Saisurya Harilokesh Parupudi

730028890

**Analysing User Sentiment and Trends of the X (Formerly Twitter) Application Using Google Play Store Reviews**

Table Of Contents

[1. Introduction 2](#_Toc175434493)

[2. Problem Statement 3](#_Toc175434494)

[3. Literature Review 3](#_Toc175434495)

[4. Methodology 4](#_Toc175434496)

[4.1 Data Collection 5](#_Toc175434497)

[4.2 Data Preprocessing 5](#_Toc175434498)

[4.3 Handling Imbalanced Data 6](#_Toc175434499)

[4.4 Sentiment Analysis 6](#_Toc175434500)

[4.5 Model Selection Justification 7](#_Toc175434501)

[4.6 Topic Modelling 8](#_Toc175434502)

[4.6 Temporal Data Analysis 9](#_Toc175434503)

[4.7 Data Analysis and Visualization 10](#_Toc175434504)

[5. Foreseeable Limitations 10](#_Toc175434505)

[6. Ethical Considerations 10](#_Toc175434506)

[6.1 User Privacy and Data Anonymization 11](#_Toc175434507)

[6.2 Data Management from a Lifecycle Perspective 11](#_Toc175434508)

[6.3 Data Cleaning and Preprocessing 11](#_Toc175434509)

[6.4 Data Destruction 12](#_Toc175434510)

[6.5 Mitigating Potential Biases 12](#_Toc175434511)

[6.6 Ethical Compliance and Integrity 12](#_Toc175434512)

[7. Project & Executive Summary Plan 12](#_Toc175434513)

[8. Conclusion 13](#_Toc175434514)

[9. References 14](#_Toc175434515)

# 1. Introduction

The X application, previously referred to as Twitter, is a very popular social networking platform with a vast user base spanning throughout the globe. Gaining insights into user sentiment and feedback is crucial for the ongoing enhancement and user contentment of such programs. The objective of this study is to perform a thorough examination of user feedback on the X application obtained from the Google Play Store. The project will utilise advanced natural language processing (NLP) tools, sentiment analysis, and ML methodologies to discover significant patterns and concerns identified by users.

The research aims to achieve three objectives: firstly, to analyse the overall sentiment of the reviews and track its changes over time; secondly, to identify particular app versions that have received the highest positive and negative feedback; and thirdly, to discover common themes and issues through topic modelling. The results of this research will offer significant perspectives for the developers of the X program, enabling them to effectively tackle user problems and improve the user experience. Moreover, this work would enhance the scholarly comprehension of user feedback analysis in social media apps.

# 2. Problem Statement

The global user base of Twitter has created different reactions in response to the rebranding of the platform to X, which was followed by considerable improvements in user experience, design, and platform functionality. The changes have resulted in a complicated terrain of user emotion, which is crucial for app developers and business strategists to comprehend in order to sustain and improve user engagement. Although there has been a lot of study on social media platforms, there is still a lack of analysis on the subtle effects of recent changes on user sentiment, particularly when using advanced sentiment analysis tools. This research tries to overcome this gap by methodically assessing user input to find important areas of discontent and satisfaction. The findings gathered from this study are meant to inspire strategic decisions that can improve the user experience, address the core causes of negative sentiment, and ultimately influence future upgrades and advancements of the X platform. By concentrating on the link between individual app features and user sentiment, this research attempts to give practical suggestions that match with the increasing requirements and expectations.

# 3. Literature Review

The study of user input, particularly through sentiment analysis, has become a key field of research in understanding and enhancing mobile applications. Sentiment analysis includes sorting user evaluations to extract relevant data for developers. This literature review investigates previous research, identifies gaps, and explains theoretical frameworks relevant to the project.

Several research have focused on sentiment analysis of user evaluations in app marketplaces. Samanmali and Rupasingha (2024) employed deep learning approaches to assess sentiment in Google Play Store evaluations, revealing the potential for neural networks to increase sentiment classification accuracy . Similarly, Hadi and Fard (2023) examined pre-trained models for user feedback analysis, emphasising the usefulness of transfer learning in sentiment analysis tasks.

Karim et al. (2020) applied ML techniques to assess and depict user attitudes in Google Play Store evaluations, creating a framework for automated sentiment analysis. Additionally, Kaur et al. (2019) studied the application of semi-supervised learning for sentiment analysis in multilingual situations, which is particularly significant for global applications like X. Moreover, research by Rahman et al. (2020) on sentiment analysis for evaluating user happiness in Android applications gives useful insights into feature extraction and sentiment classification approaches that are directly relevant to analysing mobile app evaluations.

**Gaps in Existing Literature**

Despite the substantial research on sentiment analysis, most previous research either focus on general techniques or other popular applications. There is minimal research particularly assessing user feedback for X across multiple app versions, which is vital for understanding how updates and changes effect user satisfaction. This literature review emphasises the substantial work done on sentiment analysis and user feedback in mobile applications. However, it also exposes a crucial deficiency in the deep study of X application reviews throughout multiple versions. By addressing this gap, the proposed research intends to give significant insights on user happiness and app development.

# 4. Methodology

This research seeks to extensively examine user evaluations of the X (previously Twitter) application acquired from the Google Play Store. Given the dataset's size of about 2 million reviews, the methodology recommends the use of advanced natural language processing (NLP) and ML techniques to manage and extract useful insights from the data. This section explains the data collecting, preprocessing, sentiment analysis, topic modelling, and visualization approaches that can be applied in this research.

## 4.1 Data Collection

The dataset was acquired from Kaggle and contains user reviews of the X program collected via the google\_play\_scraper tool. The dataset contains numerous variables such as review ID, author ID, review text, rating, number of likes, app version, and date. This secondary data source is freely available under the CC0: Public Domain license, assuring ethical compliance in data usage.

## 4.2 Data Preprocessing

Given the vast size of the dataset, preprocessing is a vital step to assure data quality and usability for later analysis. The preprocessing stages include:

1. **Data Cleaning:**

* Ensuring the dataset is clear of duplicate reviews to ensure consistency and dependability in the study.
* Filtering out non-English reviews using the langdetect library to focus on relevant user input.

1. **Handling Emojis and Special Characters:**

* Utilizing regular expressions to systematically remove emojis and special characters, standardizing the text data to build successful analysis.

1. **Slang Translation:**

* Leveraging a robust slang lexicon taken from trusted GitHub projects to convert popular slang terminology into regular English.

1. **Offensive Language Filtering:**

* Implementing a filter to block particular extremely offensive words, while retaining other phrases that may help to sentiment analysis.

1. **Normalization, Tokenization and Lemmatization:**

* Changing text to lowercase, removing punctuation, splitting them into tokens and reducing them to their basic word forms will be performed to normalize, tokenize and lemmatize the data.

These preparation tasks guarantee that the dataset is cleaned, standardised, and ready for further sentiment analysis and topic modelling.

## 4.3 Handling Imbalanced Data

**Algorithms Robust to Class Imbalance:**

* Algorithms that are resilient to class imbalance, such as Random Forest and Gradient Boosting, might be applied. These methods are designed to handle unbalanced data successfully by modifying the decision thresholds or introducing class weights into the analysis.

**Evaluation Metrics:**

* Standard measurements like accuracy, recall, F1 score and Area Under the Precision-Recall Curve (AUC-PR) might be used to evaluate model performance, offering a more realistic evaluation of the models ability in handling unbalanced data.

## 4.4 Sentiment Analysis

Sentiment analysis can classify the reviews into positive, negative, or neutral sentiments. The following approaches can be used:

1. **Lexicon-Based Approach**: Utilizing VADER, which awards sentiment ratings to words and combines them to get the overall sentiment of the review.
2. **ML-Based Approach**: Training models such as Logistic Regression, Random Forest, etc. using labelled data to categorise attitudes. The dataset may be separated into training and testing sets to evaluate model performance. Evaluation metrics can be used to assess model performance.

## 4.5 Model Selection Justification

The proposal employs a combination various models for sentiment analysis, each chosen for their specific strengths and suitability for the task:

**VADER:**

* VADER is a vocabulary and rule-based sentiment analysis tool. It offers instant validity of the reviews and is fast to complete. VADER can manage the informal language included in user evaluations, including slang, acronyms and emojis which is suitable for this project.

**Logistic Regression:**

* Logistic regression is a basic and interpretable ML model. It makes it simple to grasp the link between sentiment results and review text features as it gives definite coefficients for every feature Against which more intricate models may be evaluated.

**Random Forest:**

* Resilient to class imbalance and able to simulate complicated, non-linear connections in the data, Random Forest is a potent ensemble approach. It offers understanding of the relevance of various characteristics in sentiment prediction, which may be quite helpful for further research and model improvement.

**Model Comparison Criteria**

The chosen models will be evaluated and compared based on the following:

* Accuracy: How well the model predicts sentiment labels overall.
* Precision: Showing the model's ability to avoid false positives, the proportion of actual positive predictions across all positive predictions shows.
* Recall: The proportion of actual positive predictions among all genuine positives indicates the model's ability to identify every relevant situations.
* F1 Score: The harmonic mean of accuracy and recall, providing a fair evaluation of the model's performance.
* Area Under the Precision-Recall Curve (AUC-PR): Particularly in the presence of imbalanced data, this statistic presents a whole picture of the outcome of the model.

## 4.6 Topic Modelling

To detect hidden trends within the evaluations, hidden Dirichlet Allocation (LDA) might be applied.

1. Model Training: By training the LDA model on pre-processed text data, one may find clusters of words that often appear togetheras themes.
2. Subject Interpretation: Each recognised subject may be interpreted and categorised based on the most representative terms.

**LDA Model Training:**

* The data which comprises tokenised, lemmatised reviews free from stop words and pointless characters will be used to train the LDA model.
* Parameters such as alpha (document-topic density) and beta (topic-word density) will be modified to improve model performance and ensure robustness given the enormous dataset.

Interpretation of Results

**Top Words in Each Topic:**

* The model will offer a distribution of words for every found subject. The theme will be interpreted using the top words with highest likelihood.
* Every topic will be given a semantic label best fit for the top words based on underlying theme.

**Actionable Insights:**

* Common user problems, often requested features, and areas needing development will be generated from identified themes using useful data.
* These realisations will be compiled and presented to the development team and stakeholders thereby improving the user experience of the X application.

## 4.6 Temporal Data Analysis

**Time Series Analysis:**

* Review data will be combined depending on date and app version. Sentiment analysis findings will help to determine the average sentiment scores for these intervals.

**Visualization Techniques:**

* Sentiment scores will be presented over time using trend lines to visually reflect changes in user sentiment. This will assist detect times negative sentiment and associate them with specific events.
* Heatmaps will be developed to display emotion intensity over time. These visualizations will offer a clear insight of how user sentiment swings throughout different zones.

By applying these methodologies, the research will offer a complete perspective of how user sentiment changes over time and across different versions of the X. This strategy will give significant data for the development team to analyse the impact of app upgrades and user feedback over time.

## 4.7 Data Analysis and Visualization

The final step involves analysing the results of sentiment analysis and topic modelling to draw meaningful insights:

* Examining the sentiment distribution over different app versions and time periods to detect trends and patterns.
* Investigating correlations between user ratings and specified themes to determine their influence on user satisfaction.
* Utilizing data visualization tools such as Matplotlib, Plotly, WordCloud and Seaborn to build graphs, charts, and word clouds that effectively present the findings.

By concentrating on data preprocessing, sentiment analysis, topic modeling, and visualization, the study seeks to give practical insights that can influence the improvement of the X application and contribute to the business growth.

# 5. Foreseeable Limitations

Potential data duplication and inconsistencies, uncertainty in understanding text like sarcasm, and limits of lexicon-based models in capturing subtleties hamper text analysis. It also requires time-consuming preprocessing and model training and calls for great computer resources. Moreover, interpreting subjects found by the LDA model might be subjective.

# 6. Ethical Considerations

## 6.1 User Privacy and Data Anonymization

The key ethical challenges in this work are protecting user privacy and data anonymization. Although the dataset acquired from Kaggle is publicly available and anonymised, extra precautions will involve:

* Any residual identifiers will be re-anonymized to achieve perfect anonymity.
* Techniques such as deleting or masking any remaining identifiable information will be implemented to ensure user confidentiality.

## 6.2 Data Management from a Lifecycle Perspective

* Data Collection: Since the data is publicly available and anonymised, informed consent from the original users is not necessary.
* Secure Data Storage:
* Encrypted alternatives to storage will be used to safeguard data against illegal access.
* Access to the data will be restricted to approved persons only, ensuring that security.
* Data Usage:
* The study will strictly follow GDPR and university ethical guidelines, therefore guaranteeing that data is used just for the intended use in research.

## 6.3 Data Cleaning and Preprocessing

* Converting emojis to text descriptions using libraries like emoji.
* Using a slang dictionary to transform popular slang terminology into formal language.
* Implementing filtering or flagging systems for foul language using pre-built lexicons.

## 6.4 Data Destruction

* All data will be safely destroyed after study end. All copies of the dataset will be irretrievably deleted, therefore ensuring no personal information is still accessible.

## 6.5 Mitigating Potential Biases

Biases during data interpretation and analysis will be mitigated by:

* Transparency will help to reduce prejudices in data interpretation and analysis by means of which the research process is kept objective to prevent distorted findings.
* Frequent interactions with the supervisor will offer vital comments helping to fix any unintentional prejudices.

## 6.6 Ethical Compliance and Integrity

The study will uphold the highest standards of ethical conduct by following the university’s guidelines to ensure participant personal data is confidential and handled responsibly. By addressing these ethical factors fully, our research will preserve the privacy and secrecy of users.

# 7. Project & Executive Summary Plan

For my final dissertation, I plan to present the executive summary in a written format to concisely capture the key findings and recommendations of my research. Below is my intended timeline for completing the proposal.

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Milestone** | **Deadline** | **Supervisor Meeting Date** |
| June 26-30 | Proposal and Ethics Form Submission | June 28, 2024 |  |
| July 1-7 | Draft of Introduction and Literature Review | July 2, 2024 | July 2, 2024 |
| July 8-14 | Finalize Introduction and Literature Review | July 9, 2024 |  |
| July 8-14 | Data Preprocessing | July 9, 2024 |  |
| July 15-21 | Complete Sentiment Analysis | July 16, 2024 | July 16, 2024 |
| July 22-28 | Start Topic Modelling | July 23, 2024 |  |
| July 22-28 | Draft Methodology Section | July 23, 2024 |  |
| July 29-Aug 4 | Complete Topic Modelling | July 30, 2024 | July 30, 2024 |
| Aug 5-11 | Data Analysis and Visualization | August 6, 2024 |  |
| Aug 5-11 | Draft Results Section | August 6, 2024 |  |
| Aug 12-18 | Finalize Data Analysis and Visualizations | August 13, 2024 | August 13, 2024 |
| Aug 12-18 | Draft Discussion Section | August 13, 2024 |  |
| Aug 19-21 | Draft Conclusion Section | August 20, 2024 |  |
| Aug 22-Sep 1 | Compile and Review Full Report | August 27, 2024 | August 27, 2024 |
| Sep-02 | Final Report Submission | September 2, 2024 | September 2, 2024 |

# 8. Conclusion

This project will examine Google Play Store evaluations of the X (previously Twitter) app using sophisticated NLP and ML approaches. The purpose is to detect sentiment trends, analyze feedback across software versions, and identify frequent user complaints. These insights will help developers enhance user experience and solve particular problems. The research also contributes to the academic subject of user feedback analysis. Ethical issues, including data protection, shall be rigorously followed. The project is organized to guarantee steady development, with regular feedback loops to direct the study towards practical conclusions for future app upgrades.

# 9. References

1. Samanmali, P. H. C., & Rupasingha, R. a. H. M. (2024). Sentiment analysis on google play store app users’ reviews based on deep learning approach. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-024-19185-w>
2. Hadi, M.A., & Fard, F.H. (2023). Evaluating pre-trained models for user feedback analysis in software engineering: a study on classification of app-reviews. *Empirical Software Engineering*, 28, 88. <https://doi.org/10.1007/s10664-023-10314-x>
3. Karim, A., Azhari, A., Belhaouri, S.B., & Qureshi, A.A. (2020). Machine Learning Algorithm’s Measurement and Analytical Visualization of User’s Reviews for Google Play Store. *Preprints*. <https://doi.org/10.20944/preprints202003.0249.v1>
4. Kaur, G., Kaushik, A., & Sharma, S. (2019). Cooking is creating emotion: a study on Hinglish sentiments of youtube cookery channels using semi-supervised approach. *Big Data and Cognitive Computing*, 3(3), 37. <https://doi.org/10.3390/bdcc3030037>
5. Rahman, M.M., Rahman, S.S.M.M., Allayear, S.M., Patwary, M.F.K., Munna, M.T.A. (2020). A Sentiment Analysis Based Approach for Understanding the User Satisfaction on Android Application. In: Raju, K.S., Senkerik, R., Lanka, S.P., Rajagopal, V. (eds) Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing, vol 1079. Springer, Singapore. <https://doi.org/10.1007/978-981-15-1097-7_33>