News Algorithmic Trading Strategy: LLM Approach

Artem Minakov, Elena Putilova, Diana Sharafeeva

1. Introduction

Algorithmic trading, also known as algo-trading, employs computer programs that follow a defined set of instructions to place trades with the aim to generate profits at a speed and frequency that is impossible for a human trader. One of the core principles behind algorithmic trading is the systematic execution of trading orders through pre-programmed trading instructions accounting for variables such as time, price, and volume. This method of trading has grown significantly in financial markets since the 1990s, facilitating faster, more efficient, and less error-prone execution of trades compared to traditional manual trading.

One of the critical sources of information for algorithmic trading strategies is news events. News can significantly influence market sentiment and, consequently, market prices. This includes a wide array of information types, such as economic indicators, financial statements, geopolitical events, and more. Traditional algorithmic trading systems leveraged mostly numerical data and pre-defined patterns for trading decisions. However, due to the complexity and subtlety of news texts, more sophisticated approaches are necessary to harness the full potential of news-based trading strategies. Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP), demonstrating remarkable abilities in understanding, generating, and interpreting human language.

The integration of LLMs into algorithmic trading strategies based on news events represents a cutting-edge area of research. This approach not only enhances the ability to interpret and act upon news information but also opens up new avenues for developing more sophisticated and adaptive trading algorithms. For instance, Xianrong et al. (2024) demonstrate how LLMs can help in devising more nuanced and dynamic trading strategies that can better navigate market volatilities and uncertainties.

Studies have investigated how LLMs can be employed to analyze the emotional tone of textual data (financial news) and predict market movements as well as individual stock price movements. Some research examines the incredible effectiveness of LLMs in extracting relevant information from news texts and assessing their impact on market prices. Introducing the FinBERT-LSTM model, which integrates news article sentiments to predict stock prices with greater accuracy. This model analyses short-term market information, and, according to Halder (2022), it was trained on NASDAQ-100 index stock data and New York Times news articles, showing a 98% accuracy in predicting price movements for individual stocks.

Another common method for evaluating the performance of algorithmic trading strategies is the ‘Sharpe Ratio’, a metric that measures the performance of an investment such as a security or portfolio compared to a risk-free asset, after adjusting for its risk. In a recent study, 7 different LLMs were tested on historical data for stocks from the S&P 500 and news articles (headlines and full article texts). The study showcased a negative Sharpe ratio for older models like GPT-1, GPT-2, and BERT, with the best result being a 3.8 Sharpe ratio for GPT-4 (Lopez-Lira et al., 2023).

Overall, several studies have underscored the transformative potential of incorporating Large Language Models (LLMs) into algorithmic trading strategies, particularly those based on the analysis of news events. Inspired by these findings, we aim to conduct our own study to explore the feasibility of leveraging LLMs for algorithmic trading by analyzing news articles.

2. Methodology

When executing the project, we aimed to find a balance between standards of data when working with LLMs and the available resources, therefore adjusting the steps to the appropriate work and time scope.

2.1 Data

Gaining access to high-quality (historical) stock market news data is hard and expensive (both money-wise and computational and timing resources) if doing it from scratch, so we have decided to use the existing dataset. In this project we are using the ‘Daily Financial News for 6000+ Stocks’ dataset from Kaggle, which is a high-quality dataset with 843,062 unique news events concluding the following information: headline, URL, article author, publication timestamp, stock ticker symbol for the period between 2010 and 2022. All the data was scraped from Benzinga - a widely known and used financial media outlet, publishing 500+ news daily. For computational reasons, we have limited the entire dataset to 50 companies from the S&P 500 index that had the most number of connected news events, which led to 98k data points.

To conduct a comprehensive comparison of classification results, both headline-based and full-text-based approaches were explored. Full body text extraction was achieved by scraping original URLs sourced from a Kaggle dataset encompassing articles from a total of 15 web sources. However, only 7 of these sources permitted web scraping, while a significant portion of URLs in the dataset were found to be either broken or outdated. Despite these challenges, a total of 2,640 articles were successfully scraped using Python's asyncio and aiohttp libraries in conjunction with BeautifulSoup.

As the full body of articles often exceeds the token limits of models like BERT (512 tokens) and GPT (limited by total tokens), summarization techniques were employed. Given the median size of articles surpassing 700 tokens, summarization became imperative. Two primary approaches were utilized: Extractive summarization and Abstractive summarization. Extractive summarization involves selecting and extracting essential sentences or passages directly from the source text without generating new content. This method, implemented using the 'distilbert-base-uncased' model, allows for concise summaries while retaining the original context. Conversely, Abstractive summarization generates summaries containing new sentences and phrases not present in the source text, thus capturing the essence of the content more dynamically. For this approach, the 'sshleifer/distilbart-cnn-12-6/' model was employed, enabling a more comprehensive understanding of the articles' content.

As for the historical market data, we have parsed it from Yahoo finance for all the stocks from the S&P 500 list, so that in the end we could connect the news events about the company and its price on the certain date. We collected historical market data from Yahoo Finance for all S&P 500 stocks to correlate news events with stock prices. Each day's movement of a company's stock was labeled as neutral, positive, or negative relative to the S&P 500 index movement (the defined threshold is 0.4%).

2.2 Large Language Models & Prompting Techniques

To comprehensively explore news-based trading strategies, we selected five LLMs representing three main categories: encoder (BERT, RoBERTa, ELECTRA), decoder (Chat GPT-3.5 Turbo), and transformer (Flan-T5 Large). This diverse selection allows for a nuanced analysis of various model architectures' performance in news-based trading.

Prompting techniques play a crucial role in guiding LLMs to make accurate predictions based on limited examples. In our study, we adopted few-shot and zero-shot prompting techniques for Chat GPT and Flan-T5, which are commonly used in similar research endeavors. Few-shot prompting involves providing the model with a small number of examples to learn from, enhancing its understanding of the task at hand. Conversely, zero-shot prompting allows the model to make predictions without any explicit training examples, relying on its pre-existing knowledge. By employing a 3-shot mode for both headlines and body articles, we aimed to strike a balance between providing sufficient context for the models to learn from and minimizing the computational overhead. Several prompts were tested and the final one looks as followed:

"In this task you have to determine if a piece of text might lead to a positive or negative change in the price of a stock mentioned in the headline. If you are unsure whether a headline or article would lead to a change in the stock price, err on the side of caution and label it as neutral. + 3 examples, when few-shot + 'With these instructions in mind and a piece of text, please reply with either of the three options and nothing else: 1) positive, 2) negative, 3) neutral. How this headline or article might be labeled based on the past guidelines:' + given headline or body text?"

For evaluating the performance of each model and prompting technique, we employed a comprehensive set of metrics including accuracy, F1 score, and confusion matrix. These metrics, based on historical market data, offer valuable insights into the models' ability to predict stock price movements accurately based on news events.

3. Results and Analysis

3.1 Encoders: BERT, RoBERTa, ELECTRA

When fine-tuned on 80% of the dataset and tested on the remaining 20%, encoders showed disappointing results in classification. They struggled to distinguish between the three classes (positive, negative, and neutral), often classifying all data points into a single class, leading to very skewed results. Specifically, BERT performed poorly, assigning all data points to the 'negative' class, and RoBERTa exhibited similar behavior. Among the encoders, ELECTRA showed slightly better performance by managing to differentiate between positive and negative classes, yet the weighted average F1-score remained low at only 0.28. It is noteworthy that, regardless of the number of epochs, there was no improvement in the models' training loss.

3.2 Decoder: Chat-GPT-3.5

The summarization of body texts by Chat GPT-3.5 Turbo was analyzed using both few-shot and zero-shot modes for abstractive and extractive summarization methods, resulting in 10,560 responses from 2,640 data points. Due to limited computational resources, not all data could be processed with every prompting technique. For instance, processing zero-shot mode on headlines (23k data points) required 4.5 hours, leading to the exclusion of few-shot mode for headlines. The performance of the Chat-GPT model surpassed that of the fine-tuned encoders. In zero-shot mode, the accuracy was 34.20% with a weighted F1 Score of 0.309 for headlines, and 35.3% accuracy with a 0.32 weighted F1 Score for summarizations. Despite the modest performance, it was notable that the OpenAI model yielded consistent results without triggering any guardrail responses.

3.3 Transformer: Flan-T5-Large

The Flan-T5-Large model was evaluated using the same data and prompts as the ChatGPT model. Despite being an open-source model with significantly less training data compared to OpenAI's models, Flan-T5-Large demonstrated superior performance. For example, its zero-shot mode on headlines achieved an accuracy of 45.11% and a weighted F1 Score of 0.36.

3.4 Additional Findings

- Employing summarization instead of headlines led to marginally improved performance across all models, though the difference was not substantial.
- There was no significant difference between abstractive and extractive summarization methods, as shown in Figures 1 and 2.
- Contrary to our hypothesis, the analysis did not confirm better results for news events from the 2010s. Figure 3 illustrates this trend with the Flan-T5-Large model's performance, which was mirrored by the other models, showing no clear correlation between the accuracy of the model's predictions and the publication year of the news.
- The two prompting techniques, few-shot and zero-shot, did not exhibit significant differences in effectiveness.

4. Limitations

In integrating Large Language Models (LLMs) into news-based algorithmic trading strategies, several critical limitations emerged during our project. One significant limitation arises from the dependence on a single dataset, which may not capture the full spectrum of news impacts across various markets or regions, potentially limiting the generalizability of our findings. Additionally, the challenge of outdated or inaccessible full-text articles could bias the dataset towards more readily available content, possibly overlooking influential news. The necessity for summarization to manage token limitations may also simplify complex news narratives, risking the loss of crucial details.

Compounding these limitations are the inherent risks of hallucinations and biases within LLMs. Hallucinations, where LLMs generate inaccurate or fabricated information, necessitate robust validation processes to ensure reliability in financial trading decisions. Similarly, biases in LLMs, stemming from their training data, could skew predictions, requiring thorough pre-training on the relevant and valid financial data.

In conclusion, while our study offers valuable insights into the potential of LLMs in enhancing news-based algorithmic trading strategies, these limitations underscore the need for further research by incorporating a wider variety of news sources, improving methodologies for dealing with incomplete data, and continuously updating the models.

5. Conclusion

Our investigation into the application of Large Language Models (LLMs) for algorithmic trading strategies based on news events underscores the potential and limitations of current NLP technologies in financial applications. The relatively poor results across most models, can likely be attributed to the limited data used and models not being pretrained on financial data in advance. In the existing research, the best performance is shown by the LLMs trained on the combination of the financial news data and historical stock prices. Decoder and transformer models, especially Chat GPT-3.5 Turbo and Flan-T5-Large, demonstrated a more robust ability to generate predictions from financial news texts, with Flan-T5-Large showing superior performance despite its smaller size and training data volume. This suggests that the specific architecture of LLMs and their training data significantly impact their effectiveness in financial applications.

References

Cohen, G. 2022. Algorithmic Trading and Financial Forecasting Using Advanced Artificial Intelligence Methodologies. Mathematics 10(18): 3302. https://doi.org/10.3390/math10183302

Halder, S. 2022. FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis. https://arxiv.org/abs/2211.07392

Hu, G. 2023. Advancing Algorithmic Trading: A Multi-Technique Enhancement of Deep Q-Network Models. https://doi.org/10.48550/arXiv.2311.053

Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T., & Mandic, D. 2024. FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications. https://arxiv.org/abs/2403.12285

Lopez-Lira, A., & Tang, Y. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. https://arxiv.org/abs/2304.07619

Zheng, Xianrong, Gildea, Elizabeth, Chai, Sheng, Zhang, Tongxiao, and Wang, Shuxi. 2024. "Data Science in Finance: Challenges and Opportunities." AI 5(1): 55-71. https://doi.org/10.3390/ai5010004

Appendix

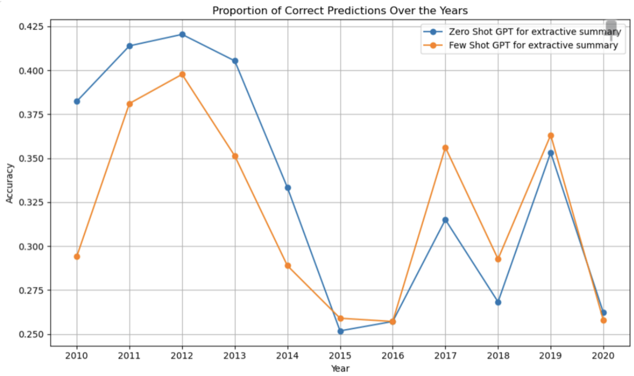

Figure 1: ChatGPT: proportion of the correct predictions for extractive summary

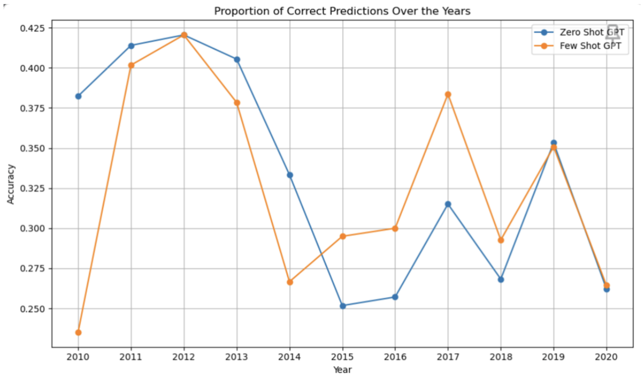

Figure 2: ChatGPT: proportion of the correct predictions for abstractive summary

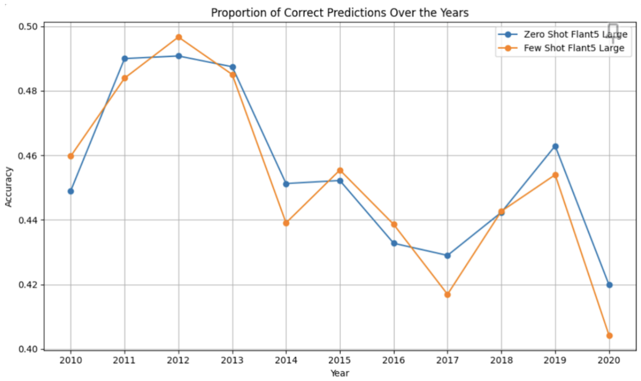

Figure 3: Flan-T5-Large: proportion of the correct predictions over the years for headlines

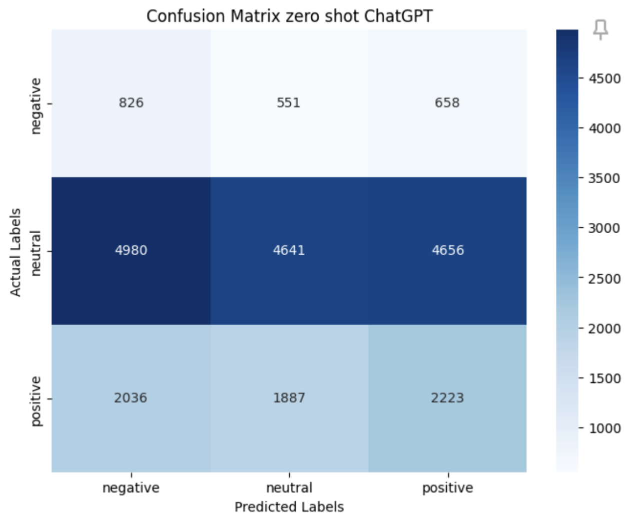

Figure 4: Figure 4 Chat-GPT model confusion matrix for zero-shot: headlines