5 epochs with varying learning rates and batch sizes to observe the model performance. Additionally, we noted the training time for each model to observe how fast each model performs evaluation. The training for each model was performed using summarized texts obtained after text summarization while data preprocessing. The VIX values used for training include log-normalized values of the VIX values to ensure that the values are not too skewed. However, the graphs used have the un-normalized ground truths and estimated values of the Volatility Indices. The results have been tabulated in Table 1.

After fine-tuning all the models, we observed that FinBERT performed better than our baseline BERT and RoBERTa. However, Roberta, even with a general dataset, was able to estimate the VIX values well.

11 Next Steps

The next steps for our project include developing models using BERT for regression and classification. Adding to this, we are also planning on implementing other language models like RoBERTa and DistilBERT to observe how these models compare with the traditional BERT. Additionally, since the value of VIX is a time series, we plan to explore models with designs that incorporate this temporal relationship.

Furthermore, we plan on making our data more expansive to include articles before 2019 giving us more data to train our model on.

12 Alternate Classification Approach (Future Step)

An alternative to the regression-based forecasting of market volatility used in our prototype is a classification approach that utilizes binning based on the VIX index values. This method can be utilized if the regression based approach fails in producing appropriate results due to the limitations of Language Models and lack of data.

In this method, the continuous scale of the VIX is discretized into categories representing different levels of volatility. The categories are delineated by Z-scores, translating to distinct VIX value ranges that are then labeled as 'Low', 'Moderate', 'High', and 'Extreme' volatility.

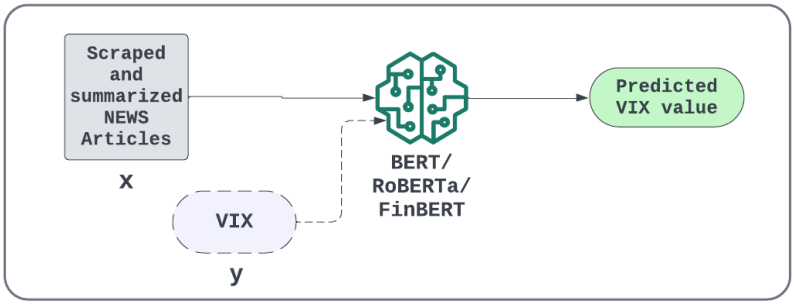

Figure 7: Model pipeline

Parameters	BERT	RoBERTa	FinBERT
Training Time	30 minutes	20 minutes	17 minutes
Training Batch Size	16	8	8
Validation Batch Size	16	10	8
Learning Rate	1e-5	1e-5	1e-4
Mean Squared Error	0.0716	0.0563	0.0055

Table 1: Comparison of Different Models

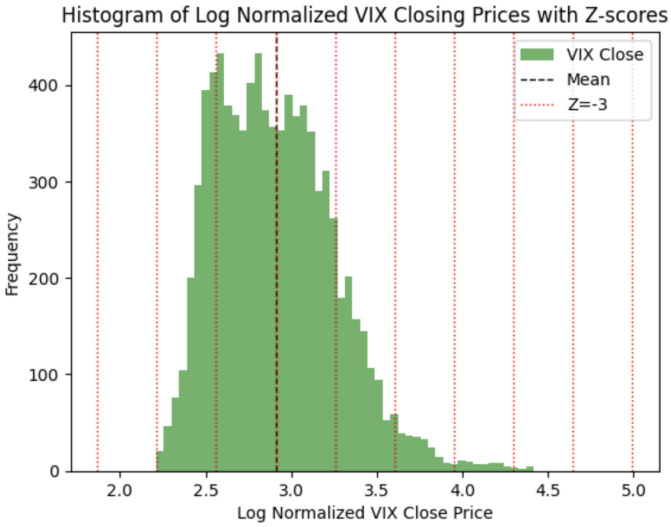

Figure 8: Histogram of Log Normalized VIX Closing Prices with Z-score based bins.

Specifically, the bins are defined as follows:

- 'Low' volatility corresponds to VIX values from 6.47 to 12.95,
- 'Moderate' volatility ranges from 12.95 to 18.32,
- 'High' volatility spans from 18.32 to 36.68,
- 'Extreme' volatility includes values from 36.68 to 73.45 and beyond.

This binning system enables the use of classification algorithms to predict volatility categories, providing a clear and actionable output that can be directly applied to trading strategies and risk management.

13 Conclusion

This study has made significant strides in advancing the predictive modeling of market volatility through the utilization of cutting-edge Natural Language Processing (NLP) techniques. By integrating BERT, RoBERTa, and FinBERT models, our approach has moved beyond traditional sentiment analysis to more accurately forecast the VIX (CBOE Volatility Index). Our findings demonstrate the robustness of these language models in interpreting complex financial narratives from news articles, which significantly impacts the predictive accuracy of market volatility.

However, our research encountered constraints related to the limited availability of historical data and computational resources. The scope of data spanning from 2019 to 2022 provided a foundational understanding but also highlighted the need for a more extensive dataset to capture a wider range of market behaviors and trends. Additionally, the computational power available limited our ability to conduct more extensive parameter tuning and experimentation with larger models or more complex architectures, which could potentially enhance prediction accuracy and model robustness.

Despite these limitations, the comparative analysis of the models underscored FinBERT's superior performance due to its specialization in financial contexts, highlighting the critical importance of domain-specific training in NLP applications. The regression-based forecasting approach we implemented has proven effective, providing a nuanced understanding of volatility based on textual data rather than just numerical indicators.

Looking ahead, our future work will focus on expanding the dataset to encompass a broader timeline and integrating more diverse language models to enhance the depth and accuracy of our predictions. We also aim to refine our models to incorporate the temporal dynamics of financial

markets, recognizing the evolving nature of news impact on market behavior. Furthermore, our results provide compelling evidence that a larger-scale study could be undertaken, provided there are enhancements in data availability and computational capabilities.

References

[1] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

[2] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. 2018. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

[5] Mark Mitchell and J Harold Mulherin. 1994. The impact of public information on the stock market. *Journal of Finance*, 49(3):923–50.

[6] Global Database of Events Language and Tone. [link].

[7] Stefan Pasch and Daniel Ehnes. 2022. Stonkbert: Can language models predict medium-run stock price movements?

[8] Karlo Puh and Marina Bagić Babac. 2023. Predicting stock market using natural language processing. *American Journal of Business*, 38(2):41–61.