Scientific Journal of Informatics
Vol. 7, No. 2, November 2020

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method

Susanti Fransiska[1], Rianto[2], Acep Irham Gufroni [3]

[1,2,3]Informatic Department, Faculty of Engineering,
Universitas Siliwangi, Tasikmalaya
Email: [1]fransiska370@gmail.com, [2]rianto@unsil.ac.id, [3]acep@unsil.ac.id

Abstract

Provider by.U is a relatively new and attractive telecommunications service with claims to be the first digital provider in Indonesia. All services are done digitally with the by.U application that offers convenience. Even so not all users are satisfied with the service, there are criticisms and suggestions, one of which is delivered through the by.U app review feature on the Google Play Store. Sentiment analysis is performed to extract information related to provider by.U. The steps taken are scrapping review data, positive and negative labeling, preprocessing data including data cleaning, data normalization, stopword removal and negation handling, sentiment classification using Support Vector Machine (SVM) and TF-IDF as feature extraction. TF-IDF+SVM with 5-Fold Validation produces pretty good accuracy with an average accuracy of 84.7%, precision of 84.9%, recall of 84.7%, and f-measure of 84.8%. The highest accuracy results in fold 2, 86.1%. The effect of TF-IDF on the measurement of model performance is not so great, but it is better.

Keywords: Classification, provider by.U, Sentiment Analysis, Support Vector Machine, TF-IDF

1. INTRODUCTION

Telkomsel in 2019 declared provider by.U as the first digital prepaid cellular service in Indonesia that provides digital experience for all telecommunications needs. All service activities are carried out in the by.U application that is installed on the smartphone, making it easy for users to control the desired cellular services according to user needs [1]. These services include the selection of numbers, determination of internet quota, sending of sim cards, up to live chat of user services. Although offering convenience with all-digital labels, of course not all users feel satisfied and well served, there are user's criticisms and suggestions that are also needed by developers to maintain the quality and development of service providers by.U. The by.U application can be downloaded at the Google Play Store. Google Play Store is equipped with a review feature where users can submit their reviews in the form of criticism, praise, suggestions, or other assessments of an application. There are quite a lot of reviews and are not structured so we need a technique to present these reviews to be more informative for users as well as application or product developers. Sentiment analysis or so-called opinion mining is done to see the trend of an opinion on a problem or object by someone, whether it tends to be a negative or positive opinion [2]. According to Chauhan C. & Sehgal S. in 2017 does sentiment analysis of product reviews because a large number of products were being discussed on the internet [3]. Sentiment analysis has also been used to help explore the opinions of customers about the product they are interested in [4]. According to Moraes, et al., in 2013 many researchers reported that Support Vector Machine (SVM) is the most accurate method for text classification and is widely used in sentiment classification [5]. The SVM classifier was chosen because it can support high-dimensional data [6]. The results of the K & Shetty experiment in 2017 show that SVM achieves high accuracy in the classification process compared to other comparative methods namely Naïve Bayes and Max-Ent [7]. SVM has an input format in the form of vector space and output in the form of 0 or 1 (positive / negative), then the text document needs to be transformed into a format that is suitable for the machine learning format, that is the use of TF-IDF (term frequency - inverse document frequency) [8]. TF-IDF so far known as the best feature extraction method for text analytics [9].

Norma Fikria on 2018 conducted sentiment analysis on 1010 KAI Access reviews and 620 reviews at Tiket.com in the Google Play Store with positive and negative labels using linear, polynomial, radial and sigmoid kernel SVM with the 5-fold validation testing. Soumik, et al. on 2019 conducted sentiment analysis on the Review of more than 100 apps (around 1000 reviews) on the Google Play Store with positive, negative and neutral labels using the TFIDF feature extraction and Naïve Bayesian Classifier, SVM, Logistic Regression, Ensemble Methods (Adaptive boosting, Gradient Boosting and extreme Gradient Boosting) with 5-fold validation and without 5-fold validation testing. Makhmudah, et al. on 2019 conducted sentiment analysis on 548 tweet data containing literature on homosexuality with positive and negative labels using TFIDF feature extraction and SVM linear kernel classification method by testing 3 different scenarios based on the ratio of training data and testing data. Pratama, et al. on 2018 conducted an aspect-based sentiment analysis on 674 South Malang coastal tourism customer opinion data on TripAdvisor with positive, negative, and neutral labels based on general aspects, cleanliness, crowd, road access, wave conditions using TFIDF and SVM linear kernel with 5 fold validation testing. Rofiqoh, et al. on 2017 conducted sentiment analysis on 300 data 3 cellular telecommunications service providers (Indosat, Telkomsel, and XL Axiata) on Twitter with positive and negative labels using TFIDF+min-max normalization+Lexicon based features and SVM polynomial kernel classification method by testing the parameters of polynomial SVM. Based on the literature review, it was concluded that sentiment analysis was carried out on various topics discussed on the internet via Twitter, reviews on websites such as TripAdvisor and Google Play Store. Many researchers have used TFIDF and SVM methods for their classification methods, but limited studies concerned on tested measure how much TFIDF affects the accuracy of SVM. Therefore, this research intends to do sentiment analysis to provider by.U based on review on the Google Play Store with the TF-IDF feature extraction and the Support Vector Machine classification method. This study aims to analyze the tendency of sentiments and measure the performance of the classification model so that it is expected to be a reference for business people to maintain the quality and development of by.U service providers and can be used as a reference for parties or other researchers who have interests with similar cases.

2. METHODS

The steps of the study are described in Figure 1.

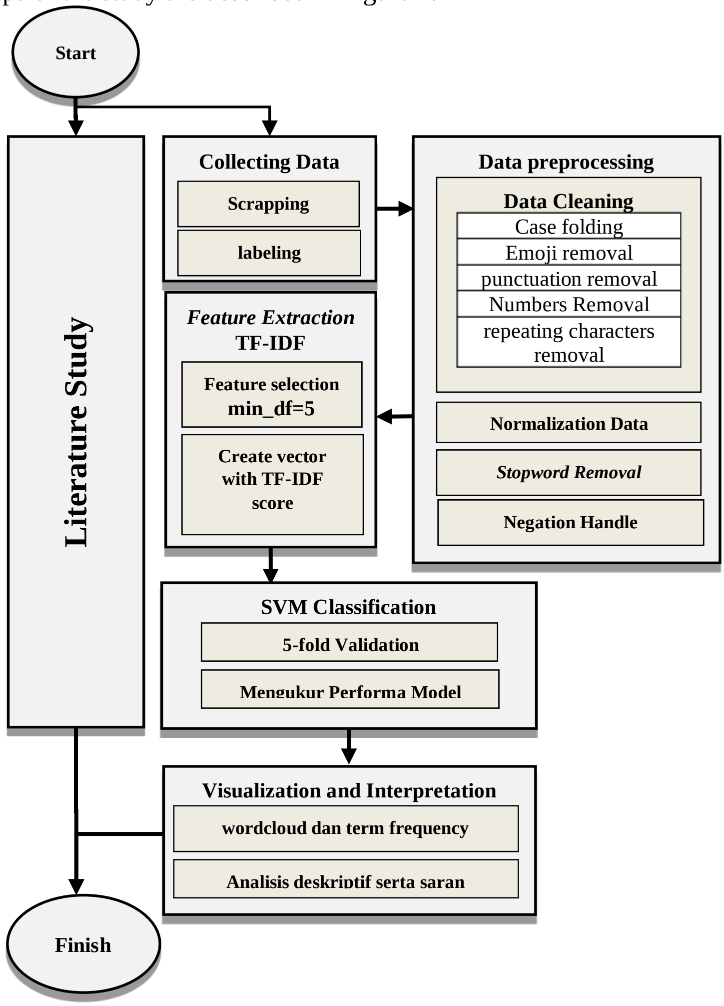

Figure 1. The steps of the study

This research was conducted through the following steps:

2.1. Literature study

Literature study is continued during the data scrapping process until the classification, this aims as a guideline in conducting research. Literature study sources conducted in the form of ebooks, national journals and several international journals.

2.2. Collecting Data

Scrapping the by.U review data from the Google Play Store site is done using the Python programming language by utilizing the google-play-scrapper library with parameters lang = 'id', country = 'id', sort = Sort.NEWEST, and count = 10000. The data is filtered according to the data needed. The data used in this study is the most recent data from March 17, 2020. The data is then labeled sentiment. The labeling process is carried out using review ratings as a sentiment reference, rating 1 and 2 are grouped into negative sentiment, ratings 4 and 5 are grouped into positive sentiment. Rating 3 is generally grouped into neutral sentiment, but in this study, neutral sentiment is considered less informative, after all reviews are dominated by complaints and compliments, therefore reviews with rating 3 are manually checked to determine which reviews are more positive or negative based on word usage or the number of positive and negative words.

2.3. Data preprocessing

Data preprocessing is performed to eliminate noise data. Steps being taken include,

1) Data cleaning such as case folding, emoji removal, punctuation removal, numbers removal, and repeating characters removal.
2) Normalizing Data, changing words into more formal standard forms using normalization dictionaries.
3) Stopword Removal, delete words with high frequency but does not have meaning using the stopword dictionary.
4) Negation Handling, combining the negation word with the words afterwards.

2.4. Feature Extraction TF-IDF

Feature Extraction with TF-IDF is done to transform text documents into a form that can be used for the classification process with SVM, in vector form. The TF-IDF weighting process uses the python programming language by utilizing the Tfidfvectorizor library with the parameters min_df = 5 and others default. TF-IDF is a weighting scheme with the following formula [9],

$$TFIDF_j = TF \ x \ IDF \tag{1}$$

Description:

TF : frequency of words appearing in the document.

IDF : value from the following calculation,

$$log \frac{total \ number \ of \ documents}{the \ number \ of \ documents \ containing \ the \ word \ j} \tag{2}$$

2.5. SVM Classification

SVM is a machine learning algorithm that separates class data by finding the most optimal hyperplane [10]. Classification with the SVM method is done by the python programming language by using the sklearn.svm library for the classification process with the SVM method, besides using the sklearn.metrics library to measure the performance of the classification model or find out the accuracy of learning with SVM. Distribution of training data and test data is done using 5-fold Validation with the KFold library on sklearn.model_selection.

2.6. Visualization and Interpretation

Visualization and interpretation are carried out to display the results of research to be more easily understood so that it becomes useful information for by.U business to maintain quality and improve services. This stage uses the wordcloud and matplotlib.pyplot library.

3. RESULT AND DISCUSSION

Scrapping data amounted to 8,925 review data, each review contains information about the date the review was made (at), the version of the application when the review was made (reviewCreatedVersion), the name of the review maker (Username), the rating given by the review maker (score) and the contents of the review (content) . But the information used is only the contents of the reviews and ratings. Review ratings are used in the labeling process. Data labeling in this study is limited to review rating, rating 1 and 2 are grouped into negative sentiment or. Ratings 4 and 5 are grouped into positive sentiment or labeled 1, while rating 3 is labeled manually. The process resulted in 4,874 reviews labeled positive and 4,051 labeled negatively. The percentage of reviews that have been labeled are visualized in the following Figure 2.

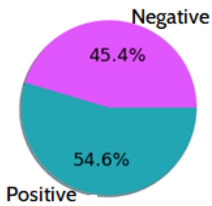

Figure 2. Sentiment percentage

Figure 2 is the percentage of the number of reviews of positive and negative sentiment grouping. The number of positive reviews was more with a percentage of 54.6%, while negative reviews were 45.4%. Visualizations related to topics often discussed in positive labeled review data are illustrated by the word frequency bar chart in Figure 3.

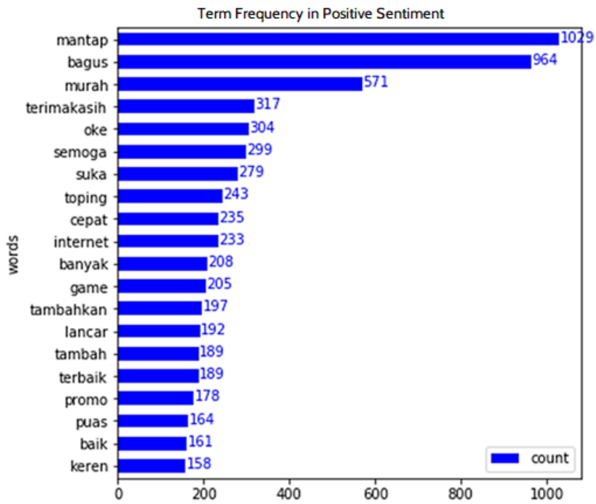

Figure 3. Frequency of positive sentiment words

Figure 3 contains information on the 20 words that appear most frequently in the positive review data and the number of frequencies. The total term is 207.838 terms. Based on the frequency of these words, it can be concluded that this positive-labeled review contains an expression of satisfaction with by.U service, discussing the price of quotas or internet packages that are considered cheap, as well as reviews of fast or speedy internet connections. In addition to expressing satisfaction with the by.U service, users also convey hopes or suggestions to maintain the price of internet packages and requests to hold promos, add package variants and internet toppings, especially topping games. Visualizations related to topics often discussed in negative labeled review data are illustrated by the word frequency bar chart in Figure 4.

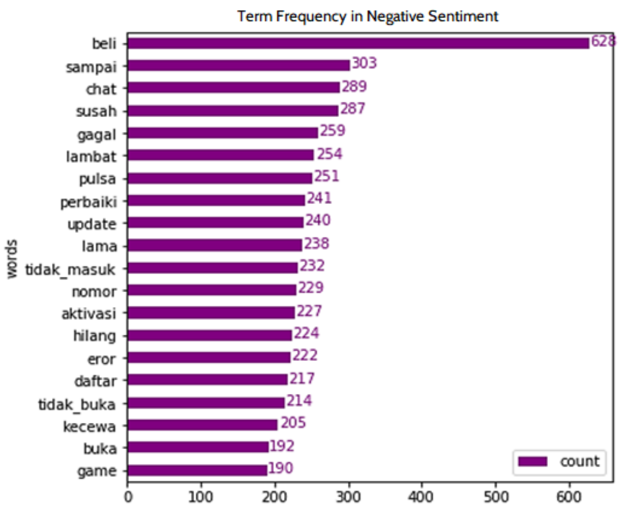

Figure 4. The Frequency of Negative Sentiment Words

Figure 4 contains information on the 20 words that appear most often in negative review data and the number of frequencies. The total term is 265,473 terms. Based on the frequency of words in negative sentiments, it can be concluded that the review contains complaints regarding problems with the purchase of internet packages and by.U sim cards which are considered long to arrive, cannot enter or fail when registering cards, complaints regarding poor service by.U applications that are difficult or unable opened after the application update, and slow or loss of internet connection and slow response to live chat customer service.

The dimensions generated from the TFIDF Feature Extraction process are 8789 x 1142. The data originally numbered 8,925 to 8,789 after selecting the preprocessed data that did not have any strings (empty) in a document. This result is used for the classification process with the Support Vector Machine (SVM) method. The process of measuring the performance of classification models uses the 5-Fold validation test. Testing using 5-Fold validation results in the distribution of training data and test data with a ratio of 80:20 with a total of 7031 training data, while the test data amounted to 1758. The results of classification performance measurements with TFIDF and SVM are described in Table 1.

Table 1. TFIDF + SVM Classification Performance Measurement Results

N-Fold	Accuracy	Precision	Recall	F-Measure
1	0,840159	0,843717	0,840159	0,840788
2	0,861775	0,862206	0,861175	0,861816
3	0,844141	0,844379	0,844141	0,844102
4	0,850967	0,850982	0,850967	0,850974
5	0,841207	0,847290	0,841207	0,842773
Average	0,84765	0,849715	0,84753	0,848091

Table 1 contains the accuracy, precision, recall and feature values of each fold of the measurement results of the TFIDF + SVM classification performance. This test produces an average value of accuracy of 84.7%, precision of 84.9%, recall of 84.7% and f-measure of 84.8%. The highest accuracy results in fold 2, which is 86.1%. The lowest accuracy results are at fold 1, with 84%. Fold 2 is described in the Table 2.

Table 2. Confusion matrix fold 2

	Prediciton		
	Class	Negative	Positive
Actual	Negative	739 (TN)	109 (FP)
	Positive	134 (FN)	776 (TP)

Table 2 is the results of testing on fold 2 which is described by the Confusion Matrix produces 739 true negative values, 134 false negative values, 109 false positive, and 776 true positive.

The vectorization process for TF weights is done by the countvectorizer library. The results of classification performance measurements with TF and SVM are described in Table 3.

Table 3. TF + SVM classification performance measurement results

N-Fold	Accuracy	Precision	Recall	F-Measure
1	0,832765	0,833084	0,832765	0,8302897
2	0,851536	0,851549	0,851536	0,851480
3	0,823094	0,823702	0,823094	0,823033
4	0,831627	0,832982	0,831627	0,831684
5	0,844622	0,844379	0,844622	0,844493
Average	0,833114	0,837139	0,836729	0,8361959

Table 3 contains the accuracy, precision, recall and feature values of each fold of the measurement results of TF + SVM classification performance. This test produces an average value of an average accuracy of 83.3%, a precision of 83.7%, a recall of 83.7% and an f-measure of 83.6%. The highest accuracy results in fold 2, which is 85.1%. The lowest accuracy results are at fold 4, with 83.1%. This test was conducted as a measure of the influence of TFIDF in the classification process with SVM.

4. CONCLUSION

Analysis of the data resulted in reviews that were more likely to be positive even though the comparison of numbers with negative reviews did not differ greatly, as many as 54.6% of the reviews were grouped into positive reviews and 45.4% of negative reviews. Positive reviews contain expressions of satisfaction with by.U services with the word 'mantap' 0.5% and the word 'bagus' 0.46% of the total term, and expressions of cheap quota or internet packages prices with the word 'murah' 0.27% of total term. Negative reviews contain complaints about problems with internet package purchases with the word 'beli' 0.23% of the total term, and by.U sim cards which are long arrives with the words 'sampai' 0.11% of the total term, and the expression difficulty and failure in performing services with the words 'susah' 0.11% and 'gagal' 0.09% of the total term. This study revealed that the TF-IDF and SVM methods can be applied to the classification process with fairly good measurement results, but the figures obtained were not better than previous studies, this is due to differences in the dataset, labeling process, preprocessing stages and feature usage. In addition, the effect of TF-IDF as a feature extraction on the measurement of model performance is not so great, but it is better to use TF-IDF weighting. Further research can use the manual labeling process or with a sentiment dictionary, it can also use other classification methods as a comparison of classification performance or can use other feature selection methods, this can affect the accuracy rate.

REFERENCES

[1] Telkomsel. (2019). *Telkomsel Luncurkan by.U, Layanan Selular Prabayar Digital End-to-end Pertama di Indonesia*. [Online]. Tersedia: https://www.telkomsel.com/about-us/news/telkomsel-luncurkan-byu-layanan-selular-prabayar-digital-end-end-pertama-di-indonesia. [Diakses 7 Februari 2020].

[2] Lorosae, T. A., Prakoso, B. D., Saifudin, S., & Kusrin, K. (2018). Analisis Sentimen Berdasarkan Opini Masyarakat pada Twitter Menggunakan Naive Bayes. *Semnasteknomedia Online*, *6*(1), 1-10.

[3] Chauhan C., Sehgal S. (2017). Sentiment Analysis On Product Reviews. *International Conference on Computing, Communication and Automation (ICCCA)*.

[4] Leung, C. W. (2009). *Sentiment analysis of product reviews. In Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1794-1799). IGI Global.

[5] Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621-633.

[6] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016, March). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 112-116). IEEE.

[7] Shivaprasad, T. K., & Shetty, J. (2017, March). Sentiment analysis of product reviews: a review. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 298-301). IEEE.

[8] Patil, G., Galande, V., Kekan, V., & Dange, K. (2014). Sentiment analysis using support vector machine. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), 2607-2612.

[9] Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches. *SMU Data Science Review*, *1*(4), 7.

[10] Kowalczyk, A. (2017). *Support vector machines succinctly*. Syncfusion Inc.

[11] Fikria, N. (2018). Analisis Klasifikasi Sentimen Review Aplikasi E-Ticketing Menggunakan Metode Support Vector Machine dan Asosiasi.

[12] U. Makhmudah, S. Bukhori and J. A. Putra. (2019). Sentiment Analysis of Indonesian Homosexual Tweets using Support Vector Machine Method. *ICOMITEE*.

[13] U. Rofiqoh, R. S. Perdana and M. A. Fauzi. (2017). Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter dengan Metode Support Vector Machine dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 1, pp. 1725-1732.

[14] Pratama, Y., Bachtiar, F., & Setiawan, N. (2018). Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2*(12), 6244-6252

[15] Soumik, M. M. J., Farhavi, S. S. M., Eva, F., Sinha, T., & Alam, M. S. (2019, December). Employing Machine Learning techniques on Sentiment Analysis of Google Play Store Bangla Reviews. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-5). IEEE.