SI 699 Final Project Report: Predict Tesla stock price trends with news articles

Jungseo Lee, Conan Wu, and **Guanghan Xi**
2023 Winter
University of Michigan

Abstract

This research project aims to develop a predictive model for stock price trends by utilizing Natural Language Processing (NLP) techniques to analyze news article data. The project focuses on identifying the relationships between stock price trends and news articles, with Tesla's stock price and surrounding news articles serving as the dataset. The study hypothesizes that a model that uncovers the correlation between stock price trends and relevant data could potentially predict stock price trends. The research presents the Binary Cross-Entropy (BCE) loss chart of the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) models, as a means of assessing their performance. The evaluation losses in the CNN models with Flair and NLTK and the GRU models did not perform well, indicating that the dataset size may be a contributing factor to the poor model performance. The results also reveal a pattern of decreasing evaluation losses as the training set size increases in both models, indicating the need for larger datasets. Overall, this research demonstrates the potential for NLP techniques to analyze news article data and predict stock price trends, but further studies with larger datasets are necessary.

1 Introduction

The objective of this project is to develop a predictive model for stock price trends by leveraging Natural Language Processing techniques to analyze news article data. The stock price is recognized to be volatile due to various factors such as the company's performance, market conditions in both macro and microeconomic areas, and competition in the industry.Therefore, predicting stock price trends is challenging due to the complex interplay of these factors. Despite numerous attempts to establish a reliable method for forecasting stock price trends, success has been limited. This project was initiated with the hypothesis that a model that uncovers the correlation between stock price trends and relevant data could potentially predict stock price trends. In theory, this appears feasible, but the pertinent question is: "Which data reflects the relevant factors?" Our idea was that news articles could provide a multi-dimensional reflection of situations that could potentially impact stock price trends, and thus, by extracting crucial information from news articles, it might be possible to predict stock price trends. This project endeavors to identify the relationships between stock price trends and news articles by leveraging cutting-edge techniques and constructing a prediction model for future stock price trends . To enable us to work with real-world data, we executed this project utilizing Tesla's stock price and the news articles surrounding it.

2 Related works

Several research papers have been published on predicting financial trends using natural language processing techniques. Cavalli and Amoretti conducted a study on predicting Bitcoin trends using a CNN-based multivariate data approach. Their technique employs a One-Dimensional Convolutional Neural Network (1D CNN) to analyze data from diverse sources, including social media, full blockchain transaction history, and several financial indicators. Guo and Li's research, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," focuses on real-time prediction of the FTSE 100 stock market price using Twitter sentiment scores (TSS) and compares their approach with traditional econometric models based on closed-end fund discounts (CEFD). These studies offer valuable insights and guidance for planning our research process and constructing a model for volatile financial data.

3 Data

In light of the limitations of available news scraping tools, we created a dataset spanning from January

1, 2021, to March 25, 2023, utilizing two primary APIs. To obtain interest data, Tesla's stock price, and other financial indicators such as competitors' stock prices, crude oil prices, and Dow Jones Automobiles Historical Data, we leveraged Yahoo Finance's API. Yahoo Finance API is a collection of libraries that allow users to obtain historical and real-time data for a variety of financial markets and products. Although the official Yahoo Finance API has been shut down, the API continues to provide the necessary data through a combination of direct API calls, HTML data scraping, and pandas table scraping. The libraries offer a remarkably wide range of data at no cost and are relatively straightforward to use.

We employed Perigon's API to gather news articles, which enabled us to obtain articles until the year 2021. The API offers access to news from over 60,000 sources, encompassing both national and international news outlets. It supports topic and entity extraction during the scraping process. The API offers full access to real-time news and allows up to 500 requests per month for testing purposes, with each request returning up to 2,000 articles, and all services are provided free of charge. To obtain an API key for Perigon's API, you must first create a free account by signing up on their website. We scraped all news articles from the top 100 news sources that included at least one of the following keywords: Tesla, Automotive, or Elon Musk. This necessitated the use of keywords and pagination in our API calls. The total number of articles collected was 19,391. Given our belief that sentiment scores derived from news articles contain concise and informative insights, we made the decision to leverage this information. To calculate sentiment scores for the title, summary, and content of the articles, we employed two distinct NLP packages, NLTK and Flair. There were some missing dates in the article data. We took the logical approach of assuming that news articles reflect market conditions not just for a single day, but for several preceding days. As a result, we imputed the missing sentiment scores using the last six days' moving average. Figure 1, 2, and 3 are the graphs of the dataset during the exploratory data analysis.

For detailed code on how to scrape each piece of data, please refer to our attached GitHub page (A.1).

4 Methodology

In the course of this project, two distinct models, namely the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU), were employed to forecast trends in Tesla's stock prices based on historical data. Subsequently, the efficacy of these models was evaluated on datasets incorporating sentiment analysis scores extracted from news articles.

4.1 Overall Structure

Input For each model, the input data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is a sequence of historical data in the past n trading days, and data for each trading day is represented as a feature vector $\boldsymbol{x} \in \mathbb{R}^d$. Within the scope of this project, we utilized historical data spanning $n = 30$ trading days to forecast stock trends for the subsequent trading day. It is important to note that the parameter d varied among the different models employed.

The financial data corresponding to each trading day is a shared component across all feature vectors. To be precise, there exist 8 distinct financial data points, which are as follows:

1. Open Price of Tesla Stock
2. High Price of Tesla Stock
3. Low Price of Tesla Stock
4. Close Price of Tesla Stock
5. Adjusted Close Price of Tesla Stock
6. Volume of Tesla Stock
7. DJH AUTO SALE Price
8. Europe Brent Crude Oil Spot Price

The initial six features in the feature vector pertain to historical information concerning Tesla, while the last two features encapsulate information concerning the global automobile market and international oil market, which can exert an influence on Tesla's stock trends. For models that do not incorporate sentiment scores, the feature vector $\boldsymbol{x}$ is defined as $\boldsymbol{x} \in \mathbb{R}^8$, and the input matrix is represented as $\boldsymbol{X} \in \mathbb{R}^{30 \times 8}$.

However, for models that incorporate sentiment scores, the dimensionality of the feature vectors can vary. Each news article is divided into three sections, namely the summary, title, and content,

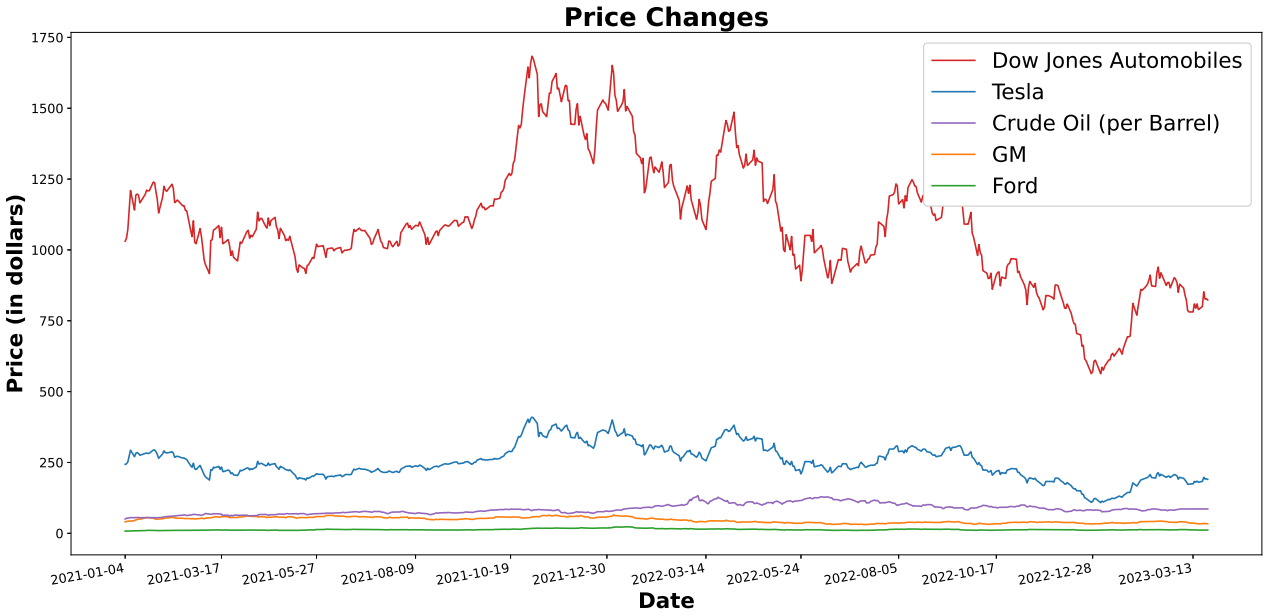

Figure 1: The close prices changes from 2021-01-01 to 2023-03-25

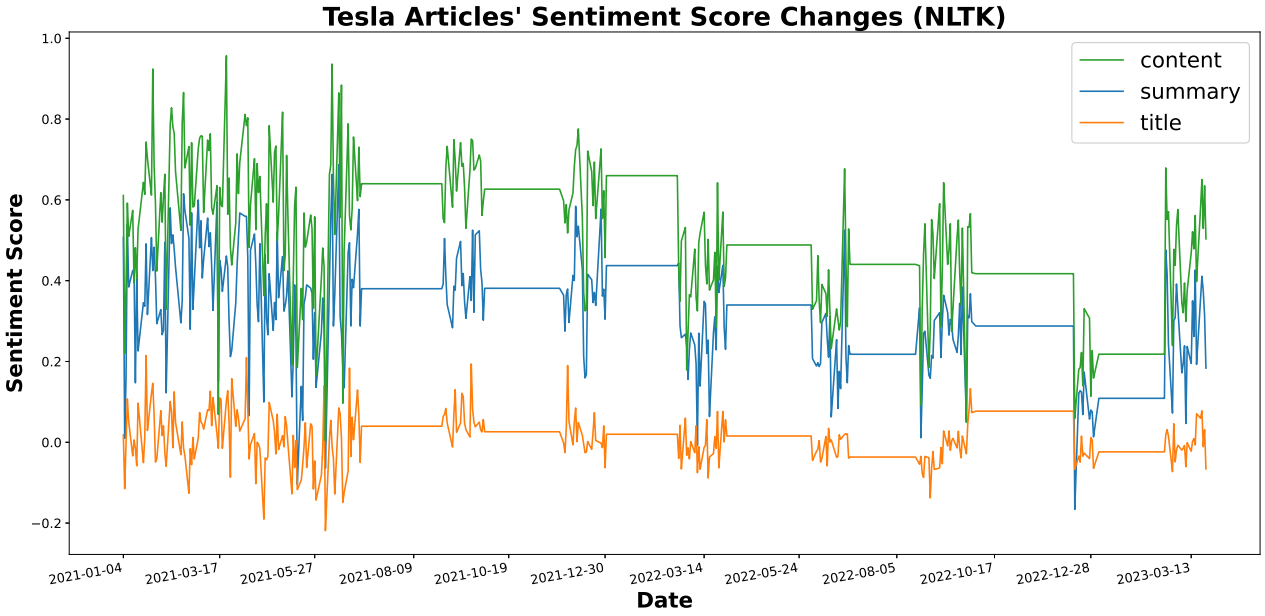

Figure 2: The Tesla Articles' NLTK sentiment score changes from 2021-01-01 to 2023-03-25

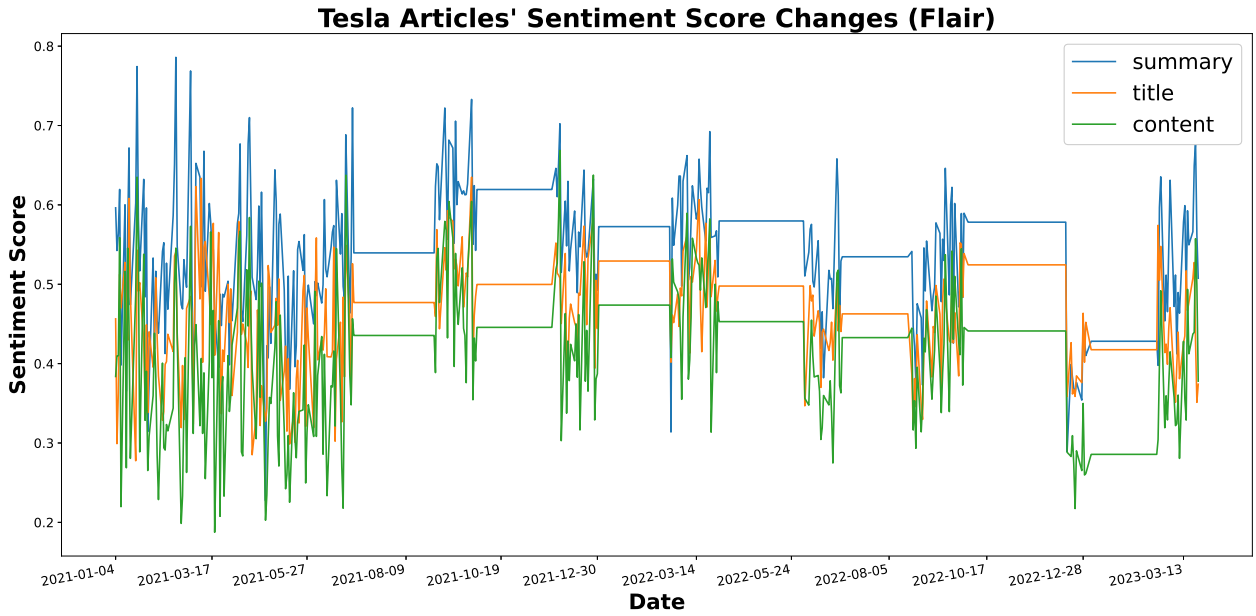

Figure 3: The Tesla Articles' Flair sentiment score changes from 2021-01-01 to 2023-03-25

and different sentiment score calculation methods can yield unique scores for each of these sections. In this project, we employed two distinct methods to generate sentiment scores, namely NLTK (Bird et al., 2009) and Flair (Akbik et al., 2019). Consequently, several types of sentiment score combinations can be added to the basic feature vectors previously described. The three primary types of combinations used in this study are as follows:

1. A single score calculated by NLTK (Bird et al., 2009) or Flair (Akbik et al., 2019) based on one of the three parts of news articles.

2. Two scores calculated by NLTK (Bird et al., 2009) or Flair (Akbik et al., 2019) based on two parts of news articles respectively, such as summary score and title score generated by NLTK (Bird et al., 2009).

3. All three scores calculated by NLTK (Bird et al., 2009) or Flair (Akbik et al., 2019)

Hence, the dimensionality of feature vectors incorporating sentiment scores of news articles can be 9, 10, or 11, depending on the specific combination of sentiment scores employed.

Output The models implemented in this project can output a probability value $p(y = 1) \in [0, 1]$, which represents the likelihood of a positive stock trend on the subsequent trading day. In particular, if $y = 1$, it implies that the closing price of Tesla stock on the day following the final day of the input sequence exceeds that of the last day in the input sequence. Conversely, if $y = 0$, it indicates the opposite scenario.

4.2 CNN Model

The One-Dimensional CNN has been established as an effective model for predicting financial data, as demonstrated by prior research (Cavalli and Amoretti, 2021). Analogous to the application of Two-Dimensional CNN models in Computer Vision, the One-Dimensional CNN can extract overall features from data sequences, which enables it to effectively identify trends and discern underlying patterns from historical data. This capability is invaluable in facilitating the accurate prediction of future trends in financial markets.

The CNN model's architecture is illustrated in Figure 4, comprising two One-Dimensional CNN modules, a fully connected layer, and a dense layer. Each One-Dimensional CNN module is composed of a one-dimensional convolutional layer, a pooling layer, and an activation function. Specifically, the step and stride in both the convolutional and pooling layers are 1, the number of filters in the first and second convolutional layers are 16 and 32, respectively, and the activation function utilized is ReLU. The filter sizes in the convolutional and pooling layers are

$$\text{conv\_filter}_1 = \text{pooling\_filter}_1 = \lfloor \frac{30 \cdot n}{100} \rfloor \quad (1)$$

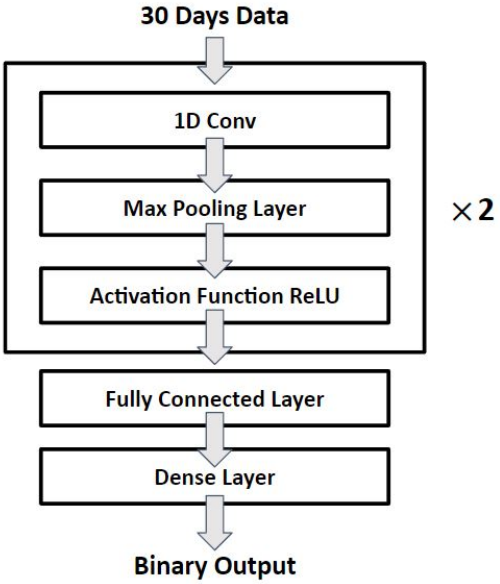

Figure 4: Structure of CNN Model

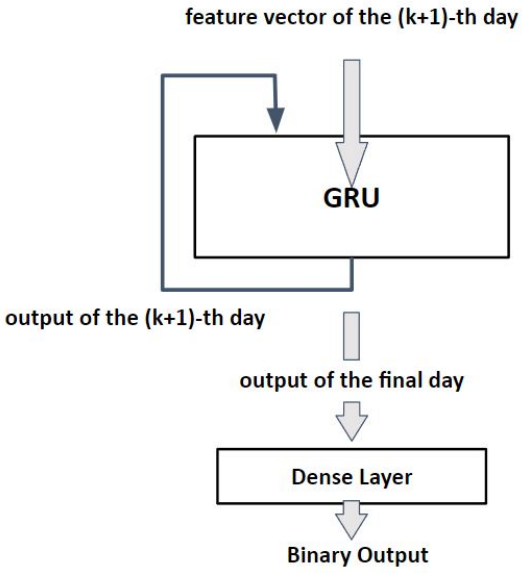

Figure 5: Structure of CNN Model

$$\text{conv\_filter}_2 = \text{pooling\_filter}_2 = \lfloor \frac{20 \cdot n}{100} \rfloor \quad (2)$$

where n represents the length of the input sequence.

Following the two One-Dimensional CNN modules, the fully connected layer aggregates the outputs of all filters into a vector, which is subsequently fed into the dense layer for prediction of the probability that the stock trend will be changed positively on the subsequent trading day. This is accomplished through leveraging the features extracted by the CNN modules.

4.3 GRU Model

The Gated Recurrent Unit (GRU) is another machine learning model that has been shown to be effective in predicting stock trends (Dey and Salem, 2017). As an enhanced version of Recurrent Neutral Neural Networks (RNNs), the GRU can process each data point in a sequence sequentially, and the output for the current input serves as a partial input for the next data point. In addition, the GRU model includes a forgetting mechanism that allows it to focus on the essential features of the sequence and disregard irrelevant details. Consequently, the GRU can efficiently extract pertinent information from data sequences and is thus suitable for predicting stock trends based on historical data sequences.

The structure of the GRU model is shown in Figure 5. The GRU model used in this project consists of two layers, each with a hidden dimension of 64. After processing the input sequence with the GRU layers, the dense layer produces the final output $p(y = 1)$ based on the output for the last input in the sequence.

5 Result

Due to the limited size of our dataset, evaluating our model's performance using statistical metrics such as the F-1 score may not yield statistically significant results. We present the BCE loss chart of our CNN model in Figures X, Y, and Z as a means of assessing its performance.

Figures 6, 7, and 8 illustrate that the evaluation losses hover around 0.69. When considering the Binary Cross-Entropy (BCE) loss range, 0 represents a perfectly predicted stock trend and 1 represents the result is totally opposite to the actual trend. Both the CNN models with Flair and NLTK and the GRU models depicted in Figures 9, 10, and 11 did not perform well. Based on these results, we infer that our dataset size may be a contributing factor of the poor performance of the model. To support this, we examine Figures 12 and 13, which show the evaluation losses in 1, 1/2, 1/4, 1/8, and 1/16 sizes of our dataset. These graphs reveal that there are patterns of decreasing evaluation losses as the training set size increases in both models. However, since the total training set size is relatively small, the numerical differences in evaluation losses do not seem to be significant.

6 Discussion

This section will outline the limitations of our project, ethical considerations, and potential prob-

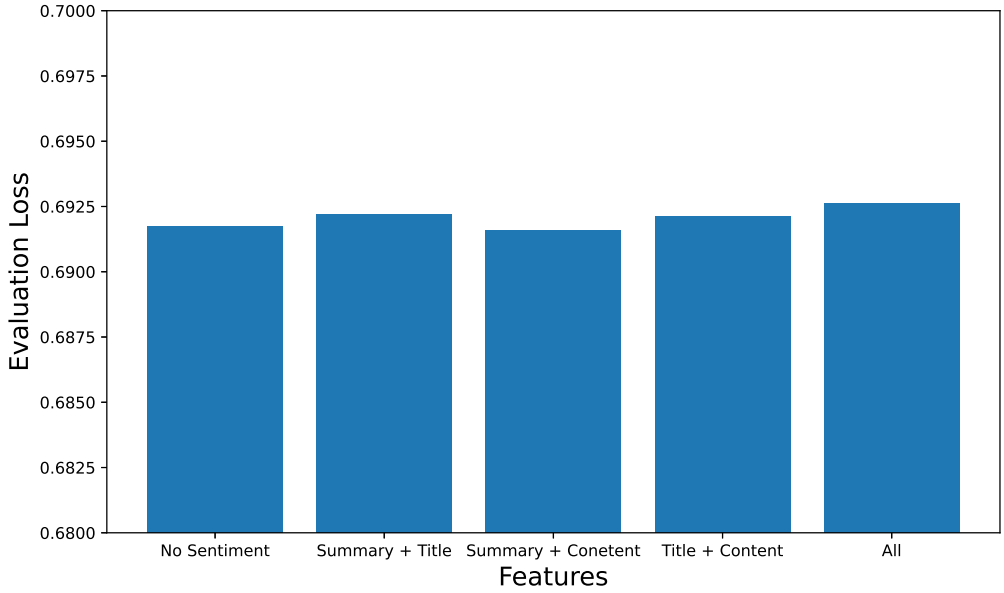

Figure 6: The CNN model with flair sentiment analysis on different pairs of textual data

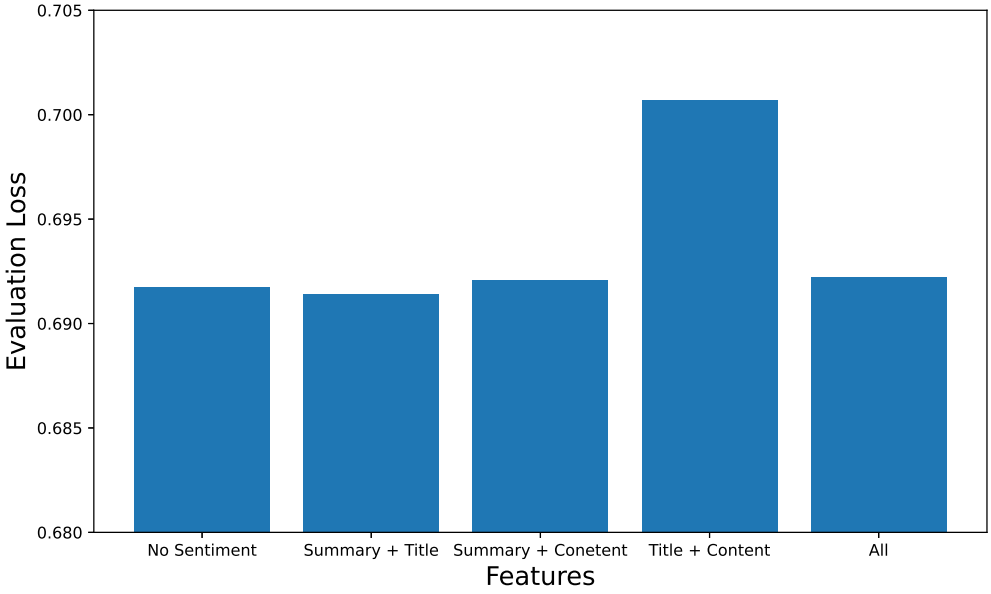

Figure 7: The CNN model with NLTK sentiment analysis on different pairs of textual data

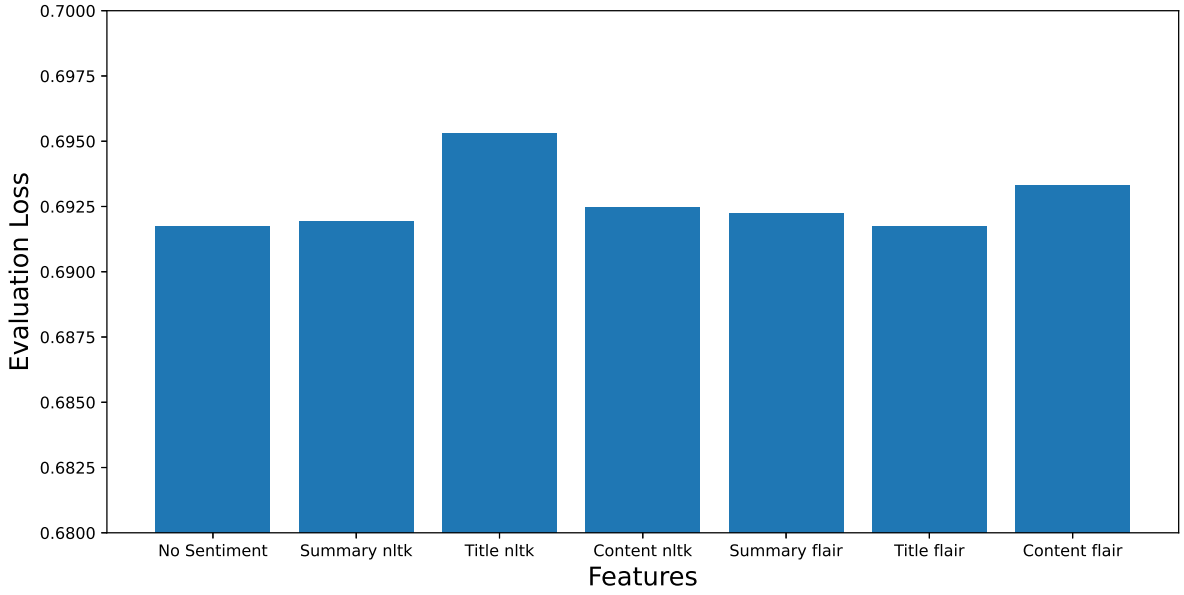

Figure 8: The CNN model with Flair and NLTK sentiment analysis on different single textual data

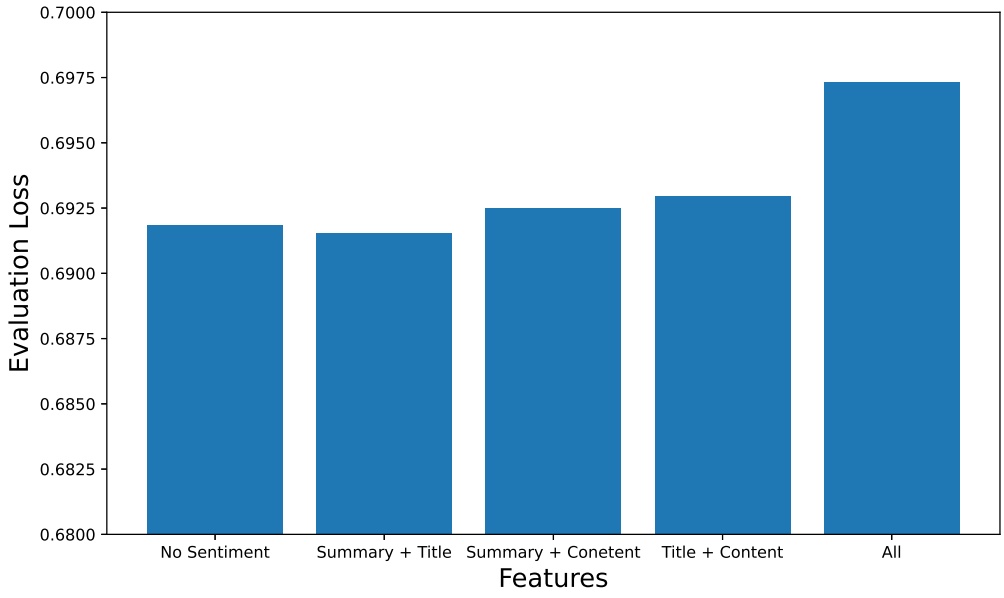

Figure 9: The GRU model with flair sentiment analysis on different pairs of textual data

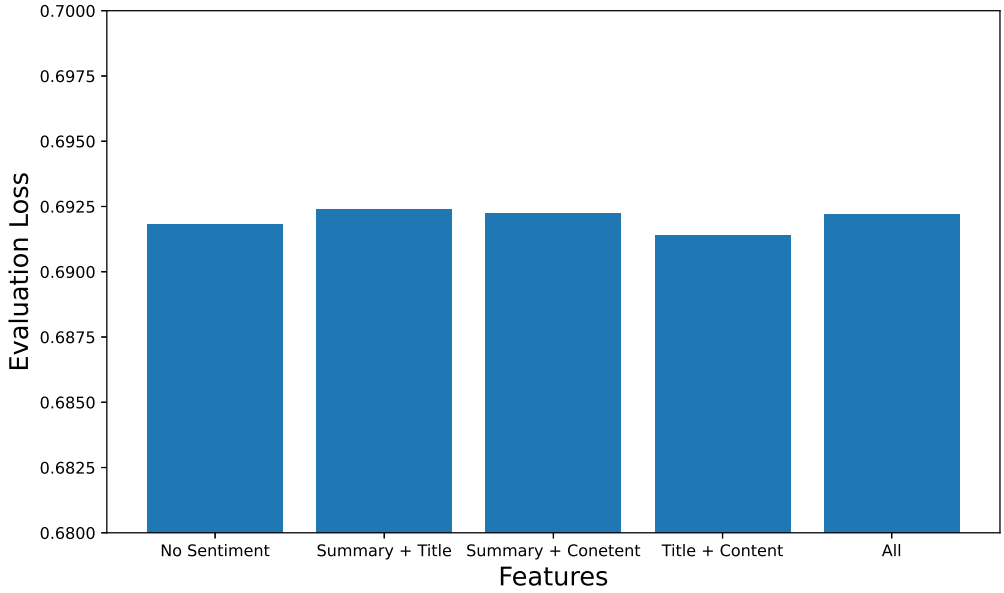

Figure 10: The GRU model with NLTK sentiment analysis on different pairs of textual data

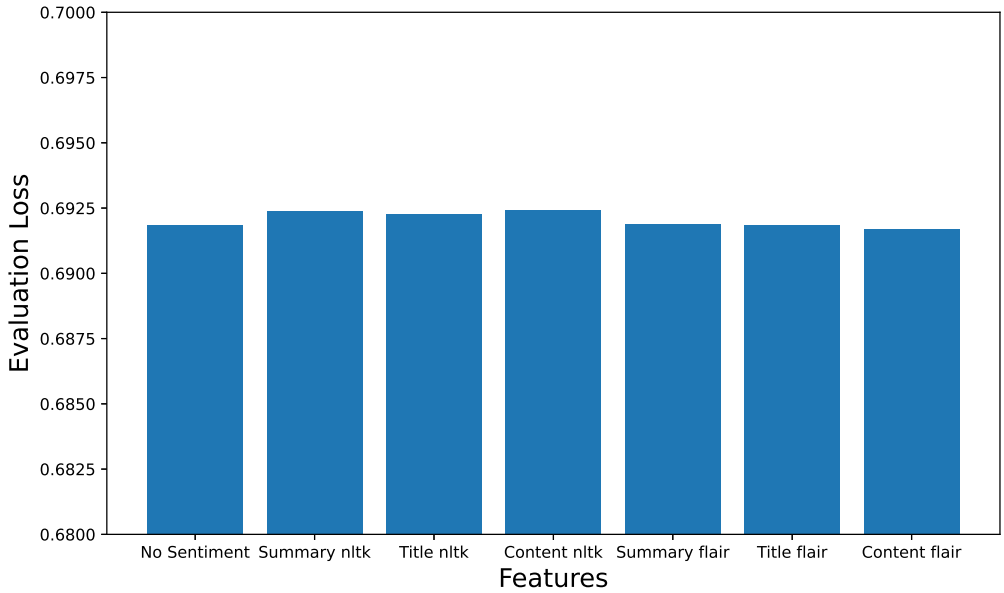

Figure 11: The GRU model with Flair and NLTK sentiment analysis on different single textual data

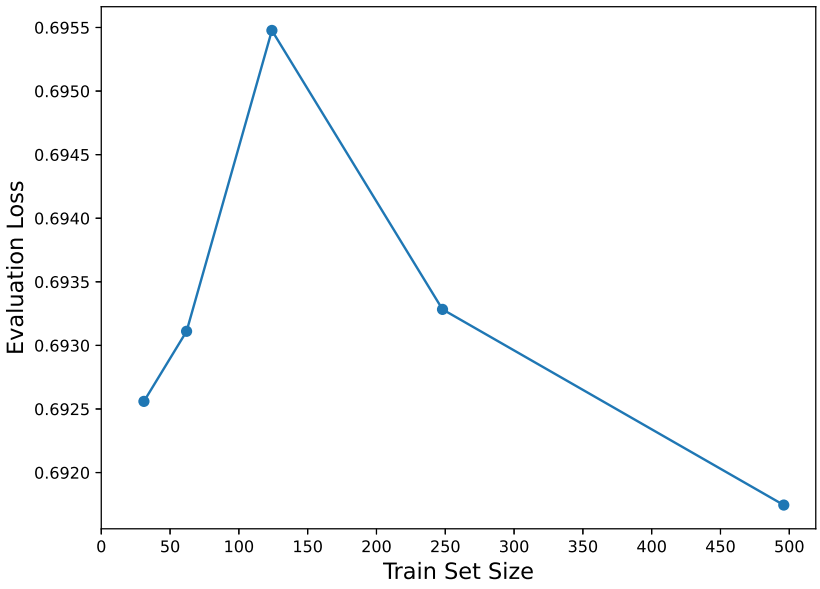

Figure 12: Evaluation loss of no sentiment CNN in 1, 1/2, 1/4, 1/8, and 1/16 sizes of the original dataset

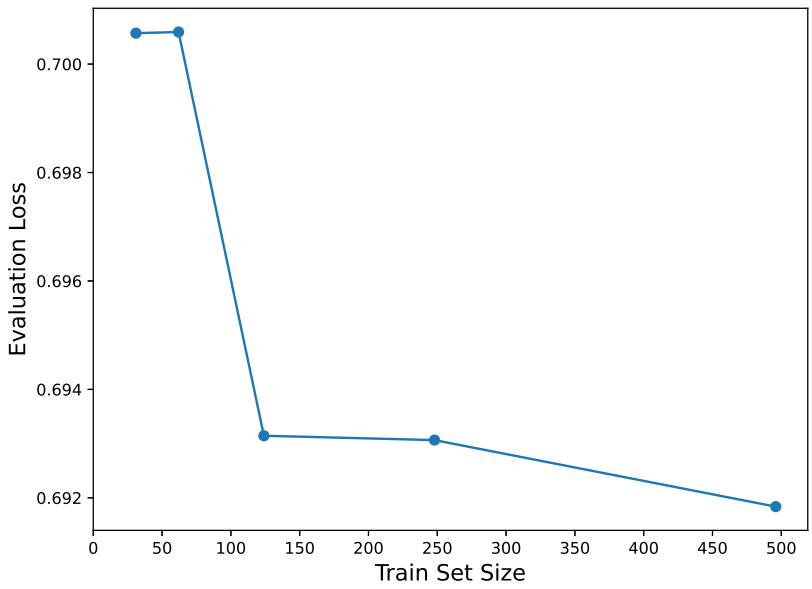

Figure 13: Evaluation loss of no sentiment GRU in 1, 1/2, 1/4, 1/8, and 1/16 sizes of the original dataset

lems , as well as future areas of improvement. It comprises three sections: Limitations, Broader Impact, and Suggestions for Future Research.

6.1 Limitations

As previously mentioned, the small size of our dataset rendered our results inconclusive regarding the impact of textual data on our models. The lack of significance of the textual data in our model predictions suggests that they were heavily reliant on the stock price dataset we used. Nonetheless, the stock market trends are not predictable like temperature trends, which follow a predictable pattern. Thus, more features need to be added to the training dataset to identify potential patterns. Even if we assume that incorporating textual data would enhance the accuracy of our models, we must address several limitations. One such limitation is the spread of misinformation. As our models cannot differentiate between reliable and fake news, they may generate inaccurate predictions based on false information. There, it is crucial to exercise caution and ensure the credibility of news sources before feeding the data into the model.

6.2 Broader Impact

Our research has significant implications for the financial industry, particularly in the field of stock prediction. However, we must consider the ethical and potential problems that may arise from the use of our models. One potential issue is the phenomenon of stock hyping. While using textual data to predict stock trends may seem reasonable, it does not necessarily follow the traditional approach of evaluating a company's actual value based on assets, cash flow, and other factors. Ignoring these factors may lead to herd behavior among investors and the promotion of companies with inflated stock prices. A similar example occurred in the cryptocurrency market when Elon Musk's tweets about Dogecoin led to dramatic fluctuations in its price. Investors using textual data to inform their investment decisions must exercise caution to avoid falling into the trap of stock hyping.

Another issue to consider is the potential for biased data. The writers of news articles and tweets may have incomplete information, misunderstand complex topics, or have their own biases that influence the way they report on a particular company or industry. Our models may not be able to account for these factors, leading to flawed or misleading analysis. To mitigate this issue, it is important to carefully evaluate the sources of data used to train models and continuously monitor for biases in the output.

6.3 Suggestions for future research

The first suggestion for future research is to explore alternative sources of textual data. Prior research in this field has typically used datasets that span over 10 years, whereas our dataset only covers just around 2 years. Additionally, approximately half of our trading dates don't have news articles, which can impact the accuracy of our models. Future research can investigate the use of Twitter data or other datasets to complement or replace our existing news article dataset. Furthermore, we could consider developing our own word lists to enhance the specificity of the sentiment analysis packages Flair and NLTK on the news article data. This would require additional time and effort, but it may improve the accuracy of our models.

The second suggestion for future research is to incorporate additional numerical features that are typically considered relevant to stock markets, such as interest rates and tax rates. While not all of these features may have a strong correlation with the stock market, techniques such as principal component analysis (PCA) can be used to identify the optimal combination of features for our models. By including these features, we can improve the accuracy of our predictions and obtain a more comprehensive understanding of the factors that influence stock prices.

Incorporating Dynamic Time Warping (DTW) Analysis into the project could be a promising direction for future research. DTW is a powerful technique for measuring similarity between two time series, even when they have different lengths or speeds. By applying DTW to the sentiment score time series and stock price time series, it could potentially uncover more nuanced patterns and relationships between the two. This could improve the accuracy of the predictions and provide more insights into the dynamics of the stock market.

7 Conclusion

In the result section, it was found that due to the insufficient data, both the CNN and GRU models performed poorly. However, it was observed that with an increase in the dataset size, the evaluation loss decreased, indicating that the models can generate more accurate results with more data. There-

fore, it can be concluded that using CNN or GRU models for stock trend prediction is not infeasible, but requires a larger dataset and more features.

However, the limitations of the study need to be acknowledged, such as the lack of verification of the authenticity of the news articles and the insufficient strength of the numerical features to find correlations with the stock trend independently. These limitations must be addressed in future research to further improve the understanding of finding stock trends.

In conclusion, although the study did not yield significant results, it provides a good starting point for research in stock trend prediction. By considering the future research suggestions, researchers can conduct more effective predictions or stock trends.

8 Group effort

In our data analysis project, we have divided the process into distinct stages, including data collection, data cleaning, data exploratory analysis, modeling, and result visualization and interpretation. Our team's allocation of work for these stages is presented in the table below. Jungseo Lee was responsible for data collection, data cleaning, and data exploratory analysis, while Conan Wu undertook data cleaning, result visualization and interpretation. Guanghan Xi took charge of the modeling and result visualization and interpretation tasks. It should be noted that some sections required contributions from multiple team members, resulting in overlapping responsibilities. As a measure to ensure grammatical correctness and accurate sentence structure, ChatGPT was employed.

Tasks	Member
Data collection, Data cleaning, Data Exploratory Analysis	Jungseo Lee
Data cleaning, Result Visualization and Interpretation	Conan Wu
Modeling, Result Visualization and Interpretation	Guanghan Xi

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Stefano Cavalli and Michele Amoretti. 2021. Cnn-based multivariate data analysis for bitcoin trend prediction. *Applied Soft Computing*, 101:107065.

Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Xinyi Guo and Jinfeng Li. 2019. A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*.

A Appendix

A.1 GitHub page

https://github.com/guanghanxi/UMSI_SI699WIN23_Tesla_Stock_and_News

A.2 Tables of detailed data

Category	CNN Evaluation Loss	GRU Evaluation Loss
summary with nltk	0.69192	0.69236
title with nltk	0.69529	0.69227
content with nltk	0.69246	0.6924
summary with flair	0.69226	0.69188
title with flair	0.69176	0.69183
content with flair	0.69331	0.69169
summary + title with nltk	0.69139	0.6924
summary + content with nltk	0.69207	0.69224
title + content with nltk	0.70067	0.69139
summary + title with flair	0.69219	0.69152
summary + content with flair	0.69158	0.69249
title + content with flair	0.69214	0.69296
All textual data with nltk	0.69222	0.6922
All textual data with flair	0.69263	0.69732

Table 1: The actual evaluation losses in different combinations of dataset

Category	CNN	GRU
1/16 of the original training size	0.6925597587052514	0.7005701339420151
1/8 of the original training size	0.6931108990136317	0.7005915573414635
1/4 of the original training size	0.6954767369172152	0.6931447993306552
1/2 of the original training size	0.6932831474963356	0.6930644703261993
The original training size	0.6917444972430958	0.6918373802128961

Table 2: Detailed evaluation loss numbers in Figure 12 and 13