COMPARISON OF SUPPORT VECTOR MACHINE(SVM), XGBOOST AND RANDOM FOREST FOR SENTIMENT ANALYSIS OF BUMBLE APP USER COMMENTS

[1]Oktafia, [2]Robertus Setiawan Aji Nugroho
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]nugroho@unika.ac.id

ABSTRACT (ABSTRACT TITLE)

Sentiment analysis is the process of classifying texts into positive or negative sentiments. This is a common process in natural language processing (NLP), and has applications in areas such as customer feedback, product reviews, and social media. In this paper, the authors compare the performance of three different machine learning algorithms for sentiment analysis namely the Support Vector Machine (SVM), XGBoost, and Random Forest. The author evaluates the algorithm on a collection of review data from the Bumble application which is the first rated dating application on the Google Play Store where this dating application allows users to swipe left or right on potential partners, as well as various other interesting features provided to communicate with partners. From the research results the authors found that the Random Forest achieved the best performance, with an accuracy of 85.76%. SVM and XGBoost achieved 85.58% and 84.14% accuracy respectively. The results of this study indicate that Random Forest is a good choice for sentiment analysis tasks, especially when data is limited.

Keyword: SVM, XGBoost, Random Forest, Stemming, Sentiment Analysis

1. INTRODUCTION

GooglePlay Store offers users the option to download and install apps on their android devices for free. There are many apps in the Google Play store, for example, games, online shops, entertainment or whatever you are looking for. Users can search for applications in categories or use the search system provided to find their own applications. In addition, a rating system for apps is available on the Google Play Store that allows users to judge and comment on their app. This rating system is very useful for determining which apps users are interested in downloading and to help them find the best ones.

Bumble is a dating app that allows users to scroll left or right on potential partners. It is the most popular dating app in the world with more than 50 million users. Bumble has a special system, where this application requires women to take the first step. It aims to provide women with a more equitable relationship experience than might be expected for those taking the first step, as well as being perfect for introverted men.The analysis of sentiment is a procedure to classify the text as either positive or negative. In the context of natural language processing, sentiment analysis has become a basic task and is being used in areas like customer feedback, product reviews as well as Social Media. It is difficult to analyse emotions in this way, because the model must be able to interpret a message and make accurate predictions of its emotion. There's a lot of ways to analyse sentiment, like machine learning, rule systems and hybrid systems.

Sentiment analysis is a useful tool for understanding people's views on various topics. It can be used to improve customer service, product development and marketing campaigns. To study machine learning systems, the author will use Support Vector MachineSVM, XGBoost and Random Forest algorithms. From previous research using SVM and Naive Bayes algorithms, I feel that doing a comparison with just two algorithms is like tossing a coin where one side will automatically be a good algorithm and the other side will be an algorithm that is not suitable for use in sentiment data types. In this study the authors used XGBoost and Random Forest as a comparison with SVM which are algorithms that have high accuracy results from previous studies, because XGBoost and Random Forest are also recommendation algorithms from several existing studies and also produce high scores for sentiment analysis such as this.

2. LITERATUR STUDY

Fajar and Iqbal [1] analyze sentiment data from Twitter posts about the COVID-19 pandemic. SVM is the most accurate, with 85% accuracy. The weakness of this paper is that it does not provide a detailed explanation of the methodology used in the research process. The authors only say that they used inexperienced SVM, Naive Bayes and KNN, but no information is provided regarding their hyperparameters or training data. Suggestions for this paper are researchers who will develop this paper can discuss the possibility that the results are affected due to the use of Twitter data. In addition, researchers can carry out follow-up studies based on other data sets such as customer reviews.

Hanna and Fahmi [2] discusses the use of Support Vector Machine (SVM) with the Naive Bayes kernel for data classification. The writer finds that data classification can be achieved satisfactorily by SVM with Naive Bayes kernel. However, the drawback of this paper is that SVM and Naive Bayes are sensitive to hyperparameter selection. The authors suggest that future research should focus on developing methods to improve SVM robustness using the Naive Bayes kernel to select hyperparameters.

Karina [3] explains how to analyze positive and negative comments on YouTube using svm and naive bayes algorithms. In terms of writing and others, this journal is very complex and clear, perhaps this journal can be further developed for further research using different media platforms and is currently trending among students and the general public.

Heart and Achmad [4] discuss the implementation of Support Vector Machine (SVM) algorithm for sentiment analysis of Twitter users' tweets on the PSBB policy. The authors found that SVM can be used to effectively classify tweets into positive and negative sentiments. However, the authors also found that SVM is sensitive to the choice of hyperparameters. The authors suggest that future research should focus on developing methods to improve the robustness of SVM to the choice of hyperparameters.

Zidna at al [5] talks about the public's response to the marketplace, namely Bukalapak from community comments on twitter. The process that the author does is almost the same as the journal I discussed above, and maybe to develop this journal we can try to compare the opinions of the public using other e-commerce platforms such as shopee, tokopedia, lazada, blibli and jd.id. The drawback in writing this report is that the author does not include examples of the twitter comment data obtained, which makes it a little difficult for ordinary people who don't really understand sentiment analysis like this.

Nur at al [6] talks about how users of the Go-Jek application respond to Go-Jek services while being a user. By using the svm algorithm the author categorizes people's opinions, especially for gojek users to find out the comparison of results between positive and negative responses from all data collected from twitter. The gap/limitation of this journal is that there is no data on the results obtained from twitter so we don't know examples of word or sentence forms that are obtained from twitter. Sentiment classification results from manual data labeling using the SVM method on Gojek produced the best overall accuracy rate of 79.19% and the best kappa accuracy of 16.52%.

Aliffia at al [7] talks about application of the SVM algorithm and information gain in the sentiment analysis process regarding the implementation of the regional elections during the pandemic. From the results of the analysis carried out by the author, it was obtained that there were many negative responses from the public regarding election policies in pandemic conditions. There are no gaps and limitations for this report because all the data has been contained in the report, the grammar is easy to understand..

Made [8] The paper provides an insight into the effect of Indonesia's stemmer in performing sentiment analysis on translated movie reviews. The authors have identified that Indonesian stem cells are able to enhance the performance of sentiment analysis. However, the authors also observed that sensitivity to choice of hyperparameters is a key factor for Indonesia stemmers' performance. The authors recommend that future research will concentrate on developing methods for improving the robustness of Indonesian stem cells when choosing hyperparameters.

Widodo and Dina [9] talk about how to give advice to students who have not submitted a thesis so that it becomes a reference for the field to be researched based on student value data. By using the k-means algorithm, I think the author is right in choosing the algorithm used, because the type of data studied is data that must be grouped to reach the desired decision. But in writing this report the author does not specify what if the student has submitted a title in the absence of a comparison of the interests of students who have submitted personally with those who have not submitted at all. The limitation in this journal is that student scores may not match their abilities.

Oman at al [10] Talks about how to optimize the selection of the final project theme for final semester students who will take the final project. By using the dataset of D3

Computer Engineering student journals at Harapan Bersama Tegal academic year 2014/2015 and the algorithms used are Text Mining, Support Vector Machine, K-Means, and Software Rapid Miner 5.0. This journal has a similar topic to the first journal that I researched, but I think these two journals can be used as a reference if you want to raise the topic of the same thesis so that the dataset studied can be more complex and produce the desired conclusions.

3. RESEARCH METHODOLOGY

3.1 DATASET COLLECTION

The dataset used in this study was taken using the google-play-scraper from the bumble application with a total of 2878 data consisting of 1199 negative data, 210 neutral and 1469 positive data. The dataset taken contains the username, when the review was sent, the contents of the review and the star value given.

The first step is to install google-play-scrapper because Google Colab is not yet available. Google scraper itself is a method of collecting data from applications provided by the Google Play Store. Next is the process of importing the library that will be used and you can see there is a library app and sort review all from Google scrapper to retrieve review data from the application as well as pandas and numpy for data processing.

Next is the data collection process. The author only takes Indonesian users, therefore, fill in 'id' from the bumble application with the following link https://play.google.com/store/search?q=bumble&c=apps&hl=id-ID with the amount of data taken as many as 3000 data with no filter score so that the resulting data will have a rating of one to five stars.At first the data taken was 3000 data but there were data whose contents were only question marks and would cause bias during the labeling process so the authors decided to delete them from the dataset before the dataset was labeled.

3.2 LABELING DATASET

The next process is dataset labeling. Dataset labeling was done by three friends. Data labeling is done using google spreadsheet which is easily accessible by the labeler. Labeling is done for approximately 3 working days. This labeling process is divided into 3 categories, namely 1 means negative, 2 is positive and 3 is neutral. Then from the three people a conclusion was drawn by having a sentiment column containing the final labeling results which were also divided into 3 categories, namely -1 means negative, 0 means neutral and 1 means positive. Drawing conclusions in this sentiment column is based on the labeling column that has been carried out by the labelers, if the values of all labelers for one review have the same value then the conclusions can be used as a dataset to be processed because there is no possibility of bias. However, if the label values one, two and three are different or there are different label values, the conclusion will be biased because it is not certain which category the review should fall into. Data like this cannot be used in research because it has a bias value which will result in low algorithm results because the data is difficult to predict. Therefore such biased data was deleted by the authors in order to reduce the risk of bias. Even so, it is possible that the resulting data still has the possibility of bias because this labeling is done by lay people and there are no linguists.

3.3 PRE – PROCESSING DATA

3.3.1 Case Folding

Case folding is the process of converting all characters to lowercase in a text. This stage will be assisted with the help of the RegEx library. In this process the characters 'A'-'Z' contained in the data are changed to characters 'a'-'z'. For example, there is a sentence "I REALLY LIKE this flower" then using the `str.lower()` method, the sentence will become "I really like this flower".

3.3.2 Tokenizing

The tokenization process divides text into individual words or phrases. This is a common pre-processing step in sentiment analysis, as it helps identify each word used to express a sentiment. At this stage it will be assisted by the NLTK library. An example of the tokenization process is if there is a sentence "I like this application, but I'm still confused about using it" then by using the word_tokenize() function from the nltk library, the sentences will become 'I', 'like', 'application', 'this', ' ,', 'but', 'I'm', 'still', 'confused', 'using it'.

3.3.3 Formalization

The formalization stage is carried out to change the use of non-standard words to become standard according to KBBI. The process will use a slangwords dataset file that contains slang words which will later be changed to default. This stage is assisted by the RegEx library.

Here is an example of how formalization can be used in sentiment analysis:

Original text: I love this movie!

Formal text: I, love, this, film

As you can see, the formalization process has removed the noise from the text and made the algorithm easier to learn. Algorithms can now use formal representations of text to predict new text sentiments with greater accuracy.

3.3.4 Stopword Removal

Stopword is the process of removing words that do not have an important contribution in a sentence. Examples of affixes like "and", "but", "a" and many more. This process will use the stopword function assisted by the NLTK library. Example: "I like to read, so I read that book" will become "like,read, read,book".

3.3.5 Stemming

Stemming is the process of removing word affixes into basic words. In the process of stemming here, we use the help of a literary library and swifter. An example of how this process works is, if there is the word "to have" then it will be the word "belong" or there are the words "to be loved" and "to love" both of which have the same root, namely "love".

3.3.6 TF-IDF

TF-IDF is a statistical measure used to evaluate how important a word is in a document or in a group of words. In this process the function will identify how important a word is in a sentence, for example there is the word "like" it will have a high TF-IDF score in the positive class and if there is a sentence "bad" it will have a high TF-IDF score in negative class.

3.3.7 Support Vector Machine(SVM)

Support Vector Machines SVM is a type of automated machine learning algorithm that can be used to solve classification and regression tasks. For tasks that allow data to be separated linearly, SVM is the most appropriate programming language because lines and fields can be drawn to separate it into two classes. The SVM method, which allows data to be divided into two groups, is the optimal way to find hyperplanes. Hyperplanes are lines or planes derived from a set of weights and biases. The aim of the SVM shall be to determine a hyperplane with the widest possible range in terms of both classes. The distance between the hyperplane and the nearest data point is the margin. The higher a margin, the more certain SVMs are of their classification. Using Twitter text or reviews as data inputs, SVM can be used to analyse sentiment. Once trained, SVMs are capable of classifying text into two classes: positive or negative.

User review classification is carried out using the Support Vector Machine algorithm which will be assisted by the Scikit-Learn library. A number of machine learning algorithms are available in the sklearn library, such as a support vector machine. The SVC class implements the SVM algorithm. Other classes in the library provide functionality for scaling data, training models, and making predictions. In this study, the authors still use the library, but not in full, alias using code whose process is similar to the sklearn library itself. The classification process uses training data and test data values 80%:20%, 70%:30%, 60%:40% and 50%:50% and is also carried out in 5x experiments and the kernel will use the rbf kernel and the linear kernel.

3.3.8 XGBoost

XGBoost is a gradient enhancement algorithm designed to be fast and accurate. It is an ensemble tree based method that builds on the idea of gradient enhancement. Gradient boosting is an iterative algorithm that starts with a simple model, such as a linear model, and then iteratively adds new models to the ensemble to increase prediction accuracy. The user review classification uses the XGBoost algorithm with the xgboost library, which is theoretically based on a decision tree but uses a gradient boosting framework. In this study the classification process will be carried out using training data and test data values of 80%:20%, 70%:30%, 60%:40% and 50%:50% and 5 experiments will be carried out.

3.3.9 Random Forest

Random Forest is a machine learning algorithm that combines the output of several decision trees to achieve a single result. The program is intended to train a large number of decision trees on different subsets of the data and then combine these decisions with their predictions for each tree to get the overall prediction. A random forest can be used when you have a lot of noisy data, or no values in it. They are also relatively simple to learn and decipher, meaning they are often more precise than most ensemble methods such as bagging or bracing. Random forests are a concept that allows the training of multiple trees for various subsets of data, thereby reducing variance in one decision tree. To this end, a random sampling of data and features applied to each tree has been performed. Random sampling shall help prevent trees from being pulled together in a manner which can lead to excessive fitting. Once that is done, a final forecast will be drawn up based on results from each tree. In that case, the allocation of classification is carried out by a majority vote and average forecasts are taken into account.

Classification of reviews using the Random Forest algorithm assisted by using the Scikit-Learn library. The classification process uses training data and test data values of 80%:20%, 70%:30%, 60%:40% and 50%:50% and also 5 experiments are carried out.

3.4 EVALUATION

The three algorithms are evaluated based on accuracy, precision, recall and F-1 Score values. In this study, comparisons were made and looking for which algorithm had better results to the review data type.

4. RESULT AND DISCUSSION

Of the three algorithms a classification process will be carried out using training data and test data values 80%:20%, 70%:30%, 60%:40% and 50%:50% and 5 experiments will be carried out, by comparing the results if using stemming and not using stemming. The following is the result of the test :

Table 4.1 SVM analysis results using stemming

Dataset	Algoritma			
	SVM			
	Akurasi	Precision	Recall	F1-score
20 %	0.86	0.95	0.88	0.91
30 %	0.82	0.93	0.86	0.89
40 %	0.83	0.90	0.87	0.88
50 %	0.82	0.92	0.85	0.88

Table 4.2 XGBoost analysis result using stemming

Dataset	Algoritma			
	XGBoost			
	Akurasi	Precision	Recall	F1-score
20 %	0.84	0.96	0.83	0.89
30 %	0.82	0.93	0.83	0.88
40 %	0.81	0.91	0.82	0.86
50 %	0.80	0.92	0.81	0.86

Table 4.3 Random Forest analysis result using stemming

Dataset	Algoritma			
	Random Forest			
	Akurasi	Precsion	Recall	F1-score
20 %	0.86	0.96	0.85	0.90
30 %	0.82	0.90	0.86	0.88
40 %	0.81	0.90	0.81	0.85
50 %	0.82	0.93	0.81	0.87

From the results of the classification above, it can be seen that of the three algorithms, the Random forest algorithm gives the highest accuracy results with a high level of precision, recall and f1 score. The results show that the best algorithm for this data set is the random forest algorithm. The results of the random forest accuracy as proof are that it has a value of 86%, 96% precision, 85% recall and 90% f1-score. This proves that to find out how many people commented positively on the bumble application review the results of the random forest algorithm have proven that the reviews that were thought to be negative turned out to be positive after being classified the results were very good, this can be seen from the precision results. But it can also be seen that of the four dataset models used in the test, the results of the three algorithms show that the more datasets used, the lower the accuracy. Next we will discuss the results of the analysis if not using stemming.

Table 4.4 SVM analysis result without stemming

Dataset	Algoritma			
	SVM			
	Akurasi	Precision	Recall	F1-score
20 %	0.85	0.96	0.88	0.92
30 %	0.85	0.93	0.88	0.91
40 %	0.84	0.93	0.87	0.89
50 %	0.84	0.93	0.86	0.90

Table 4.5 XGBoost analysis result without stemming

Dataset	Algoritma			
	XGBoost			
	Akurasi	Precision	Recall	F1-score
20 %	0.85	0.95	0.87	0.91
30 %	0.83	0.93	0.84	0.88
40 %	0.84	0.95	0.84	0.89
50 %	0.83	0.95	0.83	0.88

Table 4.6 Random Forest analysis result without stemming

Dataset	Algoritma			
	Random Forest			
	Akurasi	Precsion	Recall	F1-score
20 %	0.85	0.94	0.89	0.91
30 %	0.84	0.92	0.86	0.89
40 %	0.83	0.92	0.84	0.88
50 %	0.84	0.92	0.85	0.88

From the analysis of the results above, it can be seen that the more datasets tested, the accuracy will decrease. This is because the SVM algorithm is a linear model algorithm, which makes it less accurate and more complicated as data increases. Then XGBoost has a way of working by building a model by adding new trees repeatedly which makes this algorithm strong against overfitting, so the reduction results are not as much as SVM as well as Random Forest which has a way of working that is almost the same as XGBoost so this algorithm is also strong against overfitting. SVM is called the linear method because it uses a linear function to divide data points into two classes. The goal of SVM is to find the weight vector and the deviation period that maximizes the margin between the two classes. Margin is defined as the distance between the nearest data point and the decision boundary. The decision boundary is the line that separates the two classes of data points. The SVM algorithm finds the decision boundary with the largest margin. This means that the data points on either side of the decision boundary are as far from the boundary as possible. SVM is a linear method because the decision boundary is a linear function of the input data. This means that the decision boundary is a straight line or plane.

Next we will discuss the results of analysis using stemming, which is better and more accurate but has low recall. This is because stemming is the process of removing prefixes and suffixes from words. Reducing the number of words that appear distinct can help improve the accuracy of text analysis. For example, the words "run," "run," and "run" all come from the word "run". This will make it easier for text analysis algorithms to identify words that are related to one another, even if they are not spelled the same way.

Furthermore, by limiting the amount of vocabulary which needs to be focused on, stemming may contribute to improving text analysis algorithms. This can be a huge advantage for algorithms that train on enormous sets of data, as there is less time and more memory needed to train them.

In general, the improvement of text analysis algorithms' accuracy and efficiency can be achieved by stemming techniques. However, it is important to be aware that stemming may also lead to certain errors, because it is not always possible to determine the exact root word for a particular word because in a literary library there will definitely be words that are not in the library because this literary library has existed since 2012 so there will be many words that are currently developing which have not yet been included in the literary library because as we know that every year there will be many new syllables used by Indonesian people and from 2012 to 2023 there must have been a lot of words developed among the people .

Next, we will discuss why the random forest algorithm was decided as the best algorithm in this study. This is concluded from the resulting accuracy results and also from the resulting precision results. Why only focus on precision results? this is because the data being tested is application user reviews, the expected result is that if the model is used to identify positive product reviews, it is more important to have high accuracy, so that users do not see negative reviews that are incorrectly classified as positive. And precision is the ratio of positive correct predictions compared to the overall positive predicted results. So if you want to know whether this application is good or not, it's better to focus on precision results because we can find reviews that are likely to be positive so that the final results are more accurate because precision is more thorough in analyzing data that may be right but thought to be wrong.

From the discussion above, it can be seen that the results of this study have limitations, namely stemming still cannot be said to have a significant effect on the resulting accuracy results because the literary library is still limited in processing basic words with the review dataset used which is a review of applications used by young people where they are more many use modern terms that are not yet in the literary library to find the base word. As well as the results of labeling the data used in this study may still be lacking because the annotators are lay people and not linguists so that the resulting annotation results may not be optimal. And also this research is still limited in the program evaluation process because in this study it only focuses on comparing the results of the accuracy of the three algorithms and whether stemming can increase the resulting accuracy value.

5. CONCLUTION

The conclusions obtained from the testing process are as follows. First, it is possible to conclude that the random forest algorithm is the best algorithm for this dataset from the test process review with an accuracy of 86% (85.76%), SVM and Xgboost each have an accuracy value of 86% (85.58%) and 84%. Then the second, Stemming can increase accuracy and reduce accuracy depending on the amount of data used. And finally, if you don't use stemming, the accuracy results will decrease with each result of 85% and the results of these three algorithms are slightly different, namely SVM 84.87%, XGBoost 84.60% and Random Forest 84.71%.

And suggestions for further research are as follows. The first, Try to make the dataset labeling process carried out at least by more than two people and there are also linguists so that the annotation results are better. Second, Developing literary literature into basic words that are more in line with the times so that the stemming results are more accurate. And finally, do a more in-depth evaluation of the program according to your goals in developing this project, for example if you want to develop the stemming side, you can evaluate it more deeply by comparing the time needed to complete the sentiment analysis.

DAFTAR PUSTAKA

[1] Iqbal Kharisudin, Fajar Sodik Pamungkas. “Analisis Sentimen Dengan SVM, NAIVE BAYES Dan KNN Untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 Pada Media Sosial Twitter,” 2021, 7.

https://journal.unnes.ac.id/sju/index.php/prisma/article/view/45038

[2] Zidna Alhaq, Ali Mustopa, and Joko Dwi Santoso Sri Mulyatun. “PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER” 3, No. 1 (2021) (n.d.): 6.

https://doi.org/10.24076/joism.2021v3i2.558

[3] Heart Parasian PR Zuriel, Achmad Fahrurozi. “IMPLEMENTASI ALGORITMA KLASIFIKASI SUPPORT VECTOR MACHINE UNTUK ANALISA SENTIMEN PENGGUNA TWITTER TERHADAP KEBIJAKAN PSBB” 26 No. 2 Agustus 2021 (n.d.): 14. https://doi.org/10.35760/ik.2021.v26i2.4289.

[4] Widodo, Dina Wahyuni. “IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK MENGETAHUI BIDANG SKRIPSI MAHASISWA MULTIMEDIA PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER UNIVERSITAS NEGERI JAKARTA” VOL 1. NO.2 DESEMBER 2017 (n.d.): 11. https://doi.org/10.21009/pinter.1.2.10.

[5] Karina. “Perbandingan Support Vector Machine (SVM) Dan Naïve Bayes Pada Analisis Sentimen,” 2021.

https://repository.unsri.ac.id/54114/

[6] Intan Purnamasari, Aliffia Kulsumarwati, and Budi Arif Dermawan. “PENERAPAN SVM DAN INFORMATION GAIN PADA ANALISIS SENTIMEN PELAKSANAAN PILKADA SAAT PANDEMI” 7 No 2; September 2021 (n.d.): 9. https://doi.org/10.37012/jtik.v7i2.641.

[7] Nur Fitriyah, Budi Warsito, and Di Asih I Maruddani. “ANALISIS SENTIMEN GOJEK PADA MEDIA SOSIAL TWITTER DENGAN KLASIFIKASI SUPPORT VECTOR MACHINE (SVM)” 9, Nomor 3, Tahun 2020 (n.d.): 15.

https://doi.org/10.14710/j.gauss.9.3.376-390

[8] Oman Somantri, Slamet Wiyono, and Dairoh. “OPTIMALISASI SUPPORT VEKTOR MACHINE (SVM) UNTUK KLASIFIKASI TEMA TUGAS AKHIR BERBASIS K-MEANS” 13, No. 02, JULI, 2016 (n.d.): 10.

https://www.researchgate.net/publication/314669783_OPTIMALISASI_SUPPORT_VEKTOR_MACHINE_(SVM)_UNTUK_KLASIFIKASI_TEMA_TUGAS_AKHIR_BERBASIS_K-MEANS

[9] Hanna Willa Dhany, Fahmi Izhari. “ANALISIS ALGORITHMS SUPPORT VECTOR MACHINE DENGAN NAIVE BAYES KERNEL PADA KLASIFIKASI DATA” 6 NOMOR 2 JULI 2019 (n.d.): 6.

https://jurnal.pancabudi.ac.id/index.php/Juti/article/view/675

[10] I Made Artha Agastya. “PENGARUH STEMMER BAHASA INDONESIA TERHADAP PEFORMA ANALISIS SENTIMEN TERJEMAHAN ULASAN FILM” 2018. http://dx.doi.org/10.33365/jtk.v12i1.70