ADVANCED NLP, ANLY5800

Stock Performance Prediction Based on Annual Report

-Jingda Yang, Sukriti Mahajan, Tingsong Li

INDEX

1. Introduction

a. Objective

The primary objective of this project is to develop a robust and innovative approach for predicting stock market performance by leveraging the untapped potential of textual data contained within annual reports (10-K filings) of publicly traded companies. This project aims to bridge the existing gap in financial analysis, which often focuses predominantly on quantitative financial data, by integrating advanced natural language processing (NLP) techniques.
Key objectives include:

2. **Utilizing Advanced NLP Techniques:** Implementing state-of-the-art NLP models, particularly BERT (Bidirectional Encoder Representations from Transformers), to analyze and extract meaningful insights from the textual components of 10-K reports. This involves understanding the sentiment, tone, and thematic elements that might influence a company's stock performance.
3. **Enhancing Stock Performance Predictions**: Combining the insights gained from textual analysis with traditional financial metrics to improve the accuracy of stock performance predictions. This integrated approach aims to provide a more comprehensive understanding of a company's future financial health.
4. **Data Acquisition and Processing:** Collecting and processing extensive datasets from annual reports and financial databases, ensuring the data is clean, relevant, and formatted appropriately for analysis.
5. **Model Development and Testing**: Building and fine-tuning a predictive model that can effectively process and analyze large volumes of textual and financial data. The project will also involve a rigorous testing phase to evaluate the model's accuracy and reliability using standard metrics like precision, recall, F1 score, and accuracy.

Through these objectives, the project endeavors to make a significant contribution to the field of financial analysis, offering a novel perspective by marrying advanced NLP techniques with traditional financial analysis methods.

b. Significance of the Project

The significance of this project lies in its innovative approach to stock market analysis. Traditionally, stock performance prediction has relied heavily on quantitative financial data, such as earnings, revenue, and market trends. However, this project recognizes the substantial yet often overlooked value in the textual content of annual reports, specifically the 10-K filings.

1. **Holistic Analysis:** By analyzing the textual data in 10-K reports using advanced NLP techniques, this project introduces a more holistic approach to stock market analysis. It acknowledges that financial health and potential risks or opportunities for a company are not solely reflected in numerical data but also in qualitative disclosures.
2. **Improved Predictive Accuracy:** Integrating qualitative analysis with quantitative data potentially leads to improved predictive accuracy. Textual analysis can unveil market sentiments, strategic insights, and forward-looking statements that numbers alone might not reveal.
3. **Early Detection of Trends and Risks:** Textual analysis of 10-K reports can help in early detection of emerging trends, potential risks, and opportunities, which may take longer to be reflected in financial figures. This early detection is crucial for investors and analysts in making timely decisions.
4. **Enhanced Investor Confidence:** Providing a more comprehensive analysis can boost investor confidence, as decisions are based on a blend of quantitative data and qualitative insights, leading to more informed investment strategies.

2. Understanding 10-K

A Form 10-K is an annual report filed by publicly traded companies in the United States with the Securities and Exchange Commission (SEC). It provides a detailed summary of a company's financial performance and includes information that is not typically found in the more concise annual report.

a. Components of a 10-K

The 10-K report comprises several key components:

1. **Business Summary**: This section provides an overview of the company's main operations, products, and services.
2. **Risk Factors**: It details the potential risks and uncertainties the company faces.
3. **Selected Financial Data:** This offers a summary of financial performance over the last five years.
4. **Management's Discussion and Analysis (MD&A):** It includes management's perspective on the financial results and condition of the company.
5. **Financial Statements:** These are complete financial statements including balance sheets, income statements, and cash flow statements.
6. **Controls and Procedures:** It covers information on the company's internal control over financial reporting.
7. **Other Information**: Includes legal proceedings, market risk, and other significant data.

8. **Exhibits and Financial Statement Schedules:** Comprises supplementary legal documents, statistical data, and other important information.

b. 10-K in Financial Predictions: Qualitative Insights

Incorporating 10-K reports into financial predictions allows for an additional layer of qualitative insights:

1. **Sentiment and Tone Analysis**: By analyzing the language and tone of the 10-K reports, NLP models can gauge the sentiment towards various aspects of the company's operations, prospects, and market position.
2. **Strategic Insights:** 10-K reports often contain information about a company's strategy, goals, and market positioning, providing a deeper understanding of its future direction and potential for growth.
3. **Risk Assessment**: Analyzing the Risk Factors section can offer foresight into potential challenges and uncertainties the company might face, which could impact its stock performance.
4. **Competitive Landscape**: Information about competitors and market trends in the 10-K reports can provide valuable insights into the company's competitive position and potential market opportunities.

In conclusion, this project's integration of qualitative analysis from 10-K reports with quantitative financial data aims to revolutionize stock performance prediction, providing a more nuanced, comprehensive, and accurate analysis for investors and financial analysts.

3. Literature Review

a. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models [https://arxiv.org/abs/1908.10063]

The key findings from the research are as follows:

1. **Introduction of FinBERT**: The paper introduces FinBERT, a language model specifically tailored for NLP tasks in the financial domain, based on the BERT model. It addresses the challenges in financial sentiment analysis, such as the specialized language and scarcity of labeled data.

2. **State-of-the-Art Performance**: FinBERT achieved state-of-the-art results on two financial sentiment analysis datasets: FiQA sentiment scoring and the Financial PhraseBank. This demonstrates FinBERT's effectiveness in handling the nuances of financial language for sentiment analysis.

3. **Exploration of Model Aspects:** The research conducted experiments to explore various aspects of the FinBERT model, including the effects of additional pre-training on financial corpora, strategies to prevent catastrophic forgetting, and fine-tuning only a subset of model layers. These experiments aimed to optimize training time and model performance.

4. **Comparative Performance Metrics:** FinBERT was evaluated against other pre-trained language models like ULMFit and ELMo, as well as baseline methods. The results showed that FinBERT outperformed these models in metrics such as loss, accuracy, and F1 score on the Financial PhraseBank dataset, both on the whole dataset and in subsets with 100% annotator agreement.

b. Stock Price Prediction using BERT and GAN

[https://ar5iv.labs.arxiv.org/html/2107.09055v1]

The key findings from the research paper are as follows:

1. **Combination of Sentiment Analysis and Technical Analysis**: The paper presents a novel approach that combines sentiment analysis using BERT (a pre-trained transformer model by Google for NLP) and technical analysis using a Generative Adversarial Network (GAN) for predicting stock prices of Apple Inc. This approach integrates sentiment analysis from news and headlines with technical indicators, stock indexes, commodities, and historical prices.
2. **Innovative Use of GAN**: The GAN in the study is modified to generate sequences rather than 2-dimensional data, using a Gated Recurrent Unit (GRU) as the generator and a 1-dimensional Convolutional Neural Network (CNN) as the discriminator. The model takes sentiment scores as the input vector, which aids in early convergence and better prediction accuracy. Stock price predictions are made on 5-day, 15-day, and 30-day horizons, evaluated based on the Root Mean Squared Error (RMSE).
3. **Comparative Analysis with Other Models**: The proposed model was evaluated and compared with other baseline models such as ARIMA, LSTM, GRU, and plain vanilla GAN with noise as latent input. The comparison focused on the RMSE metric for 5-day, 15-day, and 30-day stock price predictions.

c. Stock Movement Prediction with Financial News using Contextualized Embedding from BERT

[https://ar5iv.org/abs/2107.08721]

The key findings of the research paper are:

1. **Introduction of FT-CE-RNN**: The paper introduces a novel text mining method called Fine-Tuned Contextualized-Embedding Recurrent Neural Network (FT-CE-RNN) specifically designed to predict the short-term movement of stock prices following financial news events using only the news headlines. This approach significantly differs from previous methods that relied on static vector representations of news.

2. **Use of Contextualized Embeddings**: Unlike traditional models, the FT-CE-RNN utilizes contextualized vector representations of headlines generated from the BERT model. This method allows for a more nuanced understanding of financial news content, leveraging the domain-specific knowledge embedded in the financial news data.

3. **New Evaluation Metric and Trading Simulations**: The research introduced a new evaluation metric that calculates accuracy based on various percentiles of prediction scores on the test set, rather than the entire set, aligning more closely with investors' interests. Additionally, the paper included trading simulations with different strategies to evaluate the practical application of the FT-CE-RNN model in real-world scenarios

d. BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability

[https://arxiv.org/abs/1906.09024v2]

The key findings of the research paper are as follows:

1. **Innovative Approach for Sentiment Analysis**: The paper introduces a new method for constructing a financial sentiment index using BERT.. This approach marks a departure from traditional sentiment construction in finance, which primarily relies on dictionary-based methods and simpler machine learning techniques like the Naive Bayes classifier.
2. **Enhanced Financial Sentiment Analysis**: The research demonstrates a significant enhancement in financial sentiment analysis using BERT compared to existing models. This improvement is particularly noted in the context of analyzing actively traded individual stocks in the Hong Kong market and discussions on Weibo.com.
3. **Combining BERT with Other Method**s: In addition to using BERT, the paper integrates the model with other commonly-used methods for building sentiment indices in finance, such as option-implied and market-implied approaches. This combination results in a more general and comprehensive framework for financial sentiment analysis.

4. **Predictability of Stock Returns Using LSTM**: The research also explores the predictability of individual stock returns by combining the sentiment index with an LSTM model, which features nonlinear mapping. This approach contrasts with the dominant econometric methods in sentiment influence analysis, which are typically linear in nature. The results provide convincing outcomes for predicting individual stock returns using this integrated approach.

4. Expected Workflow & Outcomes

1. **Data Collection:** Extensive collection of annual reports and vital financial data from various companies. This step is critical for building a comprehensive dataset.
2. **Textual Analysis:** Employing advanced NLP techniques, particularly the BERT algorithm, to deeply analyze and interpret the textual content within these reports, extracting key financial indicators and narratives.
3. **Data Processing**: Rigorous data cleaning and structuring are undertaken to ensure the data is optimally formatted for input into the predictive model.
4. **Model Development**: Developing and training a sophisticated model focused on predicting stock performance. This model will integrate insights derived from the textual analysis of annual reports.
5. **Evaluation**: Conducting thorough testing of the model to assess its accuracy in predicting stock trends. The evaluation phase is pivotal for refining the model and enhancing its predictive power.
6. **Actionable Insights:** Generating valuable, data-driven insights for investors and analysts. The model's predictions aim to inform strategic investment decisions and market analysis.
7. **Innovative Financial Analytics:** This project aims to push the boundaries of AI-driven financial analytics. By utilizing the BERT model, it sets a new standard in analyzing textual data for financial forecasting.

5. Potential Challenges

Several potential unknowns and challenges could impact the success of a stock performance prediction project using annual report text.

a. Text Length and Tokenization Issues

BERT has a limitation on tokens. If the length of the annual reports is beyond this threshold, it might be necessary to divide the content into sections, which could result in the loss of crucial information or context. Methods like truncation or summary could be applied, but they might leave out important details or change the text's meaning. A larger Bert model must be selected if every text is taken, but the computing time will be extremely long.

b. Report Analysis Complexity

Annual reports contain both factual information and subjective commentary. Differentiating between these and accurately capturing the sentiment can be challenging, such as management's outlook and strategic plans. Differentiating between these types of content is crucial, as they have different implications. Companies often use a formal and polished tone in their reports, potentially downplaying negative aspects or highlighting positive outcomes in subtle ways.

c. Data Sufficiency and Quality

Different industries have unique ways of reporting and discussing their performance. If the dataset lacks representation from certain sectors, the model may not learn to interpret these sectors' reports accurately, leading to biased predictions. Smaller or startup companies might report differently compared to larger, established corporations. If the dataset is skewed towards one type of company, the model's predictions may not be accurate for other types. Companies may change their reporting style, structure, or level of detail over time. Such inconsistencies can confuse the model, leading to poor learning and inaccurate predictions.

d. Computational Limitations and Resource Constraints

Analyzing annual reports using large models in stock performance forecasting brings computational challenges and resource constraints. These models, especially large models, require a lot of memory and processing power, often using high-end GPUs and large amounts of RAM. Training such models on large data sets can be time-consuming. And as data grows, the training requirements are likely to increase. We must upgrade Colab Pro, which has Faster GPUs and More memory when we choose large models.

e. Importance of Integrating Numerical Data

Combining numerical data from annual reports with textual analytics is essential for comprehensive stock price forecasting. Numerical data such as earnings, revenues, and financial metrics provide measurable evidence of a company's performance, highlight trends, and assist in risk assessment and valuation. Textual analysis, on the other hand, provides context and insights into management's views and strategic direction. By combining these two types of data, predictive models can capture a company's tangible financial health and qualitative strategy, resulting in a more detailed and accurate understanding of its potential stock performance.

f. Market Volatility and External Factors

Though annual reports provide valuable insights into a company's financial situation and management strategies, they cannot capture all the external factors that affect the stock market. These factors, which include economic indicators, political events, market sentiment, global

crises, technological change, and competitive dynamics, play a critical role in determining stock performance. The inability of models based on annual report analysis alone to predict external influences highlights the significant limitations of predicting stock market movements.

6. Data Preparation

a. Target Label Generation

In this project, we aim to predict whether a public company's stock price will rise on the nth day following the release of its annual report (10-K). Since the dataset of annual reports lacks a specific target variable for this purpose, we needed to generate it independently. Initially, we identified company tickers using a CIK-ticker map, aligning them with the CIK numbers provided in the dataset. Subsequently, we employed the Yahoo Finance API to retrieve the stock prices on the dates of the annual reports' release and for subsequent periods (7 days, 30 days, 90 days). A future stock price exceeding the price on the report release date results in labeling the corresponding time window with a 1; otherwise, it is labeled as 0. In the project's refinement phase, we began by focusing on the 30-day window, later extending the most effective model to the other time frames.

b. Label Balance

Upon analyzing the target label, we discovered that over 60% of the stocks in our dataset experienced a price increase 30 days following the release of their annual reports. To enhance the effectiveness of our model, we balanced the training set by ensuring an equal number of stocks with labels 1 (price increase) and 0 (no price increase). This approach is designed to prevent the model from disproportionately favoring class 1, thereby achieving a more balanced and accurate prediction.

c. 10-k Item Selection

According to the guidelines set by the Securities and Exchange Commission (SEC), a public company's annual report (10-K report) must comprise 14 distinct items. We hypothesize that two of these items are particularly pertinent: Item 1A, which outlines risk factors, and Item 7, which details the financial condition and results of operations. During the project's fine-tuning phase, we will initially concentrate on Item 1A.

d. Text Slicing

We chose Item 1A (Risk Factors) and Item 7 (Financial Condition and Results of Operations) from the 10-K report as the primary texts for training, given their extensive length and critical importance in assessing a company's status. However, models like DistilBERT and BERT have a limitation of processing only up to 512 tokens. To adapt to this constraint, we decided to extract the last 512 tokens from each item's text because the last part of these report items often contains summarized information, which is likely to be more helpful for predictive analysis.

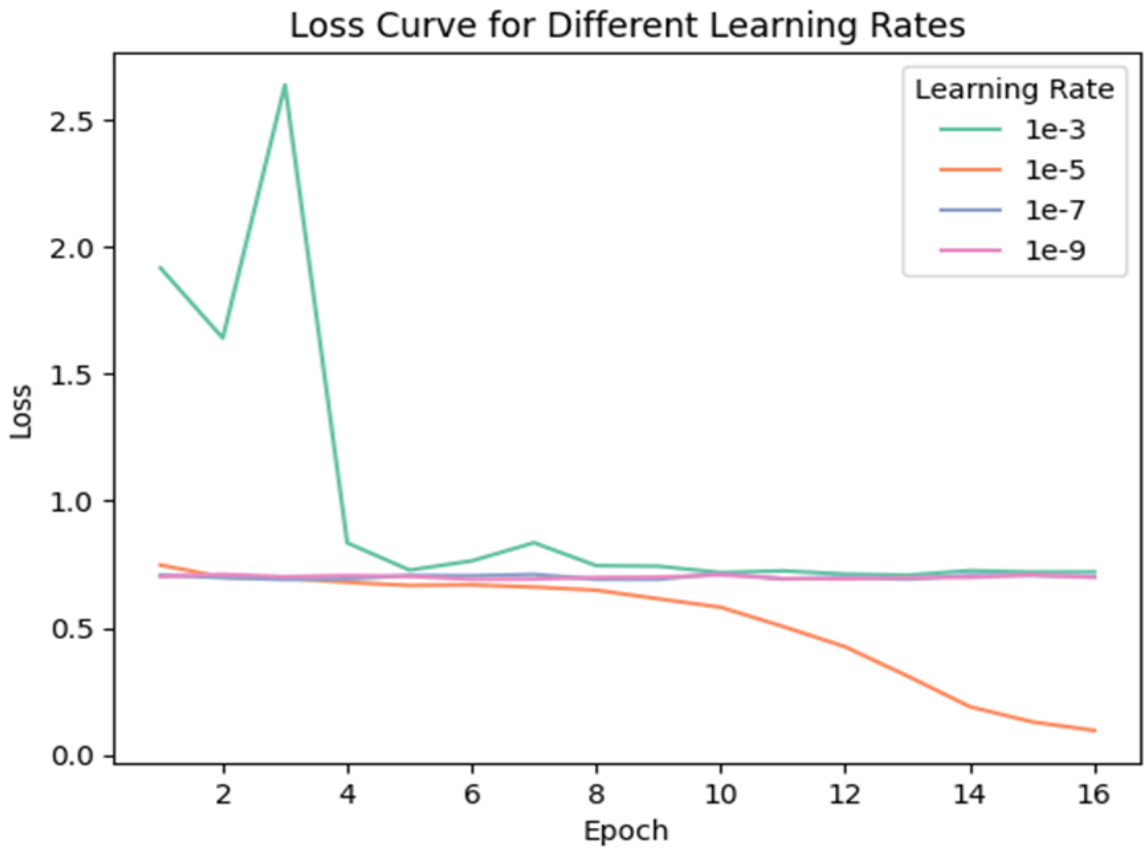
7. Model Development

a. Overview

In order to achieve decent predictions in terms of stock price change classification according to their annual reports, we decided to use a transformer model in this project. The Transformer model, a major innovation in natural language processing, significantly advances the field with its unique architecture and capabilities. Apart from its renowned efficiency in translation and text generation, it excels in classification tasks. By leveraging its self-attention mechanism, the Transformer can analyze and understand the context and relationships within text data, making it highly effective for various classification problems. This capability has been further enhanced in models like BERT, which are built upon the Transformer architecture and fine-tuned for specific classification tasks. In this case, we planned to try different variants of the transformer model: BERT (the most commonly used BERT model), DistilBERT (lighter version of BERT). Since the project is expected to be computationally expensive, we start with the lighter model, which is DistilBERT.

b. Learning rate

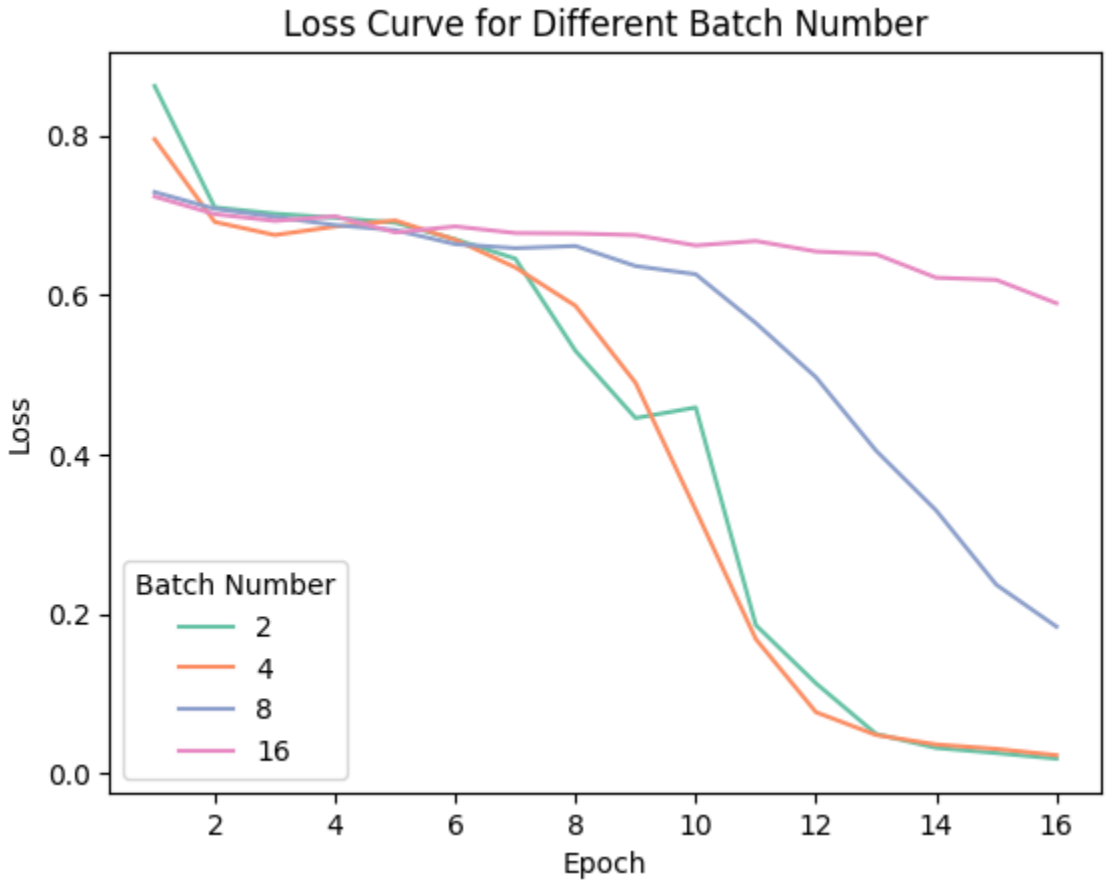
The first parameter of the DistilBERT model is the learning rate. The learning rate in machine learning is a crucial parameter that determines the size of the steps a model takes during optimization. A learning rate that's too high can cause the model to converge too quickly to a suboptimal solution, or diverge, while one that's too low can result in a long training process that might get stuck in local minima. We tried different values of learning rate and plotted some of them in the plot below.



According to the plot, the loss value is lowest when the learning rate is equal to 1e-5. The L-shape curve also shows the model benefits from the training process so the learning rate for the model is 1e-5.

c. Batch Size

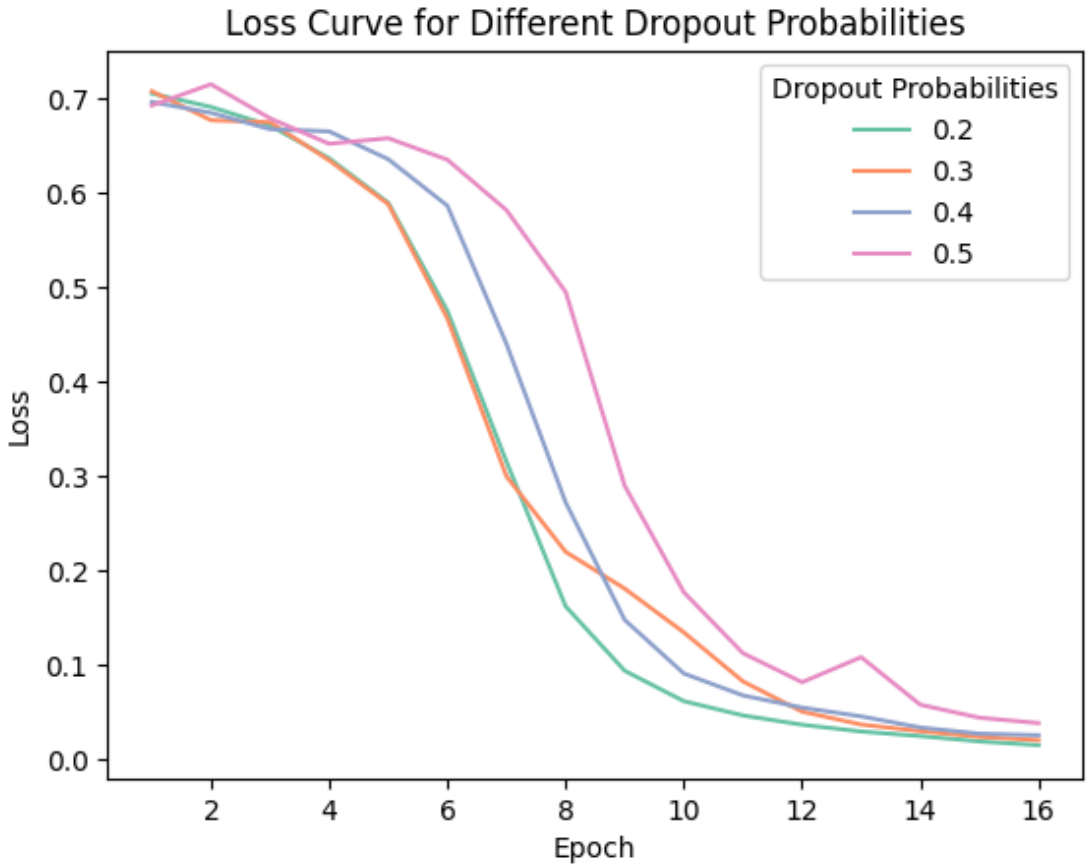
In machine learning, the batch size, limited by the machine's memory capacity and the specifics of the data and algorithm, is crucial for model performance. Smaller batch sizes, as recommended by Keskar et al. in their 2017 ICLR paper, often lead to more frequent gradient updates and better model generalization, making it advisable to use the smallest stable batch size for training. We tried different values of batch size and plotted some of them in the plot below.



According to the plot, the losses decrease relatively quickly for batch size 2 and 4, and batch 4 has a more constant decreasing rate so the ideal batch size for the model is 4.

d. Dropout Regularization

Dropout regularization is a technique used in neural networks to prevent overfitting by randomly deactivating a subset of neurons during training, effectively thinning the network. This method encourages the network to become less sensitive to the specific weights of individual neurons, thereby enhancing its ability to generalize to new data. Typically, dropout will improve generalization at a dropout rate of between 10% and 50% of neurons.



In the DistilBERT model, seq_classif_dropout is a parameter allowing users to adjust the dropout probability when the model is implemented for tokens sequence classification. We selected 0.2 since it has the lowest loss value.

e. Input Data Decisions

We decided to attempt different time windows as the target variable because we deemed that the price change could have lagged. After applying the aforementioned parameters, we found the model has better performance with a time window of 30 days.

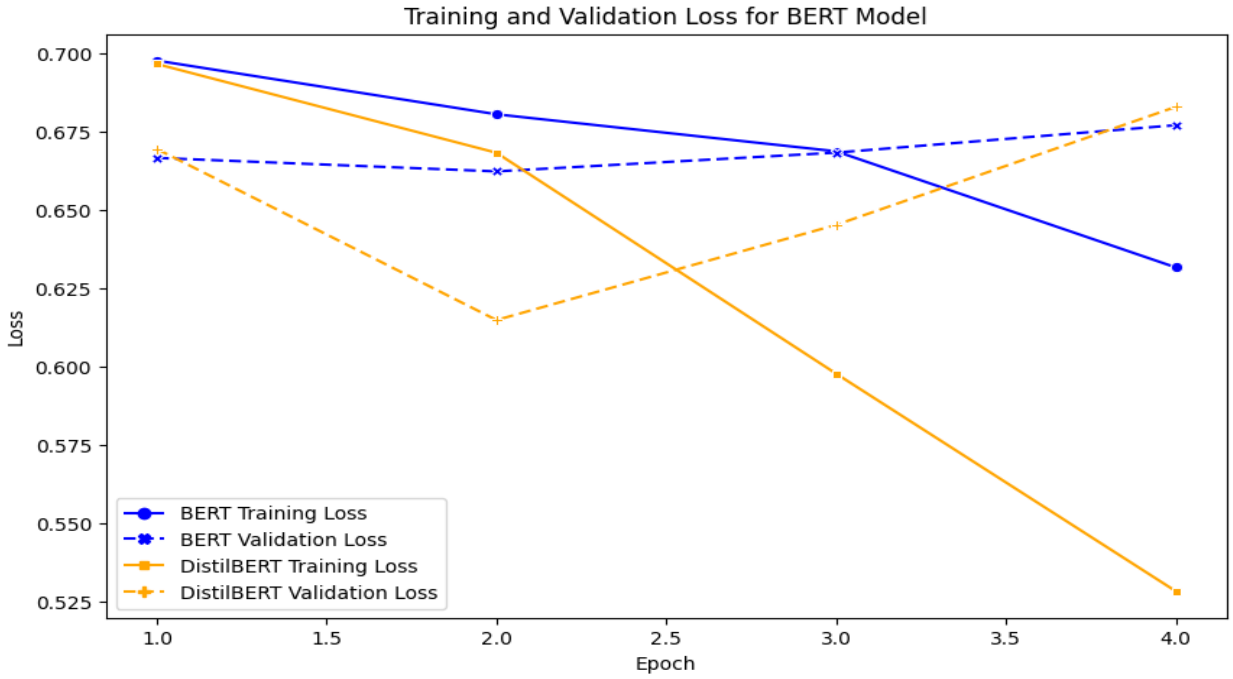
Time Window	Accuracy
7 Days	52%
30 Days	64%
180 Days	59.5%

There are 14 required items in a complete 10-K form, and some items could have more than 1000 tokens. In this case, it is not applicable to use the whole report as input data. We selected 2 most relevant items: Item 1A and Item 7. After applying the aforementioned parameters, we found the model has better performance with Item 7 as input data.

Item	Accuracy
Item 1A	64%
Item 7	71.5%

f. Final Model

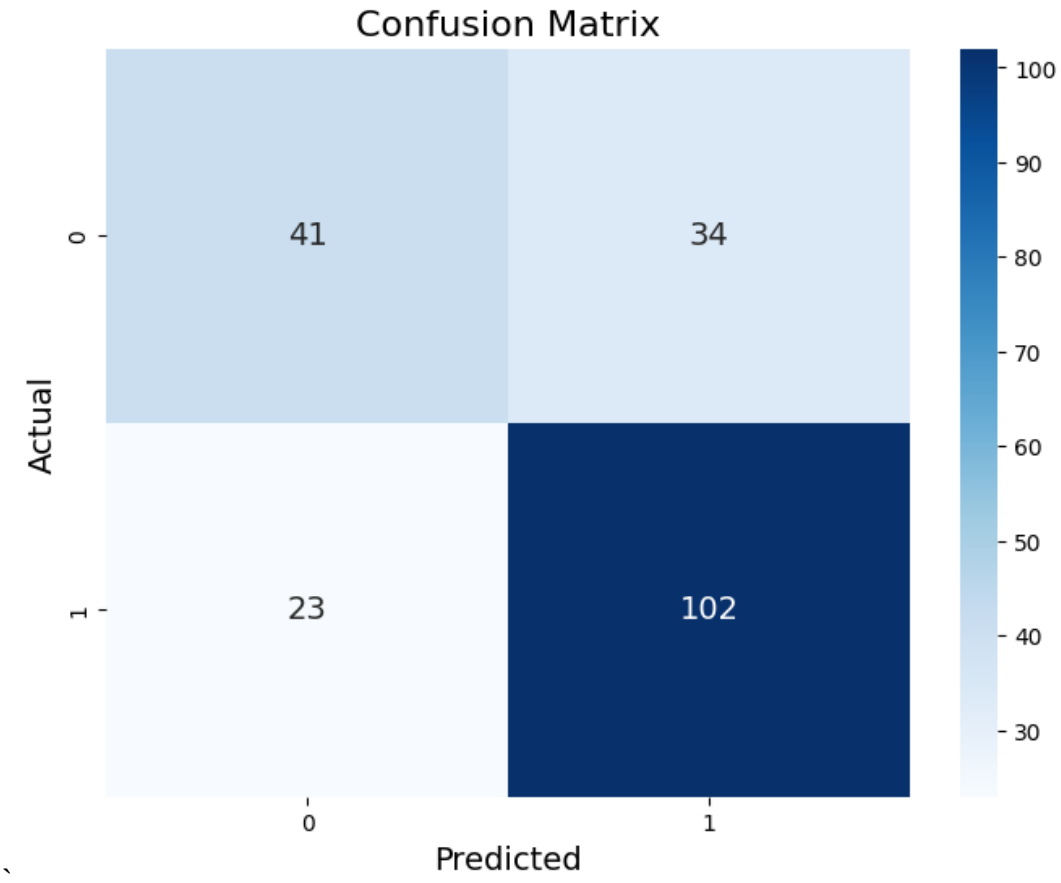
After completing fine-tuning with the DistilBERT model, the BERT model, with more layers and parameters. However, the BERT model is not performing better than the DistilBERT model despite its complexity. The loss values of both training and validation are higher for the BERT model. In this case, we will take the DistilBERT model with all aforementioned parameters as the final model for prediction.



8. Results & Discussion

The results indicate that the DistilBERT-based model, with the specified hyperparameters and analysis of Item 7 in 10-K reports, shows outstanding accuracy in predicting stock price trends. 71.5% accuracy is indeed far above our expectations because an investment strategy can be beneficial as long as the prediction accuracy is higher than 50%. It is important to note that stock price prediction is a complex task influenced by various factors, including market sentiment, external news, and macroeconomic conditions. While the model demonstrates solid performance based on the selected parameters, further refinement and the inclusion of additional features may enhance its predictive power.

Metrics	Value
Accuracy	0.7150
Precision	0.7500
Recall	0.8160
F1 Score	0.7816



An area for improvement is the exclusive use of data from 2020 in the training and validation sets. This approach presents a potential limitation for the model, as it may introduce bias by only learning from data specific to 2020, a year possibly characterized by unique circumstances due to the overall market trends. Training and validating the model with data from various years could enable it to discern more general patterns, rather than those unique to a specific year. By incorporating this broader dataset, the model would likely be more reliable for making investment decisions based on newly released 10-K reports.

Another potential area for enhancement involves understanding the accuracy of predictions as a mere "winning rate." The notion that an investor would profit if more than 50% of transactions are beneficial is misleading, as this does not guarantee overall positive returns. The issue arises when small gains are overshadowed by larger losses, leading to an overall deficit. To mitigate this flaw, one approach could be to create a more detailed system of target labels that differentiate between various extents of price changes. For example, stocks could be categorized with distinct labels if their price change exceeds 10%, thereby predicting if a stock is volatile for investors.