International Journal of Application on Sciences, Technology and Engineering (IJASTE)
Volume 1, Issue 1, 2023. ISSN 2987-2499

# Logistic Regression Method for Sentiment Analysis Application on Google Playstore

Viny Christanti Mawardi[1, a)], and Edward Darmaja[1]

[1]*Computer Science Department, Faculty of Information Technology, University Tarumanagara*

a) *Corresponding author: viny@untar.ac.id*

*Submitted: November-December 2022, Revised: January 2023, Accepted: February 21, 2023*

**Abstract**. Now days, there are a lot of social media. Each of social media has their one purpose. Each of application has their own function. We can download social media application from Playstore like Google Play or Appstore. Each Playstore gives users the opportunity to provide rating and review. Ratings and reviews from users can give conclusions about the application. Applications with a rating of 5 usually have good comments so that they are more worthy to be downloaded and used. Ratings with a small value cause other users not to use the application. Application owners need to know the results of comments from users which sometimes do not match the value given in the rating. So that the owner or application developer conducts an analysis of user comments on various social media, one of which is on the Playstore. Social media analysis is one way to get public opinion on a topic. One of them is used to analyze various applications provided in the Playstore. Various methods are used to perform social media analysis. This study uses the logistic regression method to analyze public comments into positive or negative comments. We analyze comments from social media applications on Google Play for android users. We get an accuracy value of 81% from 4 social media applications, with a total of 2268 comment data.

## INTRODUCTION

Facebook Lite, Instagram, Twitter, TikTok are various social media applications that we can get from the Playstore and are used to interact digitally via the internet. Various applications have certainly been widely used by many users. In general, when we will use an application, we will ask what the function of the application is for. In addition to the functionality of the application, we need to know how good the application is. Are there any problems in using the application or is it easy enough to use it?

How other users experience using the application is one of the opinions that helps us to see what the benefits and advantages of the application that we will download are. Currently, one way to get information regarding other people's opinions or comments on a product is to use social media. By reading comments on social media, we will read public opinion regarding a product. But when we read the comments, sometimes what we read is different from the existing rating value. In addition, this massive digital world causes a large number of comments that appear on social media and it is difficult for us to summarize them.

Social Media Analysis is the process of collecting data from social media channels and analyze the data to make conclusions. There are several things that can be analyzed from social media such as analyzing shared media, analyzing conversations or analyzing the network formed. One of the communications that can be obtained for analysis is in the form of opinions. Opinions can be in the form of sentiment sentences. This sentiment will be analyzed to find out public opinion on a product (Neri et. al, 2012). Sentiment analysis can be done by collecting reviews from users on social media and then classified using various methods (Yue et. al, 2019).

Based on the review that the makers received, the makers can find out what the users are thinking and predict what they want and find out why the product they make is not in demand or why the product is widely used. From this, the company can use it as a reference for the future to improve the quality of the products. Although humans can distinguish whether the review text contain a positive or negative response, machine cannot understand the implied meaning of the text without prior instructions, that is the reason why this research is carried out.

Conducting sentiment analysis is something that humans can do by reading, but for massive data it is very difficult for humans to do it manually. Currently, there are many applications and methods used to perform social media analysis. Machine learning methods, data mining, discourse analysis and others. Machine Learning will answer questions about how a program on a computer or machine can improve a given performance based on the

experience of tasks that have been given previously (Jordan and Mitchell, 2015). The algorithms used in Machine Learning have also been proven to have good practice values in various fields, one of which is the application of Data Mining, in this case there is a large database containing a lot of irregular data, where with the use of Machine Learning valuable data or important points can be found automatically (Jordan and Mitchell, 2015).

Google Playstore can be categorized as a social media application because there is a menu for interacting by providing rating values or comments between users on existing applications. As we can see in fig. 1, user can share review and other can reply the review and people can see each other about Instagram application. To analyze this, we can use a lot of method or application. In this research we try to create application to help user analyze application review from Google Playstore. For analyze the sentiment first we need to gather a lot of data. So, we try to create application with RPA to help gathering data easily without any other API for scraping.

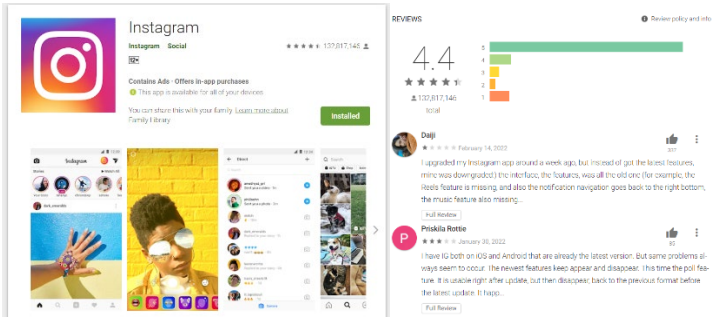


**FIGURE 1.** Example of Instagram Review on Google Play

Here are some similar studies that have been done before. In a study conducted by Pang, Lee, and Vaithyanathan, proposed the use of sentiment classification using machine learning in movie review datasets. The analysis uses the Naïve Bayes method, Max Entropy and Support Vector Machine (SVM) models on unigram and bigram data. In this experiment, the results obtained from the use of SVM with unigram with feature extraction resulted accuracy of 82.9% (Nguyen et. al., 2018).

In a study conducted by Warnia Nengsih, M. Mahrus Zein, and Nazifa Hayati, they conducted a sentiment analysis on hotel reviews. In this study, the Random Forest (RF) method was used as a review classification method. The results obtained are for a prediction accuracy rate of 90% for positive reviews of 68% and negative of 32% (Nengsih et. al., 2021).

In a study conducted by Emilie Coyne, and Jim Smit, conducted a sentiment analysis of a collection of product reviews from Amazon. In this study, a comparison of several methods such as Linear SVM, Multinominal NB, and LSTM network was used as the classification of the reviews. From the prediction results given, the three methods obtained a very large prediction accuracy value, all of which were above 90% and it proved to be very accurate and good to use (Güner et. al., 2019).

Logistic Regression is of another machine learning method that can be used for classified sentiment such as in tweet with 92% accuracy (Indra et. al., 2016). Logistic regression can be used to analyze sentiment in another conversation social media too. This relativity is considered using the logistic regression model and the accuracy of the results is found to be improved significantly (Bhargava & Katarya, 2017). Beside that, logistic regression is the most popular rapid classification method that can be used for classified pro and contra of sentiment that has been studied extensively (Zhang et. al., 2020).

## METHODOLOGY

In this research, there are three step that we used to create sentiment analyzer for classified the sentiment. First, scraping data, second is preprocessing data and classified with logistic regressing. We can see our scheme at fig. 2. The RPA with UiPath will help us to gather data from Playstore automatically and create the list of sentiment in excel. After we get the data, we preprocessing the sentiment and then we classified in two category which is positive or negative with logistic regression method.

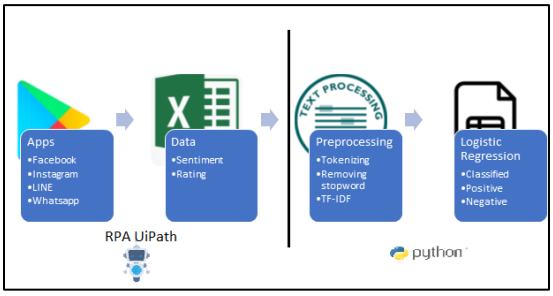


**FIGURE 2.** Scheme of our experiment

## SCRAPING DATA

Web scraping is one of method to mine information from different and unstructured websites and transform it into a comprehensible structure like spreadsheets, database or comma-separated values (CSV) (saurkar et. al., 2018). There are various technologies that can be used to collect data from the web, such as a web scraper (a web that can be used online for scraping), a stand-alone scraper application, or an API that is provided directly to access web information (Manjushree & Sharvani, 2020). Some of application, provide their own API for scrap the data such as Youtube can be used to take the comments of users from comment section (Christanti et. al., 2020). We need to learn about each of technology because there is different way to used it. Sometimes we need pay the application for used it. So, we try to scraping data with RPA, so we can get the data more flexible.

Robotic process automation (RPA) has been widely adopted in many industries to automate well-defined and repetitive tasks. We can screen scraping or web scraping with RPA like doing collecting data, synthesizing it, and putting it into some sort of document on a desktop that can be automate as much of that as possible (Tripathi, 2018). UiPath is one of RPA software that allows users to perform business processes that are run automatically with the help of robots. From this it will reduce the time and process undertaken by man. In this research, UiPath is used to collect data from Google Playstore as a required dataset to be used.

At fig. 3 we can see the UiPath code and review data from Facebook Lite. After we scrap the data with UiPath, the list of review will be saved at Excel file. All of the data from Google Playstore will automatically saved at excel file and we can arrange the quantity of the data that we will scrap.

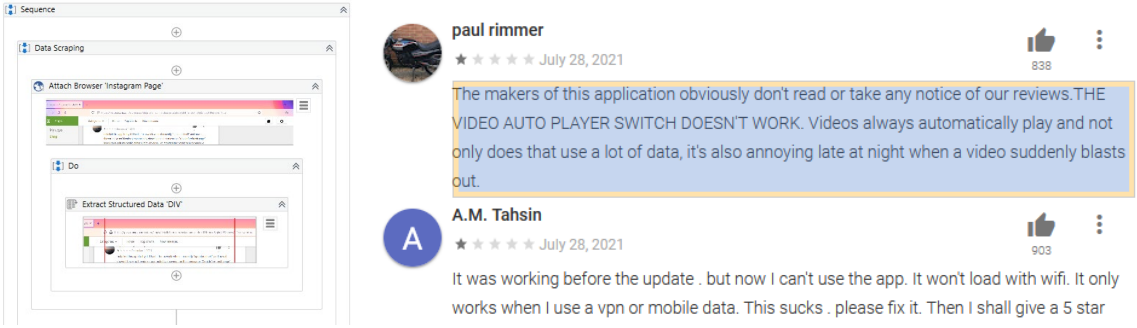


**FIGURE 3.** UiPath Scraping Data

## PREPROCESSING DATA

Sentiment Analysis, is a process in determining whether a text is positive or negative. In conducting sentiment analysis, it is necessary to carry out text data processing stages, one of which is preprocessing. There are several things that will be done at this stage such as sorting reviews, then using the Count Vectorizer which by default will lower case, tokenization, and remove stop words.

TF-IDF or it can be said "Term Frequency - Inverse Document Frequency" is a technique used to measure words in a document by calculating the weight of each word in a document. This technique is widely used in Information Retrieval and Text Mining. There are several stages that will be carried out which include (Scott, 2019):

1. Term Frequency (TF)
   This is a step to calculate how often a word (term) appears in a document. For how much the frequency of the word depends on how long the text or document you have (eq.1).

$$tf(t,d) = count\ of\ t\ in\ documents \tag{1}$$

2. Document Frequency (DF)
   Almost the same as in the previous stage, the difference is that DF will count the occurrences of each previous word based on how many occurrences of that word in a document (eq. 2).

$$df(t) = occurrence\ of\ t\ in\ documents \tag{2}$$

3. Inverse Document Frequency (IDF)
   This last stage is a stage to calculate the weight of a text or document based on the results of the TF and DF acquisition in the previous step. There are several formulas that need to be done to obtain these weights, which include eq. 3 where: t is term (word), d is document (Number of documents) and DF is document frequency by word. And eq. 4 where W is weight (word weight), TF is frequency of occurrence of words and IDF is Inverse Document Frequency.

$$idf(t) = Log(\frac{d}{DF}) \tag{3}$$

$$W = TF * (IDF + 1) \tag{4}$$

## Logistic Regression

Logistic Regression is a mathematical model whose approach can be described by the relationship from several variables X to variable D which is a dichotomous dependent variable. Dichotomous dependent variable means that a variable has only two choices, which mean one must only choose from one option when making observations or measurements (Indra et. al., 2016) (David G. Kleinbaum & Mitchel Klein, 1994).

Logistic Regression is popular for use because of the results of the logistic function f(z) where the results given are between 0 and 1. The model is designed to explain the probability which is always between 0 and 1. One example is the probability of an individual being infected. a disease. Logistic function are between values 0 and 1, even though the input value for the z variable itself has different input values (David G. Kleinbaum & Mitchel Klein, 1994).

For the system that will be designed using this method to divide the reviews into two categories, namely whether the reviews fall into the positive or negative categories, the neutral itself will not be used because the Logistic Regression results obtained are 0 and 1 for positive and negative categories, so neutral will not be used. The categorization can be obtained with the model output from the Sigmoid Function or Logistic Function used in Logistic Regression.

After obtaining the results of the weights of a review using the TF-IDF method, the results of these weights will then be used in the Sigmoid Function. Sigmoid Function is an important part for making Logistic Regression model which is defined as eq.5 where z is input data and after the function is used, the results will then be compared to determine the category with the general criteria as follows: (1) If Result > 0.5 then the prediction result is 1, (2) If Result < 0.5 then the prediction result is 0.

$$sigmoid(z) = \frac{1}{1+e^{-z}} \tag{5}$$

# RESULT

From the results of the research carried out, a desktop-based application using the Python language was created, where the application will be used as a tool for analyzing sentiment reviews of a social media application from the Google Playstore. The data that will be use is in the form of an Excel data file containing the name, review, and score given by the reviewer. At fig. 4 we can see the result of extraction data that consist of review comment from Google Playstore.

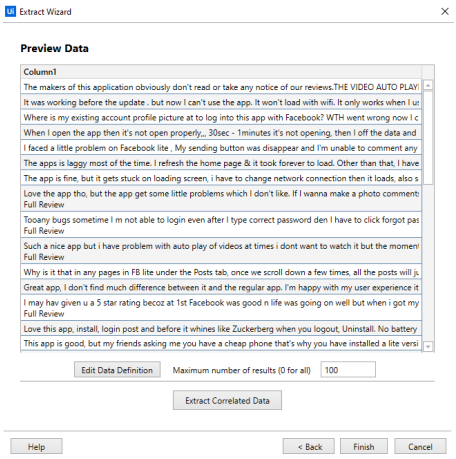


**FIGURE 4.** Review comment

We conducted experiments using 2704 data consist of reviews from social media that has been collected from Google Playstore, the accuracy obtained was around 85.76%. Next, we did the experiment 4 times for random data training and get the best accuracy at 88.13%. The experiments results can be seen in TABLE 1.

**TABLE 1.** Training Result

| Name | Data Use | Positive | Negative | TP | FP | TN | FN | Acc (%) |
|---|---|---|---|---|---|---|---|---|
| Training | 2704 | 1251 | 1453 | 1102 | 236 | 1217 | 149 | 85.76 |
| Validasi1 | 450 | 151 | 299 | 88 | 31 | 268 | 63 | 79.11 |
| Validasi2 | 450 | 102 | 348 | 51 | 19 | 329 | 51 | 84.44 |
| Validasi3 | 414 | 139 | 275 | 104 | 19 | 246 | 35 | 84.54 |
| Validasi4 | 430 | 69 | 361 | 37 | 19 | 342 | 32 | 88.13 |

The example of comment that get the true category such as: “I loved the app it saves me extra money and i can see my loved ones and talk to them for free.”, “Awesome and free for international calls and chats. It's also secure and encrypted. Love it!”. And the sentence that has been false prediction such as: “All of a sudden, the only thing I can do is send messages. Can't download or upload media. No Gifs. No calls. Nothing I searched for has helped me resolve the issues. Update: Got a brand-new phone. Now it's media friendly but I can't get notifications. All my other apps work fine.”, “Unable to send messages As of today morning, my messages to contacts are not getting delivered. My internet connection is fine - everything else works. I restarted the phone and updated the app as well but no use.”. We can see from the comment that the false comment has a lot of word compare to the true category prediction but the sentences have dotted sentences formatted (separate sentence with different meaning).

After the training is carried out and the model is formed, the next thing to do is to test 4 social media applications which are Facebook Lite, Line, WhatsApp and Instagram. The results can be seen in **TABLE 2.** From the results of the tests that have been carried out, the accuracy for Facebook Lite is 77.99%, Line is 81.81%, WhatsApp is 81.23%, and Instagram is 83.30% with an average accuracy is 81.08%.

**TABLE 2.** Testing Result

| Nama | Data Use | Positive | Negative | TP | FP | TN | FN | Acc (%) |
|---|---|---|---|---|---|---|---|---|
| Facebook Lite | 568 | 283 | 285 | 207 | 49 | 236 | 76 | 77.99 |
| Line | 572 | 283 | 289 | 218 | 39 | 250 | 65 | 81.81 |
| WhatsApp | 565 | 285 | 280 | 225 | 46 | 234 | 60 | 81.23 |
| Instagram | 563 | 225 | 338 | 189 | 58 | 280 | 36 | 83.30 |

It can be seen that the accuracy results for the Instagram application result in a more accurate classification of comments. When viewed in the confusion matrix at fig. 5, it can be seen that the number of classifications that are falser is in the positive comments. But for many negative comments get the right prediction compared to other social media applications. After re-analyzing the 4 comments from the social media application, it can be concluded that Instagram has a long and complete sentence compared to comments from 3 other social media applications. Comments on Playstore to the Instagram application, has an average sentence length of 60 words. While in other applications, it only consists of 14 words for each comment and sometimes has a truncated comment.

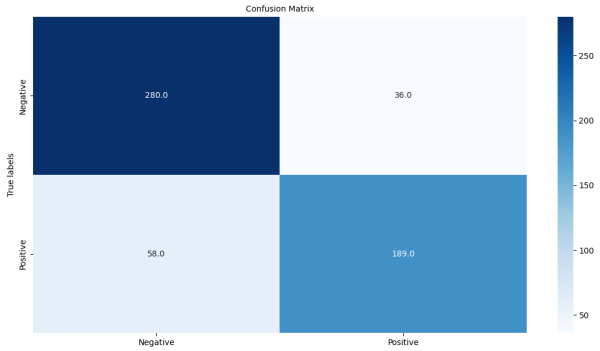


**FIGURE 5.** Instagram Confusion Matrix

## CONCLUSION

From the use of the TF-IDF and Logistic Regression methods, a result can be obtained where the reviews are positive or negative, which will then be applied to the overall reviews that have been collected. These results can then be drawn a conclusion, where the information obtained from the results of this sentiment analysis can be concluded whether a social media application being analyzed has a good or bad reputation based on the predictions of using the method in the reviews given by its users. The best accuracy result of our research is 83.3% for prediction of Instagram app comment. The disadvantage of this accuracy is the imbalance in the number of sentences in each comment, causing prediction errors. For further research, it can be developed by considering from the side of the lexicon and sentence so that there can be a balance of sentences for each category.

This research has produced a sentiment analysis application from comments on applications in the Playstore and displays the analysis in the form of pie charts and graphs. This study also uses UiPath as a tool to collect comment data from Playstore. The author hopes that from this research, can be useful for users who want to know the reputation of a social media application without having to read all the existing reviews, and also this research is a form of evaluation for the method used. the author hopes that in the future there will be research with similar case but using a different method to see a comparison between the methods used by the authors and other.

## ACKNOWLEDGMENTS

We would like to thank all those who have assisted in this research process. This research is part of a student's final project on behalf of Edward Darmadja at the Faculty of Information Technology, Tarumanagara University.

## REFERENCES

1. Kleinbaum, David G., K. Dietz, M. Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. New York: Springer-Verlag, 2002. h. 5 - 6.
2. Güner, Levent, Emilie Coyne, and Jim Smit. "Sentiment analysis for amazon. com reviews." Big Data in Media Technology (DM2583) KTH Royal Institute of Technology 9 (2019).
3. Nengsih, Warnia, M. Mahrus Zein, and Nazifa Hayati. "Coarse-grained sentiment analysis berbasis natural language processing–ulasan hotel." *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* 10, no. 1 (2021): 41-48.
4. Nguyen, Heidi, Aravind Veluchamy, Mamadou Diop, and Rashed Iqbal. "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches." *SMU Data Science Review* 1, no. 4 (2018): 7.
5. Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245 (2015): 255-260.
6. Neri, Federico, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. "Sentiment analysis on social media." In *2012 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 919-926. IEEE, 2012.
7. Yue, Lin, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. "A survey of sentiment analysis in social media." *Knowledge and Information Systems* 60, no. 2 (2019): 617-663.
8. Indra, S. T., Liza Wikarsa, and Rinaldo Turang. "Using logistic regression method to classify tweets into the selected topics." In *2016 international conference on advanced computer science and information systems (icacsis)*, pp. 385-390. IEEE, 2016.
9. Bhargava, Kunal, and Rahul Katarya. "An improved lexicon using logistic regression for sentiment analysis." In *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, pp. 332-337. IEEE, 2017.
10. Zhang, Xu, Rui Pan, Guoyu Guan, Xuening Zhu, and Hansheng Wang. "Logistic regression with network structure." *Statistica Sinica* 30, no. 2 (2020): 673-693.
11. Christanti, M. Viny, and Tri Sutrisno. "Comments Scraping Application for Review Youtube Content." In *IOP Conference Series: Materials Science and Engineering*, vol. 852, no. 1, p. 012167. IOP Publishing, 2020.
12. Saurkar, Anand V., Kedar G. Pathare, and Shweta A. Gode. "An overview on web scraping techniques and tools." *International Journal on Future Revolution in Computer Science & Communication Engineering* 4, no. 4 (2018): 363-367.
13. Manjushree, B. S., and G. S. Sharvani. "Survey on Web scraping technology." *Wutan Huatan Jisuan Jishu* 16 (2020): 1-8.
14. Tripathi, Alok Mani. *Learning Robotic Process Automation: Create Software robots and automate business processes with the leading RPA tool–UiPath*. Packt Publishing Ltd, 2018.
15. Scott, William. "TF-IDF from scratch in python on real world dataset." *Towards Data Science* 15 (2019).