EDUMATIC
Jurnal Pendidikan Informatika

e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 6 No. 1, Juni, 2022, **Hal. 143-149**
DOI: 10.29408/edumatic.v6i1.5613

Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification

Ivan Rifky Hendrawan[1,*], Ema Utami [1], Anggit Dwi Hartanto [1]

[1] Program Studi Teknik Informatika, Universitas Amikom Yogyakarta, Indonesia
[*] Correspondence: ivanrifky@students.amikom.ac.id

Received: 10 April 2022 | Revised: 16 April 2022 | Accepted: 3 June 2022 | Published: 20 June 2022

Abstract

Online reviews are critical in supporting purchasing decisions because, with the development of e-commerce, there are more and more fake reviews, so more and more consumers are worried about being deceived in online shopping. Sentiment analysis can be applied to Marketplace product reviews. This study aims to compare the two categories of Naïve Bayes and XGBoost by using the two vector spaces wod2vec and TFIDF. The methods used in this research are data collection, data cleaning, data labelling, data pre-processing, classification and evaluation. The data scraping process produced 25,581 data which was divided into 80% training data and 20% test data. The data is divided into two classes, namely good sentiment and bad sentiment. Based on the research that has been done, the combination of Word2vec + XGBoost F1 scores higher by 0.941, followed by TF-IDF + XGBoost by 0.940. Meanwhile, Naïve Bayes has an F1-Score of 0.915 with TF-IDF and 0.900 with word2vec. Classification using XGBoost proved to be able to classify unbalanced data better than Naïve Bayes.

Keywords: marketplace; naïve bayes; sentiment analysis; TF-IDF; XGBoost

INTRODUCTION

The use of internet facilities in Indonesia reached 73.7% of the total population at the beginning of 2022. This increased significantly by 2.1 million (+1.0 percent) between 2021 and 2022 (Kemp, 2022). This growth has the opportunity to make Indonesia a market for the marketplace (Rohman et al., 2020). The marketplace is one of the transaction media, one of the available features is reviews and ratings (Wang et al., 2022). In general, generating each product review is tied to the rating level, making users leave biased comments. For example, for a tolerant user, even though the user is very dissatisfied with the product, the ranking still makes him give neutral comments that cannot indicate the quality of the product. Online reviews can also be used as a data source for decision-making in both business and management (Bi et al., 2019). Customer reviews are critical to support decisions in buying and selling transactions because with the development of e-commerce, more and more fake reviews are causing consumers to fear being cheated when shopping online (Wang et al., 2022). Customer reviews are crucial because they determine user or buyer satisfaction with a product (Kevin et al., 2020). For this reason, in classifying online reviews so that they can be processed and used as strategies, there needs to be a solution, namely sentiment analysis. Sentiment analysis can answer the problem of text classification well as done in classifying data (Warsito & Prahutama, 2020).

Several techniques that can be used to find out user reviews of a product are sentiment analysis (Lestandy et al., 2021). Product research has also been carried out by Jayadi (2022) comparing several machine learning algorithms to determine the best way for sentiment analysis on product reviews contained in five E-Commerce in Indonesia. One of the algorithms often used in sentiment analysis is Naive Bayes. Naïve Bayes Classifier (NBC) is a classification method on the theorem. The classification method uses probability and statistical methods put forward by the English scientist Thomas Bayes (Permadi, 2020), which predicts future opportunities based on previous experience, so it is known as Bayes' Theorem (Yennimar & Rizal, 2019) as done by Yennimar & Rizal (2019) using the the-Nearest Neighbor (KNN) and Naïve Bayes algorithms where Naïve Bayes has a higher accuracy of 89.00% than KNN. Another study related to Naïve Bayes in classifying product reviews was also carried out by Rohman et al. (2020) where KNN produced a higher accuracy of 76.2% compared to Naïve Bayes 52.4%, for accuracy results cannot be said to be optimal because there are still some spam reviews and irrelevant. This makes the main capital for current research to add a data cleaning stage where later this stage will eliminate irrelevant reviews and spam so that the data can be classified properly. This research will also optimize in terms of the preprocessing stage so that problems such as misclassified data can be overcome.

Stage text preprocessing, it is necessary to add a word normalization, according to Rohman et al. (2020) word normalization is able to detect non-standard words and abbreviations so that they can automatically correct words to match the EYD. Preprocessing also serves to optimize the mining process because the data is not always in good condition and structured. As done by Sihombing et al. (2021) when classifying review text on Shopee using Naïve Bayes, it produces an accuracy of 85%, there is still misclassified data. This makes a reference for the author to optimize the preprocessing stages in the form of case folding, remove punctuation, stop word removal, short word removal, word normalization, stemming and tokenization so that data duplication, missing values can be resolved. Naïve Bayes is one of the efficient and fast classification methods in classifying text (Sihombing et al., 2021), the amount of data will certainly affect the selection of the right method.

The majority of previous studies only used relatively little data and had balanced data classes. However, not all of the data found have a balanced class. Currently the XGBoost method is able to classify unbalanced data, this is supported by Afifah et al. (2021) which produces an accuracy of 96.24% with a product review dataset collected from the Google Play Store. Extreme Gradient Boosting (XGBoost) is a technique in machine learning for regression analysis and classification based on the Gradient Boosting Decision Tree (GBDT). Extreme Gradient Boosting is a decision tree-based ensemble machine learning algorithm that uses a gradient enhancer framework (Afifah et al., 2021). Another study Akter et al.,(2021) discusses the comparison of the TF-IDF model on the Logistic Regression and XGBoost algorithms where XGboost produces an F1 Score value of 0.91% higher than Logistic Regression 0.90%.

The research that be carried out will use word2vec insertion, not only using TF-IDF which is mostly used in previous studies. This is because, according to Nurdin et al. (2020) the advantage of word2vec embedding is that it is able to process semantic relationships between words better than TF-IDF. The research that be carried out will also optimize the preprocessing process for unstructured text data to be cleaned and able to be processed by the classification algorithm so that classification errors such as in the study can be corrected (Sihombing et al., 2021). Thus, this study will try to compare the performance of the Naïve Bayes and XGBoost algorithms against online reviews in the marketplace, which will later be measured using a confusion matrix. This study aims to compare the performance of the Naïve Bayes algorithm model with XGBoost on product reviews in e-commerce and determine which algorithm is best for classifying unbalanced data. The results of this study are also expected to be a strategy for companies or brand owners to determine local Indonesian products that are liked and needed by the community.

METHOD

Based on the literature study that has been described, the research flow can be seen from Figure 1. Based on the research flow in Figure 1, the following is shown for an explanation of the seven stages.

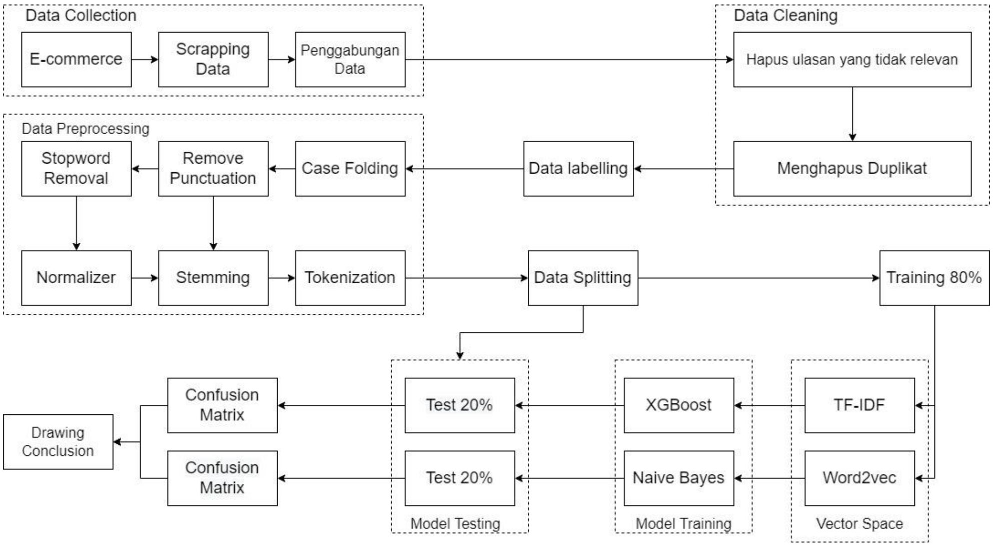

Figure 1. *Research flow*

The first stage is data collection, data obtained by scraping using the python programming language using the Google Collaboratory platform. The data taken in the amount of 25,581 raw data will then be saved in .csv format. The data is taken based on 5 local brands at Shopee Marketplaces with the bestselling apparel category. After the data has been collected, the next step is data cleaning where in this process the data is cleaned in the form of spam reviews that will be removed, irrelevant reviews will also be removed (Hidayat et al., 2022). After cleaning, the data will enter the labelling process, where for stars 1 to 3 it will be labelled as bad, while for stars 4 to 5 it will be labelled as good.

After the labelling process is complete, the data will enter the pre-processing process. For the preprocessing stage in this study, After the labeling process is complete, the data will enter the preprocessing process. For the preprocessing stage in this study, using the first 7 stages are case folding this stage processes words into lower case examples like this *Produkk ini jeleekkk BGT aku kecewaa membeli di sini hm :)* ➔ *produkk ini jeleekkk bgt aku kecewaa membeli di sini hm :)*. Next is the remove punctuation stage, this stage processes the Ascii code and removes it. The next stage is removed number and short word where sentences containing numbers and short words will be removed. Word normalizer is a step to normalize short words, words that contain the meaning of typing errors into standard words. Example of the word "bguss" is changed to "bagus". Next is the stopword removal stage, this stage is the stage to remove conjunctions such as "yang", "di", "ke". Tokenization is the stage to change sentences into words. The last stage is stemming where this stage will change words into basic words such as "membeli" to "beli". After the preprocessing process is complete the data will be divided into training data and test data. For the data to be processed will be given 80% for training data and 20% for test data.

The next stage is data training, where before entering the text classification process that is ready, it will enter the word weighting process and word embedding. Word embedding is an NLP technique that converts a basic word into a real-valued vector (Kurniawan & Maharani, 2020). In this study, two vector spaces are used, the first is TF-IDF and the second is Word2vec. For TF-IDF approach, see equation (1) (Setiawan et al., 2019):

$$W_{ij} = TF_{ij} \times \log\left(\frac{D_i}{df_i}\right) \tag{1}$$

In addition to using TF-IDF, this research will also use word2vec embedding where this model is able to represent the semantic relationship between words well.For example, if there is the word “kecil”, then after entering the wordvec model, the “kecil” word will be matched with the word that is closest to the relationship. For example, the word “Kecil” fits perfectly with a “sempit” word of 0.592 then followed by the word “cingkrang”, “ngetat” and so on. After the word has gone through the weighting process, it will then be classified using two algorithms. In this study, there are four models to be tested, namely the combination of XGBoost+Word2vec, Naïve Bayes+Word2vec, XGBoost+TF-IDF and Naive Bayes+TF-IDF

The final stage is an evaluation where the classification results of the 4 models will be evaluated using a confusion matrix. According to (Kotu & Deshpande, 2019) the confusion matrix is a tabulation of calculations based on the evaluation of the performance of the classification model based on the object of research regarding right or wrong (Shuai et al., 2018).

RESULT AND DISCUSSION

Result

The data scraping process produced 25.581 data which was divided into 80% training data and 20% test data. In the training data, it is known that the number of good labels is 8141 while for bad labels as many as 2323 this is an unbalanced classification challenge. The next stage is the text data will enter the data cleaning stage where data that is less relevant and spam will be cleaned. The cleaned data is then labeled as good and bad. Bad criteria for rating 1 to 3, while good criteria are rating 4 to 5. The next stage is preprocessing, at the preprocessing stage, case folding, eliminating punctuation, eliminating numbers, tokenization, stopword removal, stemming and word normalization. In this study, the optimization stage in the preprocessing stage is prioritized (Bi et al., 2019), one of which is word normalization (Amin et al., 2021). Normalization of this word is so important and very crucial because in previous studies that discussed the same classification, the factor that affected the difference in accuracy between research was word normalization (Yennimar & Rizal, 2019). The following is a review text that has gone through the preprocessing process and will be presented in the form of a word cloud in figure 2.

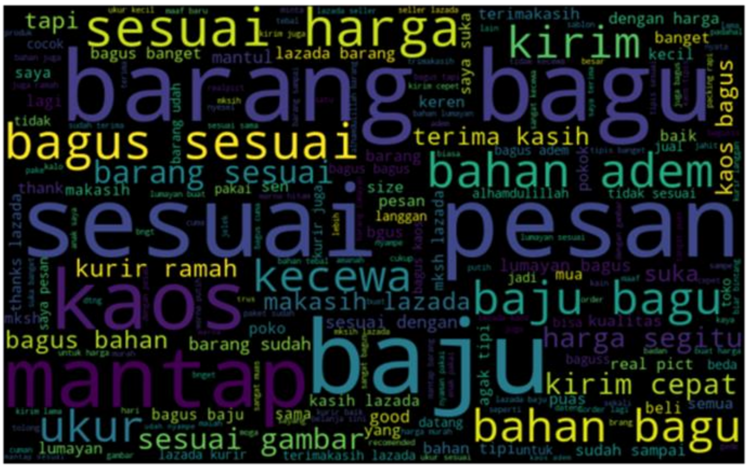

Figure 2. Word cloud product reviews

The next stage is the model evaluation stage. The four models that have been trained using training data will be measured using the Confusion Matrix which can be seen in table 1.

Table 1. Evaluation of the model using the confusion matrix

Model	Accuracy	Precision	Recall	F1 Score
XGBoost+TFIDF	0.892	0.911	0.971	0.940
XGBoost+Word2vec	0.893	0.914	0.969	0.941
NB+ TF-IDF	0.887	0.915	0.961	0.915
NB+Word2vec	0.833	0.953	0.852	0.900

Based on table 1, the results of testing the XGBoost+Word2vec algorithm have the highest accuracy of 0.893 and the value of F1 Score is 0.941 followed by XGBoost+TF-IDF with an accuracy of 0.893, F1 Score 0.940. Meanwhile, for the Naïve Bayes algorithm, the combination with TF-IDF and Word2vec has an accuracy value of 0.887 and 0.833. The XGBoost Algorithm is proven to be able to classify unbalanced datasets compared to the Naive Bayes Algorithm. This can happen because Naive Bayes only looks at word probabilities while the way the XGBoost algorithm works in machine learning competition is due to its strong handling of various data types, relationships, distributions, and various hyperparameters that can be refined. From table 1 it is known that the comparison of the XGBoost algorithm is better than naive bayes in classification. This proves from Afifah et al. (2021) research that XGboost is able to classify text data that is not well balanced. This can happen because XGBoost has the advantages of being able to perform parallel which can expect fast features, has flexibility in setting goals built in cross validation, has regularization features, and overcomes splits when downsides are negative. Where this advantage is able to make unbalanced data classified accurately.

Discussion

From the data a total of 25,581 reviews with a distribution of 80% for training data and 20% for testing data by being tested using a test scenario of 4 trials, namely using two classification algorithms, being able to classify reviews well and being able to overcome problems that have not been resolved by previous research. In this study, it is possible to find out what factors affect the accuracy of the classification algorithm, especially for Naive Bayes. The advantage of using Naive Bayes is that this method only requires a little training data to determine the parameter estimates needed in the classification process. The Naive Bayes algorithm can detect or filter spam in this experimental study. This study also proves that data cleaning at the preprocessing stage in the form of removing spam and irrelevant words can improve the accuracy of the Naive Bayes algorithm. It is proven that the results of Naive Bayes + TF-IDF in this study resulted in an accuracy of 0.892%, higher than the previous study (Rohman et al., 2020) which produced an accuracy of 52.4%. Naive Bayes is one of the popular algorithms for sentiment classification.

This can be proven by the current research results which have a relatively good level of accuracy. The accuracy results obtained are also influenced by the preprocessing stage carried out. The seven preprocessing steps can eliminate noise in the data so that the data that will enter the analysis process is better. Words such as abbreviations and typos can also be processed appropriately at the word normalization stage. The method proposed by the authors can also overcome other problems related to duplication of reviews or tests. This has proven to be able to overcome weaknesses regarding misclassified data (Sihombing et al., 2021). With the same method using TF-IDF and Naive Bayes, this study has a higher cause of 0.892% compared to the previous study which was only 0.85%. This is due to the effect of complete preprocessing so that the problem of misclassified, redundant data can be resolved.

Another result of this study is that the combination of the XGBoost+Word2vec algorithm produces an F1 Score value of 0.941% higher than the previous study (Akter et al., 2021) 0.91% using XGBoost+TFIDF. This is because the advantage of word2vec in detecting relationships between words is better than TF- IDF.Word2vec Relying on local information of the language learned semantics of a certain word is determined by the surrounding words.

CONCLUSION

Sentiment analysis can be applied to Marketplace product reviews to be used as a means of product improvement for sellers. The word normalization method at the preprocessing stage can handle the problem of data misclassification. Based on the research that has been done, the combination of Word2vec + XGBoost resulted in a higher F1 score of 0.941, followed by TF-IDF + XGBoost 0.940. Meanwhile, Naïve Bayes has an F1-Score of 0.915 with TF-IDF and 0.900 with word2vec. To handle unbalanced datasets, the XGBoost algorithm is better than Naive Bayes. This is because XGboost is capable of parallel processing, which can speed up computing and overcome splits during negative loss. Word2vec in this study is also better at representing words into vectors. This is because word2vec can represent the relationship between words better than TF-IDF. It is better to apply sarcasm detection in future research to optimize the classification again.

REFERENCES

Afifah, K., Yulita, I. N., & Sarathan, I. (2021). Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier. *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 22–27. https://doi.org/10.1109/ICAIBDA53487.2021.9689735

Akter, M. T., Begum, M., & Mustafa, R. (2021). Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors. *International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 40–44. https://doi.org/10.1109/ICICT4SD50815.2021.9396910

Amin, S., Uddin, M. I., AlSaeed, D., & Khan, A. (2021). Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches. *Complexity*, 2021, 1–12. https://doi.org/10.1155/2021/5520366

Bi, J.-W., Liu, Y., & Fan, Z.-P. (2019). Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking. *Information Sciences*, *504*, 293–307. https://doi.org/https://doi.org/10.1016/j.ins.2019.07.025

Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. *Procedia Computer Science*, *197*, 660-667. https://doi.org/https://doi.org/10.1016/j.procs.2021.12.187

Jayadi, S. F. N. H. R. (2022). Sentiment Analysis Of Indonesian E-Commerce Product Reviews Using Support Vector Machine Based Term Frequency Inverse Document. *Journal of Theoretical and Applied Information Technology*. *99*(17), 4316–4325.

Kemp, S. (2022). *Digital 2022 Indonesia :Internet use in Indonesia 2022*. Datareportal. https://datareportal.com/reports/digital-2022-indonesia

Kevin, V., Que, S., Iriani, A., & Purnomo, H. D. (2020). Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization (Online Transportation Sentiment Analysis Using Support Vector Machine Based on Particle Swarm Optimization). *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *9*(2), 162–170. https://doi.org/10.22146/jnteti.v9i2.102

Kotu, V., & Deshpande, B. (2019). Chapter 8 - Model Evaluation. In V. Kotu & B. Deshpande (Eds.), *Data Science (Second Edition)* (Second Edition, pp. 263–279). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00008-3

Kurniawan, F. W., & Maharani, W. (2020). Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec. *eProceedings of Engineering*, *7*(2), 7821–7829.

Nurdin, A., Seno aji, B., Bustamin, A., & Abidin, Z. (2020). Perbandingan Kinerja Word Embedding Word2vec, Glove, dan Fasttext pada Klasifikasi Teks. *Jurnal Tekno Kompak*, *14*(2), 74–79. https://doi.org/10.33365/jtk.v14i2.732

Permadi, V. A. (2020). Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura. *Jurnal Buana Informatika*, *11*(2), 141–151. https://doi.org/10.24002/jbi.v11i2.3769

Rohman, A. N., Luviana Musyarofah, R., Utami, E., & Raharjo, S. (2020). Natural Language Processing on Marketplace Product Review Sentiment Analysis. *2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–5. https://doi.org/10.1109/ICORIS50180.2020.9320827

Shuai, Q., Huang, Y., Jin, L., & Pang, L. (2018). Sentiment Analysis on Chinese Hotel Reviews with Doc2Vec and Classifiers. *IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 1171–1174. https://doi.org/10.1109/IAEAC.2018.8577581

Sihombing, L. O., Hannie, H., & Dermawan, B. A. (2021). Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algortima Naïve Bayes Classifier. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 233–242. https://doi.org/10.29408/edumatic.v5i2.4089

Lestandy, M., Abdurrahim, A., & Syafa'ah, L. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(4), 802-808. https://doi.org/10.29207/resti.v5i4.3308

Wang, Q., Zhang, W., Li, J., Mai, F., & Ma, Z. (2022). Effect of online review sentiment on product sales: The moderating role of review credibility perception. *Computers in Human Behavior*, *133*, 107272. https://doi.org/https://doi.org/10.1016/j.chb.2022.107272

Wang, X., Zhou, T., Wang, X., & Fang, Y. (2022). Harshness-aware sentiment mining framework for product review. *Expert Systems with Applications*, *187*, 115887. https://doi.org/10.1016/j.eswa.2021.115887

Warsito, B., & Prahutama, A. (2020). Sentiment Analysis on Tokopedia Product Online Reviews Using Random Forest Method. The 5th International Conference on Energy, Environmental and Information System (ICENIS 2020). 1-10, EDP Sciences. https://doi.org/10.1051/e3sconf/202020216006

Setiawan, E. B., & Nugraha, F. N. (2019). Implementation of Decision Tree C4. 5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter. In *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* 114–119. IEEE. https://doi.org/10.1109/IC3INA48034.2019.8949601

Yennimar, Y., & Rizal, R. A. (2019). Comparison of Machine Learning Classification Algorithms in Sentiment Analysis Product Review of North Padang Lawas Regency. *Sinkron: jurnal dan penelitian teknik informatika*, *4*(1), 268-273. https://doi.org/10.33395/sinkron.v4i1.10416