REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

TIME SERIES SENTIMENT ANALYSIS WITH LLMS

21011610 — Yusuf Taha KÖRKEM
19011006 — Mevlana Halit KAYA

SENIOR PROJECT

Advisor
Assist. Prof. Dr. Göksel BİRİCİK

December, 2023

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI	Artifical Intelligence
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
VAR	Vector Autoregression

LIST OF FIGURES

LIST OF TABLES

1
Introduction

In the era of ubiquitous connectivity and unfettered expression, individuals have unprecedented opportunities to share their opinions and sentiments regarding a vast array of products, services, and events. This wealth of emotional data, often disseminated through social media platforms and online review forums, holds immense potential for unlocking valuable insights when analyzed over time.

This project delves into the realm of time-series sentiment analysis, employing large language models (LLMs) to extract meaningful patterns from the ever-evolving tapestry of human emotions. Through this lens, we examine diverse facets of a particular topic or product, meticulously tracing the changes of sentiment over time. By discerning these complex emotional trajectories, we can uncover the underlying dynamics of public perception and attempt into the realm of predictive analytics, forecasting future trends and potential shifts in sentiment.

To conclude, this project, residing at the crossroads of technological innovation and the intricacies of human emotion, endeavors to harness the capabilities of large language models (LLMs) to illuminate the ever-shifting landscape of public sentiment. By unraveling the dynamics of sentiment over time and proactively anticipating future trends, our ultimate goal is to provide invaluable insights that not only inform strategic business decisions but also contribute to the formulation of effective policies. In doing so, we seek to amplify a deeper understanding of the collective emotional climate, offering a valuable tool for navigating the nuanced terrain of evolving public opinion.

2
Preliminary Examination

2.1 Similar Studies

In this section, similar studies previously conducted in the area where the project intends to work have been examined.

In the research by Georgoula et al, the topic discussed is factors affecting Bitcoin prices in the short and long term, considering various economic and technological indicators, including Twitter sentiment. A machine learning algorithm is employed to assess daily Twitter user sentiment towards Bitcoin. The study reveals that Twitter sentiment positively influences Bitcoin prices in the short run, indicating the potential of sentiment analysis in predicting price movements. Furthermore, increased Wikipedia search queries and hash rate positively impact Bitcoin prices, suggesting that public awareness and mining difficulty influence its value. Conversely, a negative relationship exists between Bitcoin prices and the USD-euro exchange rate. Long-run analysis indicates a positive impact of Bitcoin stock on its price, contradicting the general assumption that increased supply would lower prices. However, the S&P 500 index negatively affects Bitcoin prices, suggesting that investors view them as substitutes. The speed at which Bitcoin prices adjust to their long-run equilibrium is relatively high, validating its efficient market hypothesis. The study offers avenues for improvement, including using a larger dataset, vector autoregressive (VAR) models, and alternative sentiment indices [1].

Ali Asgarov's research, as outlined in the mentioned article, delves into develop a predictive model for stock prices by combining historical data and sentiment scores from Twitter. Focusing on major companies like Apple and Tesla, the researchers collected financial data and associated tweets using the Yahoo Finance API and Twitter API. They used the BERT model to assign sentiment scores and applied a Long Short-Term Memory (LSTM) neural network for multivariate time series forecasting. The LSTM model demonstrated moderate accuracy (MAE of 9.93), capturing general stock price trends. However, limitations, including a relatively small

dataset, suggest potential for improvement through dataset expansion, additional features, and exploring alternative model configurations. Despite discrepancies, the study highlights the potential of integrating social media sentiment analysis with traditional financial data for enhanced stock price prediction, benefiting market participants and investors [2].

2.2 Conclusion

In conclusion, Georgoula et al.'s research underscores the significant impact of Twitter sentiment, Wikipedia search queries, hash rate, USD-euro exchange rate, Bitcoin stock, and the S&P 500 index on Bitcoin prices. The study reveals the short-term positive influence of Twitter sentiment, challenging assumptions about increased supply lowering prices, and highlighting the efficiency of Bitcoin's market adjustments. Meanwhile, Ali Asgarov's study on stock price prediction through historical data and Twitter sentiment analysis showcases the potential of integrating social media analytics with traditional financial data. Despite limitations, such as a small dataset, the research points towards improved accuracy with further exploration of dataset expansion and alternative model configurations, offering valuable insights for market participants and investors.

3
Feasibility

3.1 Technical Feasibility

3.1.1 Software Feasibility

The project will leverage a versatile tech stack to fulfill its objectives, incorporating Python as primary programming language for natural language processing and artificial intelligence related works. Large Language Models (LLMs) such as GPT-4, BERT, RoBERTa, and XLNet will undergo fine-tuning using TensorFlow and PyTorch, while statistical-based AI techniques like LSTM will enhance time-series analysis. Python libraries like Matplotlib and Seaborn will be employed for data visualization. Jupyter Notebooks and Visual Studio Code will be used in the project development.

This comprehensive tech stack aims to foster an efficient, collaborative, and seamlessly integrated development process.

3.1.2 Hardware Feasibility

Considering software development environments, the system requirements required to run and develop the project are shown in Table 3.1:

CPU	8 cores or more
RAM	32 GB or more
Storage	1 TB or more
GPU	A CUDA-compatible GPU (e.g., NVIDIA RTX 3080 or higher)

Table 3.1 System Requirements

The project will use a Linux or Windows operating system. This is because both operating systems are powerful and customizable, and they are well-suited for artificial intelligence applications. Both operating systems are also free and open-source, which makes them cost-effective options.

3.2 Workforce and Time Planning

The workforce and time planning that has been done and planned to be done in our study is given in Figure 3.1.

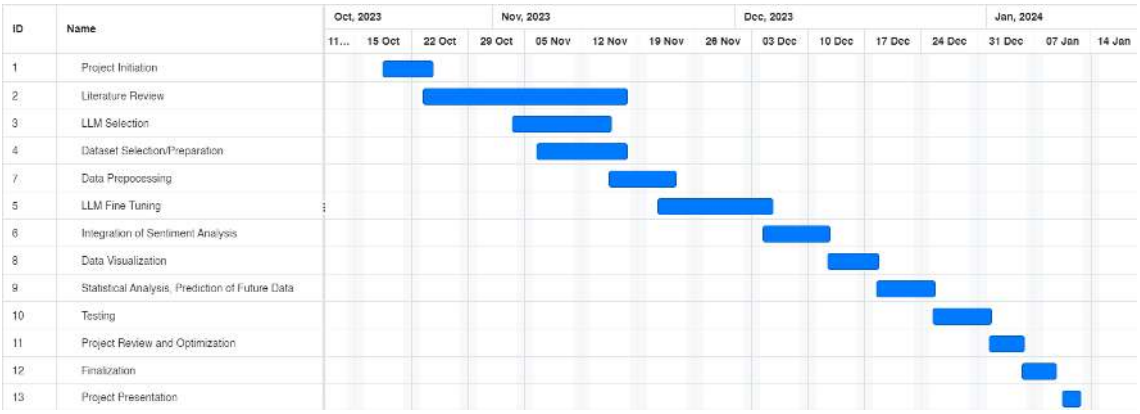

Figure 3.1 Gantt diagram

3.3 Legal Feasibility

The adoption of open-source libraries, including TensorFlow, PyTorch ensures compliance with permissive licenses. Clear terms of use and user agreements will be established to delineate responsibilities and protect both users and the project team. The project is committed to regulatory compliance, tailoring its approach to any industry-specific regulations that may apply. By navigating these legal considerations diligently, the project aims to build trust, prioritize ethical practices, and ensure legal adherence throughout its development and deployment phases.

3.4 Economic Feasibility

Development and testing environments, crucial for the project, will efficiently leverage free or low-cost platforms such as Google Colab, Jupyter Notebooks, or GitHub, eliminating unnecessary software-related expenses. Cloud services for deployment will strategically utilize cost-effective options like Heroku's free tier, resulting in minimal monthly expenditure.

The integration of open-source libraries and frameworks, inherently cost-free, will eradicate licensing fees entirely. Ongoing operational costs, covering hosting fees and maintenance, will be optimized through the judicious use of free-tier services.

Considering these adjustments, the project total cost will be confidently restricted to a nominal figure well within the range of $50-100.

4
System Analysis

The presented use case in Figure 4.1 revolves around a single user action: conducting a comprehensive sentiment analysis on a given dataset and obtaining the corresponding output.

To initiate this analysis, the user triggers the process by selecting a dataset. Preceding the sentiment analysis, pre-trained Large Language Models (LLMs) undergo an initial preprocessing phase. During this preprocessing stage, rows of text which are unsuitable for sentiment analysis are eliminated, ensuring the dataset's suitability.

Subsequently, the LLMs perform a multi-aspect sentiment analysis on remaining rows. Following these analyses, Long Short-Term Memory networks (LSTMs) are employed to conduct forecasting for each sentiment aspect. The forecasted results are then visualized and presented to the user.

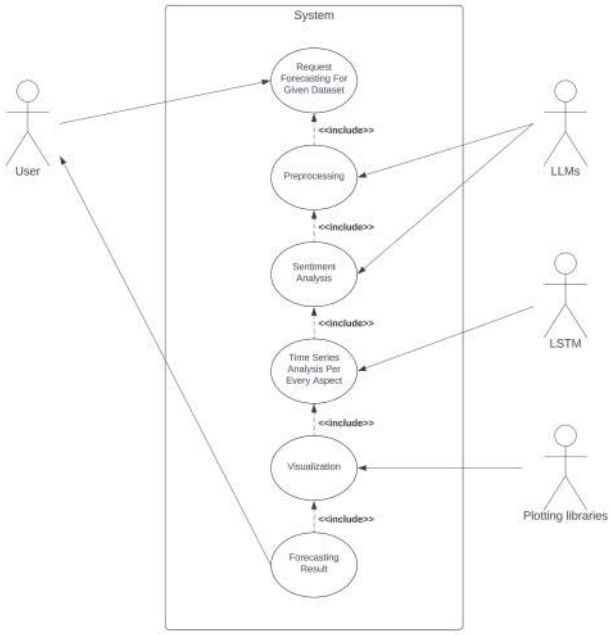

Figure 4.1 Use case diagram

5
System Design

The sequence diagram orchestrates a seamless interaction among pivotal modules, namely User, Interface, LLM (Large Language Models), Forecast Method, and Visualizer, elucidating the methodical progression of tasks. The user initiates the process by interacting with the Interface, supplying crucial inputs such as dataset selection and relevant information. These inputs are then transmitted from the Interface to the LLM, where preprocessing is executed to refine the dataset for sentiment analysis. The LLM, specializing in sentiment analysis, meticulously processes the data. Subsequently, the sentiment analysis results are conveyed from the LLM to the Forecast Method, which specializes in time series analysis and prediction. Upon completing its analysis, the Forecast Method transmits the outcomes to the Visualizer. This crucial module transforms the results into a visual format, enhancing the comprehensibility of sentiment trends. Finally, the Visualizer presents the visualized results to the User through the Interface.

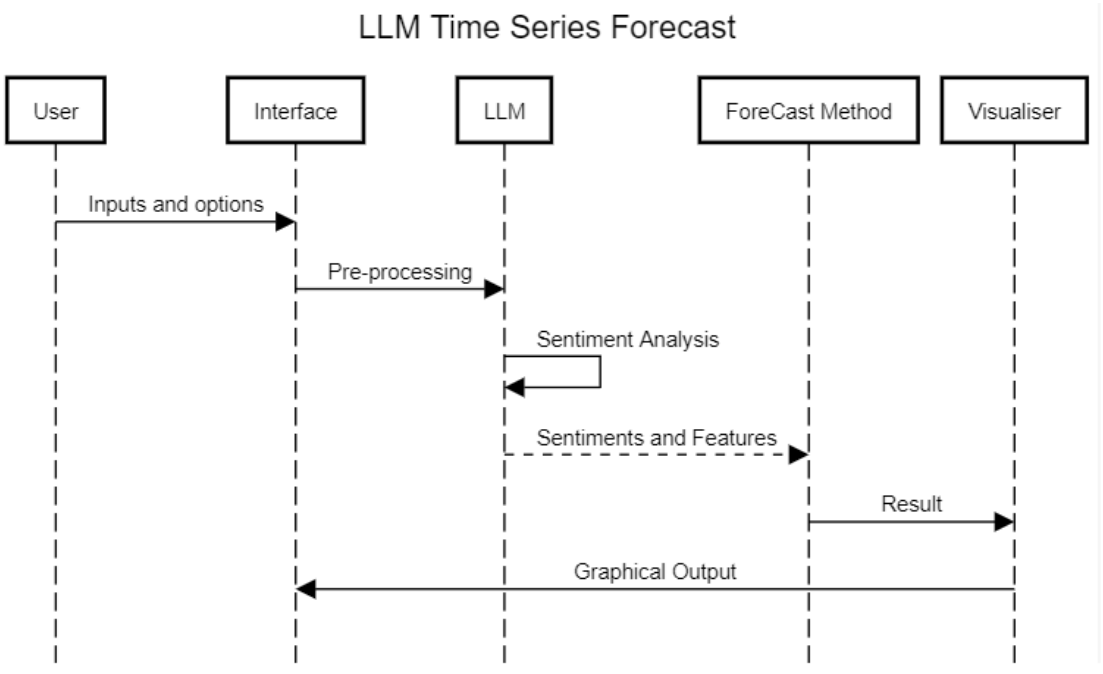

Figure 5.1 Sequence diagram

6
Application

The application interface has been thoughtfully designed to facilitate a user-friendly and intuitive experience throughout the sentiment analysis process. Users initiate the analysis by selecting the dataset of interest, setting the stage for subsequent operations. An added layer of flexibility is introduced through optional preprocessing filters, allowing users to tailor their dataset further, such as defining specific feature ranges.

Once the dataset is configured, users proceed to designate the timeframe for forecasting, specifying the duration (e.g., the number of years) over which sentiment trends will be predicted. With preferences in place, users seamlessly initiate the analysis by clicking the "Start" button. During this phase, the system autonomously executes required processes, including preprocessing and sentiment analysis.

Figure 6.1 Application interface

References

[1] I. Georgoula, D. Pournarakis, C. Bilanakos, D. Sotiropoulos, D. Sotiropoulos, and G. M. Giaglis, “Using time-series and sentiment analysis to detect the determinants of bitcoin prices,” pp. 12–13, 2015.

[2] A. Asgarov, “Ng financial market trends using time series analysis and natural language processing,” pp. 7–8, 2023.