Predict stocks from news based on multiple models

Zijian Wu
School of Computing
National University of Singapore
zijian.wu@u.nus.edu

Muhan Zhong
School of Computing
National University of Singapore
muhan@u.nus.edu

Boran Yang
School of Computing
National University of Singapore
e1101548@u.nus.edu

Xinze Lan
School of Computing
National University of Singapore
lanxinze@u.nus.edu

***Abstract*—This study investigates the potential of predicting the Dow Jones Industrial Average (DJIA) stock price movements using the top 25 daily news headlines from Reddit WorldNews Channel. The dataset, covering the period from 2008-08-08 to 2016-07-01, consists of news headlines ranked by reddit users' votes and binary labels indicating the DJIA's Adj Close value increase or decrease. To address this task, various machine learning methods such as Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting Machines, Stochastic Gradient Descent, and Support Vector Machine, FinBERT, DistilBERT, along with ChatGPT, including prompt engineering and result stabilization techniques, were applied as baseline models. More importantly, we also propose a novel approach called the Hybrid Attention Sequential Stock Model (HASS) to enhance stock price movement prediction. The SVM and naive Bayes classifiers demonstrate the best performance, and strengths and weaknesses of other models are analyzed.**

***Index Terms*—stock price prediction, text mining, machine learning, ChatGPT, prompt engineering, Hybrid Attention Sequential Stock Model**

I. Introduction

The stock market is an essential aspect of the global economy, influencing various aspects of daily life. Investors, traders, and the general public rely on various sources of information to make informed decisions about stock investments. Traditional approaches for predicting stock prices include analyzing historical data, such as price trends and trading volumes. However, with the vast amounts of news articles and social media posts, there is growing interest in exploring the potential of utilizing textual data to predict stock prices.

The dataset used in this study comprises the top 25 news headlines from the Reddit WorldNews channel and the DJIA stock data. The news headlines were sourced from the Reddit WorldNews channel and have been ranked according to users' votes. The dataset covers a period spanning from 2008-08-08 to 2016-07-01. The DJIA stock data has been obtained from Yahoo Finance and preprocessed by encoding a label of "1" to indicate that the DJIA Adj Close value either increased or remained the same, while a label of "0" was assigned to indicate a decrease in the DJIA Adj Close value.

To improve the performance of our models, we employ various pre-processing techniques, such as data cleaning with regular expressions. We also use deduplication techniques, such as clustering, to remove similar headlines and reduce noise in the dataset.

Apart from TF-IDF, we employ various machine learning methods, including logistic regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machines (GBM), stochastic gradient descent (SGD), and supported vector machine (SVM). These methods have been widely used in the analysis of textual data and have shown promising results in various studies.

We also explore the use of deep learning methods, specifically the FinBERT and DistilBERT models. The FinBERT model have been pre-trained on financial news articles and have shown promising results in financial sentiment analysis tasks. DistilBERT is a lightweight model trained by distilling the BERT base model. We also employ the ChatGPT model, exploring various prompt engineering and result stabilization techniques.

More importantly, we propose a novel approach to predict stock price movements using news articles, called the Hybrid Attention Sequential Stock Model (HASS). The HASS methodology consists of three main parts: motivation, model architecture design, and model training and evaluation. The motivation behind the HASS model is to capture the sequential nature of news articles and their relevance to stock price movements. The HASS model architecture combines attention mechanisms and sequential modeling, aiming to capture the relevant information from the news articles.

Among all the models, the SVM classifier and the naive Bayes classifier demonstrate the best performance, with an accuracy of 0.6080 and 0.6030, respectively. We analyze the strengths and weaknesses of the other models as well, highlighting their potential limitations in predicting stock price movements using textual data.

II. Related Work

Forecasting future stock prices using historical data has always been a subject of interest for researchers and financial practitioners. Early research primarily relied on conventional machine learning techniques to analyze past stock prices for predicting market trends. For example, Hegazy et al. [1] proposed a machine learning model including least square SVM (LS-SVM) and particle swarm optimization (PSO) algorithms for price forecasting of S&P 500. Yetis et al. [9] employed a feedforward artificial neural network (FFANN) on NASDAQ stock market index price data to predict stock values, and observed that the ANN demonstrated strong predictive performance for NASDAQ. Ou and Wang [2] applied ten different machine learning algorithms to stock price data to forecast movements in the Hong Kong stock market index. They discovered that SVM exhibited superior predictive performance compared to the other models.

Due to the widespread use of social media in recent years, a vast amount of news information is available on the internet. Certain types of news, such as those pertaining to natural disasters or government policies, have a close association with the stock market. Consequently, applying traditional machine learning and deep learning techniques to fit news data for predicting stock prices has become a compelling research area. The following part introduces these works from three main aspects.

A. *Disasters and terrorism*

Typically, information with an overtly negative connotation is expected to have a detrimental effect on the market. Worthington and Valadkhani [3] investigated the Australian capital market using data spanning from 31 December 1982 to 01 January 2002. Their study demonstrated that bushfires, cyclones, and earthquakes exerted a significant influence on market returns, while severe storms and floods did not have a similar impact. Chesney et al. [4] conducted a study on the impact of terrorist incidents on the capital market, bond market, and commodity market in various countries. Their findings demonstrated that roughly two-thirds of the terrorist incidents had a robust and noteworthy effect on all the markets examined. Specifically, industrial companies and airlines were found to be the most sensitive, whereas banking exhibited less sensitivity.

B. *Financial news*

Deep learning methods are more effective in processing complex financial news than in the case of disasters and terrorist attacks. Chen et al. [5] applied GRU and recurrent neural networks (RNN) to stock price data from the CSI 300 Index and news from Sina Weibo to forecast volatility in the Chinese stock market. Their proposed model outperformed the baseline methods and exhibited strong predictive performance. Dang et al. [6] introduced a new framework, Stock2Vec, which employs a two-stream gated recurrent unit (TGRU) network to predict the direction of stock prices in the S&P 500 using financial news from Reuters and BloombergFootnote6 and the Harvard IV-4 sentiment dictionary. Their results indicated that the TGRU network outperformed the baseline models, and that Stock2Vec was highly efficient when working with financial datasets.

C. *Social media*

Sentiment analysis is a primary application of social media data for predicting market trends. Users' emotions in tweets and comments can reflect their opinions regarding a company or product. Khatri and Srivastava [7] gathered tweets and comments from the Stock TwitsFootnote3 website and extracted market data from Yahoo Finance for five stock markets: Facebook, Apple, Google, Oracle, and Microsoft. They categorized the tweets and comments as up, down, happy, or rejected, and utilized this polarity index and stock data as input to an ANN for forecasting stock closing prices. Chakraborty et al. [8] acquired tweets containing the keywords 'AAPL', 'stock market', and 'stocktwits' to forecast the Apple Inc. stock index and stock market movement in the United States. They employed SVM for sentiment classification and a boosted regression tree model to predict the next day's stock difference.

III. Data

A. *Overview of the Dataset*

This report employs a dataset comprising the top 25 news headlines from the Reddit WorldNews Channel [13], along with a binary label indicating whether the stock price of the Dow Jones Industrial Average (DJIA) [14] increased on a given date. The dataset encompasses news headlines and label data spanning the period from 2008-08-08 to 2016-07-01. The news headlines, sourced from the Reddit WorldNews Channel, have been ranked according to reddit users' votes, and only the top 25 headlines for a given date are considered in this dataset. The Dow Jones Industrial Average (DJIA) stock data has been obtained from Yahoo Finance and has been preprocessed by encoding a label of "1" to indicate that the DJIA Adj Close value either increased or remained the same, while a label of "0" was assigned to indicate a decrease in the DJIA Adj Close value.

The vocabulary style typically used in headlines from the Reddit WorldNews Channel is objective and straightforward. This is likely due to the need for headlines to capture the reader's attention and convey the most important information within a limited space. The headlines often include the who, what, when, and where of the news story, with some headlines even incorporating additional details. Based on our analysis, the average headline length is approximately 17.7 words, which can be viewed as a summary of a news story. Therefore, for subsequent tasks, we do not believe it is necessary to further summarize each headline using techniques such as text summarization. In terms of content, the Reddit WorldNews Channel tends to focus on international news and current events, covering a wide range of topics such as politics and natural disasters. The headlines frequently highlight stories of global significance or those with the potential to impact people around the world.

B. *Data Cleaning with Regular Expression*

Firstly, for commonly used abbreviations, we expand them to their non-abbreviated format. For example, "hadn't" is replaced with "had not". Additionally, through exploring the dataset, we discovered that some news headlines in the dataset begin with special characters, such as "b/'" or "b/". Some news headlines also contain confusing punctuation and parentheses, which we remove using regular expressions and only retain the English alphabetic text. Finally, we remove any excess spaces from the headline text.

C. Deduplication with Clustering

After the text data cleaning in the previous step, we obtained relatively clean text data. However, for each day, we still have 25 news headlines. For different dates, these 25 news headlines may be highly repetitive. For example, they may be all about different aspects of a few events or related to ongoing coverage of these events. On the other hand, the repetition rate of 25 news headlines may be very low, with almost no significant correlation between each news item. In other words, we cannot simply cluster the news headlines for each day since we only know which news may be more similar, but we cannot select the most representative one for each news topic. We also cannot simply rank the news of each day by importance and only choose the top few, as it is entirely possible for news items from the same topic to appear more than once, while another important but lower-ranked news item may not be preserved.

As we have six years of data and manual classification is too complicated and expensive, it is almost impossible to manually evaluate how many different news topics there are for each day. Therefore, to maintain consistency in the final processing results, we implement an algorithm that takes both aspects into account. Specifically, we first try to extract keywords from the 25 news headlines of the day. It is implemented by `TfidfVectorizer` without considering stop words. We use Term Frequency-Inverse Document Frequency (TF-IDF) weights to rank all the keywords of the day and identify the top 15 most important keywords. To avoid distraction from unimportant news, we exclude news headlines that do not contain any of these 15 keywords. Statistically, after this step, we keep at least nine news items and an average of 17.5 news headlines. Then, we cluster the remaining news headlines. This step is implemented by K-means++ clustering on the previous TF-IDF encoding vectors. For each cluster, we only keep the news item with the most keywords. In this experiment, we set the hyperparameter number of clusters for the K-means++ algorithm [15] to three.

D. Text preprocessing based on word frequency

For machine learning models, embedding based on the frequency of words may be a better choice, because the structure of models makes the order have little effect on the results. We directly utilize TF-IDF to obtain the features of each data. This method is implemented using the TfidfVectorizer module from sklearn, and enables the selection of n-grams ranging from one to three.

E. Text preprocessing based on pre-training

Global Vectors for Word Representation (GloVe) is a type of word embedding method that aims to capture semantic relationships between words by analyzing the co-occurrence of words in large text corpora. It represents words as high-dimensional vectors in a continuous space, where the distances between vectors reflect the similarities or relationships between the corresponding words. To enhance the representation of text, we utilize the pre-trained glove-6B-300d for embedding, which are trained on a large corpus and can capture the semantic relationships between words.

IV. METHOD

A. Basic Methods

We utilize a variety of machine learning methods, including logistic regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machines (GBM), stochastic gradient descent (SGD), and supported vector machine (SVM). These methods are implemented based on sklearn, and follow the text preprocessing strategy in sec. III-D.

Besides, we built simple Multilayer Perceptron (MLP) and Recurrent Neural Network (RNN) through PyTorch, and compared different preprocessing methods according to sec. III-D and sec. III-E.

Since these models are relatively simple and covered in class, this section excludes method principle and are used only for comparison.

B. BERT

In this subsection, we are going to introduce the two different methods that based on the Bidirectional Encoder Representations from Transformers (BERT).

Financial domain-specific Bidirectional Encoder Representations from Transformers (FinBERT) [16] is a pre-trained transformer model for financial sentiment analysis. It is built by training the BERT model with a large financial corpus, followed by fine-tuning with FPB for sentiment classification. We fine-tune the FinBERT with our own training data to improve its performance. Early stopping was implemented and training would stop if the validation loss did not decrease further after five more steps after reaching the minimum. We also performed hyper-parameter tuning by adjusting the learning rate and weight decay.

Besides FinBERT, we also tried to fine-tune DistilBERT [17] for classification. DistilBERT model is a distilled form of the BERT model. DistilBERT is smaller, faster and cheaper than the normal BERT and it preserves over 95% of BERT's performances. The fine-tuning process was similar to the FinBERT. The best model was produced with a learning rate of $5\mathrm{e}^{-6}$ and weight decay of 0.1.

C. ChatGPT

To assess the impact of news on the stock market using ChatGPT, we employ a series of carefully designed prompting techniques. As fine-tuning the pre-trained model is not possible, we tried to resort to in-context learning, but this method is limited due to the daily volume of news articles and the increased cost of context length. Instead, we submit prompts and data to ChatGPT using the API, setting the temperature to 0 to ensure deterministic results and improve reproducibility.

Our prompting techniques include:

- JSON Formatting: We request that ChatGPT generates outputs in JSON format to maintain consistency and stability. Without this, the output may be arbitrary, making result parsing a significant challenge.
- `{"id": 1}`: For each JSON request, we include a redundant `{"id": 1}` key-value pair, and we ask ChatGPT to return the same ID in its response. This helps maintain input-output similarity, encouraging ChatGPT to adhere to our desired JSON format rather than rejecting the requirement.
- Chain-of-Thought: We prompt ChatGPT to generate an analysis before the final result. This popular technique helps ChatGPT leverage its chain-of-thought (CoT) [18] ability for more accurate answers by better understanding the problem at hand.
- Parameter documentation: We include parameter documentation for ChatGPT's JSON output format to enforce adherence to a specific format.
- Marking scheme: We incorporate a marking scheme in the prompt, requesting a binary 0/1 response from ChatGPT, even when information may be insufficient to determine stock changes. This approach aligns with our evaluation method, which is based on binary labels and does not accommodate intermediate values.

We found that placing prompts and data in the system prompt enhances generation quality, and we tested the performance of five distinct prompt variations.

We input raw data into ChatGPT models without preprocessing techniques like stemming, stop-word removal and text normalization, as ChatGPT is trained on natural language texts without preprocessing. We tested ChatGPT's performance on four different kind of inputs: all 25 news of a specific day, top-3 news, top-1 news, and top-1 news without CoT.

The prompt used for analysing the all 25 news is shown in table I.

TABLE I
FULL CHATGPT PROMPT

Based on several news articles from a certain day (given in JSON format below, taken from that day's top headlines), use a number to indicate the overall impact on that day's Dow Jones Industrial Average (DJIA) index (if the news articles suggest that the index is expected to rise, give 1, if it is expected to fall, give 0. DO NOT give other values, such as "neutral").

```
{
"id": 1,
"news": [
<inject news here>
]
}
```

Output in this JSON format:

```
{"id": ..., "analysis": ..., "impact": ...}
```

Parameters:

- 'id' ('int'): Same as the given 'id' value.
- 'analysis' ('str'): An analysis no more than 200 words. You must analyse before you reach a conclusion. You must not draw conclusions at the beginning.
- 'impact' ('int'): The overall impact suggested by the news. This value can only be 1 or 0. If the news articles suggest that the index is expected to rise, give 1, if it is expected to fall, give 0. This value can only be 1 (representing a rise) or 0 (representing a fall). This is because providing an uncertain judgment would be meaningless. Regardless of whether stocks rise or fall, even a small fluctuation can have a significant impact due to leverage. You are not held responsible for your judgment results, but you must provide a reasonable and accurate analysis for your 1 (rise) or 0 (fall) judgment. You must consider only news that has a significant impact on the stock market and disregard other news and other irrelevant information.

Marking scheme:

- Always give a clear conclusion (rise/fall). If your 'analysis' leads to an unclear conclusion, you will receive no point;
- If your 'impact' value (0/1) is correct, you will receive full points (1000);
- If your 'impact' value (0/1) is incorrect, you will be given 0-20 points based on your 'analysis';
- If your 'impact' value is not 0 or 1, you will receive no point (therefore, you must not give any value other than 0 or 1);
- If your response is not a single JSON object, you will receive no point.

D. Hybrid Attention Sequential Stock Model

In this study, we propose a novel approach to predict stock price movements using news articles, called the **H**ybrid **A**ttention **S**equential **S**tock Model (HASS). Our methodology consists of three main parts: motivation, model architecture design, and model training and evaluation. In the following sections, we detail each step and discuss the rationale behind the design choices.

1) Motivation: The primary motivation behind the development of the Hybrid Attention Sequential Stock Model (HASS) is to effectively capture and incorporate information from both recent and more distant news articles to predict stock price movements. By combining the strengths of LSTM [10] and attention mechanisms [12], our model aims to achieve a more comprehensive understanding of the impact of news articles on stock market movements. In this section, we discuss the key motivations behind the design choices for the HASS model.

2) Capturing Sequential Patterns from Recent Days: One important aspect of stock market behavior is the influence of recent events on stock prices. News articles from the latest days can often contain valuable information about current trends, events, or market sentiments that directly impact stock prices. To capture the sequential patterns from recent days, we employ an LSTM-based model. LSTM networks are well-suited for capturing temporal relationships in sequential data, making them an ideal choice for modeling the dependencies between news articles from the latest days. Let X be the input tensor with dimensions [batch size, number of days, max length], and let E be the BERT embeddings. We calculate the embeddings as follows:

$$E = BERT(X) \tag{1}$$

By extracting the contextualized embeddings from the pre-trained BERT [11] model and feeding them into an LSTM layer, our model is able to effectively capture the sequential patterns in the recent news articles and incorporate them into the prediction process. Let L be the LSTM output, and let P_L be the pooled LSTM output. We calculate the LSTM output and pooled LSTM output as follows:

$$L = LSTM(E) \tag{2}$$

$$P_L = AdaptiveAvgPool(L) \tag{3}$$

3) Attending to Important Patterns from More Distant History: While recent news articles play a significant role in stock market movements, it is also crucial to consider the impact of more distant historical events. Some events or trends from the past may continue to influence stock prices, and certain patterns in the news articles might only become apparent when considering a longer time frame. To attend to these important patterns from more distant history, we employ an attention-based model. Attention mechanisms are designed to identify and emphasize relevant information within a given context.

Let A be the attention weights, W be the attention-weighted embeddings, and P_A be the pooled attention output. We calculate the attention weights, attention-weighted embeddings, and pooled attention output as follows:

$$A = mean(AttentionWeights(E)) \tag{4}$$

$$W = A \cdot E \tag{5}$$

$$P_A = mean(W) \tag{6}$$

By computing attention weights for each token in the news articles from the past N days, our model can focus on the most significant patterns from more distant history. These attention-weighted embeddings are then pooled and combined with the embeddings from the LSTM-based model, allowing the HASS model to take into account both recent and more distant historical information when making predictions.

Let F be the concatenated embeddings from the LSTM-based and attention-based models, and let Y be the final output. We calculate the concatenated embeddings and the final output as follows:

$$F = concat(P_L, P_A) \tag{7}$$

$$Y = FC2(ReLU(FC1(F))) \tag{8}$$

Here, FC1 and FC2 represent the two fully connected layers in the fusion model.

In summary, the primary motivation behind the HASS model is to capture and incorporate both recent and more distant historical information when predicting stock price movements. By combining the strengths of LSTM and attention mechanisms, our model aims to provide a more comprehensive understanding of the impact of news articles on stock market behavior, ultimately leading to better predictive performance. Through the use of the described formulae and model components, the HASS model effectively captures sequential patterns from recent days and attends to important patterns from more distant history, resulting in a robust and versatile approach to stock price prediction based on news data.

E. Model Training and Evaluation

In the following section, we describe the optimization process for the Hybrid Attention Sequential Stock Model (HASS).

1) Loss Function: For the HASS model, we use the Binary Cross-Entropy Loss (BCELoss) as the loss function since we are dealing with a binary classification problem. The BCELoss is defined as:

$$\text{BCELoss}(Y,T) = -\frac{1}{N}\sum_{i=1}^{N}[T_i \cdot \log(Y_i) + (1-T_i)\cdot\log(1-Y_i)] \tag{9}$$

Here, Y represents the predicted output of the model, T denotes the target labels, and N is the total number of samples in the batch.

2) Model Training and Evaluation: During the training phase, the model's parameters are updated iteratively using Adam, to minimize the BCELoss. The model is evaluated on the validation set periodically to monitor its performance and prevent overfitting. The training process is stopped when the validation loss stops improving, and the model with the best validation performance is saved for final evaluation on the test set.

By employing the BCELoss as the loss function and monitoring the model's performance on the validation set, we ensure that the HASS model is effectively trained to capture the underlying patterns in the news data and make accurate predictions for stock price movements.

V. Experiment

1) Results of Different Models: The results of different models is shown in table II.

TABLE II
RESULTS OF DIFFERENT MODELS

Approach	***Accuracy***
Logistic Regression (token counts)	0.4623
Logistic Regression (bigram)	0.5729
Naive Bayes	**0.6030**
Random Forest	0.5678
Gradient Boosting	0.4773
Stochastic Gradient Descent Classifier	0.5126
SVM classifier	**0.6080**
MLP classifier	0.5980
RNN + Glove	0.5276
FinBERT	0.4975
DistilBERT	0.5327
DistilBERT (top 3 news)	0.5377
ChatGPT (all news)	0.4724
ChatGPT (top-3 news)	0.4925
ChatGPT (top-1 news)	0.4774
ChatGPT (top-1 news w/o CoT)	0.5025
HASS (Hybrid Attention Sequential Stock Model)	0.5450

2) Analysis of ChatGPT: Our ChatGPT approach has its limitations, which are discussed below.

One major drawback of ChatGPT is its unstable output, even after using stabilization methods. In some cases, ChatGPT still outputs "it's impossible to determine the impact" or "the news articles do not provide a clear indication", which is not ideal for our use case. This is due to the fact that ChatGPT has been trained with reinforcement learning from human feedback (RLHF) to avoid providing definitive answers to uncertain questions [19]. However, this approach is not suitable for our scenario. To address this issue, we can fine-tune separate ChatGPT-like language models, such as LLaMA [20], for specific use cases.

Another limitation of ChatGPT is that it has been trained on a general dataset, and may not capture financial analysis or investment expertise as well as human experts. This means that its ability to extract relevant information from news articles may be limited in some cases.

Moreover, ChatGPT may suffer from information leakage, as it is aware of past events. However, the low prediction results demonstrate that this is not a major concern, possibly due to the need to predict fine-grained data at the daily level, which ChatGPT may not be able to achieve.

Despite these limitations, ChatGPT has several advantages over traditional machine learning models. For example, although deep learning models are often depicted as "black boxes", ChatGPT can output explanations and analyses alongside its answers, as demonstrated in table III, allowing users to judge the sufficiency of the reasoning behind the model's predictions. This makes it more suitable for real-world applications where trust is critical.

TABLE III
AN ANALYSIS GENERATED BY CHATGPT

The release of the Panama Papers, which reveal the offshore holdings of politicians, celebrities, and criminals, has caused a stir in the financial world. The leak implicates several major banks, legal firms, and asset management companies in the management of these offshore accounts. The fallout from the leak has already led to calls for the resignation of Iceland's Prime Minister and investigations into tax evasion by the Australian Tax Office. However, Panama has vowed to cooperate with any legal fallout from the leak. The impact on the DJIA is likely to be negative, as the leak has the potential to cause a loss of confidence in the financial system and lead to increased regulation and scrutiny. Additionally, the report claiming that the 'war on drugs' has harmed public health could lead to a decrease in pharmaceutical stocks. However, the news that half of Scotland's energy consumption came from renewables last year could have a positive impact on energy stocks.

Interestingly, using only the top-3 news headlines produced better results than using all 25 news headlines. This suggests that too many news headlines can be a distraction for the model.

Surprisingly, the best results were achieved when the CoT was not used, contrary to our expectations. We believe this is because predicting stock prices is a challenging task for ChatGPT, and its performance is affected by randomness.

3) Analysis of HASS: The HASS model combines two distinct approaches for capturing and incorporating information from news articles to predict stock price movements: an LSTM-based model to capture sequential patterns from recent days and an attention-based model to attend to important patterns from more distant history. The integration of these two mechanisms within a single model allows the HASS model to effectively leverage both recent and more distant historical information when making predictions. The HASS model offers several notable advantages:

- Comprehensive understanding of news data: By combining the strengths of LSTM and attention mechanisms, the HASS model can capture and incorporate a wide range of information from news articles, leading to a more comprehensive understanding of the impact of news on stock market movements.
- Robustness to varying time frames: The HASS model can effectively adapt to different time frames, as it is designed to capture both recent sequential patterns and important patterns from more distant history.
- Flexibility: The model is highly flexible, allowing for adjustments to the number of days considered in the LSTM-based and attention-based models, as well as modifications to the model architecture and hyperparameters for better performance on different datasets and tasks.
- Transfer learning: The HASS model leverages pre-trained BERT embeddings, allowing for effective transfer learning and reduced training time.

Despite its strengths, the HASS model also has some limitations:

- Computational complexity: The HASS model's combination of LSTM and attention mechanisms, along with the use of BERT embeddings, can result in increased computational complexity and longer training times.
- Susceptibility to overfitting: Due to the large number of parameters in the model, the HASS model can be prone to overfitting, especially when trained on small datasets. Regularization techniques and early stopping can be employed to mitigate this issue.
- Interpretability: The HASS model's complex architecture and the combination of LSTM and attention mechanisms may make it challenging to interpret and understand the model's predictions.

The HASS model's ability to effectively capture and incorporate information from news articles makes it suitable for various applications related to stock market prediction and financial analysis. Some potential applications include:

- Predicting stock price movements based on news articles.
- Identifying the impact of specific news events on stock market behavior.
- Analyzing the influence of market sentiment and news trends on stock prices.
- Investigating the relationship between news articles from different time frames and stock market movements.

In conclusion, the Hybrid Attention Sequential Stock Model (HASS) presents a robust and versatile approach for predicting stock price movements based on news data. By effectively combining LSTM and attention mechanisms, the HASS model captures both recent and more distant historical information, resulting in a comprehensive understanding of the impact of news articles on stock market behavior. While there are some limitations, such as computational complexity and susceptibility to overfitting, the HASS model offers a promising direction for further research and applications in stock market prediction and financial analysis.

4) Advantages of machine learning models: Although we utilize many methods including pre-trained embedding, Bert, ChatGPT and a new model based on hybrid attention, the traditional maching learning method still shows its effectiveness. We believe there are several main reasons, as follows:

First, due to the small amount of data, it is difficult for our deep learning model to achieve the best performance. In such a case, machine learning methods can achieve higher indicators due to the simple fitting process.

Second, different data pre-processing methods has a greater impact on the final results. For machine learning, the input data is embedded according to the word frequency. But for other models, the entire sentence is embedded directly following a common way. This illustrates that for stock predictions, specific words are more important than the meaning of the entire sentence. For example, most of the time, regardless of the details of a terrorist attack, as long as there is a terrorist attack, the stock market will inevitably react negatively. This is also in line with our common sense.

Third, stocks are inherently unpredictable. Clearly, if stocks can be predicted that easily, we'd be richer than Elon Musk. Despite all the processing, we only focus on the news, but the actual stock market is much more complicated than that. It is very likely that our model can only learn the bias rather than the actual knowledge. In this case, machine learning models can quickly learn this bias, which makes the result better.

VI. Conclusion

This paper presents a comprehensive analysis of using news articles to predict stock price movements. We employed a dataset of top news headlines from the Reddit WorldNews Channel and Dow Jones Industrial Average stock data to train and evaluate various machine learning models. In addition, we proposed a novel approach, the HASS model, which outperformed several other models. Our study provides insights into the strengths and weaknesses of different models and demonstrates the potential of using news articles as a valuable source of information for stock price prediction. Further research could focus on exploring other datasets and incorporating more advanced text mining techniques.

References

[1] Hentschel M, Alonso O (2014) Follow the money: a study of cashtags on Twitter. First Monday 19(8)

[2] Ou P, Wang H (2009) Prediction of stock market index movement by ten data mining techniques. Mod Appl Sci 3(12):28

[3] Worthington A.C. & Valadkhani A, 2004, 'Measuring the impact of natural disasters on capital markets: An empirical application using intervention analysis', Applied Economics 36(19), 2177–2186.

[4] Chesney M., Reshetar G. & Karaman M, 2011, 'The impact of terrorism on financial markets: An empirical study', Journal of Banking and Finance 35(2), 253–267.

[5] Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS (2017b) Stock market prediction using neural network through news on online social networks. In: IEEE international ISC2, pp 1–6

[6] Dang LM, Sadeghi-Niaraki A, Huynh HD, Min K, Moon H (2018) Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access 6:55392–55404

[7] Khatri SK, Srivastava A (2016) Using sentimental analysis in prediction of stock market investment. In: IEEE 5th international conference ICRITO, pp 566–569

[8] Chakraborty P, Pria US, Rony M, Majumdar MA (2017) Predicting stock movement using sentiment analysis of Twitter feed. In: IEEE 6th international conference ICIEV-ISCMHT, pp 1–6

[9] Yetis Y, Kaplan H, Jamshidi M (2014) Stock market prediction by using artificial neural network. In: IEEE WAC, pp 718–722

[10] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[11] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[12] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[13] Reddit WorldNews Channel (2023). Retrieved 16 April 2023, from https://www.reddit.com/r/worldnews/

[14] Dow Jones & Company (2023). Dow Jones Industrial Average [DJIA]. Retrieved from https://finance.yahoo.com/

[15] Arthur, D., & Vassilvitskii, S. (2007, January). K-means++ the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027-1035).

[16] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

[17] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

[19] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. arXiv preprint arXiv:2304.01852.

[20] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.