Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM

Huda Mustakim*[1], Sigit Priyanta[2]
[1]Undergraduate Program of Computer Science; FMIPA UGM, Yogyakarta Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]hudamustakim@mail.ugm.ac.id**, [2]seagatejogja@ugm.ac.id

Abstrak

Keberadaan aplikasi KAI Access milik PT. KAI merupakan salah satu bentuk usaha mereka dalam melayani konsumen di era modern ini. Namun, banyak ditemukan ulasan negatif di kolom ulasan Google Play Store. Sudah ada penelitian pada ulasan tersebut, namun tahap klasifikasi masih pada tingkat dokumen sehingga aspek pada aplikasi belum diketahui secara gamblang dan terstruktur. Maka, diperlukan analisis sentimen berbasis aspek untuk mengekstrak aspek-aspek yang diulas dan mengetahui sentimennya. Pada penelitian ini dilakukan analisis sentimen berbasis aspek pada ulasan pengguna KAI Access dengan metode Naive Bayes Classifier (NBC) dan Support Vector Machine (SVM), dengan 3 skenario. Skenario 1 menggunakan NBC dengan Multinomial Naive Bayes, skenario 2 menggunakan SVM dengan parameter default library Sklearn, dan skenario 3, dengan SVM dan hyperparameter tunning, dan data diambil dari Google Play Store. Hasil penelitian menunjukkan mayoritas sentimen pengguna bernilai negatif di setiap aspek, dengan aspek errors yang terbanyak dibahas menunjukkan tingginya kesalahan sistem. Hasil pengujian memberikan model terbaik dari skenario 3 dengan rata-rata skor akurasi 91,63%, f1-score 75,55%, presisi 77,60%, dan recall 74,47%.

Kata kunci— *analisis sentimen berbasis aspek, kai access, support vector machine, naive bayes classifier*

Abstract

The existence of KAI Access from PT. KAI prove their sincerity in serving consumers in this modern era. However, many negative reviews found in Google Play Store. There has been research on the review, but the analysis stage still at document level so the aspect related to the application is not known clearly and structured. So it is necessary to do an aspect-based sentiment analysis to extract the aspects and the sentiment. This study aims to do an aspect-based sentiment analysis on user reviews of KAI Access using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), with 3 scenarios. Scenario 1 uses NBC with Multinomial Naive Bayes, scenario 2 uses SVM with default Sklearn library parameter, and scenario 3, uses SVM with hyperparameter tunning, while the data scrapped from Google Play Store. The results show the majority of user sentiment is negative for each aspect, with most discussed errors aspect shows the high system errors. The test results gives the best model from scenario 3 with an average accuracy 91.63%, f1-score 75.55%, precision 77.60%, and recall 74.47%.

Keywords—*aspect-based sentiment analysis , kai access, support vector machine, naive bayes classifier*

Received September 3rd,2021; Revised April 30th, 2022; Accepted April 30th, 2022

1. INTRODUCTION

The web has become an indistinguishable piece of regular daily existence, including the use of ticketing online system in public transportation. One of them is KAI Access application that launched by PT. KAI to make easier for consumers in ordering tickets. However, there have been many negative reviews of the app on google play store. So the application developers need to pay attention to it.

Mean while, there have been studies that conducted sentiment analysis on these reviews [1], but the classification carried out still at the document level and has not been able to produce aspects related to the application in structured form and it's sentiment. Even though the user's sentiment to the application aspect is very important for developers to know to make it easier for them to make improvements.

Sentiment analysis or also known as opinion mining is a method to analize text in the form of opinions or attitudes towards a topic whose analysis results are in the form of information about the polarity of the analyzed text or document, whether it is positive or negative [2].

Previous research related to aspect-based sentiment analysis in user reviews has been carried out in several studies, such as the research by Astuti [3], which applies Latent Dirichelet Allocation (LDA) to extract the aspects and the NBC to classify the polarity of sentiment in each aspect with TF-IDF features extraction. There was also a research by Ailiyya [4], who applies the SVM to perform aspect classification and also sentiment polarity classification, using TF-IDF feature extraction on Tokopedia user reviews. Then, research by Al-Smadi [5], who conducted research on Arabic hotel review data, which compared the SVM and Recurrent Neural Network (RNN) methods, with the conclusion that SVM had a better performance in classifying aspect and it's sentiments, and also Rodrigues [6], who conduct an aspect-based sentiment analysis using SVM on product reviews.

Naive Bayes Classifier (NBC) is a probabilistic classification model that measures the relationship between feature variables and target variables as probabilities, which is built based on Bayes' theorem for conditional probabilities [7]. Naive Bayes has several advantages such as little training data required, simple algorithm, easy to implement, efficiency, able to handle large data or missing values, and not sensitive to noise [8]. While Support Vector Machine is a method that uses a hypothetical space which is linear functions in a high-dimensional feature space, trained with a learning algorithm based on optimization theory of learning bias derived from statistical learning theory [7], and its has been proved in many research that SVM have good performance in classifying text data, such as research by [4], [5], and [6].

Based on the background above, this study conduct an aspect-based sentiment analysis on user reviews of the KAI Access and compare between SVM and Naive Bayes Classifier performance in classifying aspect and it's sentiment.

2. METHODS

In this segment, we talk about the design and technique used to do the aspect-based sentiment analysis. The stages used in this study are: data collection, data preparation, preprocessing and feature selection, training data, testing and evaluation, and finally, result analysis and word cloud. The portrayal of the stages used in this study are shown on Figure 1 bellow:

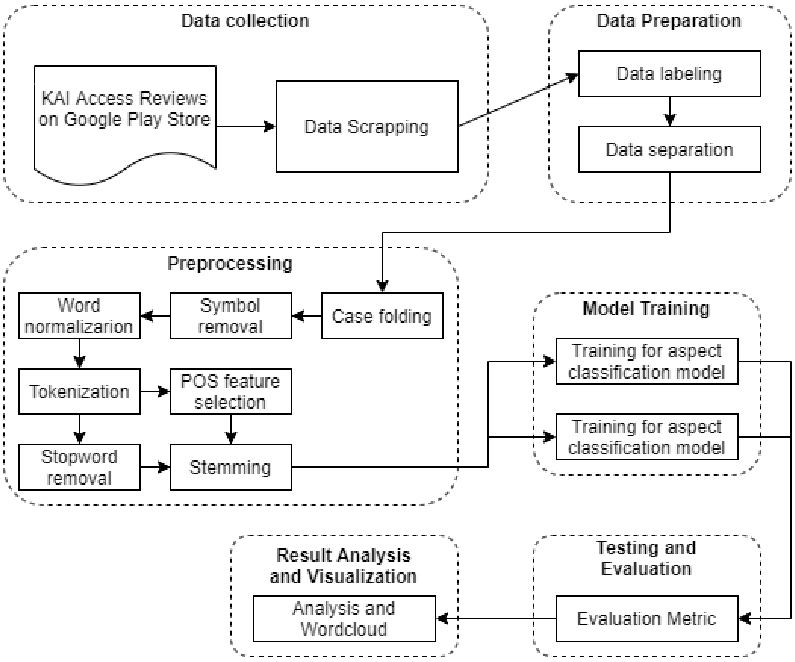

Figure 1 Research workflow

As shown on Figure 1, the first stage is data collection. Here, the reviews is scrapped from Google Play Store for KAI Access using a scrapping tools, WebHarvy. After that, the data need to be prepared, which consist of data labeling and data separation. The labeling was done for labeling the aspect and the sentiment polarity of each review. While data separation is separating the data, which is duplicating the data for each aspect category, and splitting the training set and test set. The next stage is preprocessing, which include case folding, symbol removal, word normalization, tokenization, feature selection based on part-of-speech (POS) for sentiment labeled data, stopword removal, and stemming. After that, the clean data used to train the model classification. The aspect labeled data were used to train the aspect classification model, while the sentiment labeled data were used to train the sentiment classification model. In this research, this training process was conducted using 3 skenario. And then, each model were tested and evaluated by calculating the evaluation metric such as accuracy, f1-score, precision, and recall. And finally, the final stage is analyzing and visualizing the result of the best model among 3 skenario which used on the test set using word cloud.

2.1 Data collection

The data used in this research was scrapped from Google Play Store for the user reviews of KAI Access for last two years, which is from October 2018 to October 2020. While the scrapping process was done using a scrapping tools, which is WebHarvy.

2. 2 Data preparation

The data collected from scrapping process need to be prepared. First, the data need to be labeled. The labeling process was done manually by two final year Information System student of Universitas Amikom Yogyakarta, which work together on labeling the aspect category and the sentiment polarity, based on usability aspect of Nielsen (1994) which is learnability, memorability, efficienty, and errors [9]. But here, the memorability aspect are treated as the same aspect as learnability since is has similar meaning. The review may consist of 1 or more aspect (multi-label). After being labeled, the data were categorized for each labeled aspect category. Then, each data set splitted into train set and test set. The illustration of this process are shown on figure 2.

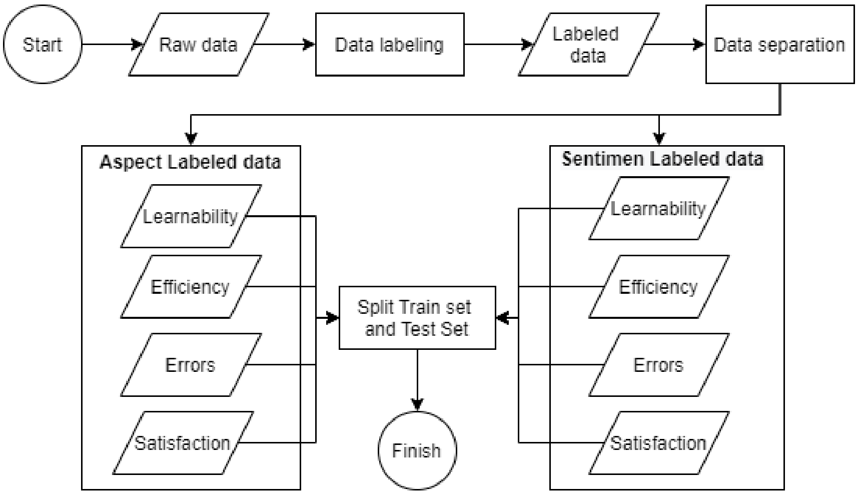

Figure 2 Data preparation workflow

2. 3 Preprocessing

After being prepared, each data set for both train set and test set were preprocessed to make the data cleaner and optimize the classification model performance. The stages used are case folding, symbol removal, word normalization by creating a slang word dictionary manually based on the observation of the data used, feature selection based on part-of-speech (POS) which tagged using POS-tagging method or word anotation [10]. This process conducted with stanza library from [11], with label *verb*, *noun*, *adjective*, and *adverb*, since this label is the type of word that mostly contains the sentiment [12]. After that, stopword removal, and finally stemming.

2. 4 Model Training

After the data preprocessing was done, the the clean train set is used to train the model classification. There are two types of model classification here, they are aspect classification, and sentiment classification. Each type will be built to make 4 models for each aspect category, so there will be 8 models in total. The model training was conducted by implementing 5-fold cross validation in order to validate the model performance, and select the best model among each fold to be tested in test data set. The workflow of model training are shown as Figure 3.

As seen on Figure 3, for each fold after data partition, the new train set and validation set need to be procesed by feature extraction using TF-IDF, and the extracted feature of new train set are oversampled. The oversampling are conducted because the data in this research is not balanced, and it’s important to make it balance since the model created from unbalaced data mostly will be poor on classifying data from minor class [13]. Meanwhile, there are 3 model training skenario, they are shown at Table 1.

Table 1 Model training skenarios

Skenario	Method	Special treatment
1	Naive Bayes	Multinomial NB
2	SVM	Sklearn default perparameter
3	SVM	With Hyperparameter tuning

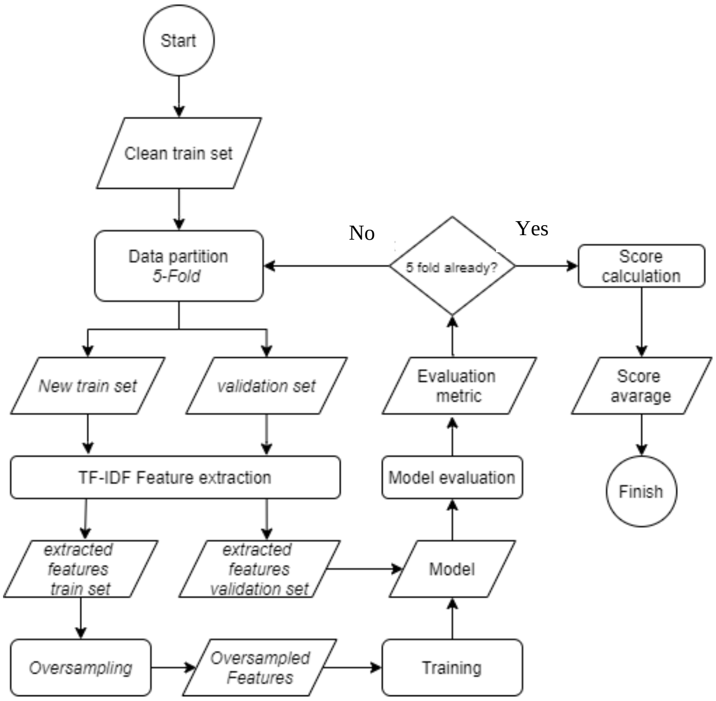

Figure 3 Model training with 5-fold cross validation Enough

The training process for aspect classification was conducted as shown on Figure 3, where each training set on each fold are resampled to avoid overfitting [14]. Each set were trained for different model since each data may contain more than 1 aspect. So, the classifier for each aspect can be built. While the model training for sentiment classification also separated for each aspect category, as shown as Figure 4.

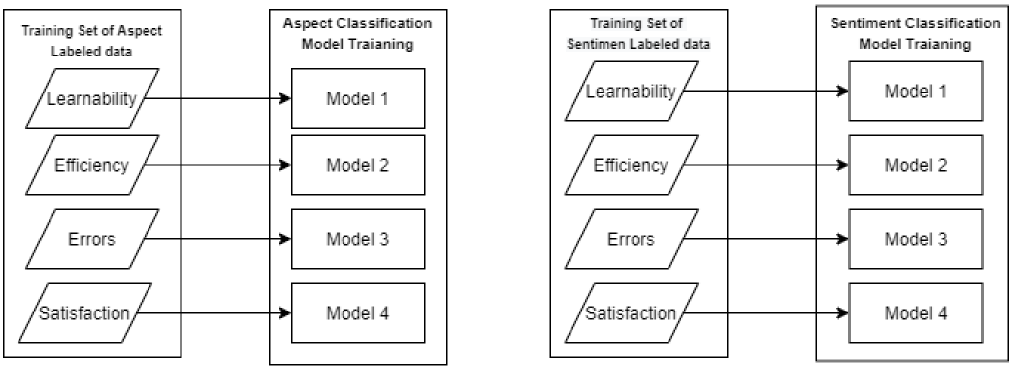
Training Set of Aspect Labeled data
Learnability
Efficiency
Errors
Satisfaction
Aspect Classification Model Traianing
Model 1
Model 2
Model 3
Model 4
Training Set of Sentimen Labeled data
Learnability
Efficiency
Errors
Satisfaction
Sentiment Classification Model Traianing
Model 1
Model 2
Model 3
Model 4

Figure 4 The data set used for aspect and sentiment model training

2. 4.1 Naive Bayes Classifier

In this research, the type of naive bayes used was Multinomial Naive Bayes which is one type of Naive Bayes classification model that is most often used in document classification problems [3]. The basic theorm of Multinomial Naive Bayes was still follow the same rule as Naive Bayes classifier that can be seen at equation (1). Hypothesis *H* is a class that acts as a target in the classification, while evidence *E*, is a term or words used as input to the classification model. Thus, if C represents a class variable. So the conditional probability $P(C|t_1, t_2, t_3, \ldots, t_n)$ can be interpreted as the probability of class C, based on the features *t*.

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(E)} \tag{1}$$

While Multinomial Naive Bayes used the frequency of words contained in the document as a predictor or feature, with the calculation formula are as follows [3]:

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c) \tag{2}$$

$$c_{map} = \arg\max P(c \mid d) = \arg\max P(c) \prod_{1 \le k \le n_d} P(t_k \mid c) \tag{3}$$

Based on equation (2), *P(*t_k *| c)* is the conditional probability of the term *t* in class *c,* and *P(c)* is the prior probability of documents in class *c*. While on equation (3), it’s goal is to find the maximum a posteriori (MAP) or *cmap.* The value of *P(c)* is the probability of the document in the data with *c* label , for all documents in the data. While *P(tk|c)*, is the comparison of the number of occurrences of terms in class *c* (*Tct'*), with the product of the total number of terms in topic *c* and the number of all terms in the data (∑ *t V*).

2. 4.2 Support Vector Machine

Meanwhile, SVM is a classification method that can be used on linearly separable data and nonlinearly separable data, which works by transform the training data into a higher dimension, and using the best linear plane (hyperplane) for each class. SVM find the best linear plane based on the support vector that is in the boundary plane, and also the margin, which is the distance between the boundary planes [3].

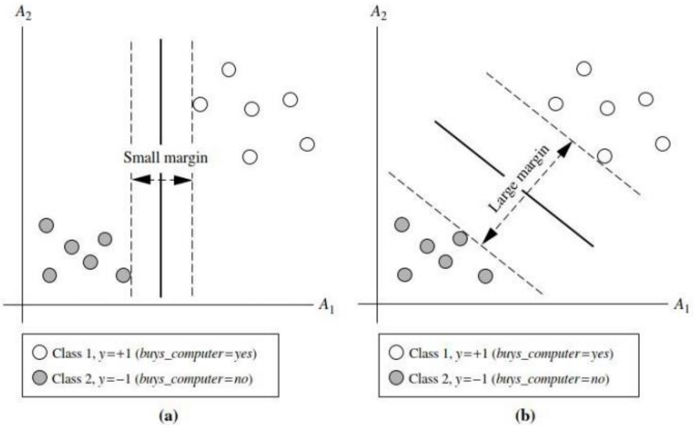
A_2
Small margin
A_1
○ Class 1, y=+1 (buys_computer=yes)
● Class 2, y=−1 (buys_computer=no)
(a)
A_2
Large margin
A_1
○ Class 1, y=+1 (buys_computer=yes)
● Class 2, y=−1 (buys_computer=no)
(b)

Figure 5 Hyperplane with margin

Based on Figure 5 which taken from [15], can be summarize that picture (b) show the best hyperplane, since it has the maximum margin when dividing the classes. While picture (b) have a small margin.

2. 5 Testing and Evaluation

The testing and the evaluation of the model in this research are conducted by testing the best model produced from k-fold cross validation on the test set. While the calculation of the performance are done by calculating the accuracy, f1-score, precision, and recall.

3. RESULTS AND DISCUSSION

In this section, the result of this reseach are discussed. Starts form data collection result, data preparation result, model training result, testing result, and analysis result.

3.1 Data collection result

From the scrapping process, there are 3792 reviews which is the review data. The sample of the data scrapped are shown on Table 2.

Table 2 Sample of scrapping result

No	username	date	rating	review
1	Tivanwt	31-Oct-2020	5.0	Aplikasi ini sangat membantu saya saat pandemi…
2	Hery	31-Oct-2020	5.0	Memesan perjalanan lebih mudah
3	Sugiyo triyoga	31-Oct-2020	5.0	KAI smakin bagus pelayanannya, Krena situasi Pa…

As shown on Table 2, the data scrapped are username, date, rating, and the reviews it self.

3.2 Data preparation result

As discussed before on sub topic 2.2, the collected data are labeled by 2 expert. The sample of the manual labeling result are shown on Table 3.

Table 3 Sample of manual labelling result

No	Learnability	Efficiency	Errors	Satisfaction	Review
1		pos		pos	Sy banyak merasa terbantu
2			neg		Kenapa gk bisa log on aneh
3			neg	neg	Aplikasi ga guna. Ga bisa milih daftar penumpang, pembayaran error

As shown on Table 3, each review are labeled for each aspect using it's sentimen (positive or negative). An aspect which not exist on a review did not get any label. So, based on the manual labeling result, the data are separated for each aspect. For aspect classification, the data with no label are labeled as 0 and data with label 'pos' or 'neg' are labeled as 1, and for sentiment classification, the data with no label will be omitted and give label 1 for positive reviews and 0 for negative reviews. This is conducted for every data set on each aspect category.

3.3 Model training result

To simplify the result analysis, this section discusses the result of model training by avaraging the score of 5-fold cross validation for each model training on 3 scenarios. There are 8 model training in total for each scenario. First, 4 aspect classification models for each aspect category (learnability, efficiency, errors, and satisfaction) and 4 sentiment classification models for same aspects.

Table 4 Avarage score of validation result

No	Scenario	Metric Evaluation			
		Accuracy	F1-Score	Precision	Recall
1	1 (Multinomial NB + add one)	82,23%	59,06%	51,69%	74,87%
2	2 (SVM + default parameter)	89,96%	62,92%	89,04%	52,67%
3	3 (SVM + best parameter)	90,74%	**73,15%**	75,00%	72,73%

From Table 4, it can be seen that the results of the average score for each evaluation metric calculated for each model made. The high recall obtained from scenario 1, high precision in scenario 2, and high scores with good balance of precision and recall from scenario 3.

Meanwhile, by doing this validation, an overview of the performance of the model has been obtained when the model is applied to the actual situation or new data that was not previously included in the training process or data training.

In addition in this study, a model without oversampling was carried out to see the effect of the oversampling on the performance of the model in each modeling scenario. The comparation between them are shown on Figure 6.

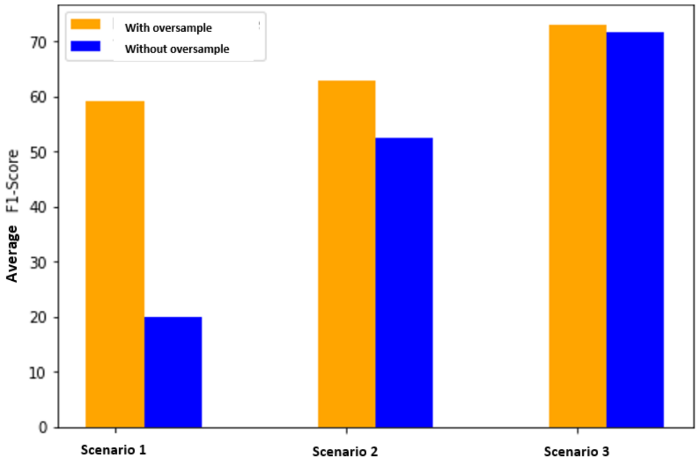

Figure 6 Comparison score of validation with oversampling and without oversampling

Based on figure 6, it can be seen that oversampling method can improve the performance of the model, especially for skenario 1 which is using Multinomial Naive Bayes. So, it can be conclude that Naive bayes is very sensitive of data balance. While the model built by SVM are perform better than Naive Bayes even without oversampling, especially on skenario 3, which is implemented using best parameter. This prove that SVM with right parameter can perform well on unballance data.

3.4 Model testing result

The testing stage on this research are conducted on the best model between model training in each fold during 5-fold cross validation for each skeanrio. Each model are tested on their respectively test set. The result of the test are shown at Table 5.

Table 5 Avarage score of model testing result

No	Scenario	Metric Evaluation			
		Accuracy	F1-Score	Precision	Recall
1	1 (Multinomial NB + add one)	83,47%	61,92%	54,73%	76,59%
2	2 (SVM + default parameter)	90,85%	66,03%	88,96%	55,89%
3	3 (SVM + best parameter)	91,63%	**75,55%**	77,60%	74,47%

From Table 5 above, it can be seen that the result of each skenario are quiet similar to the validation result, with skenario 1 get the highest recall, skenario 2 get the highest precision, and skenario 3 get the highest f1-score, which is the best balance between precision and recall.

To make it easier in comparing between validation result and test result, here is the comparison in bar graph for the avarage of f1-score.

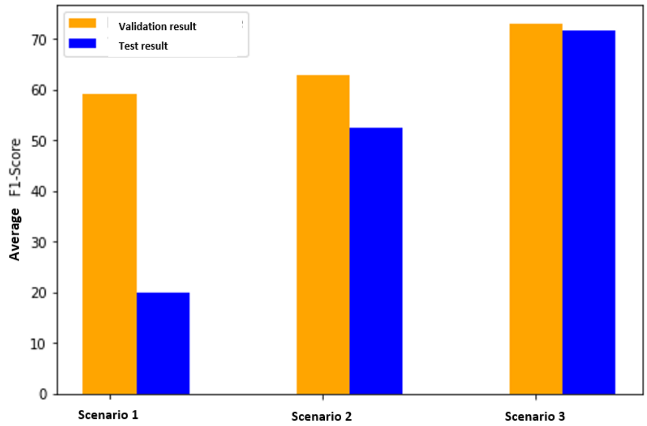

Figure 7 F1-score comparison of validation result and test result

Based on Figure 7, it shown that the test result is not too far from the validation result. This indicate that the validation are pretty accurate. While the better score got from testing indicate that the models are perform well on new data.

3.5 Result analysis and visualization

This section discusses the analysis of each result from the best model implementation on the test set for each aspect category.

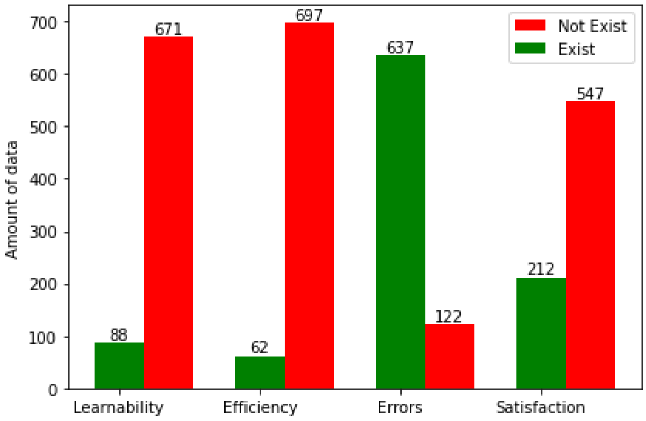

Figure 8 Aspect classification result

From Figure 8, it can be conclude that only a small number of reviews with learnability and efficiency aspect were discussed, and a greater number can be found for satisfaction aspect. While errors aspect are the most discussed topic on the reviews, which is 637 of 759 reviews or 83,93% of data. This indicate that there are something that make most of users talking about the system of the application.

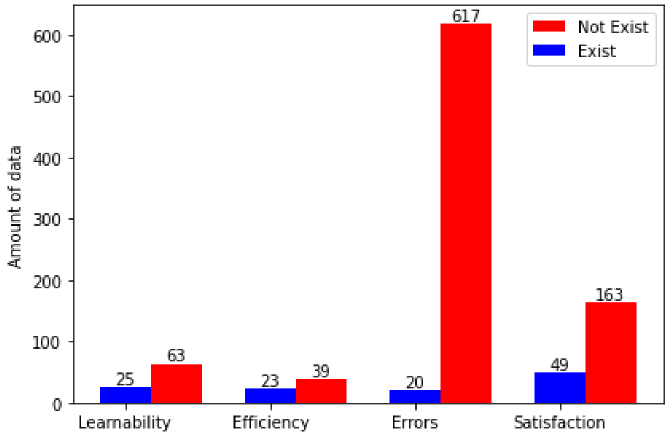

Figure 9 Sentiment classification result

Meanwhile, Figure 9 shows the sentiment classification result for each data which have corresponding aspect. It can be seen that all of the aspect mostly have negative review than positive review, especially on errors aspect which most of the reviews give a negative opinion.

While the word cloud for each sentiment for every aspect are shown on the figures 10.

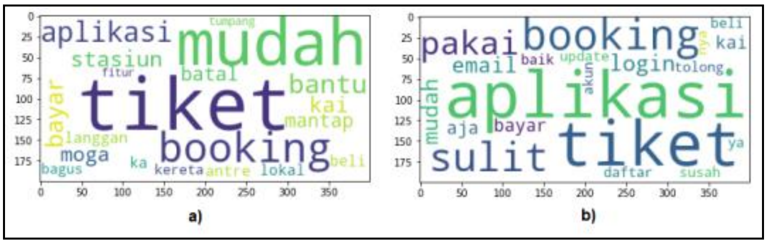

Figure 10 Word cloud for positive sentiment (a) and negative sentimen (b) on aspect learnability

On the Figure 10, it shows that the most popular words for positive class on learnability aspect are “tiket”, “mudah”, and “booking”. That means the user think the apps are easy to use especially on booking the ticket. While on the negative class, the most popular words are “pakai”, “aplikasi”, “tiket”, “sulit”, and “booking”, which means the user felt that the apps are hard to use, which also on ticket booking.

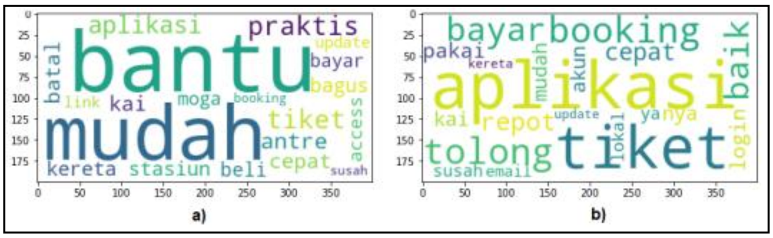

Figure 11 Word cloud for positive sentiment (a) and negative sentimen (b) on aspect efficiency

On the Figure 11, it shows that the most popular words for positive class on efficiency aspect are “bantu” and “mudah”, which means the user think the apps are easy and helpfull. While on the negative class, the most popular words are “aplikasi”, “tiket”, “booking”, with small appearance of "repot", "susah" and "tolong", which indicate that the ticket booking system of the apps have a certain problem and make the apps hard to use and not efficient.

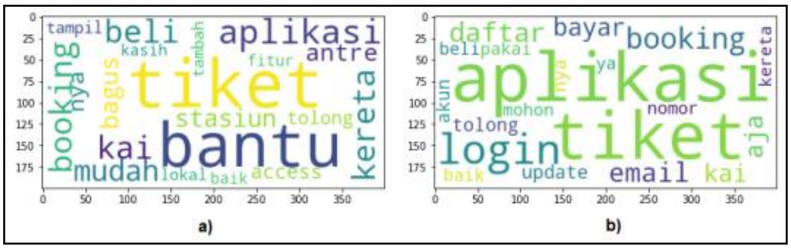

Figure 12 Word cloud for positive sentiment (a) and negative sentimen (b) on aspect errors

On the Figure 12, it can be seen that the most popular words for positive class on errors aspect are "bantu" and "tiket". It means the user think the apps are helpfull and indicate the ticketing system are working fine. While on the negative class, the most popular words are "aplikasi", "login", "tiket", followed by some small appearance words, which is the part of the system it self such as login, update, registration, payment, and booking. This indicate that those system are the main problem that make the user give negative feedback on errors aspect.

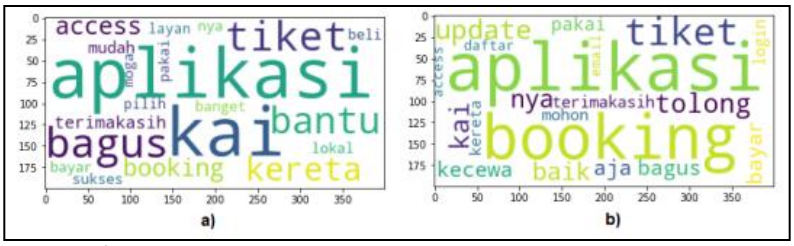

Figure 13 Word cloud for positive sentiment (a) and negative sentimen (b) on aspect satisfaction

Finally, for Figure 13, it shows that on positive class for aspect satisfaction, the most popular words are "aplikasi", "kai", "bantu", "bagus", which means they think the apps are helpfull and good. While on the negative class, the most popular words are "aplikasi" and "booking", and followed by small sized word "kecewa". The positive word like "bagus" still can be found here since it might be a negation which not being handled on this research yet.

4. CONCLUSIONS

From the test that has been conducted, the result implies that classification model that built using Naive Bayes are good at achieving high score on recall, while SVM with default parameters from Sklearn library which is rbf kernel with 1 alpha and 0.1 gamma, are good at resulting high precision. Meanwhile, from the third skenario which implementing SVM using best parameters from hyperparameter tunning are able to produce the best result with good balance on precision and recall which got average score on accuracy 91.63%, f1-score 75.55%, precision 77.60%, and recall 74.47%. This happened because the good performance of SVM in processing high dimensional data especially with the right parameter.

The result of the sentiment analysis shows that the most discussed topic on the reviews is errors aspect which reach 83,93% of data and most of them have negative sentiment, and the others aspect is not that much, but still have high percentage of negative reviews. This result indicate that the system of the application have high errors rate. Based on the word cloud, the most disscussed system are login, update, registration, payment, and booking.

This research still need more improvement. Such as handling the negation on the data which may affect the features membership on each class. So, the future research are suggested to handle the negation and try different feature extraction and classification method for comparation.

REFERENCES

[1] N. Fikria, "Analisis Klasifikasi Sentimen Review Aplikasi E-Ticketing Menggunakan Metode Support Vector Machine Dan Asosiasi", Undergraduate, Universitas Islam Indonesia, 2018.

[2] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, pp. 1-167, 2012.

[3] S. Astuti, "Analisis Sentimen Berbasis Aspek Pada Aplikasi Tokopedia Menggunakan LDA dan Naïve Bayes", Undergraduate, UIN Syarif Hidayatullah, 2020.

[4] S. Ailiyya, "Analisis Sentimen Berbasis Aspek Pada Ulasan Aplikasi Tokopedia Menggunakan Support Vector Machine", Undergraduate, UIN Syarif Hidayatullah, 2020.

[5] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta. “Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews”. Journal of Computational Science, 27, 386–393, 2018. https://doi.org/10.1016/j.jocs.2017.11.006

[6] A. P. Rodrigues and N. N. Chiplunkar, "Aspect Based Sentiment Analysis on Product Reviews," 2018 Fourteenth International Conference on Information Processing (ICINPRO), 2018, pp. 1-6, doi: 10.1109/ICINPRO43533.2018.9096796

[7] C. Aggarwal, Data Mining, 1st ed. Springer, Cham, 2015. https://doi.org/10.1007/978-3-319-14142-8

[8] A. Wibawa, A. Kurniawan, D. Murti, R. Adiperkasa, S. Putra, S. Kurniawan, and Y. Nugraha. “Naïve Bayes Classifier for Journal Quartile Classification”. International Journal of Recent Contributions from Engineering, Science & IT (IJES), vol. 7, no. 2, 2019. https://doi.org/10.3991/ijes.v7i2.10659

[9] S. R. Wardhana, "Analisis sentimen pada opini pengguna Aplikasi Mobile untuk evaluasi faktor kebergunaan", Postgraduate, Institut Teknologi Sepuluh Nopember, 2017.

[10] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, “Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus,” in Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014, Oct. 2014, pp. 66–69, doi: 10.1109/IALP.2014.6973519.

[11] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.

[12] G. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter", INTEGER: Journal of Information Technology, vol. 2, no. 1, pp. 32-41, 2017. https://doi.org/10.31284/j.integer.2017.v2i1.95

[13] O. Heranova. “Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring”. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 3, no. 3, pp. 443-450, 2019. https://doi.org/10.29207/resti.v3i3.1275

[14] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, “Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier],” IEEE Computational Intelligence Magazine, vol. 13, no. 4, pp. 59–76, 2018.

[15] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. 3rd ed. Waltham: Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.