İstanbul
Bilgi Üniversitesi

AI-Driven Stock Market Forecasting: A Hybrid Approach of News Analytics and Time-Series Data

by

ASLI GİZEM ULUSOY, 121200107
UMUTCAN ADIGUZEL, 120200087
RAMİZ KADAYIFÇI, 120200027

Supervised by

Dr. Emel Küpçü

Submitted to the
Faculty of Engineering and Natural Sciences
in partial fulfillment of the requirements for the

Bachelor of Science

in the
Department of Computer Engineering

—June, 2024

Table of Contents

Table of Contents **ii**

1 Realistic Constraints **1**
- 1.1 Data Diversity . . . 1
- 1.2 Date Mismatch . . . 1
- 1.3 Limited Compute Power . . . 1
- 1.4 GDELT Access Failure . . . 1

2 Risk Analysis and Precautions **2**
- 2.1 Risk 1 . . . 2
- 2.2 Risk 2 . . . 2
- 2.3 Risk 3 . . . 2
- 2.4 Risk 4 . . . 2

3 Cost **3**
- 3.1 Design and Implementation Cost: . . . 3
- 3.2 Computational Resources: . . . 3
- 3.3 Data Costs: . . . 3
- 3.4 Other Expenses: . . . 3

4 Standards **4**
- 4.1 Data ethics and Legal compliance: . . . 4
- 4.2 IEEE Standards . . . 4
- 4.3 NSPE Code of Ethics: . . . 4

5 Contributions **5**
- 5.1 Ramiz Kadayıfçı (FrontEnd and Mobile Development) . . . 5
- 5.2 Umutcan Adiguzel (Backend Development) . . . 6
- 5.3 Aslı Gizem Ulusoy (Hybrid Artificial Intelligence Model; Numerical and News Sentiment Analysis) . . . 6

References **7**

6 Conclusion **7**

1 Realistic Constraints

1.1 Data Diversity

We encountered several limitations during the project development process. Since the datasets used (GDELT news archive) are a very diverse news archive, we spent a long time processing and cleaning the data. Since GDELT data contains news published worldwide, there are news in various languages. This situation caused difficulties in the natural language processing process.

1.2 Date Mismatch

Correctly matching the news data with the financial movement data is a big challenge. Because there may be time differences between the date the news was published and the date of the financial movement.

1.3 Limited Compute Power

Working with deep learning models such as LSTM requires powerful computer performance and long training times. Due to the lack of resources at our disposal, the training process was limited to free platforms such as Google Colab, which led to inefficiency.

1.4 GDELT Access Failure

Our access to the GDELT site was completely cut off 1 week before the project delivery. The site crashed and is still not accessible. Due to the inconsistency of the data received from different sources, we are currently using the last 6 months of data from GDELT. This temporarily prevented us from seeing the performance of the model we produced in the long term.

2 Risk Analysis and Precautions

2.1 Risk 1

The biggest risk is that LSTM is a model prone to overfitting. This tendency means that the model loses its general analysis ability due to its excessive focus on the training data. Therefore, dropout layers were added and early stopping techniques were applied.

2.2 Risk 2

Another risk is the misclassification of news during sentiment analysis. Some news headlines can be short and disconnected from the meaning. Therefore, BERT-style models that grasp this better were preferred.

2.3 Risk 3

Time shifts are also a risk when matching between news and stock market movement data. In order to prevent this situation, news was analyzed in a limited area around certain local points.

2.4 Risk 4

Since the financial world is affected by various situations, the project is sensitive to financial shocks. Since it was observed that the performance of the model decreased due to such reasons, anomalies were labeled and the model was allowed to make separate evaluations on the data.

3 Cost

The cost analysis of the project was developed by estimating both labor and technical expenses required in the research and development processes. Although no external financial support was allocated, we calculated the actual infrastructure usage and costs assuming the conditions in which the project developed were professional.

3.1 Design and Implementation Cost:

Although physical components were not used in this project, a digital infrastructure is still required for the design process. We used languages such as Python, PHP, Javascript and free open source packages that do not require licensing costs.

3.2 Computational Resources:

Model training was performed on paid platforms such as Colab Pro due to data redundancy and model size. A fee of 165.60 was paid for Colab Pro. In scenarios where there will be more intensive workloads, if these costs were to be done in a physical GPU environment, it would have caused a monthly cost of approximately 200-300 USD.

3.3 Data Costs:

The GDELT project provides the use of a real-time global news archive for free. Although no subscription is required, processing them requires continuous internet use and temporary storage since the data volume is very high. An indirect cost can also be evaluated from here. In case of using cloud-based storage, such data operations may have a cost band that will increase up to 100-200 BD dollars depending on the frequency.

3.4 Other Expenses:

Additional expenses such as internet usage, computers that remain uninterrupted for days for model training, data backup contribute marginally to the total cost, with an estimated approximate 100-150 USD throughout the project.

4 Standards

During the development of the project, we followed various engineering and technology standards affecting application design, data processing, model training and ethical responsibilities. It was a fundamental goal to continuously ensure that the system also complied with professional principles.

4.1 Data ethics and Legal compliance:

Since the project used publicly available data, no private or personal user data was used. We complied with Turkey's KVKK regulations regarding data privacy. These principles were especially important due to the multilingual nature of the data.

4.2 IEEE Standards

- **IEEE/ISO/IEC 16085:2020** This standard provides a systematic view to identify, analyze and monitor risks throughout the software lifecycle. In our project, we used principles to evaluate overfitting risks, sentiment misclassification risks and data reliability issues, especially during the model development and data processing stages.
- **IEEE 70001-2021** Since our project involved complex prediction and sentiment analysis processes, this standard guided us in ensuring that AI decisions were explainable and traceable. We ensured that model predictions were reproducible and accurately recorded.
- **IEEE 7002-2022** All data used in the project is publicly available through the GDELT archive. However, we followed this standard to understand how to manage data flow, storage, and access in the most reliable way. Even when using publicly available data, we implemented best practices to avoid data misuse.
- **IEEE 829** During testing of the LSTM and BERT models, we kept a log of the training output. This standard influenced how we documented model performance, including precision, recall, F1 scores, and losses.

4.3 NSPE Code of Ethics:

Throughout the development of the project, we adhere to the principles set forth in the NSPE Code of Ethics towards society, the client, and engineering

itself.

- Engineers should avoid deceptive behavior and misleading actions.
- Professional responsibility requires engineers to act ethically and honestly to preserve the integrity and purpose of the profession.
- Engineers should prioritize honesty and transparency.
  - Errors should be acknowledged and facts should not be altered or misrepresented.
- Any behavior that misleads or confuses the public is unethical and should be avoided.
- Professional responsibility requires engineers to act ethically and honestly to maintain the integrity and purpose of the profession.

Engineers should prioritize honesty and transparency. a. Errors should be acknowledged and facts should not be changed or misrepresented.

Any behavior that misleads or confuses the public is unethical and should be avoided.

5 Contributions

Authors in the project are Ramiz Kadayıfçı, Aslı Gizem Ulusoy and Umutcan Adıgüzel.

5.1 Ramiz Kadayıfçı (FrontEnd and Mobile Development)

- Creating the cross platform mobile application Finvisor using React Native (Javascript), and Tailwind CSS.
- Designed the UI from stratch.
- Working with his teammates Umutcan and Aslı to create the report and presentation.
- Implementing and testing API connections with using Redux Toolkit, a package of React(Javascript).

5.2 Umutcan Adiguzel (Backend Development)

- Developing the backend infrastructure of the Finvisor mobile application using Laravel (PHP).
- Designing and implementing RESTful APIs to support user authentication, post creation, comments, likes, and AI prediction endpoints.
- Managing database architecture, migrations, and Eloquent relationships to ensure scalable and maintainable data flow.
- Collaborating with teammates Ramiz and Aslı to prepare the final report and presentation.
- Ensuring API security and performance by integrating Sanctum authentication, request validation, and optimized query structures.

5.3 Aslı Gizem Ulusoy (Hybrid Artificial Intelligence Model; Numerical and News Sentiment Analysis)

- Conducted an in-depth and comparative literature review on the use of artificial intelligence in financial forecasting. Performed comparative analysis of various methodologies applicable to financial prediction.
- Developed a custom financial forecasting methodology combining both traditional and modern approaches.
- Designed and executed experiments using conventional prediction techniques and analyzed their outcomes.
- Implemented sentiment analysis using the GDELT v2 dataset and FinBERT to extract meaningful insights from global news sources.
- Performed numerical analysis using historical market data obtained from Yahoo Finance.
- Trained and evaluated multiple baseline models, including Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron (MLP) neural networks.
- Developed, tested, and optimized LSTM and Bidirectional LSTM models under various configurations for improved sequential modeling performance.

6 Conclusion

This project successfully demonstrated a forecasting system that effectively applied AI techniques for market analysis in the financial world and integrated news analysis into it. Integrating GDELT data with the models created a platform that predicted stock movements based on global news.

The system effectively processed global news from GDELT, performed sentiment analysis using NLP models, and was capable of making predictions about the movement of the financial market. The risk of overfitting was minimized with dropout layer and early stopping techniques, adopting BERT-based models and increased the accuracy of sentiment classification for complex news.

IEEE Standards (16085:2020, 7001-2021,7002-2022,829) and NSPE Code of Ethics were followed. This allowed us to maintain professional integrity throughout the development process. Risk management techniques effectively addressed the issues of overfitting and misclassifying sentiment, temporal mismatch, and financial shock sensitivity.

As a result, this project can be seen as an important step towards combining AI-driven sentiment and movement analysis with financial forecasting, and can be presented as a practical solution for market analysis. The potential to adapt to local markets such as Borsa Istanbul also shows the project's potential to contribute to Türkiye's financial technology.