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

January, 2024

© All rights reserved to Yıldız Technical University, Computer Engineering Department.

ACKNOWLEDGEMENTS

We would like to thank Assist. Prof. Dr. Göksel BİRİCİK, who brought this project to us and found us suitable, who helped us using his experience throughout the project.

Yusuf Taha KÖRKEM
Mevlana Halit KAYA

TABLE OF CONTENTS

LIST OF ABBREVIATIONS **v**

LIST OF FIGURES **vi**

LIST OF TABLES **viii**

ABSTRACT **ix**

ÖZET **xi**

1 Introduction **1**

2 Preliminary Examination **2**
2.1 Similar Studies 2
2.2 Conclusion 3

3 Feasibility **4**
3.1 Technical Feasibility 4
3.1.1 Software Feasibility 4
3.1.2 Hardware Feasibility 4
3.2 Workforce and Time Planning 5
3.3 Legal Feasibility 5
3.4 Economic Feasibility 5

4 System Analysis **6**

5 System Design **7**
5.1 Overview 7
5.2 Software Design 8
5.2.1 GPT 3.5 Turbo 8
5.2.2 Dataset 8
5.2.3 Preprocessing The Data 9
5.2.4 Fine Tuning 10
5.2.5 Training Format 11
5.2.6 Prediction 11

5.3 Input-Output Design 12

6 Application **13**

7 Experimental Results **14**
7.1 Fine Tuning 14
7.2 Fine-tuned LLM Sentiment Analysis Results 15
7.2.1 Content 15
7.2.2 UI/UX 17
7.2.3 Bugs/Stability 18
7.2.4 Customer Service 19
7.2.5 Subscription/Payment 21
7.2.6 Downloading/Connection 22
7.2.7 Any 23
7.3 Prediction with LSTM 25

8 Performance Analysis **28**
8.1 Sentiment Analysis Test 28
8.1.1 Concept of Binary Classification 28
8.1.2 Evaluation Metrics 28
8.2 Accuracy Analysis 29
8.2.1 GPT 3.5 Turbo 0613 29
8.2.2 GPT 3.5 Turbo 1106 31
8.3 Prediction Test 33
8.3.1 Test Results of Dissatisfaction Categories 33

9 Conclusion **36**

References **37**

Curriculum Vitae **38**

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

Figure 3.1 Gantt diagram 5
Figure 4.1 Use case diagram 6
Figure 5.1 Sequence diagram 7
Figure 5.2 Dataset 9
Figure 5.3 Reviews Frequency 9
Figure 5.4 LSTM Architecture 11
Figure 6.1 Sentiment analysis interface 13
Figure 6.2 Result filtering and listing interface 13
Figure 7.1 GPT-3.5 Turbo 0613 Learning Curve 14
Figure 7.2 GPT-3.5 Turbo 1106 Learning Curve 15
Figure 7.3 Daily sentiment analysis results for content 16
Figure 7.4 Weekly sentiment analysis results for content 16
Figure 7.5 Monthly sentiment analysis results for content 16
Figure 7.6 Daily sentiment analysis results for UI/UX 17
Figure 7.7 Weekly sentiment analysis results for UI/UX 17
Figure 7.8 Monthly sentiment analysis results for UI/UX 18
Figure 7.9 Daily sentiment analysis results for Bugs/Stability 18
Figure 7.10 Weekly sentiment analysis results for Bugs/Stability 19
Figure 7.11 Monthly sentiment analysis results for Bugs/Stability 19
Figure 7.12 Monthly sentiment analysis results for Customer Service 20
Figure 7.13 Daily sentiment analysis results for Customer Service 20
Figure 7.14 Weekly sentiment analysis results for Customer Service 20
Figure 7.15 Daily sentiment analysis results for Subscription/Payment . . . 21
Figure 7.16 Weekly sentiment analysis results for Subscription/Payment . . 21
Figure 7.17 Monthly sentiment analysis results for Subscription/Payment . 22
Figure 7.18 Daily sentiment analysis results for Downloading/Connection . 22
Figure 7.19 Weekly sentiment analysis results for Downloading/Connection 23
Figure 7.20 Monthly sentiment analysis results for Downloading/Connection 23
Figure 7.21 Daily sentiment analysis results for Any 24
Figure 7.22 Weekly sentiment analysis results for Any 24
Figure 7.23 Monthly sentiment analysis results for Any 24

Figure 7.24 Content basis future prediction results 25
Figure 7.25 UI/UX basis future prediction results 25
Figure 7.26 Bugs/Stability basis future prediction results 26
Figure 7.27 Customer Service basis future prediction results 26
Figure 7.28 Subscription/Payment basis future prediction results 26
Figure 7.29 Downloading/Connection basis future prediction results 27
Figure 7.30 Any category basis future prediction results 27
Figure 8.1 Confusion Matrix of GPT 3.5-0613 30
Figure 8.2 Confusion Matrix of GPT 3.5-1106 32
Figure 8.3 Prediction test results for Content 33
Figure 8.4 Prediction test results for UI/UX 34
Figure 8.5 Prediction test results Bugs/Stability 34
Figure 8.6 Prediction test results for Customer Service 34
Figure 8.7 Prediction test results for Subscription/Payment 35
Figure 8.8 Prediction test results for Download/Connectivity 35
Figure 8.9 Prediction test results for Any 35

LIST OF TABLES

Table 3.1 System Requirements 4
Table 8.1 GPT-0613 Accuracy Table 29
Table 8.2 GPT-0613 Classification Report 29
Table 8.3 GPT-1106 Accuracy Table 31
Table 8.4 GPT-1106 Classification Report 31

ABSTRACT

TIME SERIES SENTIMENT ANALYSIS WITH LLMs

Yusuf Taha KÖRKEM
Mevlana Halit KAYA

Department of Computer Engineering
Senior Project

Advisor: Assist. Prof. Dr. Göksel BİRİCİK

The project's multi-aspect sentiment analysis on time series data and precise future prediction skills allowed it to successfully accomplish the intended objectives. By utilizing an intricate technology stack including of Python, LSTM, OpenAI's GPT-3.5 Turbo, and visualization tools, the system showcased its capacity to execute complex time series analysis and natural language processing tasks.

High accuracy was demonstrated by the sentiment analysis component, which was made possible by optimizing GPT-3.5 Turbo. Following testing, two models—the GPT-3.5 Turbo 0613 and the GPT-3.5 Turbo 1106—were shown to be superior in terms of accuracy, precision, recall, F1 score, and Hamming loss. In order to guarantee reliable sentiment analysis, the project demonstrated a painstaking process of training and evaluating these models.

Dissatisfaction in the following areas is covered by multi-aspect sentiment analysis: content, UI/UX, bugs/stability, customer service, subscription/payment, downloading/connection. Only these particular categories were employed to identify dissatisfaction. Because, based on the majority of evaluations, they suggest the most common complaint topics.

The project's success was further highlighted by the prediction testing, which was carried out utilizing LSTM for time series forecasting. Mean Squared Error (MSE), one of the evaluation criteria, showed how well the model predicted future values based on historical data. The graphs showing the 20-day forecast gave users a visual

understanding of the system's prediction power.

In conclusion, the project demonstrated its ability to produce precise future predictions and sophisticated sentiment analysis. The system's ability to achieve its goals was confirmed by careful model training, extensive testing, and the application of state-of-the-art technology. For complex language processing and prediction applications, the improved GPT-3.5 Turbo and LSTM-based time series forecasting proven to be a potent combo.

Keywords: LLM, GPT 3.5 Turbo, time series sentiment analysis, LSTM, neural networks, future prediction, multi-aspect sentiment analysis

ÖZET

X

Yusuf Taha KÖRKEM
Mevlana Halit KAYA

Bilgisayar Mühendisliği Bölümü
Bitirme Projesi

Danışman: Dr. Ögr. Üyesi Göksel BİRİCİK

Proje, zaman serisi verileri üzerinde çok yönlü duygu analizi ve isabetli gelecek tahmini yapabilmesi sayesinde istenen sonuçları başarıyla elde etmiştir. Python, OpenAI'nin GPT-3.5 Turbo LLM'si, LSTM ve görselleştirme araçlarını içeren komplike bir altyapıdan oluşan sistem, gelişmiş doğal dil işleme ve zaman serisi analizi gerçekleştirme yeteneğini göstermiştir.

GPT-3.5 Turbo'ya ince ayar yapılarak efektif ve hızlı hale getirilen duygu analizi süreci yüksek doğruluk sergilemiştir. GPT-3.5 Turbo 0613 ve GPT-3.5 Turbo 1106 olmak üzere iki model test edilmiş ve ikincisinin doğruluk, kesinlik, geri çağırma, F1 puanı ve Hamming kaybı açısından daha yetenekli olduğu kanıtlanmıştır. Projede, iyi bir duygu analizi sağlamak için bu modelleri eğitmek ve doğrulamak için titiz bir süreç izlenmiştir.

Çok yönlü duygu analizi şu kategorilerde memnuniyetsizlik olup olmadığını kapsar: İçerik, UI/UX, Hatalar/stabilite, Müşteri hizmetleri, Abonelik/ödeme, İndirme/bağlantı. Memnuniyetsizlik tespiti için sadece bu spesifik kategoriler kullanılmıştır. Bu ise incelemelerin çoğu dikkate alındığında en sık karşılaşılan şikayet konularının bunlar olduğu kanaatine varıldığından böyle yapılmıştır.

Zaman serisi tahmini için LSTM kullanılarak uygulanan tahmin testi, projenin başarısını daha da vurgulamıştır. Ortalama Karesel Hata (MSE) dahil olmak üzere değerlendirme metrikleri, modelin geçmiş verilere dayanarak gelecekteki değerleri tahmin etmedeki başarsını göstermiştir. 20 günlük tahmin grafikleri, sistemin tahmin

yetenekleri hakkında görsel bilgiler sağlamıştır.

Özetleyecek olursak; proje, çok yönlü duygu analizi ve doğru gelecek tahminleri elde etme konusunda yeterliliğini ortaya koymuştur. En son teknolojilerin kullanımı, özenli model eğitimi ve kapsamlı testler, sistemin hedeflerine ulaşmadaki başarısını doğrulamıştır. İnce ayarlı GPT-3.5 Turbo ve LSTM tabanlı zaman serisi tahmininin, gelişmiş doğal dil işleme ve gelecek tahmini için sağlam bir kombinasyon olduğu kanıtlanmıştır.

Anahtar Kelimeler: LLM, GPT 3.5 Turbo, zaman serilerinde duygu analizi, LSTM, sinir ağları, gelecek tahmini, çok yönlü duygu analizi

1
Introduction

People have never had greater chance to share their opinions and feelings about a wide range of products, services, and occasions than they do in this age of ubiquitous connectedness and unfettered communication. This wealth of emotional data, often disseminated through social media and online review boards, can provide a wealth of illuminating information when analyzed over time.

Large language models (LLMs) are used in this project to investigate the field of time-series sentiment analysis and find patterns in the dynamic fabric of human emotions. From this vantage point, we examine many facets of a certain topic or product, meticulously recording changes in mindset throughout time. By identifying these profound emotional trajectories, we may go into the realm of predictive analytics and uncover the underlying dynamics of public opinion, forecasting future trends and potential shifts in attitudes.

In essence, the project combines technological innovation with the complexity of human feeling in order to use large language models (LLMs) to shed light on the constantly shifting public mood environment. Our ultimate goal is to provide invaluable insights that, through evaluating sentiment dynamics over time and proactively projecting future trends, not only inform strategic business decisions but also assist in the construction of successful policies. By doing this, we seek to advance understanding of the overall emotional climate and offer a practical tool for navigating the challenging terrain of shifting public opinion.

2
Preliminary Examination

2.1 Similar Studies

Similar studies that have been done in the area where the project plans to operate have been reviewed in this section.

Georgoula et al.'s study examines the short- and long-term factors that affect Bitcoin values, accounting for a variety of technological and economic indicators, such as Twitter sentiment. We apply a machine learning technique to assess the attitude of common Twitter users toward Bitcoin. The study found a favorable short-term correlation between Twitter mood and Bitcoin values, indicating the potential value of sentiment analysis in price prediction. Additionally, an increase in hash rate and Wikipedia search inquiries positively impacts the value of Bitcoin, suggesting that mining difficulty and public awareness have an impact on the cryptocurrency's value. However, there is a negative relationship between the USD/EUR exchange rate and the price of Bitcoin.Unlike the common assumption that decreased pricing would follow from increased supply, long-term research indicates that the price of Bitcoin shares is positively impacted.However, Bitcoin values are negatively impacted by the S&P 500 index, suggesting that investors view them as alternatives. The efficient market hypothesis demonstrates how swiftly Bitcoin prices return to their long-term equilibrium. A larger dataset, alternative sentiment indices, and the use of vector autoregressive (VAR) models are just a few of the enhancement strategies suggested in the study. [1].

As per the article cited, Ali Asgarov's research delves into the creation of a stock price prediction model through the amalgamation of historical data and emotion ratings obtained from Twitter. The researchers collected financial data and related tweets about well-known companies like Apple and Tesla using the Twitter and Yahoo Finance APIs. They used the BERT model for sentiment scoring and a Long Short-Term Memory (LSTM) neural network for multivariate time series forecasting. The LSTM model demonstrated respectable accuracy in recognizing broad trends in

stock prices, with an MAE of 9.93. Even with some limitations, like a relatively small dataset, there is still potential for improvement through the addition of more variables, the enlargement of the dataset, and the exploration of alternative model configurations. Notwithstanding these discrepancies, the study demonstrates how stock price forecasting can be enhanced and investors and market participants can benefit by combining social media sentiment analysis with traditional financial data. [2].

2.2 Conclusion

In summary, the research by Georgoula et al. emphasizes how significant a factor emotion on Twitter, Wikipedia search queries, hash rate, USD-euro exchange rate, Bitcoin stock, and the S&P 500 index are in determining the price of Bitcoin. The analysis demonstrates the effectiveness of Bitcoin's market adjustments and refutes the idea that an increase in supply will lead to a reduction in price. It also demonstrates Twitter sentiment's beneficial short-term effects. Meanwhile, Ali Asgarov's study on stock price prediction using historical data and Twitter sentiment analysis illustrates the possibility for combining social media analytics with traditional financial data. Despite certain limitations, such as a small dataset, the research indicates that more research into larger datasets and different model configurations will improve accuracy and give investors and market participants useful information.

3
Feasibility

3.1 Technical Feasibility

3.1.1 Software Feasibility

The project will employ a flexible technology stack to meet its objectives, with Python acting as the primary programming language for activities related to natural language processing and artificial intelligence. OpenAI's GPT-3.5 Turbo will be improved as a Large Language Model (LLM) with the use of its API, and time-series analysis will be enhanced by statistically-based AI techniques like LSTM. The Python graphical package Matplotlib will be used for data visualization. Google Colab and Visual Studio Code will be used during the project development process.

The goal of this all-inclusive tech stack is to promote a productive, cooperative, and smoothly integrated development process.

3.1.2 Hardware Feasibility

Table displays the system requirements needed to execute and develop the project, taking software development environments into consideration. 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 employ Linux or Windows as its operating system. This is because both operating systems are robust, flexible, and perfect for artificial intelligence applications. Because these operating systems are free and open-source, they are also reasonably priced 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.

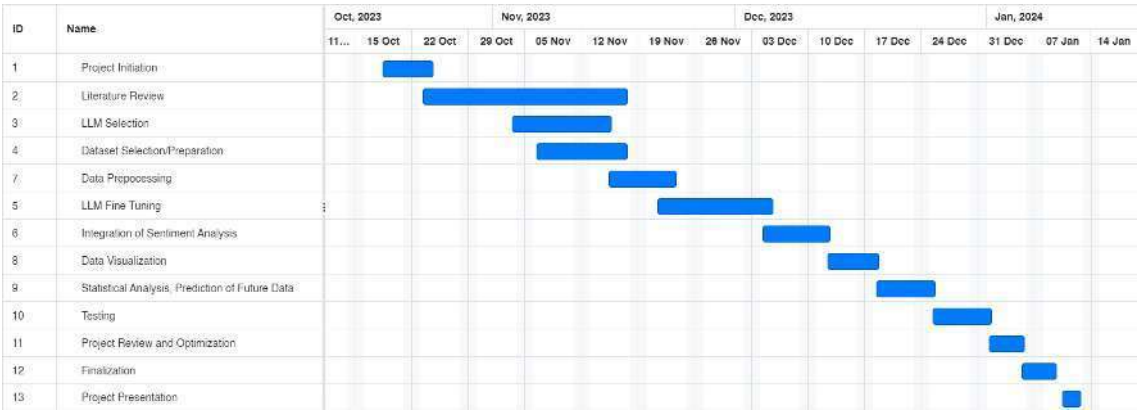

Figure 3.1 Gantt diagram

3.3 Legal Feasibility

Using open-source libraries like Pandas and Keras ensures adherence to permissive licensing. Explicit terms of use and user agreements will be developed in order to specify roles, protect users, and protect the project team. The initiative is committed to following regulations, and it will adjust its plan to abide with any legislation that may be relevant to the business. The project aims to build trust, prioritize ethical behavior, and ensure legal compliance throughout the planning and implementation phases by carefully navigating these legal factors.

3.4 Economic Feasibility

Critical project settings such as development and testing will make good use of low-cost or free platforms like Google Colab or GitHub. This will save monthly expenses and cut down on unnecessary software-related charges.

Using free and open-source frameworks and libraries will reduce the cost of licensing. The economical use of free-tier services would maximize continuous running expenses, such as maintenance and hosting charges.

The primary cost of the project is represented by OpenAI billings, which are the result of fine-tuning services through its API. This fee is determined by the utilization of the API, which is associated with token counts[3]. The project's scope may restrict this cost to a range of roughly $50 to $100.

4
System Analysis

Figure 4.1 presents a use case that centers around a single user action: conducting a comprehensive sentiment analysis on a specific dataset and obtaining pertinent outcomes.

To begin the analysis process, the user selects a dataset. Prior to sentiment analysis, the updated Large Language Models (LLMs) undergo preliminary preprocessing. By eliminating unsuitable rows, this preprocessing step guarantees that the dataset is appropriate for sentiment analysis.

The LLMs then perform a multi-aspect sentiment analysis on the remaining rows. Following these investigations, Long Short-Term Memory networks (LSTMs) are used to forecast each sentiment feature. Next, a depiction of the expected results is presented to the user.

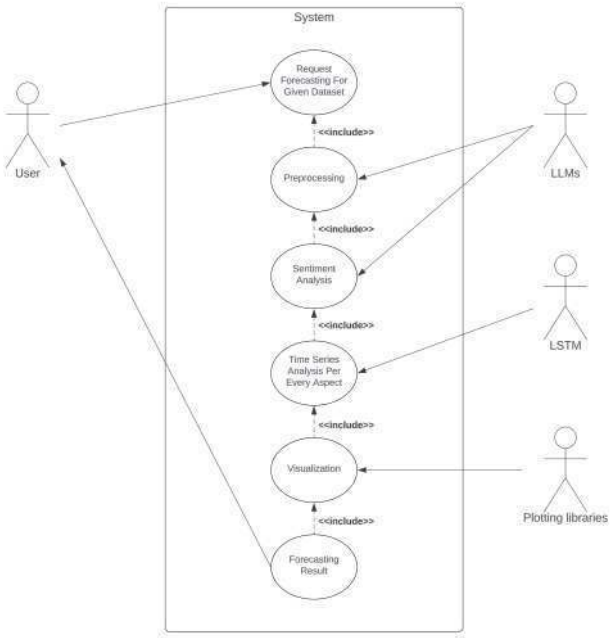

Figure 4.1 Use case diagram

5
System Design

5.1 Overview

By coordinating a seamless connection between crucial modules—User, Interface, LLM (Large Language Models), Forecast Method, and Visualizer—the sequence diagram clarifies the systematic flow of tasks. The user initiates the process by interacting with the interface and giving necessary inputs, such as relevant data and dataset selection. The LLM then preprocesses the dataset in order to get it ready for sentiment analysis, receiving these inputs from the Interface. The Forecast Method, a time series analysis and prediction expert, receives the sentiment analysis results from the LLM after that. Upon completion of its analysis, the Forecast Method passes its findings to the Visualizer. This crucial module transforms the results into a legible visual format that facilitates comprehension of sentiment trends. Finally, the Visualizer shows the User the visualized findings via the Interface.

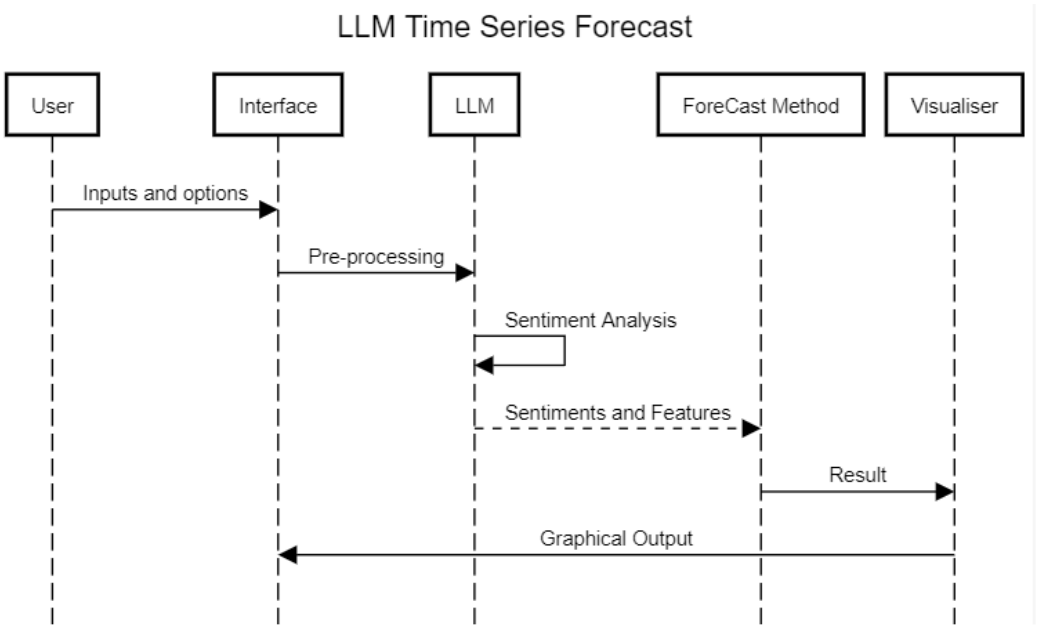

Figure 5.1 Sequence diagram

5.2 Software Design

The following Python libraries were used in this project: Pandas for effective data manipulation; **Plotly** and **Matplotlib** for interactive data visualization; **Keras** for neural network-based predictive analytics, specifically for LSTM; and **OpenAI** for advanced natural language processing with its **GPT 3.5 Turbo** LLM and its Fine Tuning API. The software design is modular, with distinct modules for data processing, natural language processing, visualization, and predictive analytics. Each module uses certain libraries to perform its role, resulting in a cohesive and scalable design.

5.2.1 GPT 3.5 Turbo

GPT-3.5 Turbo is an advanced language model developed by OpenAI that was pre-trained on a range of linguistic inputs. Its adaptability, which allows adaptation to specific tasks like question answering or sentiment analysis, is its outstanding quality. A task-specific dataset is used to train the model initially, and previously learnt weights are then used to fine-tune it. Performance is evaluated on a validation set using this iterative process until the intended results are achieved.

5.2.2 Dataset

In this project, time series sentiment analysis is applied to reviews of Netflix on the Google Play Store[4]. since both of these evaluations include timestamp values and have a suitable structure for multi-aspect sentiment analysis. This collection actually includes 1.5 million reviews from 2011 to 2023. We took approximately the last 250,000 reviews out of it for our analysis. In Figure 5.2, it is depicted.

	Unnamed: 0	review_id	pseudo_author_id	author_name	review_text	review_rating	review_likes	author_app_version	review_timestamp
0	1186307	18fc539a-0689-4577-b43e-5c222d4ccf08	286334437144453538609	VA**********AD	Superb Application	5	0	7.94.0 build 8 35372	2021-03-22 14:19:22
1	1186308	9900e631-53fd-4c03-b1a0-e560ba4007ff	404047194685663652375	Ma**********na	Me encanta esta app lo mejor	5	0	6.13.0 build 29940	2021-03-22 14:21:42
2	1186309	796a156f-9e97-431d-8783-fce427757167	212258163415595624141	Mo**********an	Very good	5	0	NaN	2021-03-22 14:22:03
3	1186310	fa658f14-fc82-4259-a20c-15bff84f4374	257232535785602248880	je********rs	Best ever	5	0	7.76.1 build 9 35139	2021-03-22 14:22:44
4	1186311	8b31bdbb-0fec-4551-8021-1d444a9885cc	337128270308445015308	Du*******bu	„ 0	5	0	NaN	2021-03-22 14:30:40
...	...	...	...	...	...	...	...	...	...
344782	1531121	5b819b4a-f49f-4012-b1cc-146b581aec6e	517084783367708002209	Az*********er	Bad app	1	0	NaN	2023-11-15 22:34:37
344783	1531122	afe340b9-68df-4df9-8a86-7e9304e1e271	217585066694826156159	Ma**********ey	What more do you want from me tf? BRING BACK P...	2	0	8.94.0 build 10 50546	2023-11-15 22:44:59
344784	1531123	3015ab73-75e8-4f17-8377-4757abbb8f0c	268385941811343301666	Em**********ey	I will love this app	5	0	NaN	2023-11-15 22:45:05
344785	1531124	25b4b68e-a432-4f21-bf1c-68835f88b56e	259993922622854778058	****	The content is great but they keep adding more...	2	0	8.94.0 build 10 50546	2023-11-15 22:48:54
344786	1531125	7c8e755c-0505-4e0c-ac89-719f8f5fff50	527925027675131929486	Ra**********on	I promise you on this, I WILL NOT RESUBSCRIBE ...	1	1	8.94.0 build 10 50546	2023-11-15 22:54:42

339638 rows × 9 columns

Figure 5.2 Dataset

Also, reviews frequency over timestamp is shown in Figure 5.3.

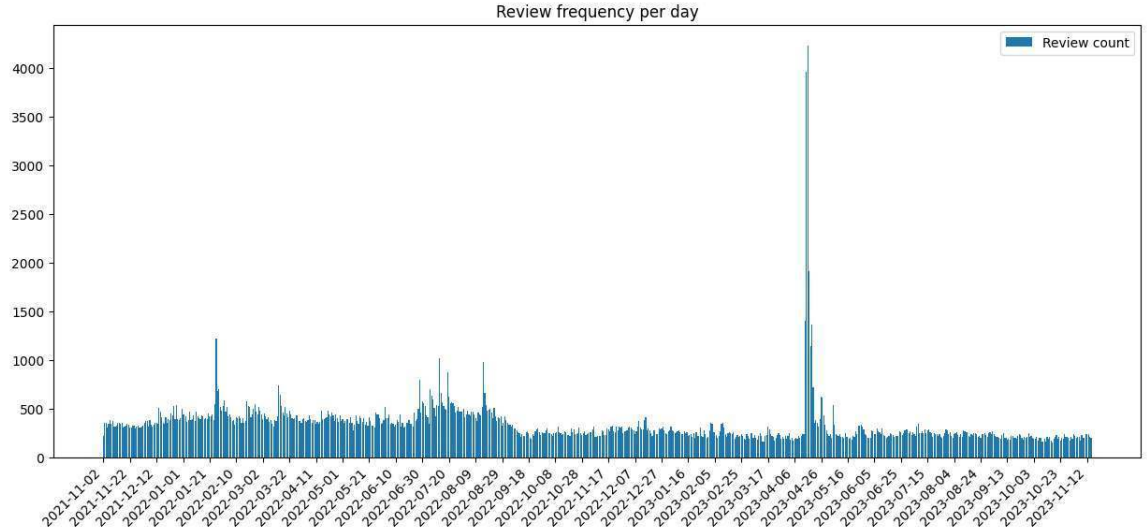

Figure 5.3 Reviews Frequency

5.2.3 Preprocessing The Data

Every review within the dataset undergoes preprocessing procedures. Our main preprocessing goal is to eliminate duplicate and superfluous tokens before analysis. Reviews that are incomplete are eliminated from the dataset after preprocessing.

These preprocessing steps are followings:

- The review has been updated with no emojis. Emojis don't convey any significance regarding the aspects of our discontent.
- A single character is used in place of repeated ones. Since the OpenAI API prohibits the usage of repetitive patterns, this results in a bad request exception. Furthermore, repetitive characters have no significance for us.
- Ultimately, this review will be removed if it contains only blank text.

5.2.4 Fine Tuning

We used a collection of data to train our LLM, enabling it to understand the relationship between the input and output for our sentiment analysis. A few discontent categories serve as the foundation for our sentiment analysis. These categories are designed with the majority of the dissatisfaction-related subjects in the evaluations in mind.

The categories are followings:

- **Content:**Complaints regarding available content, visibility, selection, deletion, removal, difficulty in finding, recommendation, leakage of new content, location-based content issues, subtitles, localization, quality adjustment, or poor video or audio quality may fall under this category.
- **UI/UX:** Customization, navigation, user interface, user experience, and app usage difficulties are some of the topics that may be covered in this section.
- **Bugs/stability:** Bug reports, screen brightness complaints (high or low), black screen issues, and other technical issues may fall under this category.
- **Customer service:** Complaints pertaining to a range of customer service-related issues may fall under this category.
- **Subscription/payment:** This topic may include complaints about pricing policies, rising subscription fees, overpriced services, invoicing problems, payment problems, cancellation of subscriptions, or problems with a monthly subscription cost.
- **Downloading/connection:** This topic may include complaints about proxy errors, login issues, connectivity issues, casting issues, auto-download capabilities, and downloading issues.

5.2.5 Training Format

Our training data include user inputs and their completions. Reviews are considered to be user-provided content, whereas completions consist of a variety of related dissatisfaction categories. The following is an example of a training format:

Training Formatting
Prompt: Return the categories of discontent in each review in JSON format. Content, UI/UX, bugs/stability, customer service, subscription/payment, downloading/connection, and customer service are the categories where one of them must be present. If there is no issue with these categories, do not return any results for a review. A review may simultaneously pertain to more than one category. An object called reviews is handed to you; its value is review, and its key is id. The object you must return has a string array as its value and an id as its key. It is important to assess each review separately from the others.
User Input: {{Review id}}: "{{A review text given by user}}"
Assistant Response: {{Review id}}: {{Dissatisfaction categories array of the review}}

5.2.6 Prediction

For future prediction, we used Long Short-Term Memory (LSTM), an efficient deep learning-based architecture. The LSTM neural network design is suitable for time series forecasting applications like future prediction because it can handle long-term dependencies, sequential data processing, and effective training on small datasets. Modeling patterns in sequential data is a great use case for the long short-term memory (LSTM) cell structure, which makes it easier to capture and store data over a broad range of time steps. Originally designed to outperform classical RNNs in vanishing gradient issues, it has evolved into a powerful tool for time series data prediction. LSTM architecture is presented in Figure 5.4

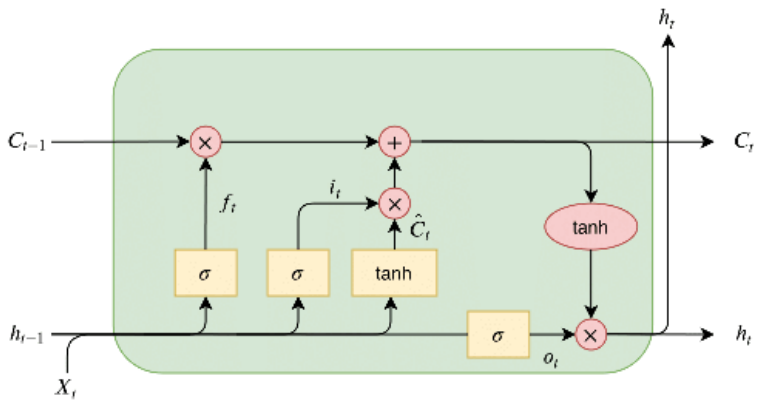

Figure 5.4 LSTM Architecture

5.3 Input-Output Design

The input and output of the system is handled by Google Colab forms. There are two interfaces as follows:

1. An interface for sentiment analysis from review: The user enters id and text of a review into inputs. After that, the user selects which LLM model to use in the sentiment analysis. Inputs filled by the user are generated to input format as mentioned in Section 5.2.5. In the same way, output is given from LLM and is shown to the user. The interface can be seen in Figure 6.1.

2. An interface for filtering and listing results: The user enters two dates for the filter range and may select suitable filter type as can be seen in Figure 6.2.

6
Application

Figure 6.1 Sentiment analysis interface

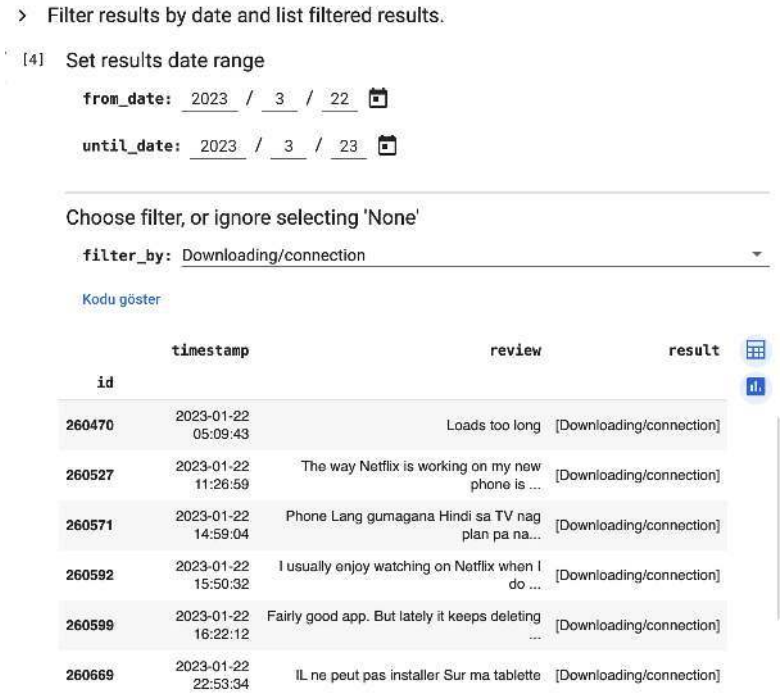

Figure 6.2 Result filtering and listing interface

7
Experimental Results

7.1 Fine Tuning

We used two variant of GPT-3.5 Turbo: GPT-3.5 Turbo 0613 and GPT-3.5 Turbo 1106 for fine tuning. After fine tuning, we used both of them to generate sentiment analysis results. Result of testing both fine-tuned models, we decided that GPT-3.5 Turbo 1106 is more capable and suitable for our job.

Training and validation loss curves for both fine tuning job are given in Figure 7.1 and Figure 7.2.

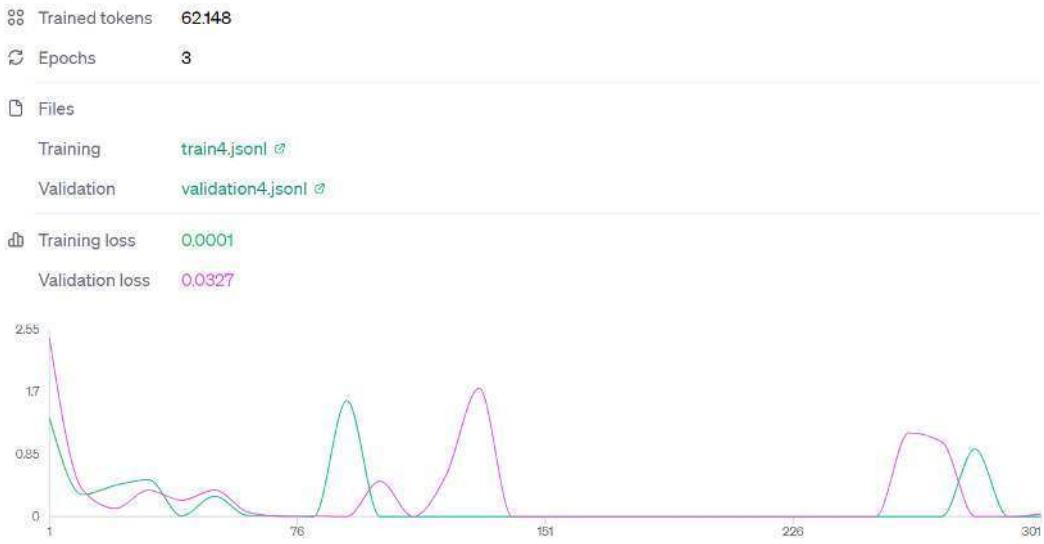

Figure 7.1 GPT-3.5 Turbo 0613 Learning Curve

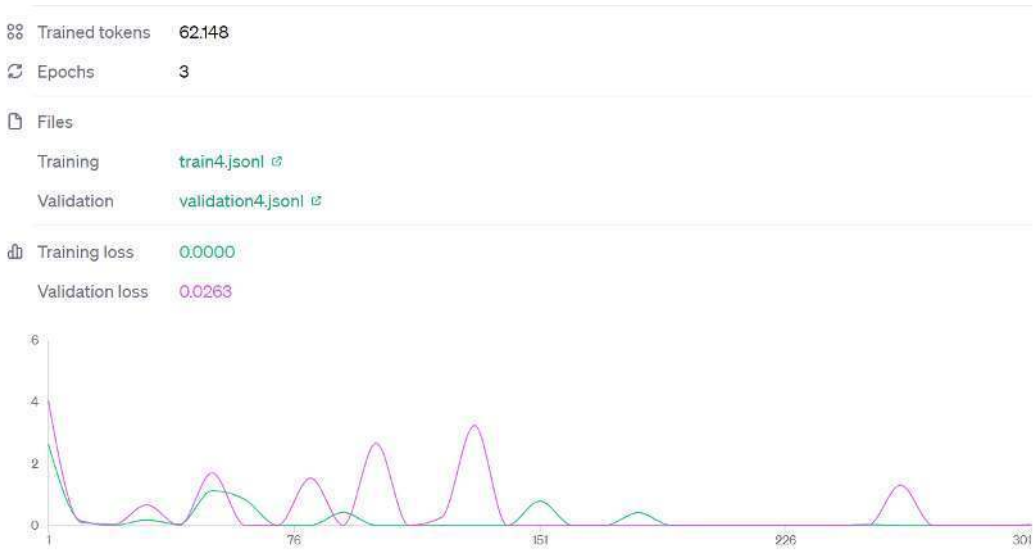

Figure 7.2 GPT-3.5 Turbo 1106 Learning Curve

As mentioned above, we have selected GPT-3.5 Turbo 1106 because of its capability considering its testing results. Testing results of both fine-tuned models will be given in Performance Analysis section.

7.2 Fine-tuned LLM Sentiment Analysis Results

Our results obtained based on separate experiments we conducted for each topic. All reviews examined by fine-tuned model and then clustered as it needed. All clusters handled one by one themselves and one final value obtained. Every topic handled in daily, weekly and monthly basis. All outputs are given below seperate.

7.2.1 Content

Below graphs are related to content availability, removal, recommandation or similar issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.3, 7.4 and 7.5.

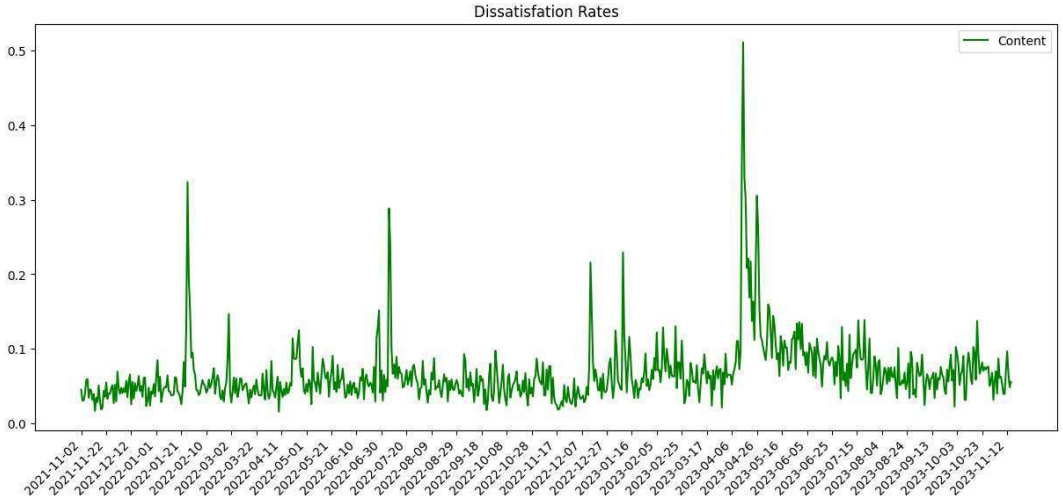

Figure 7.3 Daily sentiment analysis results for content

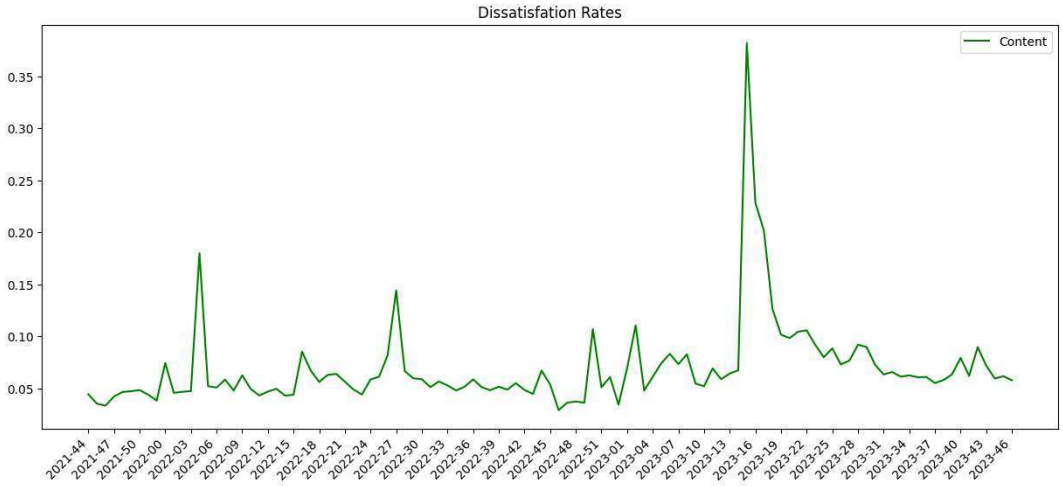

Figure 7.4 Weekly sentiment analysis results for content

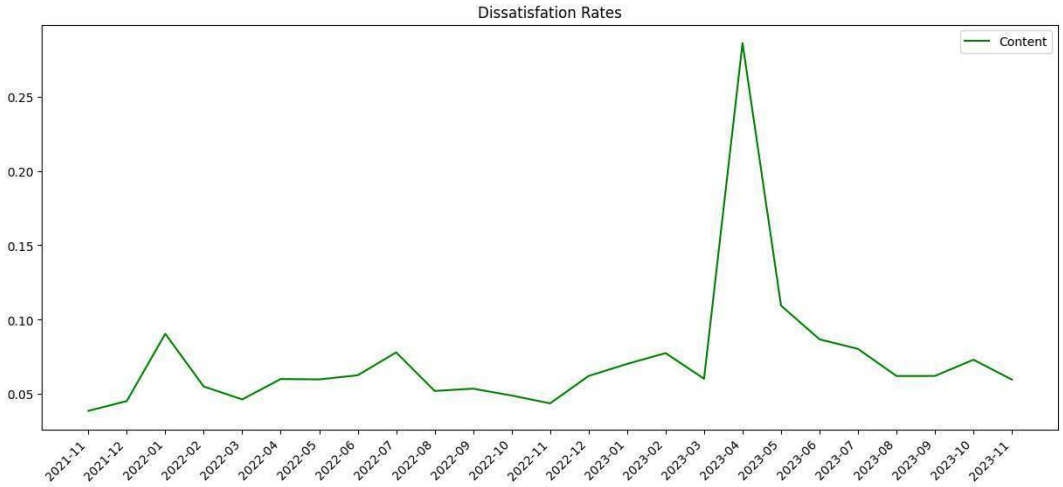

Figure 7.5 Monthly sentiment analysis results for content

7.2.2 UI/UX

Below graphs are related to personalization, navigation, user interface, user experience, or challenges with using an app issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.6, 7.7 and 7.8.

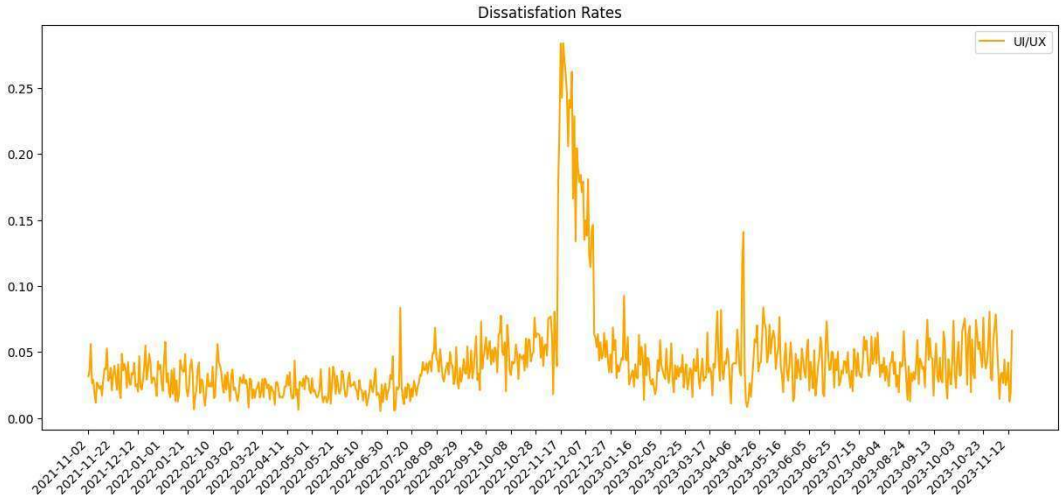

Figure 7.6 Daily sentiment analysis results for UI/UX

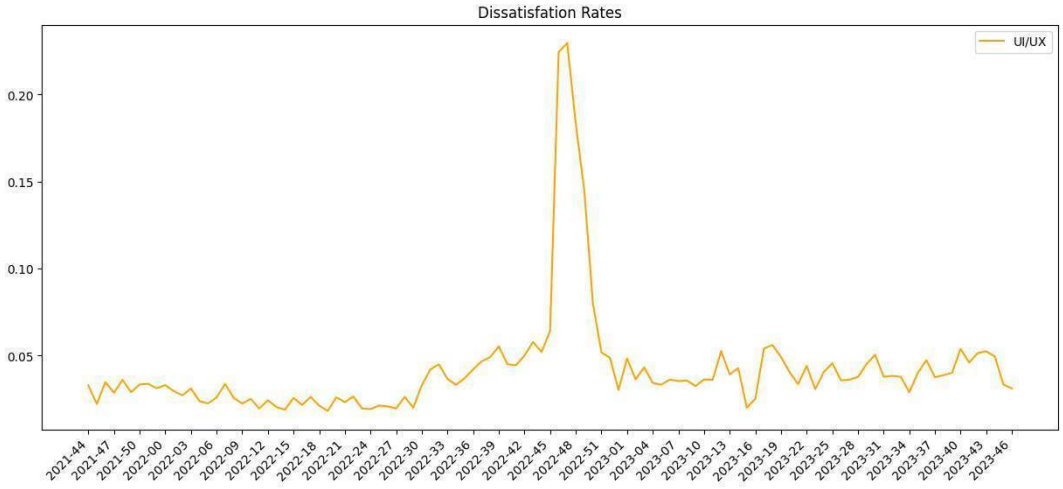

Figure 7.7 Weekly sentiment analysis results for UI/UX

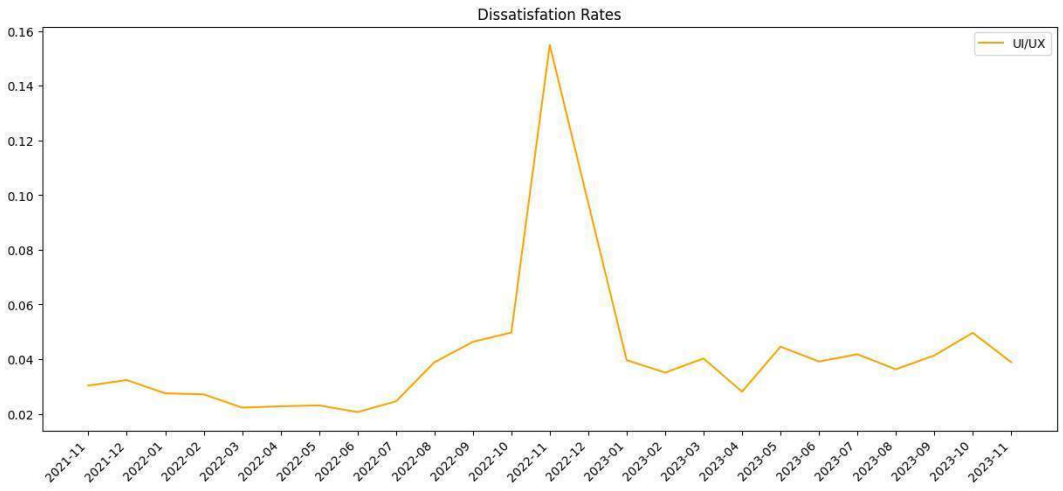

Figure 7.8 Monthly sentiment analysis results for UI/UX

7.2.3 Bugs/Stability

Below graphs are related to application bugs, black screen, brightness and other technical issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.9, 7.10 and 7.11.

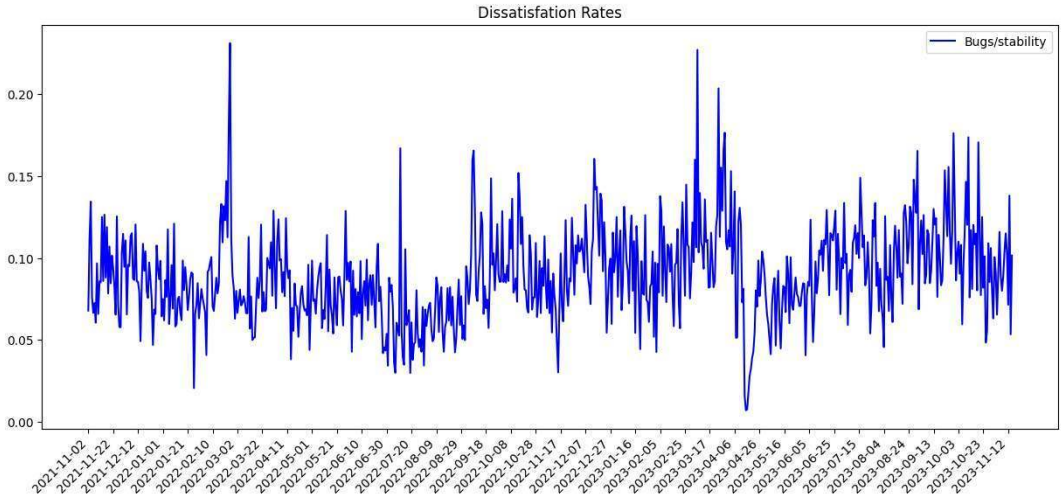

Figure 7.9 Daily sentiment analysis results for Bugs/Stability

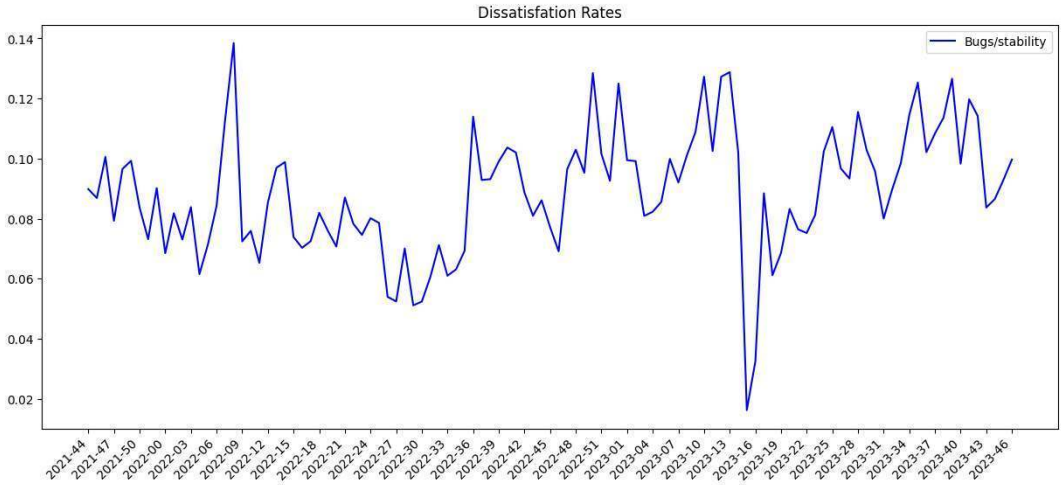

Figure 7.10 Weekly sentiment analysis results for Bugs/Stability

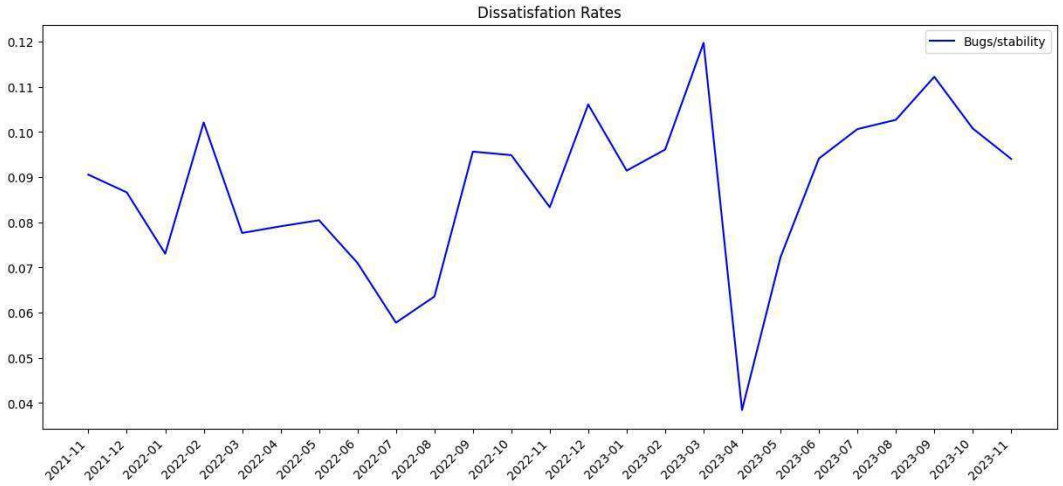

Figure 7.11 Monthly sentiment analysis results for Bugs/Stability

7.2.4 Customer Service

Below graphs are related to variety of customer service issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.12, 7.13 and 7.14.

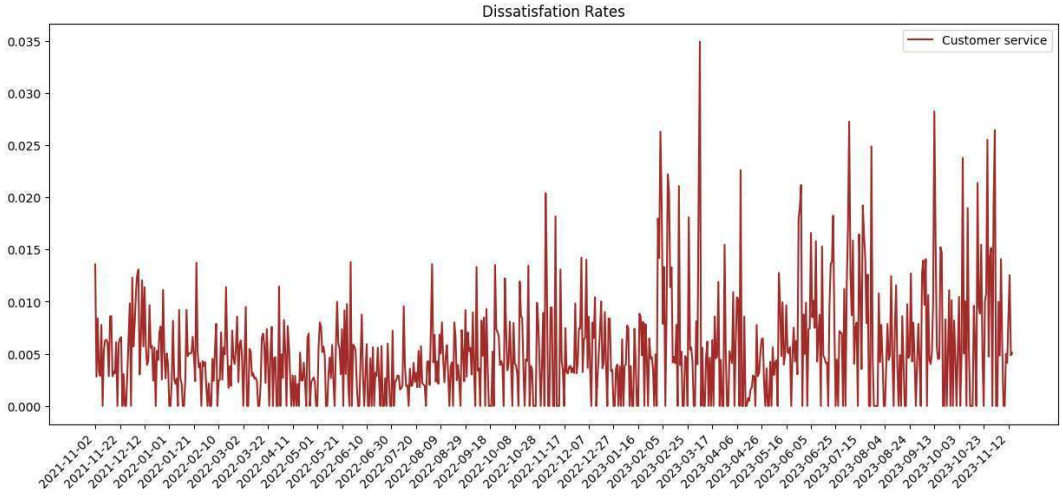

Figure 7.12 Monthly sentiment analysis results for Customer Service

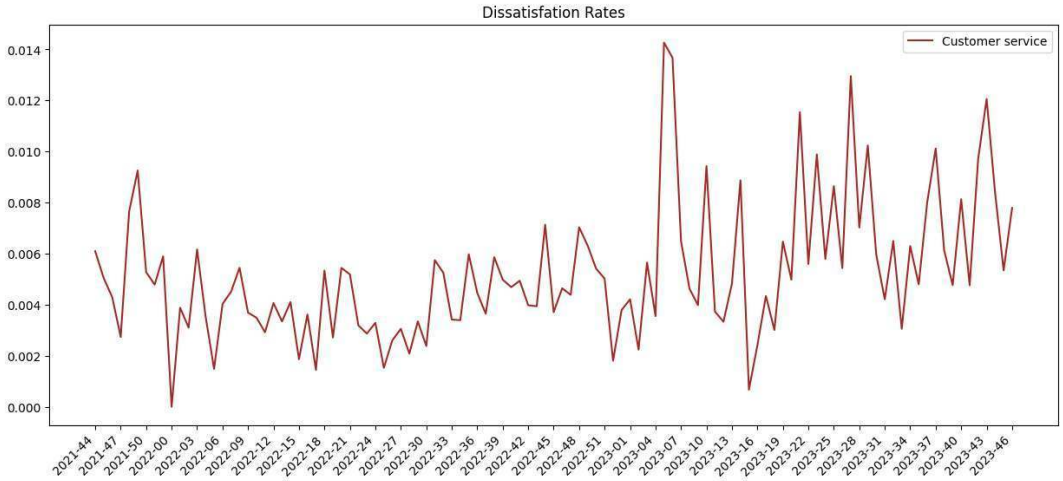

Figure 7.13 Daily sentiment analysis results for Customer Service

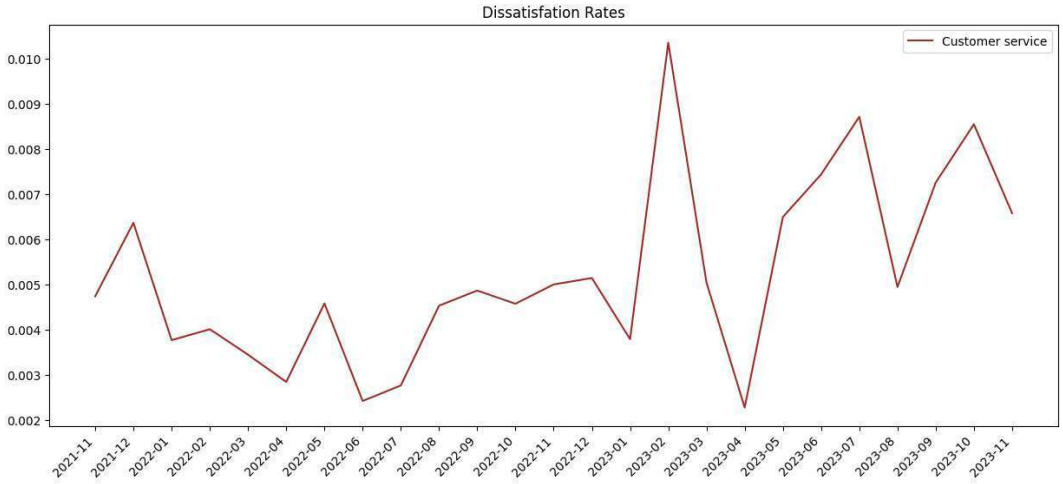

Figure 7.14 Weekly sentiment analysis results for Customer Service

7.2.5 Subscription/Payment

Below graphs are related to subscription, pricing or payment or similar issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.15, 7.16 and 7.17.

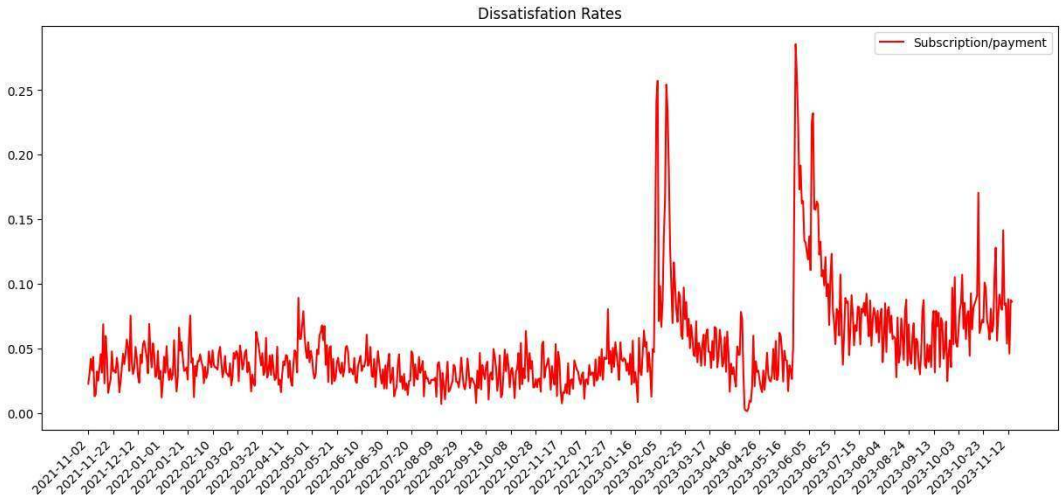

Figure 7.15 Daily sentiment analysis results for Subscription/Payment

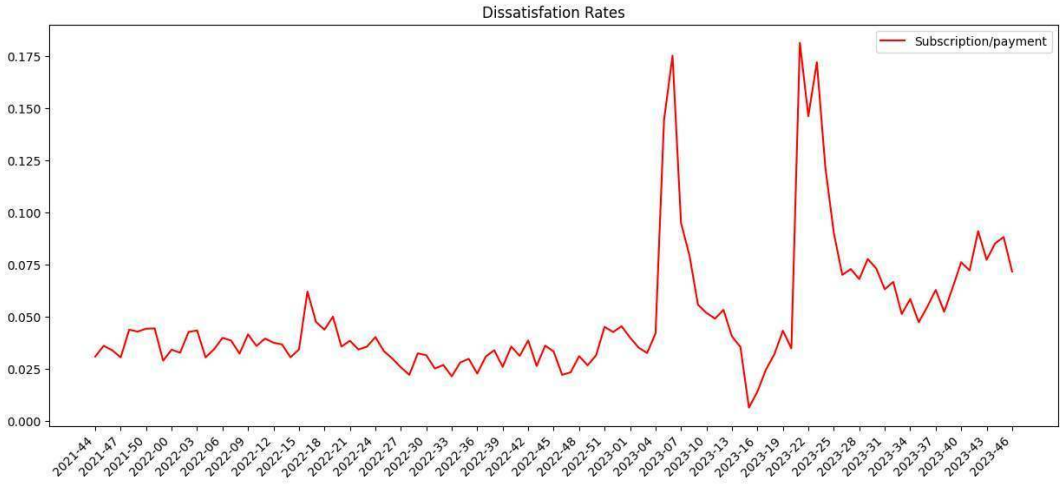

Figure 7.16 Weekly sentiment analysis results for Subscription/Payment

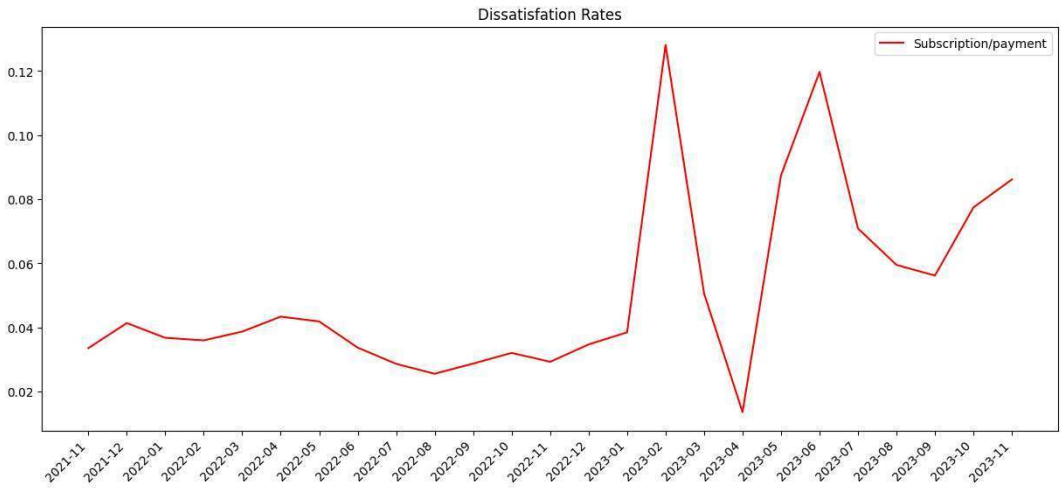

Figure 7.17 Monthly sentiment analysis results for Subscription/Payment

7.2.6 Downloading/Connection

Below graphs are related to downloading, login or connectivity issues. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.18, 7.19 and 7.20.

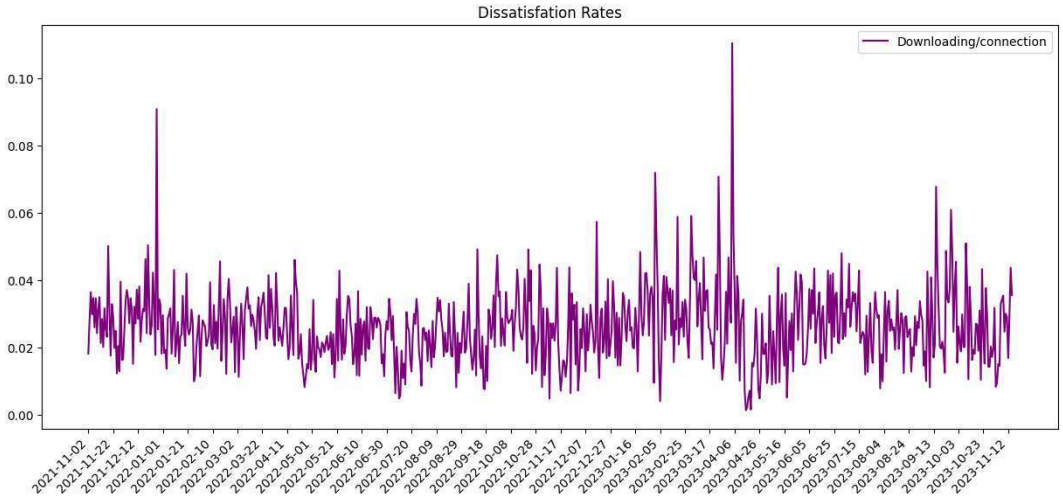

Figure 7.18 Daily sentiment analysis results for Downloading/Connection

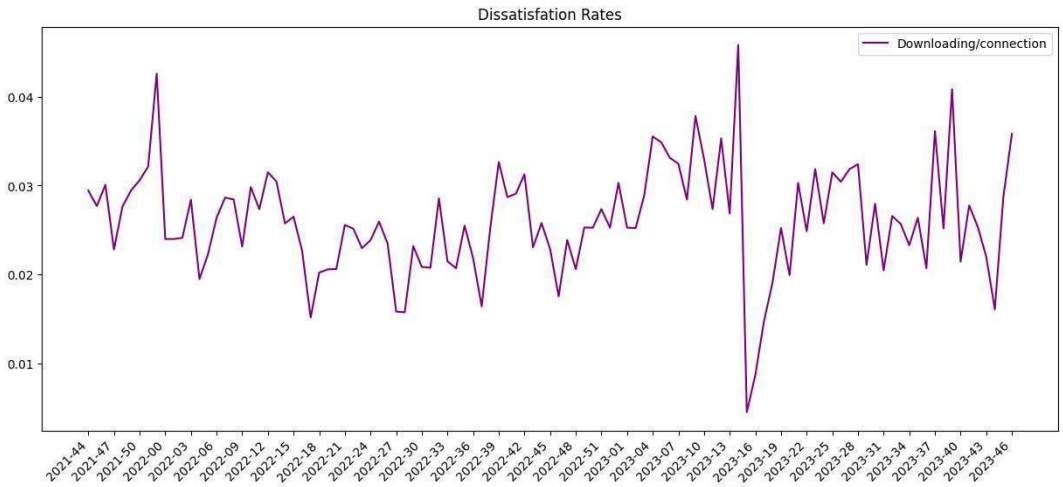

Figure 7.19 Weekly sentiment analysis results for Downloading/Connection

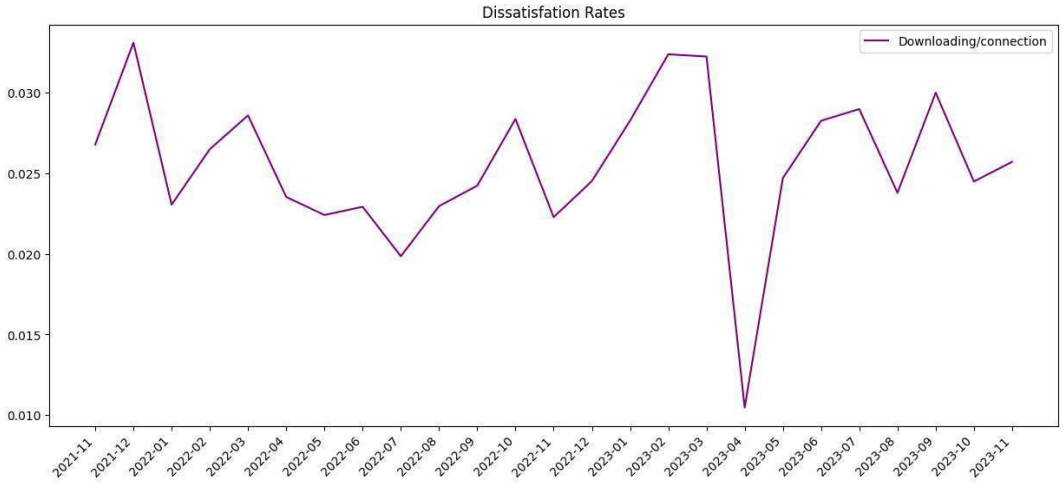

Figure 7.20 Monthly sentiment analysis results for Downloading/Connection

7.2.7 Any

In this section whether any dissatisfaction occured is taken into account. Regardless of what it is any topic from above is considered dissatisfaction and reflected on the graph. Results obtained on a daily, weekly and monthly basis are given respectively in Figure 7.21, 7.22 and 7.23.

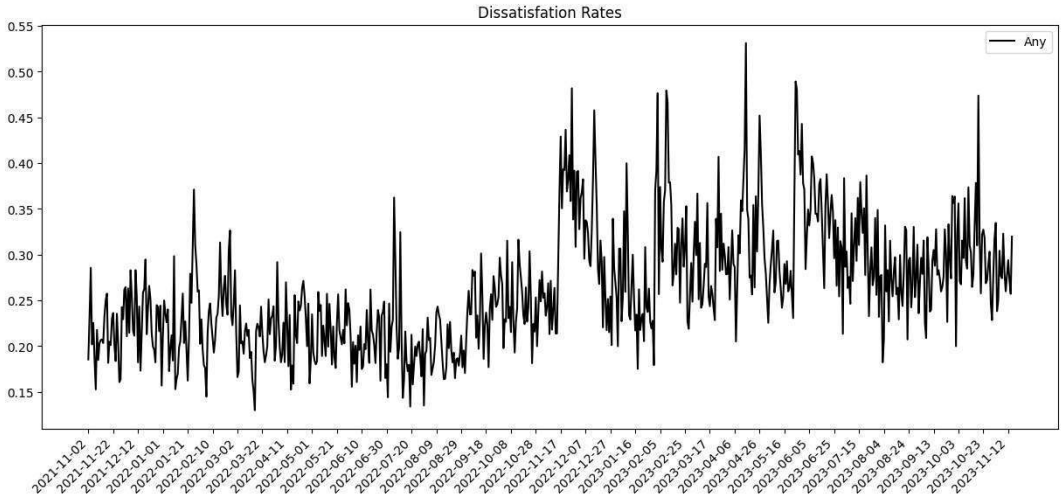

Figure 7.21 Daily sentiment analysis results for Any

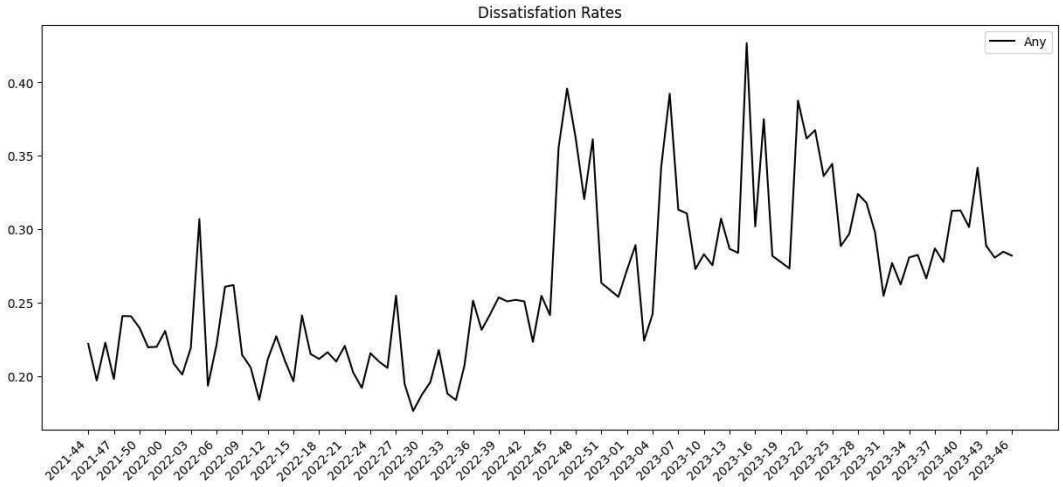

Figure 7.22 Weekly sentiment analysis results for Any

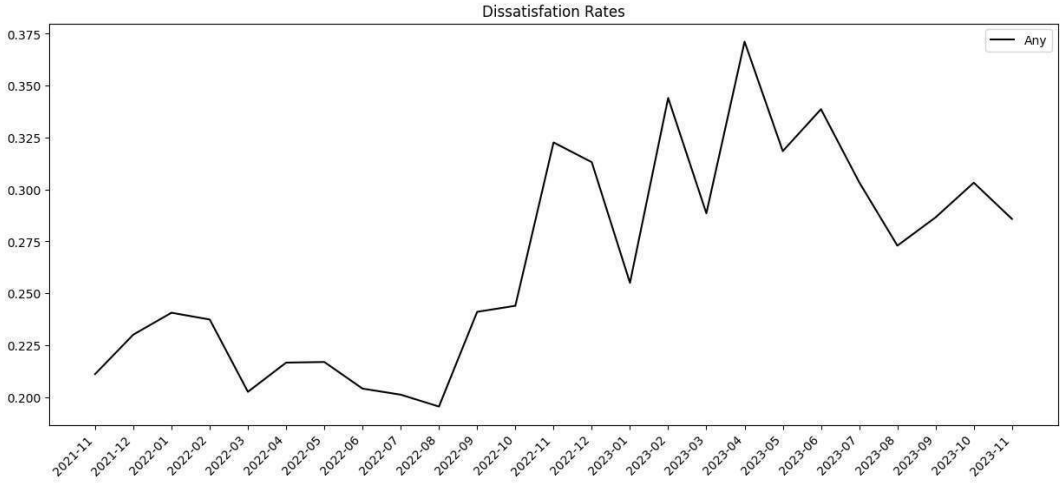

Figure 7.23 Monthly sentiment analysis results for Any

7.3 Prediction with LSTM

We used a Long Short-Term Memory (LSTM) model, a kind of recurrent neural network that is well-suited for time-series data, to predict future patterns in user feedback. The following configuration was used to train the Keras-built model on the processed data:

50 neurons in the LSTM layer with ReLU activation function. dense layer for forecasting output. Mean Squared Error (MSE) loss function and Adam optimizer. Using a dataset that was divided into 80% training and 20% testing segments, the model was trained across 100 epochs. Predicting customer discontent rates for future periods was the main goal of the forecasting of future trends based on previous data.

Below are the dissatisfaction forecasts graphs for 20 days in different aspects.

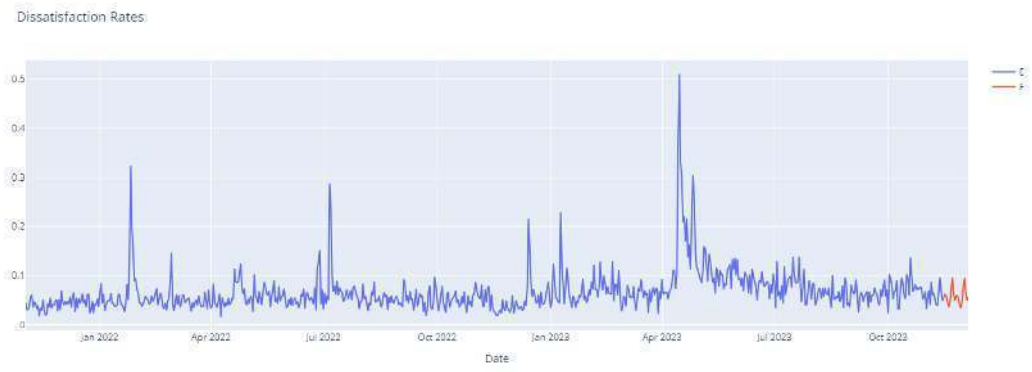

Figure 7.24 Content basis future prediction results

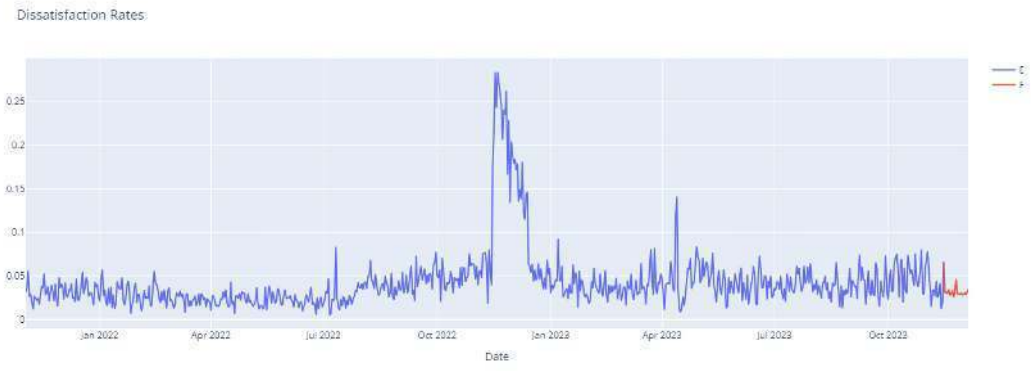

Figure 7.25 UI/UX basis future prediction results

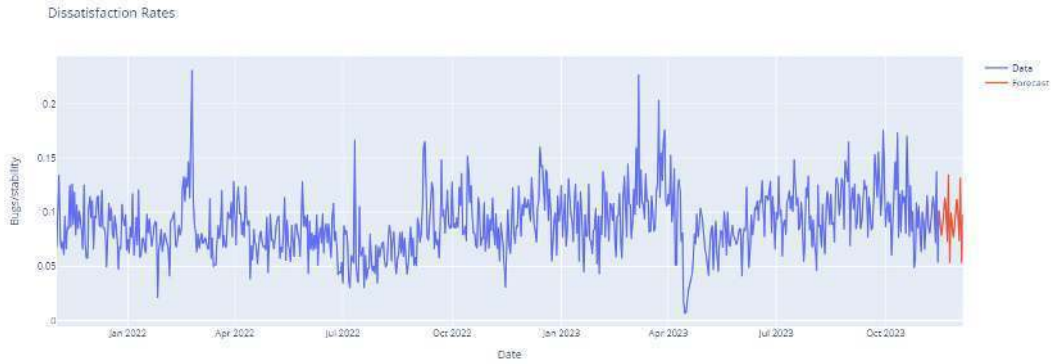

Figure 7.26 Bugs/Stability basis future prediction results

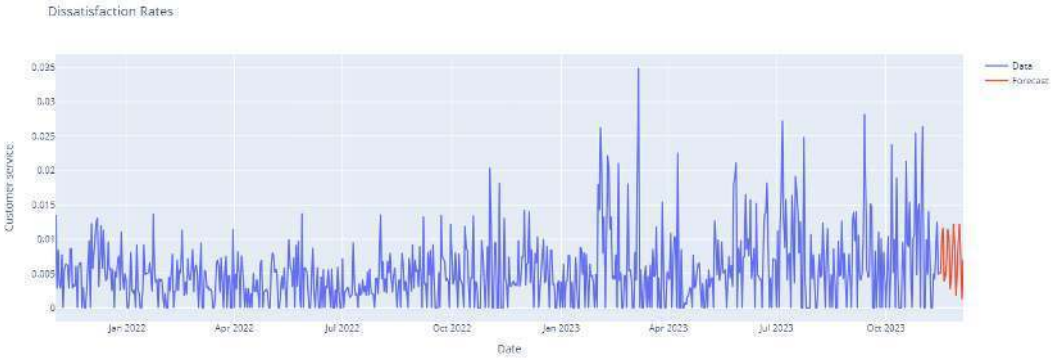

Figure 7.27 Customer Service basis future prediction results

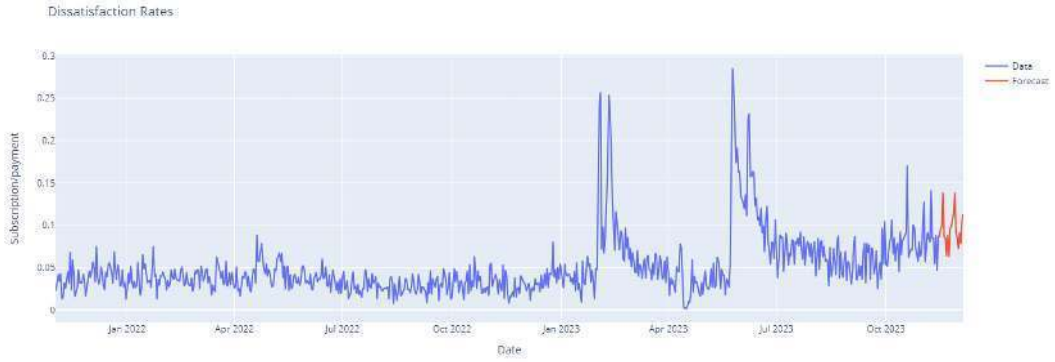

Figure 7.28 Subscription/Payment basis future prediction results

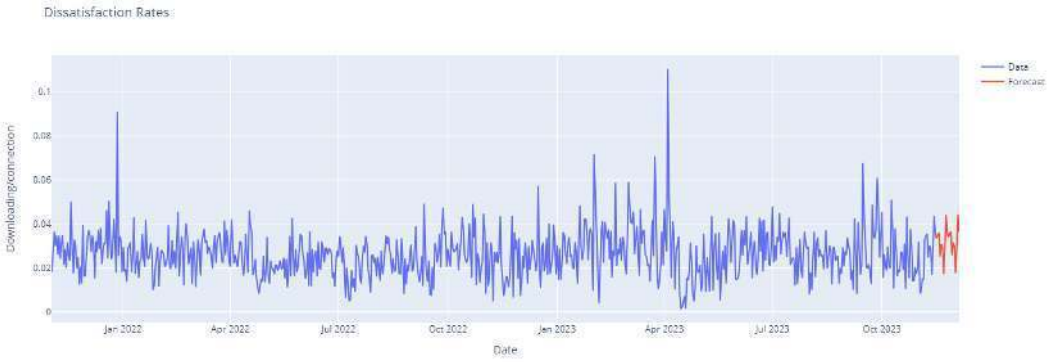

Figure 7.29 Downloading/Connection basis future prediction results

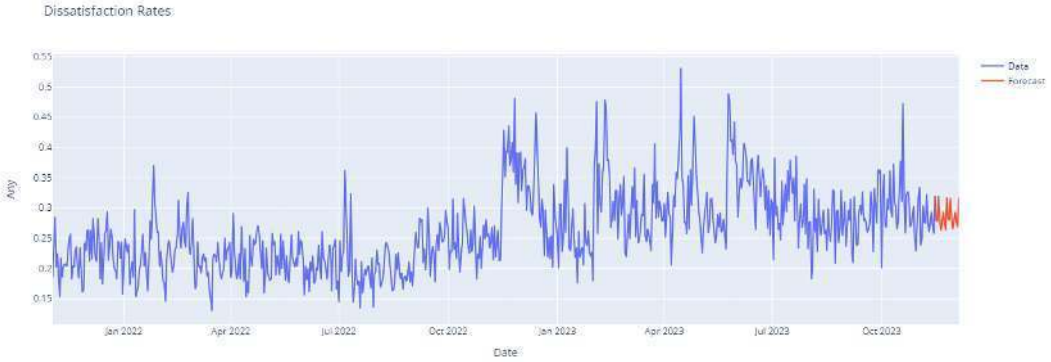

Figure 7.30 Any category basis future prediction results

8
Performance Analysis

In this section, analysis of testing results is examined using both score-based and graph-based outputs. For fine-tuned models accuracy, Multi Label Model Evaluation used, while their confusion matrixes are listed. For prediction testing with LSTM,

8.1 Sentiment Analysis Test

In our experiment multi label model evaluation binary classification used in comparing and evaluating arrays in accuracy.

8.1.1 Concept of Binary Classification

Binary classification involves predicting whether an instance belongs to one of two categories. Multi-label binary classification extends this concept, requiring the model to independently decide the applicability of each label to an instance.

8.1.2 Evaluation Metrics

Essential unique metrics for assessing multi-label models are followings:

- **Accuracy:** Measures correct predictions but may be unsuitable for imbalanced datasets.
- **Precision and Recall:** Assess correct predictions of positive instances, considering both false positives and false negatives.
- **F1 Score:** Balances precision and recall, particularly valuable for uneven class distributions.
- **Hamming Loss:** Specific to multi-label classification, gauges the fraction of incorrect labels.

8.2 Accuracy Analysis

We used 2 models for testing our fined tuned model. GPT 3.5-0613 and GPT 3.5-1106. Result are in tables given below.

8.2.1 GPT 3.5 Turbo 0613

GPT 3.5 Turbo 0613 accuracy scores and classification report of testing data are given in the below. Also, confusion matrix for each categories is subsequently shown.

Table 8.1 GPT-0613 Accuracy Table

Total item count	130
Average accurate count	94.0
Prediction accuracy	0.723077
Hamming loss	0.071795

Table 8.2 GPT-0613 Classification Report

	precision	recall	f1-score	support
Content	0.86	0.83	0.84	23
Customer service	0.73	0.79	0.76	14
Bugs/stability	0.68	0.68	0.68	25
UI/UX	0.67	0.95	0.78	19
Subscription/payment	0.77	0.95	0.85	21
Downloading/connection	0.92	0.58	0.71	19
micro avg	0.76	0.79	0.77	121
macro avg	0.77	0.80	0.77	121
weighted avg	0.77	0.79	0.77	121
samples avg	0.67	0.67	0.66	121

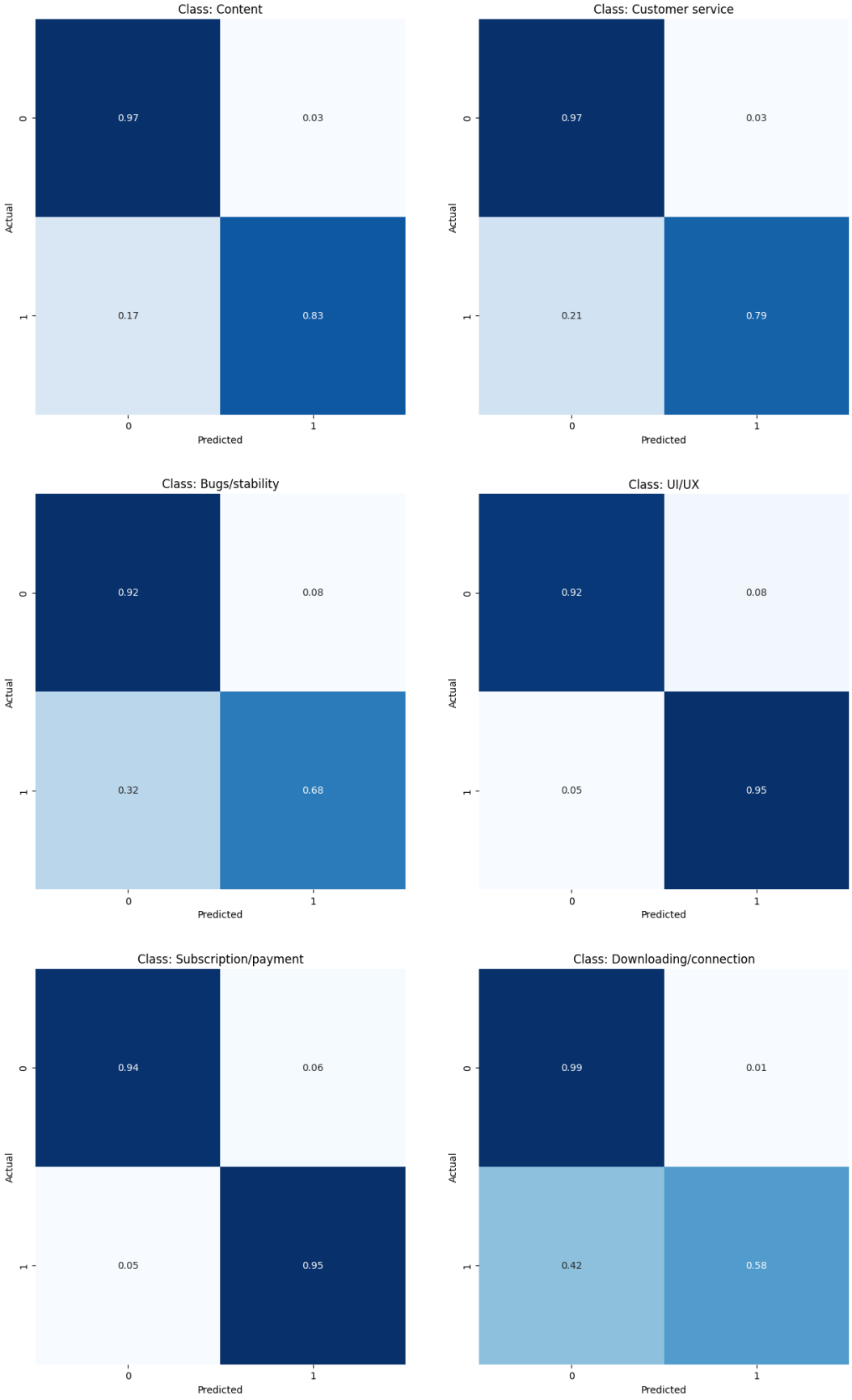

Figure 8.1 Confusion Matrix of GPT 3.5-0613

8.2.2 GPT 3.5 Turbo 1106

GPT 3.5 Turbo 1106 accuracy scores and classification report of testing data are present in the below, meanwhile confusion matrix of every single category is shown just after the tables.

Table 8.3 GPT-1106 Accuracy Table

Total item count	130
Average accurate count	95.0
Prediction accuracy	0.730769
Hamming loss	0.060256

Table 8.4 GPT-1106 Classification Report

	precision	recall	f1-score	support
Content	0.82	0.61	0.70	23
Customer service	0.75	0.86	0.80	14
Bugs/stability	0.78	0.84	0.81	25
UI/UX	0.84	0.84	0.84	19
Subscription/payment	0.94	0.76	0.84	21
Downloading/connection	0.93	0.68	0.79	19
micro avg	0.84	0.76	0.80	121
macro avg	0.84	0.77	0.80	121
weighted avg	0.85	0.76	0.79	121
samples avg	0.65	0.64	0.64	121

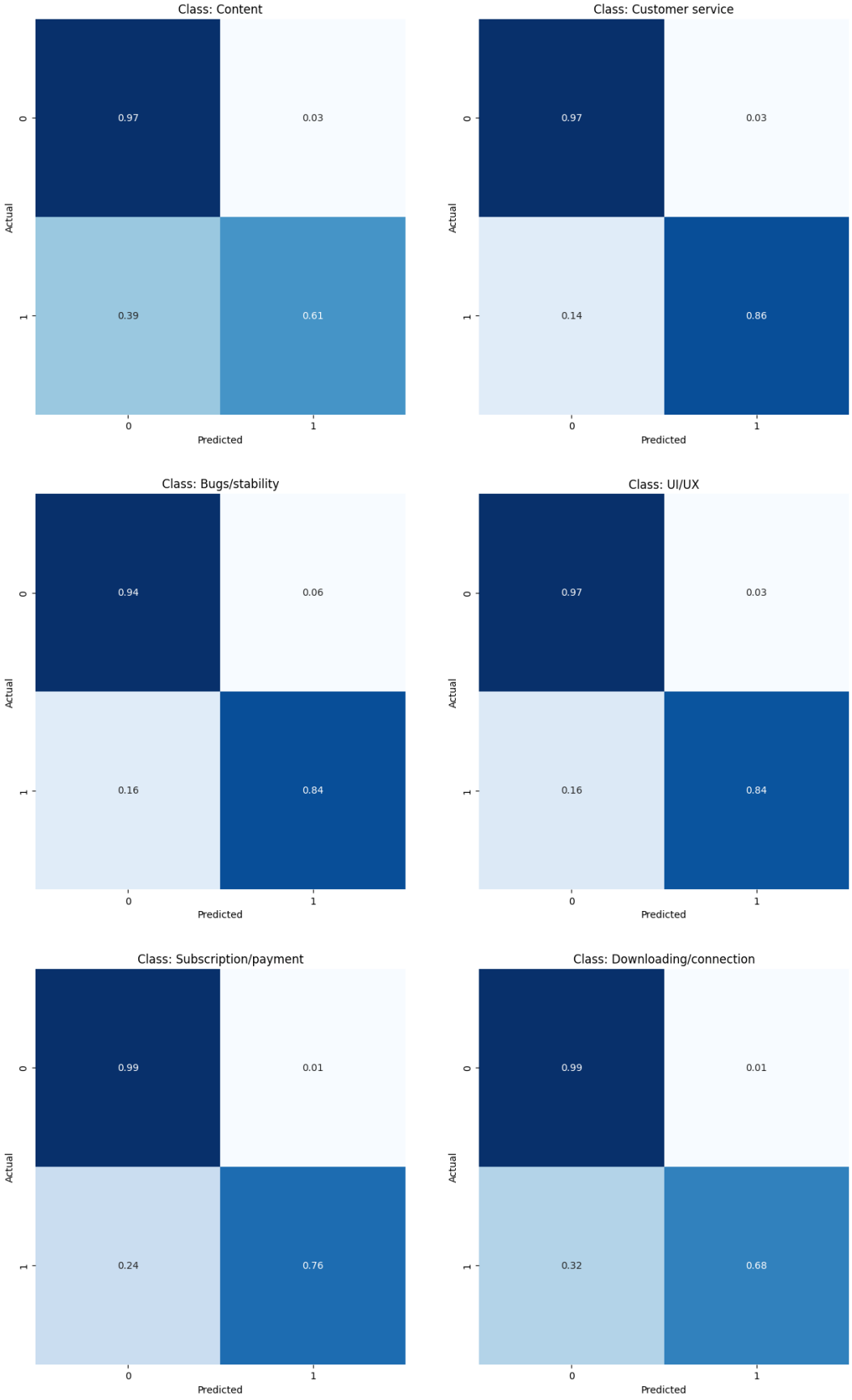

Figure 8.2 Confusion Matrix of GPT 3.5-1106

8.3 Prediction Test

The model is trained to predict future values based on a specified look-back window. The subsequent evaluation of the model's performance is visualized using Plotly, and relevant metrics for performance analysis are discussed.

- **Loss Function:** Mean Squared Error (MSE). Calculates the mean squared deviation between the expected and actual data. Regression activities are frequently performed with it.
- **Optimizer:** Adam Optimizer. A popular optimization technique that is effective at training neural networks due to its adjustable learning rates.
- **Model Architecture:** LSTM Layer. 50 units including an activation function for ReLU. When the input shape is given as (look_back, 1), it means that the model is taking into account a series of look_back time steps with a single feature.
- **Dense Layer:** One unit, typical for problems involving regression.
- **Number of Epochs:** The model is trained for 100 epochs, controlling the number of times the model sees the entire training dataset.
- **Predicton Test Results:** Below are prediction test graphs of different aspects using lstm while look_back=10.

8.3.1 Test Results of Dissatisfaction Categories

In the below, test result of each category is plotted. In this testing job, we splitted the result data of each category into 0.8 and 0.2 for training and test respectively. Each test part predicted using LSTM and visualized. In order to get optimal results, we adjusted hyperparameters of LSTM such as epoch count, and unit count.

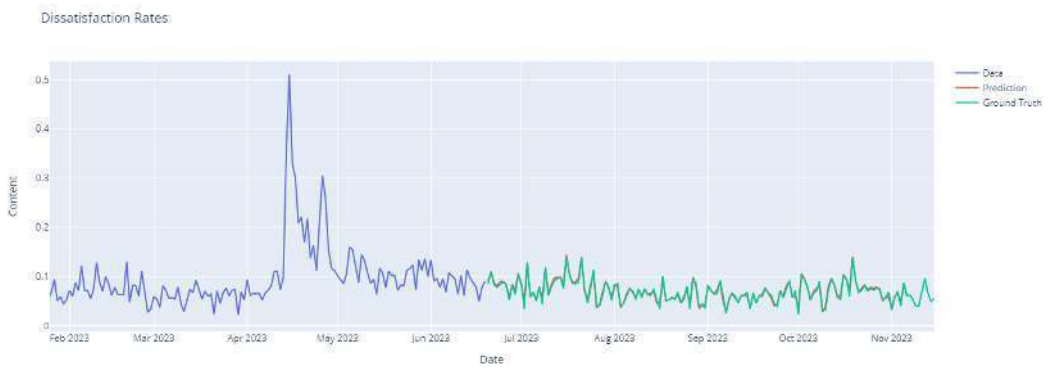

Figure 8.3 Prediction test results for Content

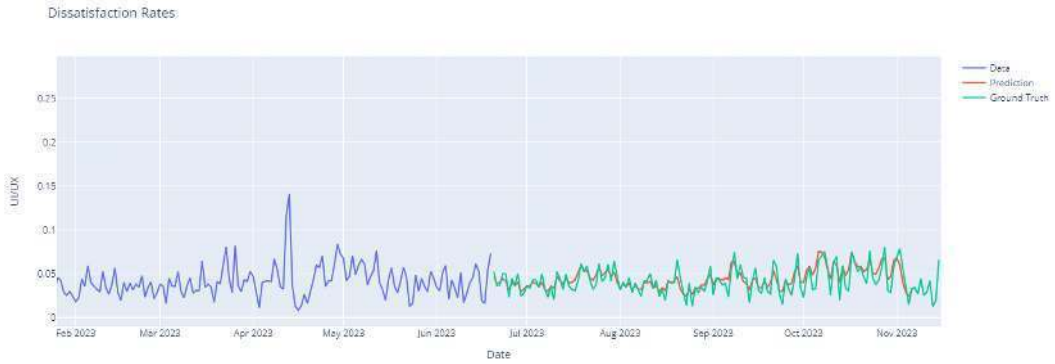

Figure 8.4 Prediction test results for UI/UX

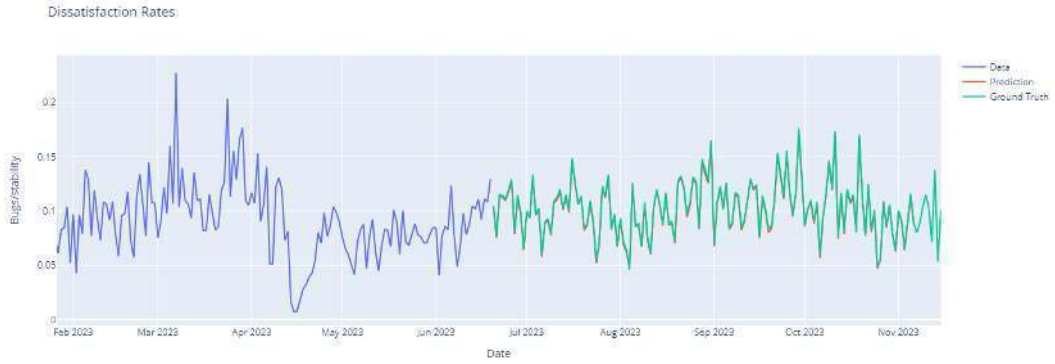

Figure 8.5 Prediction test results Bugs/Stability

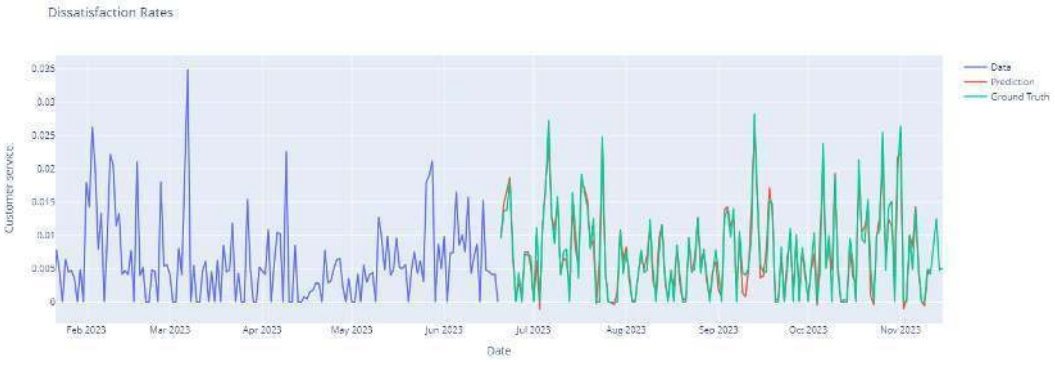

Figure 8.6 Prediction test results for Customer Service

Figure 8.7 Prediction test results for Subscription/Payment

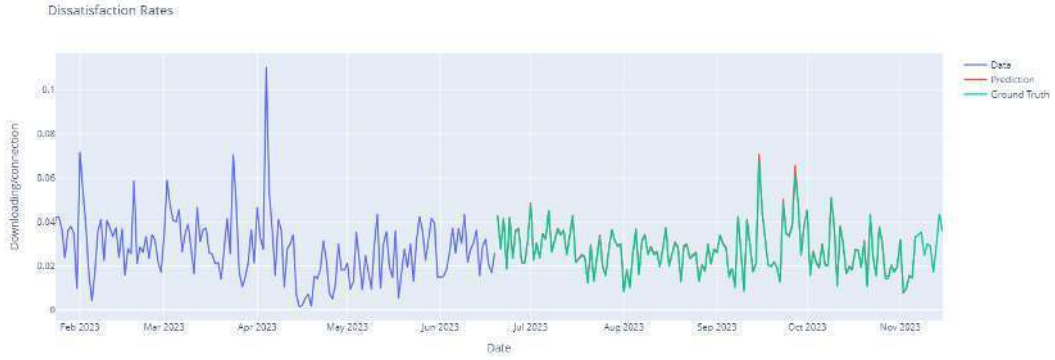

Figure 8.8 Prediction test results for Download/Connectivity

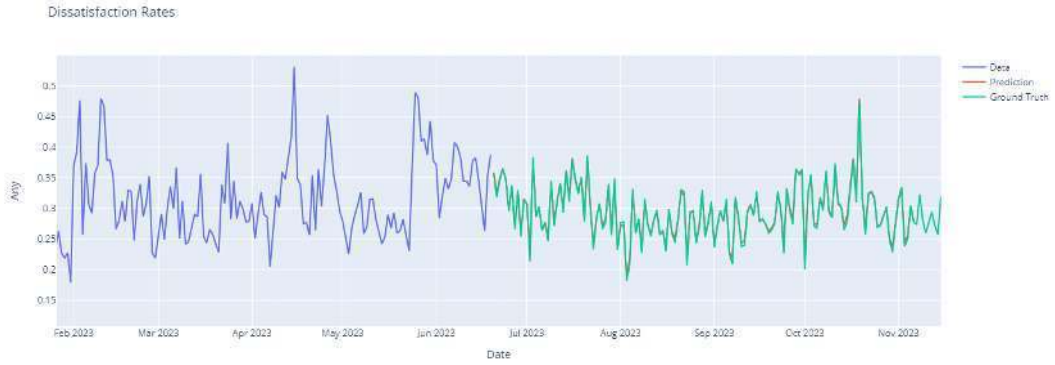

Figure 8.9 Prediction test results for Any

9
Conclusion

In summary, this research study has effectively demonstrated how to combine time series analysis and sophisticated natural language processing for sentiment analysis and future prediction. The optimized GPT-3.5 Turbo models, most notably GPT-3.5 Turbo 1106, showed an impressive 73.07% accuracy in sentiment classification. The system's capabilities were further confirmed by using Long Short-Term Memory (LSTM) for time series forecasting; a low mean squared error (MSE) attested to the accuracy of future value forecasts.

The successful combination of Python, GPT-3.5 Turbo, LSTM, and visualization tools, which promoted a flexible and potent system, demonstrated the project's technological viability. The project's affordability was highlighted by the wise utilization of open-source and free resources, which further strengthened its economic viability.

All things considered, the research is evidence of how well neural network designs and sophisticated language models work together to accomplish challenging tasks. The system's competency and effectiveness in achieving its goals are highlighted by the graphs and assessment metrics, which demonstrate the accuracy of sentiment analysis and the dependability of future predictions.

Although the project's goals were met, there is still room for growth and development. More varied datasets and iterative model training could improve the sentiment analysis accuracy. Furthermore, increasing the LSTM model's refinement with a larger dataset and adjusting its hyperparameters may help future predictions be more accurate.

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.

[3] OpenAI. "Pricing." (), [Online]. Available: https://openai.com/pricing (visited on 01/22/2024).

[4] "1.5 million netflix google store reviews." (), [Online]. Available: https://www.kaggle.com/datasets/bwandowando/1-5-million-netflix-google-store-reviews (visited on 01/22/2024).

Curriculum Vitae

FIRST MEMBER

Name-Surname: Yusuf Taha KÖRKEM
Birthdate and Place of Birth: 10.12.1999, İstanbul
E-mail: taha.korkem@std.yildiz.edu.tr
Phone: 0542 917 55 73
Practical Training: -

SECOND MEMBER

Name-Surname: Mevlana Halit KAYA
Birthdate and Place of Birth: 10.10.1999, İstanbul
E-mail: mevlana.kaya@std.yildiz.edu.tr
Phone: 0544 681 20 40
Practical Training: -

Project System Informations

System and Software: Windows Operating System, Python
Required RAM: 2GB
Required Disk: 256MB