International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04122
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

"Sentiment Analysis and NLP On Google Play Store Reviews"

Aniket A. Waghmare, Prof. Bisweswar Thakur

[1]*Aniket Waghmare Master of Computer Application & Trinity Academy of Engineering, Pune*
[2] *Prof. Bisweswar Thakur Master of Computer Application & Trinity Academy of Engineering, Pune*

Abstract - The Google Play Store sees thousands of new apps regularly, developed by individuals or teams competing globally. Most apps are free, making revenue models like ads and in-app purchases unclear. As a result, app success is often judged by installation counts and user ratings. However, ratings can be biased or inconsistent, and there's often a gap between number of ratings and written reviews. This study uses machine learning to predict app ratings based on a dataset from Kaggle, analysing features such as app type, user reviews, and ratings. The analysis merges app data with user reviews and performs sentiment analysis using Sentiment Polarity. The sentiment distribution is visualized for free and paid apps using bar charts. Review texts are summarized using the Sumy library, and spam is detected by flagging reviews with fewer than three words or common spam keywords. Summaries and spam labels are saved to new CSV files for further analysis. Reviews are then cleaned, lemmatized, and stripped of stop words before being fed into a logistic regression model trained on a small labelled dataset. The model classifies reviews into topics like 'bug', 'UI', or 'feature request' and is evaluated on a test set. Finally, it predicts categories for unlabelled reviews, and a pie chart shows how many summaries contain meaningful content versus empty ones.

Keywords: - *Google Play Store Apps, Ratings Prediction, Data Analysis Library, Sentiment Analysis, NLP, Machine Learning Algorithm*

1.INTRODUCTION

The Google Play Store is home to millions of mobile applications, with thousands of new apps being introduced regularly. As the competition grows, developers face the challenge of standing out in a crowded market. While many apps are free, monetization strategies such as in-app purchases, advertisements, and subscriptions are not always transparent. As a result, an app's success is more commonly measured by the number of installations and user ratings rather than direct revenue generation. However, ratings can often be biased, as they are voluntary and may not fully reflect the quality of the app due to limited participation or inconsistent user feedback. Reviews, though valuable for understanding user experiences, often present a more complex picture, as they can be lengthy, inconsistent, or difficult to interpret. Analysing these reviews manually is time-consuming and impractical, especially when dealing with millions of apps. This project aims to utilize machine learning algorithms to predict app ratings based on multiple features, including app type (free or paid), user reviews, and other attributes. Additionally, the project explores the potential of sentiment analysis to gain deeper insights into user satisfaction and emotions expressed through reviews.

2. Methodology

2.1 Technology Stack

The project uses **Python** as the main language due to its strong libraries for data analysis and machine learning. **pandas** and **NumPy** are used for data cleaning and transformation, while **matplotlib** and **seaborn** help visualize distributions and trends across app categories, ratings, installs, and more. For natural language processing, tools like **TextBlob**, **VADER**, **NLTK**, and **spaCy** are used for sentiment analysis, lemmatization, and tokenization. **sumy** is applied for automatic text summarization. Machine learning models such as **logistic regression** and **Naive Bayes** are implemented using **scikit-learn** to classify sentiment and predict app installs. **TfidfVectorizer** and **CountVectorizer** convert text data into numerical format. Data storage is managed through **CSV** files, and development is done in **Jupyter Notebook** or **Google Colab** for easy experimentation and visualization.

2.2 Data Collection and Management

The data for this project was collected from two public datasets related to Google Play Store apps and their user reviews. The app dataset includes details like app name, category, rating, installs, price, and content rating, while the review dataset provides user-written reviews along with sentiment labels. These datasets were loaded into **pandas DataFrames** for processing. Missing values were handled by imputing the mode for categorical data and the median for numerical data. Data cleaning involved formatting text fields, removing unwanted characters, converting units (e.g., MB to KB), and handling inconsistent entries such as "Varies with device." Outliers in numerical columns like 'Rating' were removed using the **IQR method**. Duplicate entries were identified and dropped to maintain data integrity. Cleaned and processed data was stored in **CSV files** for further analysis and model training. This

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04122
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

structured approach ensured reliable, consistent, and analyzable datasets for the project.

2.3 Application Features

- **Data Visualization**: Visualizes app distribution across categories, average ratings, install trends, and app size statistics.
- **Sentiment Analysis**: Analyzes user reviews to gauge public sentiment and compare sentiments between free and paid apps.
- **Text Summarization**: Automatically summarizes lengthy user reviews, providing concise insights.
- **Spam Detection**: Flags potential spam reviews based on predefined keywords and review length.
- **Review Categorization**: Uses machine learning to categorize reviews into topics such as bugs, UI issues, and feature requests.
- **Sentiment Classification**: Classifies user sentiments into positive, negative, or neutral based on review content.
- **Install Prediction**: Predicts the number of app installs based on app ratings using machine learning models.

2.4 Middleware Logic

Data Preprocessing Handler: Handles cleaning tasks such as removing nulls, formatting fields (e.g., price, installs), and transforming size units before analysis.

Review Processing Module: Cleans review text, performs lemmatization, removes stop words, and prepares text data for sentiment and topic analysis.

Sentiment Analyzer: Applies polarity scoring using NLP libraries (like TextBlob or VADER) to determine sentiment for each review.

Text Summarization Engine: Uses the Sumy library to generate brief summaries from lengthy reviews for improved readability.

Spam Detection Filter: Scans review text for patterns such as overly short content or known spam keywords and flags them accordingly.

Review Categorization Model: Employs trained logistic regression to classify reviews into topics like bugs, UI feedback, or feature requests.

Sentiment Classification Model: Uses Naive Bayes to classify the emotional tone (positive, negative, neutral) of each review.

Install Prediction Logic: Applies logistic regression to predict whether an app has over 100K installs based on features like rating.

Data Routing & Caching: Routes cleaned data and predictions between backend modules and stores intermediate results for faster access.

CSV Output Generator: Compiles processed results (summaries, flags, predictions) and exports them into structured CSV files for further use.

3.Results and Discussion

The analysis revealed several key insights into Google Play Store apps. Most apps belonged to the "FAMILY," "TOOLS," and "GAME" categories, while "EVENTS" had the highest average rating. Sentiment analysis showed that free apps generally received more mixed reviews, whereas paid apps had slightly more positive sentiments. The summarization module effectively reduced long reviews into concise summaries, and the spam filter successfully flagged irrelevant or low-quality content. The topic classification model categorized reviews into actionable topics such as bugs or feature requests, aiding developers in understanding user priorities. The sentiment classifier achieved reasonable accuracy, highlighting the emotional tone of user feedback. The install prediction model also performed well, correctly identifying apps with high user engagement based on rating patterns.

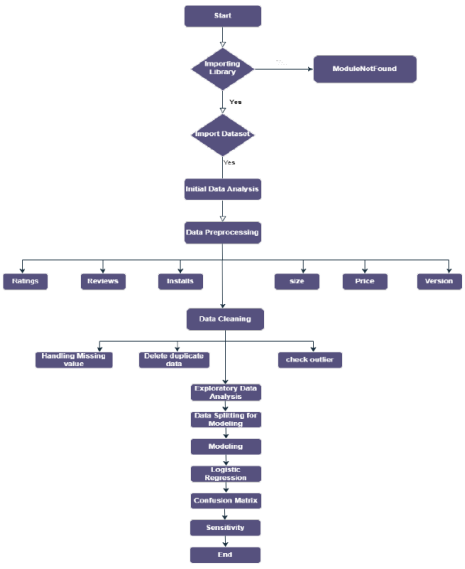

Fig -1: Flow Chart Diagram

- The sentiment analysis and classification models successfully identified user opinions, with paid apps generally receiving more positive sentiments than free apps.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04122
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

- The review categorization model effectively grouped user feedback into actionable topics like bugs, UI issues, and feature requests, aiding targeted improvements.
- The install prediction model accurately estimated whether an app would reach 100K+ installs based on its rating, supporting marketing and development strategies.

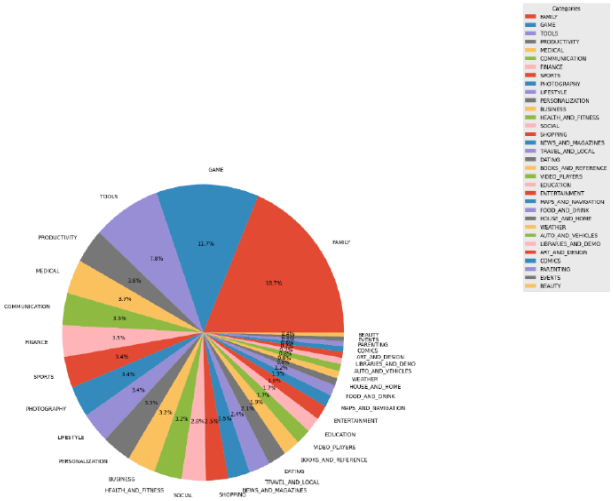

Fig -2: Pie-chart on Categories of Apps

The image is a pie chart that shows the distribution of different categories of apps in the Google Play Store. The largest slice of the pie chart is for the Family category, followed by the Game category. Other categories include Tools, Productivity, Medical, Communication, Finance, Sports, Photography, Lifestyle, Personalization, Business, Health and Fitness, Social, Shopping, News and Magazines, Travel and Local, Dating, Books and Reference, Video Players, Education, Entertainment, Maps and Navigation, Food and Drink, House and Home, Weather, Auto and Vehicles, Libraries and Demo, Art and Design, Comics, Parenting, Events, and Beauty. The pie chart provides a visual representation of the popularity of different app categories.

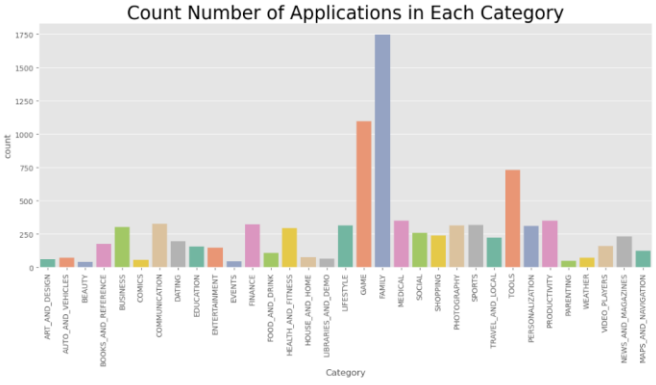

Fig -3: Countplot for category Column

The FAMILY category has the highest number of applications in the Google Play Store, making it the most dominant. In addition, the GAME and frequent development. These three categories and TOOLS categories also feature a large number of apps, reflecting their popularity represent a significant portion of the overall app distribution on the platform.

3. CONCLUSIONS

The analysis of Google Play Store data highlights the significant impact of user reviews on app success. Through data cleaning, the datasets were prepared for reliable analysis by addressing missing values, correcting data formats, and removing outliers. Exploratory Data Analysis (EDA) uncovered meaningful trends across categories, ratings, installs, and app sizes. Categories such as "FAMILY," "TOOLS," and "GAME" were found to be the most popular, while "EVENTS" had the highest average rating, and "DATING" had the lowest. Sentiment analysis played a crucial role in understanding user satisfaction by examining the polarity of reviews and comparing sentiments between free and paid apps. Review summarization, supported by the Sumy library, extracted essential insights from lengthy feedback, while spam detection filtered irrelevant or repetitive content, enhancing the quality of data for analysis. Machine learning models were used effectively to classify review sentiments and categorize user feedback into specific topics such as bugs, UI problems, and feature requests. A logistic regression model also helped predict whether an app would achieve high install numbers based on its rating. These models provided developers with tools to better interpret user opinions and improve their applications accordingly.

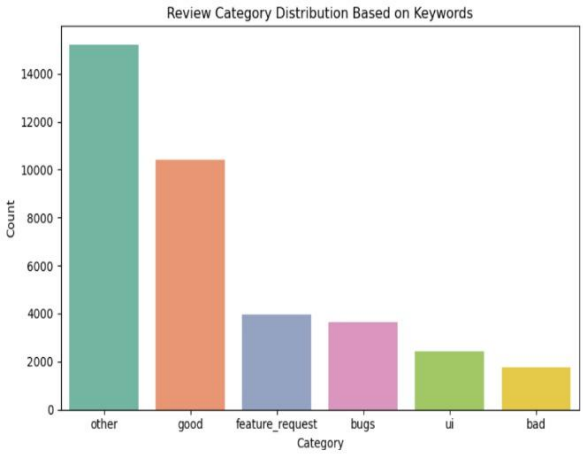

Fig – 4 Review summary distributed on keywords

The diagram is a bar chart that visually represents the distribution of user reviews across several categories identified through keyword matching: good, bad, bugs, UI, feature request, and other. Each bar corresponds to a category and its height indicates the number of reviews that fall into that group. The" good" category typically has the tallest bar, suggesting that a large portion of users left positive feedback. The" bugs" category often follows, highlighting the frequency of technical issues mentioned by users. Categories like" bad"," UI", and"

feature request" show smaller but still meaningful counts, pointing to user dissatisfaction, interface concerns, or suggestions for new features.

ACKNOWLEDGEMENT

The author thanks Prof. Bisweswar Thakur and Trinity Academy of Engineering for their guidance and support throughout the project.

REFERENCES

[1] A. Martin, B. Kuppusamy, and G. K. S. Prabaharan, "App review analysis using sentiment and topic modeling," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5052–5061, 2022.
[2] M. S. Rana and S. K. Sharma, "Opinion mining of mobile app reviews using supervised and unsupervised approaches," Procedia Computer Science, vol. 132, pp. 632–639, 2018.
[3] P. Pandey, N. S. Bhadoria, and R. P. Yadav, "Review rating prediction using machine learning and nlp techniques," Materials Today: Proceedings, vol. 37, pp. 401–406, 2021.
[4] G. V. Kassab, J. M. Mancebo, and C. F. R. Gimenes, "Mining user reviews for software product improvement," Journal of Systems and Software, vol. 170, p. 110758, 2020.
[5] S. Gao, H. Zhang, and Y. Yang, "Analyzing google play app reviews for identifying user concerns," Empirical Software Engineering, vol. 24, no. 6, pp. 3266–3300, 2019.
[6] L. D. Sorbo, M. Linares-V´asquez, G. Bavota, and D. Poshyvanyk, "What's in a name?a study of human-assigned labels to mobile apps," IEEE Transactions on Software Engineering,
1067–1087, 2018.

Aniket Waghmare is a postgraduate student pursuing a Master of Computer Applications (MCA) degree at Trinity Academy of Engineering, Pune, India. His research interests focus on data analysis, natural language processing, and machine learning applied to mobile app ecosystems. He has practical experience working with Python libraries such as pandas, NumPy, scikit-learn, and NLP tools like TextBlob and spaCy. His current project involves analyzing Google Play Store reviews to extract meaningful insights through sentiment analysis, text summarization, spam detection, and review classification. Aniket is passionate about leveraging data-driven techniques to improve user experience and support app developers in making informed decisions based on comprehensive app store analytics.