Study of sentiment on Google Play Store Applications

Aayush Gupta[1] and Deepali Kamthania[2*]

[1]Les Transformation Learning Pvt. Ltd., #367/8, Urban Estate, Karnal, Haryana 132001

[2]School of Information Technology, Vivekananda Institute of Professional Studies, New Delhi 110034

E-mail: [1]aayushgupta2398@gmail.com and [2]deepali102@gmail.com

This paper proposes an approach to perform sentiment analysis on Google play store's applications considering Kaggle dataset. The dataset includes 10,000 applications with their ratings, reviews, number of download and other application related parameters. The dataset is pre-processed using a number of data cleaning steps, including data reduction, tokenization, stemming etc. For classification of sentiment polarity, a logistic regression model has been proposed. The logistic regression model has been extended to a tri-polar classifier. Accuracy of 81.1% was achieved for document-level classification of reviews. The top two applications, 'Candy Crush Saga' and 'Clash of Clans', have been selected for the analysis based on popularity. Even though the 'family' category has a higher number of applications on the store, games are much more popular on Google's play store. Results show that these apps, though popular, receive a higher number of negative reviews. The sentiment of play store users gives us an idea of how the applications market reacts, what are the needs and how to succeed in the android market. The opinion of the audience is an essential component for every business. These reviews prove to be useful for further development and improvement of the applications.

Keywords---Google Play Store, Sentiment analysis, Logistic Regression, natural language processing (NLP)

1. Introduction

Sentiment analysis (SA) is an application for natural language processing (NLP) aimed at analyzing and recognizing sentiment within a piece of text (Abu Farha, I., & Magdy, W., 2014). The development of sentiment analysis coincides with the growth of social media (i.e., reviews, forum discussions, and blogs) on the web (Zhang, L. and Liu, B., 2017). Mining useful knowledge from these corpora gives rise to the task of sentiment analysis. Advancements in sentiment analysis and opinion mining have developed new possibilities to improve information gathering interests (Islam, M.R., 2014). Since the early 2000s, it has been one of the most active research areas in (NLP) (Pang, B.and Lee, L., 2008) (Liu, B., 2012). In the past years, numerous papers have been proposed on sentiment analysis on topics on micro-blogging sites (Martin, W., Sarro, F., Jia, Y., Zhang, Y. and Harman, M., 2016) (Gomez, M. Martinez, M. Monperrus, M. and Rouvoy, R., 2015) (Chen, N., Lin, J. Hoi, S. C., Xiao, X. and Zhang, B., 2014) (Iacob, C. and Harrison, R., 2013). The opinion of the audience is an essential component for every business. Google play store is a common platform for Android users to download and install mobile applications. With about 2.1 million applications, it is the world's leading application store. The users can freely rate these applications and submit their reviews and suggestions about the applications. These reviews prove to be useful for further development and improvement of the applications (Martin, W., Sarro, F., Jia, Y., Zhang, Y. and Harman, M., 2016). Google Android is already a well-established community. It is tough for new application developers to sell their products on Google play store. There is an unremitting requirement for mobile software companies and application developers to develop products that meet users' requirements (Rizk, N.M., Ebada, A. and Nasr, E.S., 2015). For new developers, it is important to know about the successful applications, about the features that people like, and the latest trends in the Android sector. User's feedback plays a crucial role in application development industry. (Aldabbas H., Bajahzar A., Alruily M., Ali Adil Q., Rana M. Amir Latif, Farhan M., 2020). Information from user's feedback can't be extracted directly.(Sakshi Ranjan, Subhankar Mishra, 2020). The sentiment analysis and word weighting techniques on Google play store reviews help in assessing the factors essential for the improvement of future applications. Many Android apps and observed that by taking care of user reviews increases the chance of apps' success (Palomba, F., Linares-Vasquez, M., Bavota, G., Oliveto, R., Di Penta, M. Poshyvanyk, D. and De Lucia, A., 2015). SVM (Support Vector Machine) has been applied to classify VOOT app review into two categories to help the app owner to improve their ratings in the industry (Yadav, S. and Yadav, S., 2018). The two variations of the Naıve Bayes classifier has been built for categorization of Android app published on Google Play Store. It was observed that the Multinomial Naive Bayes algorithm performed better than the Bernoulli Naıve Bayes algorithm (Olabenjo, B., 2016). CLAP (Crowd Listener for releAse Planning) has been introduced for categorizing reviews into bug reports, clustering related reviews, and prioritizing reviews (Villarroel, L., Bavota, G., Russo, B., Oliveto, R. and Di Penta, M., 2016) (Mohammad, S. M., Kiritchenko, S. and Zhu, X., 2013). Rosenthal et al. (Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A. and Stoyanov. V., 2015) proposed methods with high accuracy for sentiment analysis of tweets. Guzman et al. (Guzman, E., Alkadhi, R. and Seyff, N., 2017) proved the SVM method proves to more precise as compared to Decision Trees (C4.5 algorithm) for filtering irrelevant tweets.

This paper proposes an approach to perform sentiment analysis on Google play store's applications using Kaggle dataset including 10,000 applications with their ratings, reviews, number of download and other application related parameters. On the basis of the popularity, a document-level classification has been performed on the top two applications.

2. Methodology

The following steps have been applied for sentiment analysis on Google play store's applications.

2.1 Data Collection

In the current study, dataset has been fetched from Kaggle.com [https://www.kaggle.com/lava18/google-play-store-apps]. The dataset gives the information on 10,000 applications available on the Android play store. The analysis has been performed using Python and NLTK (Natural Language Toolkit). NLTK uses NLP libraries to facilitate text processing for python programs including classification, tokenization, stemming, etc.[https://www.nltk.org/]

2.2 Data Preprocessing

After retrieving the data, it was preprocessed to be fit for further analysis. Data preprocessing methods used are as follows (Agarwal, V., 2015):

- Data cleaning - Firstly, data was cleaned, i.e., missing values were filled, redundant tuples, null/unidentified values or any sort of inconsistency was removed.
- Data reduction - After cleaning, all the irrelevant attributes were removed so that only relevant information is available. The unnecessary attributes like content rating, price, current version, etc. were removed.

- Sorting - The concise dataset was then sorted in a way to facilitate further analysis. In this case, the dataset was sorted according to the application categories from A to Z.
- Noise removal - All the stop words, punctuations, numbers and special characters are considered as noise that was removed from each review.
- Tokenization - Tokenization refers to splitting each string into a single word/token. Each of the reviews from the database under study was reduced to tokens.
- Stemming - Stemming refers to striping a token's suffix. For example, love, loving, loved, loves were all considered and stored as 'love'.

2.3 Logistic Regression Model

The logistic regression model has been used for classification. The bag of words features has been extracted using sklearn for model development. The logistic regression is similar to linear regression but used for prediction of categorical data. Logistic regression uses the logit link function.

Logit link function:

$$\ln\left(\frac{P}{1-P}\right) = \beta + \beta_1 X \qquad (1)$$

where,

P = probability of positive occurring,

$\frac{P}{1-P}$ = odds

ln = natural log

X is the predictor variable

β (beta coefficient) = standard regression coefficient and

β0 = True regression coefficient

While logistic regression is mainly used for binary classification, it has been extended to a tri-polar classifier. In logistic regression, a threshold is set. Values greater than or equal to the threshold are positive, values less than that are marked as negative.

In the experiment, the classifier's threshold was set to be 0. Our extended classifier segregates the data using this threshold. Each token in a document is then classified as negative/positive. Then using the equation (1), the polarity of the document is calculated. Documents with polarity score greater than 0 are marked as positive, documents with polarity score less than 0 are marked as negative and those with polarity score equal to 0 are marked as neutral. Nearly all the results were found to be correct.

The study is based on document level sentiment analysis considering 70% data for training and 30% for testing. The extracted features have been ranked on the basis of the Maximum Score algorithm. The frequency of each review is calculated with the word list. Each review is considered as a document with three polarities, i.e., positive, neutral and negative. The neutral refers to no opinion.

3. Results and Discussions

After data preprocessing, applications were grouped together on the basis of their categories. It was noted that about a quarter among all the apps on the Google play store was from the categories 'family' and 'games' as shown in Fig 1.

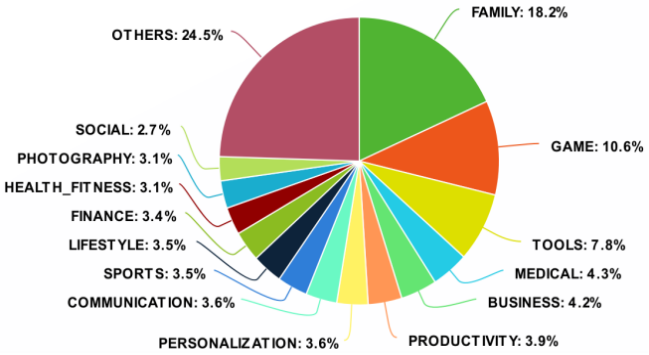

Figure 1: Pie chart illustrating the distribution of apps among various categories.

Fig1. shows that the Games account for 10% of the Google play store and attract most of the users.

Table 1. Number of applications per category and their respective shares.

Application category	Number of applications	Category's share on google play store (based on number of applications)
FAMILY	1829	18.2
GAME	1068	10.6
TOOLS	784	7.8
MEDICAL	433	4.3
BUSINESS	420	4.2
PRODUCTIVITY	390	3.9
PERSONALIZATION	361	3.6
COMMUNICATION	364	3.6
SPORTS	357	3.5
LIFESTYLE	358	3.5
FINANCE	340	3.4
HEALTH AND FITNESS	311	3.1
SOCIAL	315	3.1
OTHERS	2459	24.5
TOTAL	9789	100

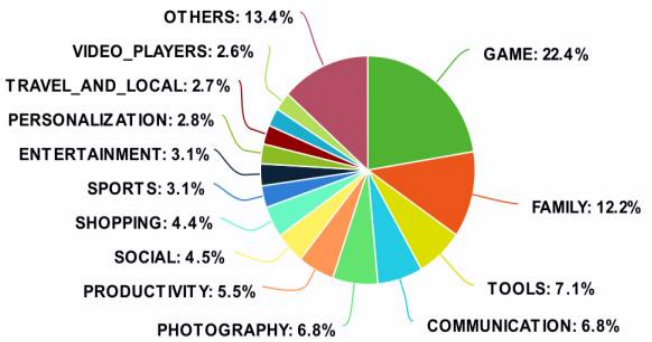

Figure 2: Pie chart illustrating the distribution of top applications (more than a million downloads)

Fig2. shows that among top applications based on the number of downloads, games top the charts with more than 22% of the applications. The top two games namely, 'candy crush saga' and 'clash of clans', both with more than 500 million+ downloads each, has been selected for the analysis.

Table 2. Number of applications per category and their respective shares based on the number of downloads.

Application category	Number of applications	Category's share on Google play store (based on number of applications)
GAME	466	22.4
FAMILY	254	12.2
TOOLS	148	7.1
COMMUNICATION	141	6.8
PHOTOGRAPHY	140	6.8
PRODUCTIVITY	114	5.5
SOCIAL	94	4.5
SHOPPING	92	4.4
SPORTS	65	3.1
ENTERTAINMENT	64	3.1
PERSONALIZATION	58	2.8
TRAVEL AND LOCAL	56	2.7
VIDEO PLAYERS	54	2.6
OTHERS	279	13.4
TOTAL	2025	100

From Table 1 and 2, it can be seen that the applications under the 'games' category are much more popular among the users than the 'family' category, which has the biggest share on the play store in terms of number of applications.

Figure 3(a): Word cloud for candy crush saga's reviews

Figure 3 (b): Word cloud for clash of clans' reviews

Figure 3(a) and 3(b) give us the word clouds for Candy crush saga and Clash of clans' reviews. Words like "good", "game", "love", "like", "playing", "treat", "time", "hours", "open", "support" were among the most used words. Some of the negative words include, "distracting", "annoying", "can't", "back", "problem", "never".

Opinion is calculated based on the documents. Out of 240 reviews of candy crush saga, negative reviews were 126, while positive and neutral reviews were 102 and 12 respectively. This suggests that top applications not only receive positive reviews, but a high number of negative reviews, that can further help in improvement of the application.

Table 3 - Confusion matrix.

		Predicted outcome	
Actual value		**P**	**N**
	P	82 (T_p)	18 (F_n)
	N	26 (F_p)	108 (T_n)

Table 3 gives the confusion matrix that is used to calculate the accuracy scores of the model. The four statistical parameters, namely precision, recall, accuracy and F-score, have been used for evaluation of the proposed approach Precision shows how much relevant information is identified correctly and

Recall shows how much extracted information is relevant.

F-score is the harmonic mean of precision and recall.

The formula to calculate precision, recall, accuracy and F-score:

$$Precision = \frac{T_p}{(T_p+F_p)} \quad (2)$$

$$Recall = \frac{T_p}{(T_p+F_n)} \quad (3)$$

$$Accuracy = \frac{(T_p+T_n)}{(T_p+T_n+F_n+F_n)} \quad (4)$$

$$F - Score = \frac{(2\times Preceision\times Recall)}{(Precision+Recall)} \quad (5)$$

where, Tp is true positive, Tn is true negative, Fp is false positive, and Fn is false negative. Table 4 gives us the calculated scores based on the equations (2), (3), (4), (5).

Table 4 - Performance evaluation of proposed method

Evaluation Metric	Value
Accuracy	81.1%
Precision	82
Recall	43
F-Score	56.416

4. Conclusions

From the present study, it has been observed that among all the apps on the Google play store, 'family' and 'games' categories are most popular. When categorizing on the number of downloads, games top the charts with more than 22% of the applications. The top two games are 'candy crush saga' and 'clash of clans', both with more than 500 million+ downloads. The document-level sentiment analysis has been performed on the two games, with the logistic regression with 81.1% accuracy. The analysis can be used for further development by the game development community. The study can also be extended for the applications in other categories to improve their popularity by comparing them with the ones mentioned.

REFERENCES

Abu Farha, I., & Magdy, W. (2021). A comparative study of effective approaches for Arabic sentiment analysis. Information Processing & Management, 58(2), 102438. doi:10.1016/j.ipm.2020.102438

Zhang, L. and Liu, B., "Sentiment analysis and opinion Mining", In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA, 1152–1161, (2017). doi:10.1007/978-1-4899-7687-1_907

Islam, M.R., "Numeric rating of apps on google play store by sentiment analysis on user reviews", 1st International Conference on Electrical Engineering and Information Communication Technology (ICEEICT 2014) At: Military Institute of Science and Technology, Dhaka, Bangladesh. DOI: 10.1109/ICEEICT.2014.6919058

Pang, B.and Lee, L., "Opinion mining and sentiment analysis. Foundations and trends in information retrieval", Foundations and Trends in Information Retrieval. 2(1-2), 1–135, 2008.

Liu, B., "Sentiment Analysis and Opinion Mining", Morgan &. Claypool, San Rafael Publishers, May 2012.

Martin, W., Sarro, F., Jia, Y., Zhang, Y. and Harman, M., "A survey of app store analysis for software engineering", RN. 16:02, 2016.

Gomez, M. Martinez, M. Monperrus, M. and Rouvoy, R., "When app stores listen to the crowd to fight bugs in the wild". In Proceedings of the 37th International Conference on Software Engineering (ICSE), Vol. 2, 567–570. IEEE Press, 2015.

Chen, N., Lin, J. Hoi, S. C., Xiao, X. and Zhang, B., "Ar-miner: mining informative reviews for developers from mobile app marketplace". In Proceedings of the 36th International Conference on Software Engineering, 767–778. ACM, 2014.

Iacob, C. and Harrison, R., "Retrieving and analyzing mobile apps feature requests from online reviews". In Mining Software Repositories (MSR), 10th IEEE Working Conference, 41–44, IEEE, 2013.

Rizk, N.M., Ebada, A. and Nasr, E.S., "Investigating mobile applications' requirements evolution through sentiment analysis of users' reviews". 11th International Computer Engineering Conference (ICENCO), 123-130, 2015.

Aldabbas H., Bajahzar A., Alruily M., Ali Adil Q., Rana M. Amir Latif, Farhan M., "Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews", Journal of Intelligent Systems, Volume 30: Issue 1, 2020

Sakshi Ranjan, Subhankar Mishra, 11th ICCCNT, IIT Kharagpur, arXiv:2006.09739v1 [cs.IR], 2020

Palomba, F., Linares-Vasquez, M., Bavota, G., Oliveto, R., Di Penta, M. Poshyvanyk, D. and De Lucia, A., "User reviews matter! tracking crowdsourced reviews to support evolution of successful apps". In Software Maintenance and Evolution (ICSME), 291–300. IEEE, 2015.

Yadav, S. and Yadav, S., "Text mining of VOOT application reviews on google play store", International Research Journal of Engineering and Technology (IRJET), 05 (01), 1024- 28, 2018, e-ISSN: 2395-0056

Olabenjo, B., "Applying Naıve Bayes Classification to Google Play Apps Categorization", (2016), https://arxiv.org/pdf/1608.08574

Villarroel, L., Bavota, G., Russo, B., Oliveto, R. and Di Penta, M., "Release planning of mobile apps based on user reviews". In Proceedings of the 38th International Conference on Software Engineering, 14–24. ACM, 2016.

Mohammad, S. M., Kiritchenko, S. and Zhu, X., "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets". arXiv preprint arXiv:1308.6242, 2013.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A. and Stoyanov. V., "Sentiment analysis in twitter". Proceedings of SemEval-2015, 2015

Guzman, E., Alkadhi, R. and Seyff, N., "An exploratory study of twitter messages about software applications. Requirements Engineering", 22(3):387–412, 2017.

Agarwal, V., "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis", International Journal of Computer Applications, (0975-8887), Volume 131- No. 4, 2015.