TRANSFORMERS IN FINANCIAL NLP: MARKET MOVEMENTS FROM 8-K REPORTS

David Laziuk
Stevens Institute of Technology
Hoboken, NJ
dlaziuk@stevens.edu

ABSTRACT

In this study, we harness the BERT transformer model to predict stock market reactions to 8-K reports from S&P 500 companies. My approach included meticulous data curation, ensuring accurate ticker alignment and the exclusion of irrelevant currency entries. We then computed market movement targets using the "yfinance" tool for specific post-report time frames. Following extensive textual data preprocessing and exploratory data analysis, we established baseline models and proceeded to fine-tune BERT for multi-target regression. The results demonstrate BERT's superior predictive power for immediate post-report market openings, while its advantage tapered off for predictions at later time frames. This investigation highlights the strengths and potential limitations of transformer models in financial NLP, offering a pathway for future research in enhancing predictive financial analytics.

1 Introduction

1.1 Background

The exploration of financial market movements through the lens of Natural Language Processing (NLP) is a burgeoning field, offering new insights into how market sentiment and investor behavior can be influenced by corporate disclosures. Among the various documents analyzed in this domain, the 8-K reports filed by publicly traded companies in the United States have emerged as a critical source of information. These reports, mandated by the U.S. Securities and Exchange Commission (SEC), contain significant events or corporate changes that could be material to investors and thus, are believed to have a substantial impact on stock prices. This study aims to extend the current understanding of the relationship between 8-K reports and market movements by employing advanced NLP techniques, particularly focusing on transformer-based models.

Previous studies in this field have adopted diverse methodologies with varying degrees of success. For instance, the research titled "Forecasting Stock Excess Returns with SEC 8-K Filings" (Han, Wu, Ren, & Diane, 2022) approached this forecasting challenge by framing it as an imbalanced learning problem. It utilized Support Vector Machines (SVMs) with specially tailored Gaussian kernels and introduced an innovative dimension reduction stacking method. This study highlighted certain preferences in methodology, notably the use of TF-IDF vectorization over BERT embeddings, and pointed out potential drawbacks of dimensionality reduction in forecasting effectiveness.

Conversely, another notable study, "Attention-based Stock Price Movement Prediction Using 8-K Filings" (Masoud, n.d.) shifted the focus to deep learning techniques. It employed a parallel neural network architecture, incorporating dual attention mechanisms integrated with Bidirectional Long Short-Term Memory (BiLSTM) layers. This approach underscored the effectiveness of attention-based models in extracting pertinent linguistic signals from 8-K reports, suggesting their potential superiority in understanding complex financial narratives.

These preceding studies set the stage for our research, which aims to build upon and expand the existing knowledge in this area. By leveraging the latest advancements in NLP, specifically the transformer model BERT, this study seeks to provide a more nuanced and effective analysis of how 8-K reports influence market movements. The objective is to not only advance the academic understanding of financial NLP but also to offer practical insights that could aid investors and analysts in making more informed decisions based on corporate disclosures.

1.2 Task Description

The primary objective of this study is to investigate the impact of 8-K reports on the stock market movements of companies listed in the S&P 500 index. 8-K reports are significant regulatory filings that provide investors with timely information about major corporate events, such as mergers, acquisitions, bankruptcies, executive changes, and more. These reports are crucial as they often contain information that can materially affect a company's stock price. The task at hand is to analyze these reports using advanced NLP techniques and to assess their impact on market prices at various time intervals following the release of the report.

To accomplish this, the study involves several key steps. First, a comprehensive dataset of 8-K reports for companies in the S&P 500 is curated. This involves aligning the ticker symbols of these companies with the requirements of the "sec_edgar_downloader"(Chaar, n.d.) library and ensuring the exclusion of non-USD and GBP entities to maintain consistency in the financial analysis. Following this, historical 8-K reports for these companies are downloaded, forming the primary dataset for analysis.

The next critical step is the collection of target data, which in this context refers to the stock price movements of these companies. Using the "yfinance"(Aroussi, n.d.) tool, stock price data is gathered for specific intervals around the release of each 8-K report. These intervals include the stock price at the market open, the adjusted close, two days post-report, and one week post-report. The percentage change in the stock price relative to the price before the report release is calculated and standardized to a 0-based scale for consistency.

Once the data is collected, extensive preprocessing is conducted. This includes eliminating reports that lack essential information such as the time of release or corresponding price data, as well as those released during trading hours. Over 50 textual modifications are made to enhance the quality of the text data in the reports.

The study then progresses to exploratory data analysis (EDA) and the development of baseline models for comparison. These models include naive predictions based on mean and median price changes, a linear regression model using TF-IDF vectorization, and an SVM model employing TF-IDF and truncated SVD. The core of the study involves fine-tuning the BERT model for a multi-target regression task, aiming to predict the stock price movements at the specified intervals post-report release.

The findings of this study are expected to contribute to the growing field of financial NLP by providing insights into the effectiveness of transformer-based models, like BERT, in analyzing the impact of corporate disclosures on stock market movements.

1.3 Applications

The applications of this research are broad and multifaceted, spanning both the academic and practical realms of financial analysis and investment strategy. By harnessing the power of advanced NLP techniques, particularly transformer-based models like BERT, to analyze 8-K reports, this study opens up new avenues for understanding and predicting stock market movements. The insights derived from this research can be applied in various ways:

- **Investment Strategies:** Investors and financial analysts can use the insights from this study to refine their investment strategies. By understanding the impact of 8-K reports on stock prices, investors can make more informed decisions about buying, holding, or selling stocks. This research provides a deeper understanding of how specific types of corporate events reported in 8-K filings affect the stock market, enabling investors to anticipate market reactions to certain types of news.
- **Risk Management:** This study's findings can aid in risk assessment and management for portfolio managers and individual investors. By predicting potential market movements post-8-K report releases, investors can better gauge the risk associated with certain stocks or sectors. This can lead to more effective hedging strategies and risk mitigation techniques.
- **Algorithmic Trading:** The methodologies and models developed in this study can be integrated into algorithmic trading systems. These systems can use the predictive insights from the analysis of 8-K reports to automate trading decisions, potentially capitalizing on market movements immediately following the release of these reports.
- **Regulatory Analysis and Compliance:** Regulatory bodies and compliance officers can use the findings from this research to understand the market impact of 8-K reports better. This can assist in evaluating the

effectiveness of current disclosure requirements and in shaping future regulatory policies to ensure fair and transparent market practices.

- **Academic Research:** This study contributes to the field of financial NLP by exploring the effectiveness of transformer-based models in financial analysis. It opens up new research questions about the applicability of NLP techniques in financial markets and sets the groundwork for further studies that can explore different aspects of corporate disclosures and their market implications.

In summary, the practical implications of this study are significant, offering valuable tools and insights for investors, traders, regulatory bodies, and academics. The integration of advanced NLP techniques in financial analysis represents a promising frontier in the quest for more nuanced and effective market analysis tools.

1.4 Challenges

The initiation of this project was marked by the essential task of compiling a comprehensive dataset of 8-K reports from S&P 500 companies. The absence of an existing consolidated dataset necessitated the creation of one, crucial for an all-encompassing analysis of market responses to corporate disclosures. This step was fundamental to ensure that the analysis spanned a diverse and representative array of corporate sectors.

The most formidable challenge in applying Natural Language Processing to these reports was not the deployment of sophisticated models like BERT, but the intricate task of text data cleaning. The initial dataset, an overwhelming 140GB of unstructured data, was replete with noise, primarily due to HTML codes embedded in the text, posing significant hurdles in extracting relevant information.

This necessitated a thorough preprocessing of the text data to ensure its accuracy and applicability. Over fifty textual modifications were vital, necessitating innovative approaches that transcended conventional parsing methods like BeautifulSoup(Richardson, n.d.). A critical aspect was ensuring the start of the reports aligned with the onset of substantial content, rather than being buried in technical preambles. This task was particularly challenging, considering the dataset's expansive scale, encompassing over 100,000 documents.

Additionally, the task of extracting stock price data was fraught with its own complexities. Initially aimed at intra-day change analysis, the prohibitive cost of detailed historical data led to the adoption of less granular Yahoo Finance data. This necessitated meticulous alignment of the data with specific time intervals, adding further complexity to the project.

The extensive size of the textual corpus, encompassing billions of words, presented significant challenges in model fitting. The sheer volume of data tested the limits of baseline modeling techniques, highlighting that managing such scale was an integral part of the challenge, on par with advanced modeling.

In this context, while BERT's token limitation was a significant constraint, it was secondary to the foundational challenges of data cleansing and preparation. Tailoring the data to fit within BERT's input restrictions was just one aspect of the comprehensive data preprocessing undertaken. Ultimately, the project navigated through these complexities, yielding valuable insights into the dynamic relationship between corporate disclosures and market reactions.

2 Problem Formulation

In this study, we address the task of predicting stock market movements in response to 8-K reports as a multiple target regression problem. Two sets of models are developed: baseline models using TF-IDF vectorization and a fine-tuned BERT model for regression.

2.1 Baseline Models

For the baseline models, linear regression and SVM, the TF-IDF vectorized text of the 8-K reports, denoted as $\mathbf{X_{tfidf}}$, is used as the input. The TF-IDF vectorization is limited to a maximum of 1000 features. The SVM model additionally employs truncated Singular Value Decomposition (SVD) to reduce the feature dimensions to 80. The models predict the percentage changes in stock prices at different time intervals after the report's release, represented as $\mathbf{Y} = [\text{open}, \text{close}, 2\text{d}, 1\text{w}]$. The formal representation of the baseline models is given by the equations:

$$\mathbf{Y} = f_{\text{linear regression}}(\mathbf{X_{tfidf}}) \tag{1}$$

$$\mathbf{Y} = f_{\text{SVM}}(\text{SVD}(\mathbf{X_{tfidf}}, 80)) \tag{2}$$

2.2 BERT Fine-Tuning

The core of the study involves fine-tuning BERT for regression tasks, without freezing any layers. The inputs for BERT are the tokenized text and attention masks, represented as $\mathbf{X_{input\_ids}}$ and $\mathbf{X_{attention\_mask}}$, respectively. The output is the same set of target variables $\mathbf{Y}$. For regression, we modify BERT to use the pooled output of the [CLS] token. The BERT regression model can be formulated as:

$$\mathbf{Y} = f_{\text{BERT}}(\mathbf{X_{input\_ids}}, \mathbf{X_{attention\_mask}}) \tag{3}$$

2.3 Dataset and Model Architecture

The dataset comprises tuples of input_ids, attention masks, and target values, formally represented as:

$$\text{Dataset} = \{(\mathbf{X_{input\_ids}}[\mathbf{i}], \mathbf{X_{attention\_mask}}[\mathbf{i}], \mathbf{Y}[\mathbf{i}])\}_{i=1}^{N} \tag{4}$$

where N is the number of samples in the dataset. The BERT model architecture for this regression task, denoted as 'BERTForRegression', includes a dropout layer and a linear layer for output, defined as:

$$\text{BERTForRegression} = \text{BERTModel} \rightarrow \text{Dropout}(0.2) \rightarrow \text{Linear}(H, 4) \tag{5}$$

where H is the hidden size of the BERT model.

2.4 Evaluation Metrics

The model's performance is assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Directional Accuracy. These metrics provide a comprehensive evaluation of the model's accuracy and predictive capabilities in the context of financial market movements.

3 Methods

3.1 Data Aggregation

The in-house data aggregation phase was integral to the study, entailing two primary tasks: sourcing 8-K filings for S&P 500 companies and obtaining related stock market data.

3.1.1 Collection of 8-K Filings

The foundational phase of this project entailed the meticulous assembly of an up-to-date S&P 500 company list (Hunkar, n.d.). This step was not without its nuances, as it involved fine-tuning the ticker symbols to align with the requirements of the "sec_edgar_downloader" tool (Chaar, n.d.), a crucial component for accessing the SEC EDGAR database (*SEC EDGAR Database*, n.d.). Noteworthy adjustments included modifying "BRKB" to "BRK-B" and "BFB" to "BF-B", while "ATVI" (Activision) was omitted post its acquisition by Microsoft on October 12, 2023. Additionally, entries such as GBP and USD were removed, being currency-related rather than company tickers.

This preparatory stage set the stage for the subsequent bulk downloading of 8-K filings, a task that was both data-intensive and time-consuming. The download process itself was significant, encompassing approximately 140GB of data, a volume constrained by the SEC's API rate limits.

3.1.2 Collection of Target Stock Price Data

Concurrent to the collection of 8-K filings, a critical task was the extraction of stock price data corresponding to these reports, a process facilitated by the "yfinance" package (Aroussi, n.d.), which sources data from Yahoo Finance (Yahoo, n.d.). This involved an intricate process of retrieving stock prices for timeframes surrounding each report's release to analyze the impact of 8-K filings on the stock market.

Given the complexity of varying trading days throughout historical data, and the necessity for precise time intervals, my approach involved pulling data for 8 days prior to and 16 days following each report's release. This strategy ensured the availability of specific time intervals needed for the analysis, irrespective of the trading day configurations. It allowed for a more comprehensive assessment of the stock price movements pre- and post-report release, enhancing the depth and relevance of our analysis.

While this methodology was effective in capturing a broad view of market reactions, it is important to note that ideally, intra-day data would have been more insightful for a granular analysis of market responses. Such detailed data, though preferable for capturing immediate market reactions, was not feasible due to its prohibitive cost and accessibility challenges.

3.2 Data Pre-processing

The data pre-processing stage was instrumental in refining the raw data into a format suitable for analysis. This stage encompassed several key tasks, each designed to enhance the quality and relevance of the data.

3.2.1 Cleaning and Processing 8-K Filings:

The initial and crucial phase of this project was dedicated to the meticulous cleaning and enhancement of the 8-K filings dataset, originally a vast collection of 140GB. This process was pivotal in ensuring the data's suitability for the nuanced demands of NLP analysis in financial contexts.

Key steps in this cleaning process involved rigorous HTML cleaning to strip away web formatting, thus extracting the raw textual content. This step was essential to remove the extensive noise presented by HTML code inaccurately labeled as text. We also removed non-printable characters and normalized the text, such as converting all letters to lowercase, to maintain consistency across the dataset.

A crucial aspect of this process was identifying and excising specific unwanted elements like hyperlinks, numeric HTML entities, and various special characters. This filtration was vital to distill the core content of each report, focusing on the substantive information crucial for our analysis.

The cleaning process also involved sophisticated text enhancements. These included adjusting punctuation and spacing, rectifying abbreviations, and meticulously removing redundant or irrelevant parts of the text. For instance, special attention was paid to correctly spacing out words that were joined together, separating monetary amounts from text, and ensuring the start of each document was aligned with the beginning of the actual report content. These enhancements were not just about cleaning the data but refining it to a level where it became a high-quality input for advanced NLP models.

The result of these rigorous preprocessing efforts was a significant reduction in the dataset's size to about 10GB, a testament to the effectiveness of our cleaning and enhancement strategies. This meticulous approach to data preparation laid the groundwork for the accurate and contextually rich analysis that followed, demonstrating the importance of quality data in the realm of NLP and financial market analysis.

3.2.2 Aggregation into a Unified Dataset:

Post-cleaning, the individual text files were aggregated into a single CSV file, creating a consolidated dataset. This involved reading each processed text file, extracting relevant metadata (like ticker, accession number, and datetime), and compiling the information into a structured format. The dataset then underwent additional cleaning steps, such as removing documents with invalid datetimes or those filed during trading hours to ensure the reliability of the analysis. New columns were added to the dataset to facilitate the integration of stock price data in subsequent stages.

3.2.3 Integration with Stock Price Data:

The refined dataset of 8-K filings was then prepared for integration with the corresponding stock price data. This involved adding columns to the dataset for stock prices at specific intervals (prior, open, close, 2d, and 1w) relative to the report release. The combined dataset, now containing both the cleaned 8-K report text and relevant stock price information, formed the basis for the ensuing analysis. This comprehensive pre-processing phase was crucial in transforming the raw data into a format conducive to effective analysis, ensuring the accuracy and quality of the insights derived from the subsequent modeling and evaluation stages.

3.3 Model Overview

The model development for this study was designed to predict market movements from the text of 8-K reports using both traditional machine learning approaches and advanced neural networks. This section outlines the model design, loss function formulation, and training and inference processes.

3.3.1 Model Design:

Baseline models such as naive predictors, linear regression, and Support Vector Machines (SVM) were first established to set a performance benchmark. For the linear regression and SVM models, TF-IDF vectorization was employed to transform the text data into a suitable format for model input. A fine-tuning approach was adopted for the BERT (Bidirectional Encoder Representations from Transformers) model, leveraging its pre-trained embeddings and neural network architecture to capture the complex relationships within the text data.

3.3.2 Loss Function Design:

The Mean Squared Error (MSE) loss function was chosen for its effectiveness in regression tasks. It measures the average squared difference between the estimated values and the actual value, providing a clear gradient for model optimization.

3.3.3 Training and Inference:

The training process involved iteratively passing the tokenized text data through the BERT model, calculating the loss, and adjusting the model parameters via backpropagation. The AdamW optimizer, known for its adaptive learning rate capabilities, was used to update the weights. Training was conducted over multiple epochs with the intention to minimize the validation loss, and performance metrics such as Mean Absolute Error (MAE), MSE, Root Mean Squared Error (RMSE), and Directional Accuracy were monitored. Performance metrics were calculated for each epoch to track and compare the model's predictive power over time. Special attention was given to the BERT model's output, specifically utilizing the pooled output of the [CLS] token, which is often used as the aggregate representation of the input text for classification tasks.

4 Dataset and Experiments

The experimental framework of this study was designed to rigorously evaluate the predictive performance of NLP models on stock market movements from 8-K report texts. The experiment was conducted with a carefully curated dataset, detailed data statistics, and a series of exploratory data analysis to understand the data distribution and characteristics.

4.1 Experimental Settings:

- **Model Parameters:** For the BERT model, the 'bert-base-uncased' version was fine-tuned with an added linear regression layer to predict the target variables. The model's parameters were optimized to handle the specificity of the financial text data contained within the 8-K reports.
- **Hardware/Software Environment:** The training and inference processes were carried out on a single NVIDIA A6000 in a Lambda Labs cloud notebook environment.
- **Training Settings:** The models were trained over 30 epochs with a batch size of 16. The AdamW optimizer was utilized as it is a standard for BERT fine-tuning, due to its ability to adapt learning rates based on the training dynamics. A ReduceLROnPlateau scheduler was applied to adjust the learning rate upon the plateauing of validation loss.

4.2 Exploratory Data Analysis (EDA):

4.2.1 Report Statistics:

The final dataset consisted of 104,545 samples, encompassing 502 unique companies from the S&P 500 index. The distribution of samples per company was analyzed to ensure a comprehensive representation across the dataset. The distribution is skewed, with a few companies having a significantly higher number of reports than others. The EDA commenced with an analysis of the distribution of samples per company, highlighting disparities in reporting frequencies. Key statistics included:

- Minimum Samples: The minimum number of samples for a single company was 4.
- Percentiles: The 25th, 50th (median), and 75th percentiles stood at 143, 196, and 252 samples, respectively.
- Mean: The mean number of samples per company was approximately 208.
- Maximum Samples: The maximum number of samples for a single company was 859, belonging to a major financial institution.
- The temporal range of the reports spanned from March 1995 to November 2023, providing a broad historical context for the study.

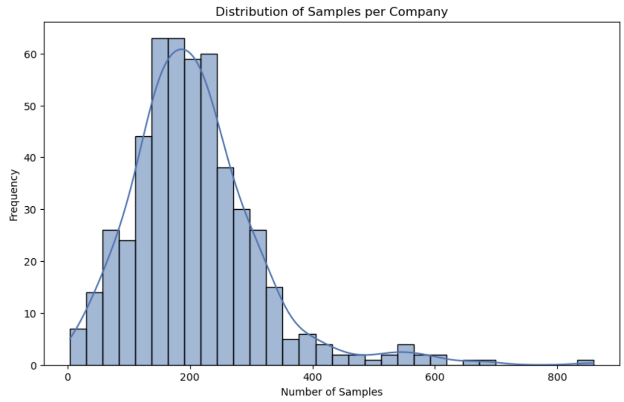

Figure 1: Samples Per Company

As seen in Figure 1, the distribution of samples per company exhibits a right-skewed normal distribution, with a mean of 208. This skewness indicates that while most companies have a moderate number of reports, a few companies have a disproportionately high number, which could affect the training of the predictive models.

4.2.2 Text Statistics:

An extensive analysis of the text data was performed to assess the variability in the length and composition of the 8-K report texts. This examination was crucial in understanding the dataset's structure and ensuring the models could handle texts of varying lengths and complexities.

- **Text Length in Words:** The lengths of the reports varied significantly, with the shortest report containing 48 words and the longest containing over 27 million words. The distribution of word counts exhibited a wide range, with the mean notably higher than the median due to the presence of some exceptionally long documents. Note, this may be due to edge-case failures of text cleaning, and HTML artifacts being represented as words.
- **Text Length in Characters:** Character counts also displayed a broad spectrum, from a minimum of 314 characters to a maximum of approximately 75 million characters.
- **Word Lengths:** The analysis of word lengths revealed that the corpus contained words ranging from a single character to 1,763 characters. The mean word length was slightly above 4 characters, indicating a corpus with a typical mix of short and long words. These long "words" are also attributed to cleaning failures.
- **Corpus Vocabulary:** The corpus was rich in vocabulary, containing nearly 9.4 million unique words out of over 1.4 billion total words, which amounted to more than 8 billion total characters. This extensive vocabulary is attributed to punctuation and symbols being included.

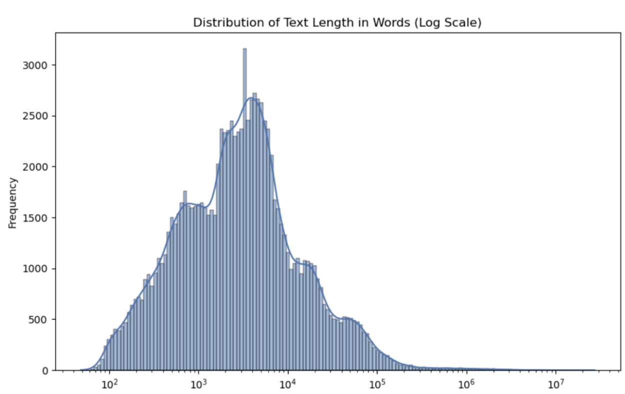

Figure 2: Text Length Distribution

4.2.3 Price Statistics:

The exploratory analysis of stock price changes revealed notable trends post the release of 8-K reports. Key statistical insights included a pronounced increase in stock prices at the market open, followed by a general stabilization in subsequent timeframes. Specifically:

- **At Open:** The average increase in stock price was approximately 21.07%, with some stocks reaching an impressive surge of up to 1187.5%.
- **At Close:** The median change hovered around 0.06%, indicating a regression toward baseline levels.
- **Two Days Post-Report (2d)** and **One Week Post-Report (1w):** Both time points showed minimal median changes, suggesting that the initial impact observed at the open had dissipated.

The standard deviation of stock price changes at the open was notably high at 40.35%, reflecting significant variability and potential volatility immediately following report releases. In contrast, the standard deviation at the close and beyond was much lower, demonstrating less fluctuation and a return to typical market conditions.

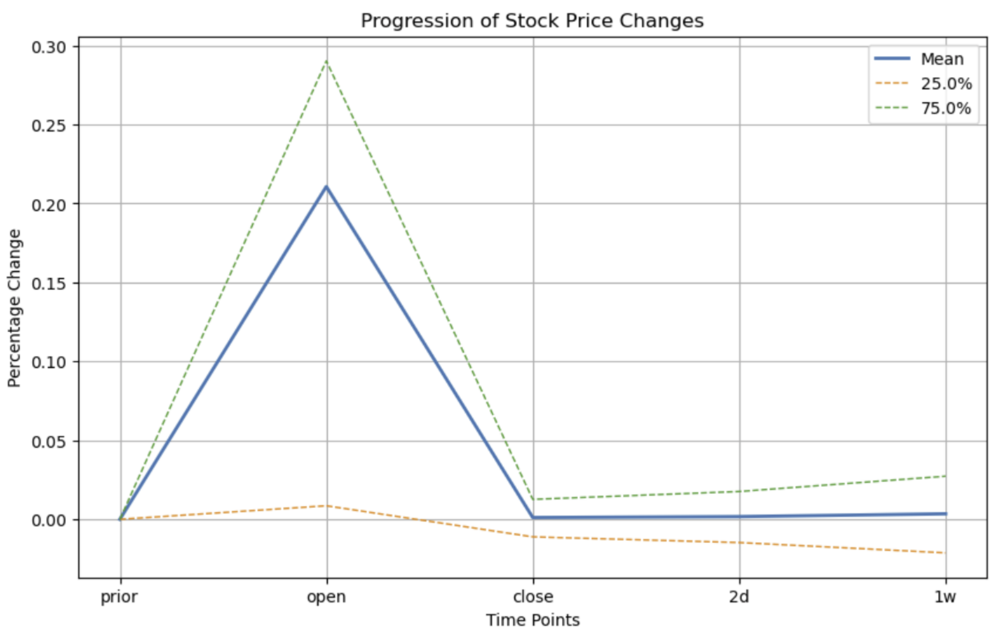

Figure 3: Mean Price Progression

Figure 3 depicts these dynamics, showing the initial spike in stock price changes at the open with a rapid decline towards the close, which then levels off in the days following. The figure illustrates the progression of stock prices, capturing the immediate market reaction and the gradual stabilization over time. This pattern emphasizes the short-lived effect of 8-K reports on stock prices, which could be attributed to the market's assimilation of the new information and subsequent price correction mechanisms. Understanding this behavior is vital for modeling strategies that aim to predict immediate and short-term market movements.

4.3 Experimental Results:

4.3.1 BERT Model Tuning and Training Curve:

The fine-tuning of the BERT model was meticulously tracked, with key performance metrics recorded at each epoch. The Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Directional Accuracy were all monitored to gauge the model's predictive proficiency over time. Models were saved at the points of best MAE, MSE (loss), RMSE, and Directional Accuracy, which represent the states where the model achieved its superior performance by those measures.

Throughout the training process, a notable trend was observed:

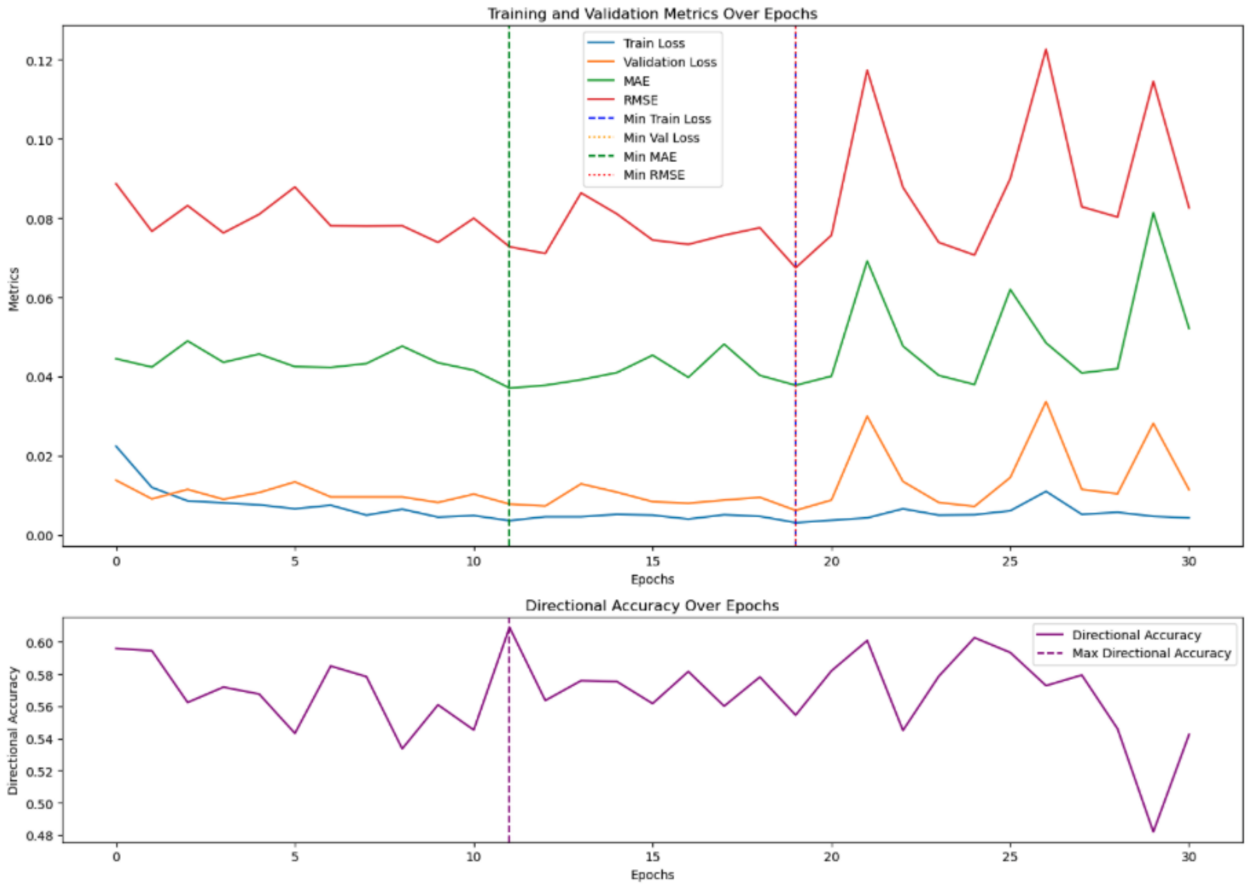

Figure 4: BERT Training Performance

Figure 4 illustrates the training and validation performance across epochs. Note, 'MSE' is not included as it is the same as Validation Loss. There is some overlap with best performing models across metrics.

- The MAE for stock price changes at the opening fluctuated but showed a trend of decrease, highlighting the model's improving precision in predicting immediate post-report stock price movements.
- The Directional Accuracy started strong and generally maintained a level above 50%, indicating the model's consistent capability to correctly predict the direction of stock price movements.
- The validation loss initially decreases, indicating that the model is learning generalizable patterns from the data.
- However, fluctuations in validation loss and the performance metrics like MAE and RMSE suggest the complexity of the task and the model's sensitivity to the nuances of the financial text data.

The dashed lines in the graph mark the epochs where the models achieved their best loss and metric values, signifying the points where models were saved. This approach was instrumental in capturing the most efficient state of the model for each metric. Noticing a decline in training performance around epoch 20, the training was concluded at epoch 30, instead of the originally allocated 100 epochs.

4.3.2 Test Set Performance

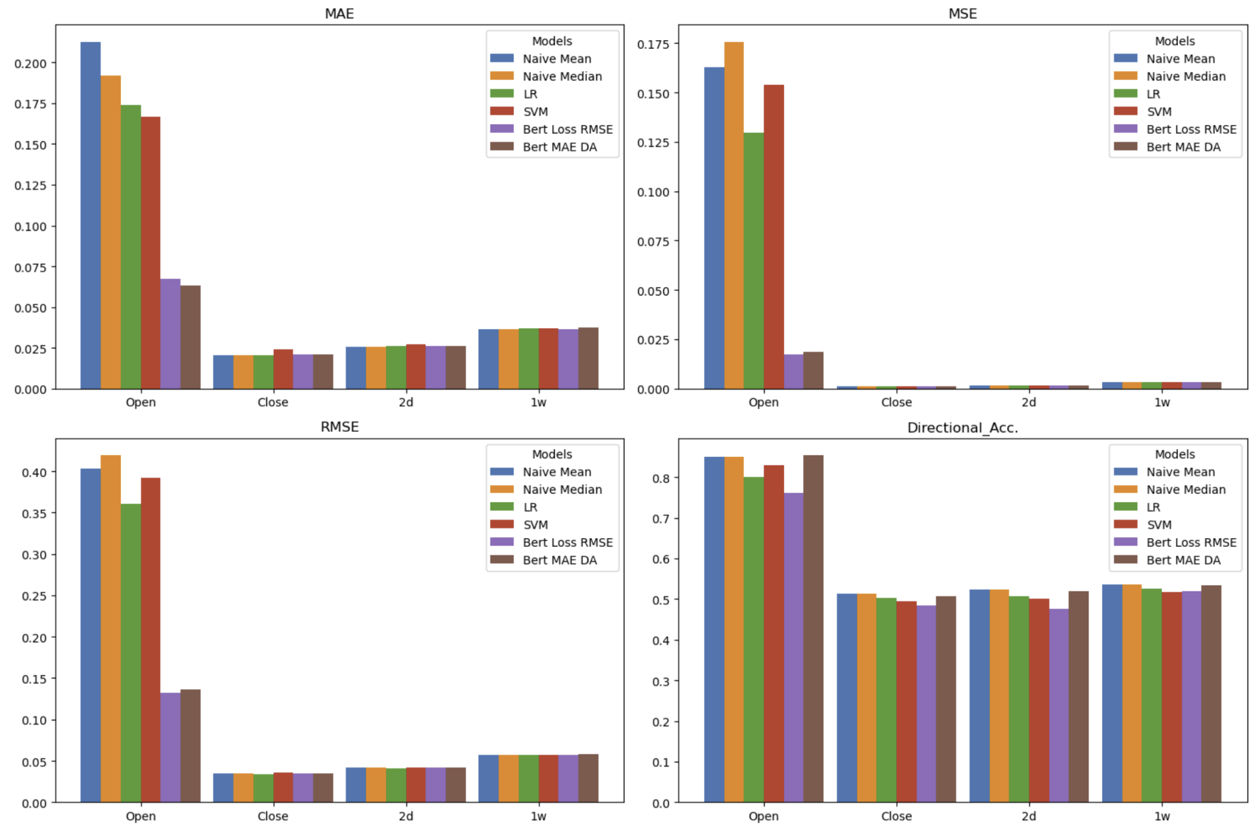

Figure 5: Test Set Performance

The performance of various models on the test set provided significant insights into the effectiveness of each approach in predicting stock price movements from 8-K reports. The BERT model, fine-tuned on the financial text corpus, was evaluated against baseline models, including Naive Mean, Naive Median, Linear Regression (LR), and Support Vector Machine (SVM) models. The BERT model variants optimized for different metrics showed distinct performance characteristics. The visual representation of model performance in Figure 5 reflects this diversity. It can be seen that:

- The BERT model fine-tuned for Mean Absolute Error (MAE) and Directional Accuracy (DA) achieved the lowest MAE for the 'Open' stock prices, substantially outperforming the Naive baselines and traditional machine learning approaches.
- For 'Close', '2d', and '1w' predictions, the performances of the models were more closely matched, with the Naive Median performance being particularly competitive.
- In terms of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), the BERT model fine-tuned for loss (which corresponds to MSE) showed superior performance on the 'Open' predictions, but this advantage was less pronounced for 'Close', '2d', and '1w' predictions.

The Directional Accuracy of the models varied, with the BERT model optimized for RMSE showing the best performance for 'Open' prices, suggesting a strong ability to predict the direction of price movement at market open. However, for 'Close' and subsequent time points, this model's advantage was less evident. The results indicate that while the BERT model's complex features extraction capabilities offer a distinct advantage for immediate post-release predictions, the chaotic nature of the stock market may dilute this advantage over time. Furthermore, the comparable performances of the Naive Median and traditional models at later time points suggest that the market's response to 8-K reports may be inherently difficult to predict with high accuracy, emphasizing the need for sophisticated models capable of capturing subtle linguistic cues.

The performance analysis not only underscores the potential of NLP in financial applications but also highlights the challenges in making predictions in a highly volatile environment. The fine-tuned BERT model, despite its sophistication, encounters the complexities of financial markets, where many factors beyond the text of 8-K reports influence stock prices.

5 Project Management

The execution of this project was undertaken as an individual effort, with all roles and responsibilities fulfilled by myself. This encompassed the entire gamut of research activities, from the initial proposal through data collection, preprocessing, modeling, and up to the final report submission.

5.1 Roles and Responsibilities:

As the sole member of this project, the responsibilities included:

- Research: Conducting literature reviews and determining the project's scope based on the latest findings in the field of NLP and financial market prediction.
- Data Collection: Gathering historical 8-K reports for S&P 500 companies and the corresponding stock price data.
- Data Preprocessing: Cleaning and preparing the text data for modeling, which involved handling large datasets and ensuring data quality.
- Feature Engineering: Extracting features from the text data that could be predictive of market movements.
- Model Development and Tuning: Implementing various machine learning models, including advanced techniques like BERT, and tuning them for optimal performance.
- Evaluation: Assessing the models' performance using various metrics and adjusting strategies accordingly.
- Documentation and Reporting: Writing the proposal, midterm report, final report, and preparing the project presentation, along with maintaining comprehensive code documentation.

5.2 Project Timeline

The project was structured around key milestones, which were met according to the following timeline:

- Oct 10: Submission of the project proposal, which laid out the research question, methodology, and expected outcomes.
- Oct 24: Completion of data collection, which provided the raw material for the subsequent analysis and model training.
- Nov 9: Finalization of data preprocessing and the initial round of feature engineering, setting the stage for model development.
- Nov 10: Submission of the midterm report, detailing the project's progress and preliminary findings.
- Nov 20: Advanced model tuning and experimentation phase, where the BERT model was fine-tuned, and various experiments were conducted to refine the predictive models.
- Dec 4: Presentation of the project, summarizing the research process, methodologies employed, and the results obtained.
- Dec 11: Final report and code submission, culminating the project with a comprehensive set of documents and artifacts that encapsulate the research findings and technical work.

6 Conclusion

This project aimed to combine Natural Language Processing with financial market analysis, using Transformer-based models to analyze 8-K reports and predict market movements. It demonstrated the capabilities of NLP in financial contexts and provided insights into how stock prices react to corporate disclosures.

Key steps in the project included:

- Collecting and preprocessing a large dataset of 8-K reports from S&P 500 companies.

- Comparing baseline models to a fine-tuned BERT model, with BERT showing better performance for immediate post-release price predictions.
- Addressing challenges in financial NLP, from data processing to model evaluation, within a tight timeline.

The project's findings highlight the complexities of using machine learning in financial markets, where unpredictability and numerous influencing factors make accurate prediction challenging. The BERT model showed promise, especially for immediate market responses, suggesting that 8-K reports hold valuable information, though its impact may be short-lived.

Potential improvements include:

- Access to intra-day data for more immediate market reaction analysis.
- Utilization of Transformer models capable of handling larger input sizes or splitting the input for a voting system from different runs.
- Enhanced text cleaning to further improve data quality.

In summary, while accurately predicting financial market movements remains challenging, this project shows that NLP can provide significant insights into market-moving events and paves the way for future research in this area.

References

Aroussi, R. (n.d.). *yfinance – python package.* `https://pypi.org/project/yfinance/`. (Accessed: 2023-10-26)

Chaar, J. (n.d.). *sec-edgar-downloader – python package.* `https://pypi.org/project/sec-edgar-downloader/`. (Accessed: 2023-10-24)

Han, H., Wu, Y., Ren, J., & Diane, L. (2022). Forecasting stock excess returns with sec 8-k filings. In H. Han & E. Baker (Eds.), *The recent advances in transdisciplinary data science* (pp. 3–18). Cham: Springer Nature Switzerland.

Hunkar, D. (n.d.). *The complete list of constituents of the sp 500 index.* `https://topforeignstocks.com/indices/components-of-the-sp-500-index/#google_vignette`. (Accessed: 2023-10-24)

Masoud, M. (n.d.). *Attention-based stock price movement prediction using 8-k filings.* Retrieved from `https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15751328.pdf`

Richardson, L. (n.d.). *beautifulsoup4 – python package.* `https://pypi.org/project/beautifulsoup4/`. (Accessed: 2023-10-24)

Sec edgar database. (n.d.). `https://www.sec.gov/edgar.shtml`. (Accessed: 2023-10-24)

Yahoo. (n.d.). *Yahoo finance.* `https://finance.yahoo.com`. (Accessed: 2023-10-26)