SENTIMENT ANALYSIS OF PRODUCT REVIEWS DATA ON TOKOPEDIA BY COMPARING THE PERFORMANCE OF CLASSIFICATION ALGORITHMS

Dwi Widiastuti[1], Isram Rasal[2,] Dessy Wulandari Asfary Putri[3]

[1]Information Systems, Faculty of Computer Science and Information Technology, Gunadarma University
[3]Informatics Engineering, Faculty of Industrial Technology, Gunadarma University
[2]Computer Systems, Faculty of Computer Science and Information Technology, Gunadarma University

dwidiastuti@staff.gunadarma.ac.id, isramrasal@staff.gunadarma.ac.id
, dessywap@staff.gunadarma.ac.id,

Abstract

Article Info
Received : 10 May 2022
Revised : 30 May 2022
Accepted : 05 June 2022

Social media is a medium where people can express their opinion on something. Opinion mining or sentiment analysis, which is studying people's sentiments towards certain entities. This can be used by companies to find out people's responses to a sales product. Sentiment analysis has received a lot of attention in recent years. Sentiment analysis is one of the main tasks of NLP (Natural Language Processing). In this paper, sentiment polarity categorization becomes the basis for sentiment analysis problems in product reviews. A general process for sentiment polarity categorization is proposed with a detailed description of the process. The data used in this study is an online product review collected from the Tokopedia application. Classification is carried out on sentence level categorization and star rating level categorization. There are three models used to compare the classification process, namely SVM, Random Forest, and Naïve Bayes models. The results of this research paper are in the form of a comparison of the performance of the three models against the polarity categorization of product review sentiment on Tokopedia.

Keywords: Sentiment analysis, Sentiment polarity categorization, Nave bayes, Product reviews, Random Forest, SVM

1. INTRODUCTION

According to the Big Indonesian Dictionary (KBBI), sentiment is an opinion or view that is based on an exaggerated feeling of something. According to some researchers, sentiment analysis [1], or also known as opinion mining, is the study of people's sentiments towards certain entities. Marketplace is an e-commerce media that connects sellers and buyers, as well as a place to conduct business activities [2]. Referring to data compiled from the iPrice website, Tokopedia ranks first as the most visited marketplace, reaching 157 million visitors per month until the 3rd quarter of 2021. This encourages researchers and application developers to take advantage of the collection and analysis of product review data so that promos are offered. offered on target. Therefore,

Online data has weaknesses that can hinder the sentiment analysis process [3], among others, everyone can freely post content, so the quality of their opinions cannot be guaranteed. For example, online spammers, they post spam in forums, while their opinions are irrelevant to the related topic. Another weakness is opinion labels that are not always available, such as the absence of giving emoticons as positive or negative opinions, or not responding by not giving a rating in the form of a star scale.

Sentiment polarity category is one of the fundamental problems in sentiment analysis, namely categorizing text into one particular sentiment polarity (positive, or negative, or neutral). Based on the scope of the text, researcher Tan et al [4] divides into three levels of sentiment polarity categorization, including document level, sentence level, and entity and aspect levels. Document level is whether a document as a whole expresses a negative or positive sentiment. Sentence level relates to sentiment categorization in each sentence. While the entity and aspect levels are targeting what people actually like or dislike in their opinion.

Several studies have stated that product reviews have a positive influence on product purchase intentions, and even become a factor of consideration to ensure the quality of the product to be purchased. Researchers Lutfi and Permatasari [5] analyzed product reviews on the Bukalapak marketplace to find out whether the sentiment of user reviews was positive or negative by using the Support Vector Machine approach with an accuracy of 93.42%.

Research conducted by Muljono and Dian [6], describes a sentiment analysis of opinion data on Twitter on marketplace site services in Indonesia using the Naive Bayes algorithm with an accuracy of 93.33%.

Sentiment categories are the basis for classification problems, where features containing opinions or sentiment information must be identified before classification. The difficult thing for the system is the use of abbreviations and non-standard language. In Hu and Liu's research [7] has summarized a list of words based on customer reviews, where there are 2006 positive words and 4783 negative words. The study also included some misspelled words that are often present in social media content. In Pang and Lee's research [8], feature selection was carried out by extracting subjective sentences to eliminate objective sentences. They propose a text categorization technique that is able to identify subjective content using the minimum cuts method.

This paper will discuss the sentiment analysis of product review data with the keyword “fashion” collected from Tokopedia by comparing the SVM (Support Vector Machine), Naïve Bayes, and Random Forest models. The analysis was carried out on the product comments section consisting of review contents in free text format (sentences) and a rating of 1 to 5 stars. The use of non-standard language and abbreviated writing became a problem in sentiment analysis, so the Natural Language Processing (NLP) approach was used to improve language in the review content so as to achieve maximum performance. The data is processed using R programming tools and RStudio (IDE), then tested with three models and the same object value.

2. METHODS

This study applies the schematic in Figure 1 [9] which shows the process flow diagram for categorization as well as an outline of this paper. Phase 1 describes the flow of data collection and extraction of sentiment sentences and then marks them. Phase 2 is the implementation of the algorithm to identify phrases and the sentiment score process, then vector features are carried out. In phase 3, sentiment polarity categorization is carried out using three models, namely SVM, Naïve Bayes, and Random Forest. Then the results are evaluated and compared.

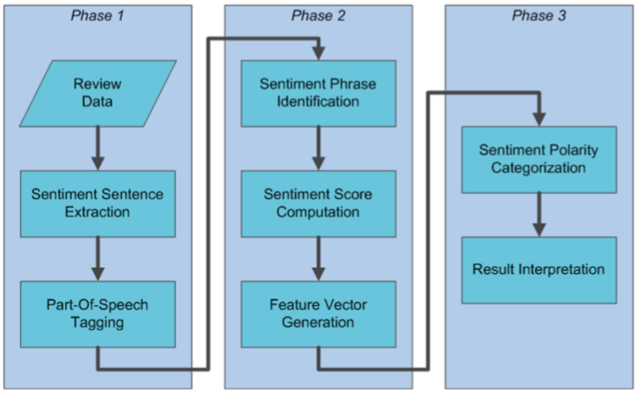

Phase 1
Review Data
Sentiment Sentence Extraction
Part-Of-Speech Tagging
Phase 2
Sentiment Phrase Identification
Sentiment Score Computation
Feature Vector Generation
Phase 3
Sentiment Polarity Categorization
Result Interpretation

Figure 1. Sentiment Polarity Categorization Process Flow

The data used in this paper are product reviews with the keyword "fashion" collected from Tokopedia between January and March 2022. Reviews on Tokopedia were posted by more than 3.2 million reviewers (customers) on 225,663 products and more than 5.1 million product reviews from four main categories, namely: beauty, sports, electronics, and fashion, can be seen graphically in Figure 2. The keyword “fashion” was chosen because it has the highest number of reviews on Tokopedia with 2.3 million reviews. Information on reviews includes, among others: reviewer id, product id, rating, review time, helpfulness (snippets of information provided by Tokopedia can be added as part of the content of comments), and review text. Every review taken must have a rating that can be used as ground truth. The rating is based on a star scale system, where the highest rating has 5 stars and the lowest rating has only 1 star. Examples of reviews used can be seen in table 1.

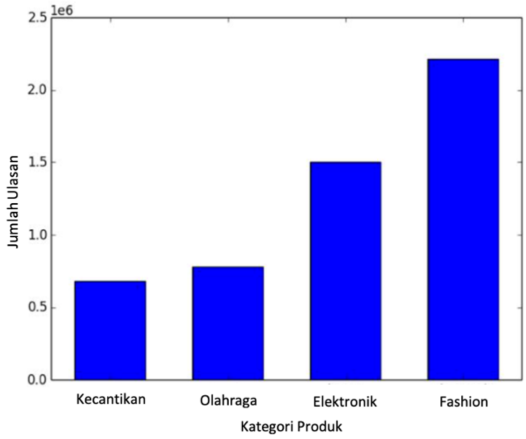


Figure 2. Collection of Review Data by Product Category on Tokopedia

Phase 1 of the flow of sentiment polarity categorization begins with the collection of reviews carried out using the Scraper application made of the Python language [2]. The Scraper application workflow is shown in figure 3. The Scraper application will download the product reviews that appear and save to a database based on the number of star ratings. The download starts from the number of 1 star ratings and will stop when the 5 star rating review page is complete. Reviews in the range of January to March 2022, this process generates 198,671 reviews which are stored in a MySQL database.

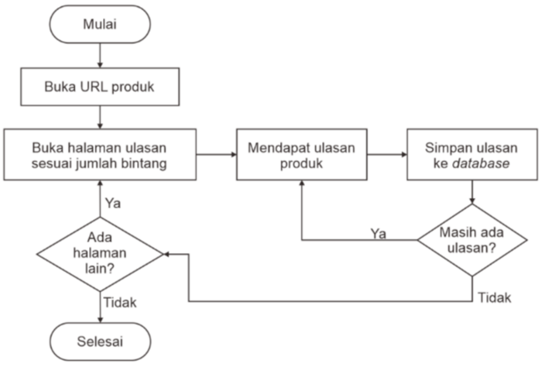


Figure 3. Scraper Application Flowchart

Then the dataset goes through pre-processing to clean and prepare the data for analysis. Pre-processing is carried out in two conditions [3], first the dataset without using the NLP approach to remove emoticons and symbols, and lowercase folding (converting all letters to lowercase). Both datasets use an NLP approach to improve non-standard language and abbreviations in the review (word normalizer). From the results of this pre-processing selection, the 198,671 reviews collected were reduced to 21,586 reviews. The results of this dataset are divided into training data and test data, then the weighting calculation of Term Frequency Inverse Document Frequency (TF-IDF) is applied and the results are compared whether the review has a relationship value of a word or phrase to the whole review or not.

$$W_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right) \qquad (1)$$

Where is the word weight (ty) to the document (dx). Meanwhile, the number of occurrences of the term (ty) in the document (dx). N is the number of all documents in the dataset and is the number of documents containing the term (ty), at least one word, namely the term (ty).$W_{x,y} tf_{x,y} df_x$

Table 1. Example of Product Reviews

No	reviewer id	product id	Rating	Review Time	Helpfulness	Review Text
1	Denny	ZHFU MAGSAFE CASE IPHONE 12 SERIES	5/5	3 weeks ago	Very satisfied	Fast response seller, get a quota bonus too. Thank you Continued success
2	Johannes	Electrical Power Meter Watt Ampere Electronic Power Meter	5/5	More than 1 year ago	Very satisfied	I have received the goods, in good condition, good packaging, fast delivery, the seller is quick to respond, for the next shopping subscription here again, thank you 🙏

3	V***a	Calculator / Calculator Joyko CC-38 / 12 Digits / Check Correct	2/5	3 months ago	Less satisfied	the number 7 can't be pressed
4	R***a	Safe Safe Electronic Safe Box Key Password Anti Fire - 23X17X17PASWORD	4/5	3 weeks ago	Satisfied	Neat and Comfortable House. Good Product Material. it's just that for the opening the next day why can't it? the password is correct. even tried with the original password, it didn't work. is there any other way than using the key? Thank you
5	Temmy	electronic pcb ro dispenser	5/5	More than 1 year ago	Very satisfied	Items are suitable and good, well packaged, recommended 👍

Phase 2 of the flow of sentiment polarity categorization begins with the identification of words/phrases. Each word of a sentence has a syntactic role that defines how it is used. According to the KBBI, syntax is also known as the science of sentence structure. Of the 21,586 reviews collected, all sentences were tokenized into separate words. Separated words are words that contain verbs, nouns, adjectives, adverbs, prepositions, conjunctions, interjections, and pronouns. Sentiment sentences are sentences that contain at least one positive or negative word. In sentiment sentence extraction, it was suggested by Pang and Lee [10] that all subjective and objective words should be removed using the POS (part-of-speech) tagger developed for the Penn Treebank Project [9]. Each sentence is then tagged using a POS tagger. Then the words that are not used will be filtered out, such as nouns and pronouns which usually do not contain sentiments, while the words that are used, such as adjectives, adverbs, and verbs, contain expressions of sentiment. Of the 21,586 sentences in this review, processed using the POS tagger application made in Python and the results obtained are more than 15 thousand adjectives, about 12 thousand adverbs, and more than 46 thousand verbs.

Words such as adjectives and verbs can contain opposite sentiments in addition to negative words, an example sentence found in electronic device reviews: “Internal speakers have their uses too but so far no updates”. The word "update" is a positive word according to Pang and Lee's research [10], but the phrase "no update" makes the sentence contain negative sentiments. So to identify these phrases, two types of phrases have been identified [11], namely: negation-of-adjective (NOA) and negation-of-verb (NOV). In table 2, it can be seen the phrases that often appear in the review sentence data.

Table 2. Most Negative Sentiment Phrases Found

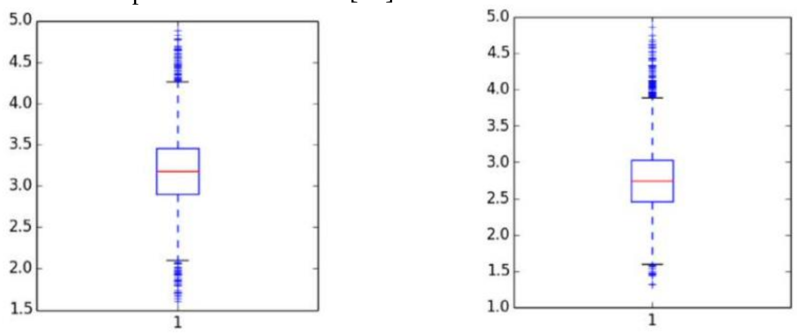
PHRASE	TYPE	INCIDENT
not feasible	NOA	269
not wrong	NOA	154
not bad	NOA	152
no happier	NOA	142
not good	NOA	129
do not like	NOV	425
not successful	NOV	387
Doesn't work	NOV	216
don't work	NOV	101
don't recommend	NOV	96

A sentiment token is a word or phrase that contains a sentiment. This feature will be used for sentiment categorization. If there are more positive tokens than negative ones, the sentence will be marked as positive, and vice versa. This approach applies a bag-of-word model that only counts the appearance of positive or negative tokens (words) in each sentence. To train a classifier, each training data entry needs to be transformed to a vector containing that token (a feature vector). Gan et al's research [12] based on Twitter data selected 6799 tokens, where each token was assigned a sentiment score, namely TSI (Total Sentiment Index), presenting itself as a positive token or a negative token. In this study, using TSI for certain tokens is calculated by equation 2.

$$TSI = \frac{p - \frac{tp}{tn} \text{ x } n}{p + \frac{tp}{tn} \text{ x } n} \qquad (2)$$

where p is the number of times the token appears in positive tweets and n is the number of times the token appears in negative tweets. And is the ratio of the number of positive tweets per-number of negative tweets.$\frac{tp}{tn}$

In this study, there is an imbalance of review data due to the tendency of reviews towards positive labels. To overcome this problem, the Synthetic Minority Oversampling Technique (SMOTE) method is used, which is an effective and good oversampling technique approach used to deal with imbalanced datasets because it can handle overfitting during the oversampling process for the minority class. The SMOTE approach is carried out by adding a lot of minority class data so that it is balanced with the majority class, this is done by generating synthetic data based on the nearest neighbors (k-nearest neighbors) so that it is hoped that this technical approach can have an impact on the classification performance results [13]. .



(a)Median Token Positive Word (b)Median Token Negative Word

Figure 4. Sentiment Score for Kata Token

Figure 4 shows the sentiment score for word tokens based on the star ratio, for positive word tokens the median sentiment must exceed a score of 3, and the median negative word token must be less than 3. From this study, the statistical results of positive word tokens with a mean of 3.18 and median 3.16. While the statistical results of the negative word tokens obtained a mean of 2.75 and a median of 2.71.

In this study, data that has been labeled sentiment class will be partitioned into training and testing with a distribution of 80:20. To evaluate the performance of the model, it is calculated from the estimated level of model accuracy using the 10-fold cross validation method. The dataset is partitioned into 10 subsets of the same size, each of which consists of 10 positive class vectors and 10 negative class vectors. Of the 10 subsets, 1 subset is used as validation data for testing the classification model, and the remaining 9 subsets are used as training data. The cross-validation process was then repeated 10 times, with each of the 10 subsets used exactly once as validation data. The 10 fold results are then averaged to produce the accuracy value of the model. Because the process of categorizing sentiment polarity there are 2, namely: sentence level categorization and star rating categorization, then the training data are labeled under two classes (positive and negative) for sentence level categorization. And also ROC (Receiver Operating Characteristic) curve is plotted for better performance comparison. ROC is a kind of tool that measures performance for classification problems in determining the threshold of a model [2]. To select the model, it is seen from the F1-Score value, to see how many examples are classified correctly and how appropriate the proposed model is. F1-Score is the Harmonic Mean between precision and recall. The F1-Score range is (0, 1). The F1-Score mathematical formula can be seen in equation 3. then the training data were labeled under two classes (positive and negative) for sentence level categorization. And also ROC (Receiver Operating Characteristic) curve is plotted for better performance comparison. ROC is a kind of tool that measures performance for classification problems in determining the threshold of a model [2]. To select the model, it is seen from the F1-Score value, to see how many examples are classified correctly and how appropriate the proposed model is. F1-Score is the Harmonic Mean between precision and recall. The F1-Score range is (0, 1). The F1-Score mathematical formula can be seen in equation 3. then the training data were labeled under two classes (positive and negative) for sentence level categorization. And also ROC (Receiver Operating Characteristic) curve is plotted for better performance comparison. ROC is a kind of tool that measures performance for classification problems in determining the threshold of a model [2]. To select the model, it is seen from the F1-Score value, to see how many examples are classified correctly and how appropriate the proposed model is. F1-Score is the Harmonic Mean between precision and recall. The F1-Score range is (0, 1). The F1-Score mathematical formula can be seen in equation 3. ROC is a kind of tool that measures performance for classification problems in determining the threshold of a model [2]. To select the model, it is seen from the F1-Score value, to see how many examples are classified correctly and how appropriate the proposed model is. F1-Score is the Harmonic Mean between precision and recall. The F1-Score range is (0, 1). The F1-Score mathematical formula can be seen in equation 3. ROC is a kind of tool that measures performance for classification problems in determining the threshold of a model [2]. To select the model, it is seen from the F1-Score value, to see how many examples are classified correctly and how appropriate the proposed model is. F1-Score is the Harmonic Mean between precision and recall. The F1-Score range is (0, 1). The F1-Score mathematical formula can be seen in equation 3.

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right) \qquad (3)$$

3. RESULTS AND DISCUSSION

In the final stage of phase 1, a test is carried out to see the match between words or phrases in the review data based on the accuracy value of the classification with the NLP approach and without the NLP approach, the results are shown in table 3. The test results show that the classification through pre-processing without the NLP approach produces the accuracy value is 67.52%, while with the NLP approach the accuracy value is 76.98%. From the value of the test results, overall has increased. This shows that the application of pre-processing with the NLP approach makes it possible to change non-standard words and abbreviations so that word weighting and calculation of the frequency of occurrence of words that are already uniform with TF-IDF become more effective. The results of this process can classify 21,586 review sentences with positive labels 17. 989 review sentences and negative label 3,597 review sentences. It can be seen that the number of positive label reviews is more than the negative label with a ratio of 5:1. This shows an imbalance in the positive and negative classes. Then synthetic data (oversampling) was made in the minority class (negative class) using the SMOTE method, the results can be seen in table 4.

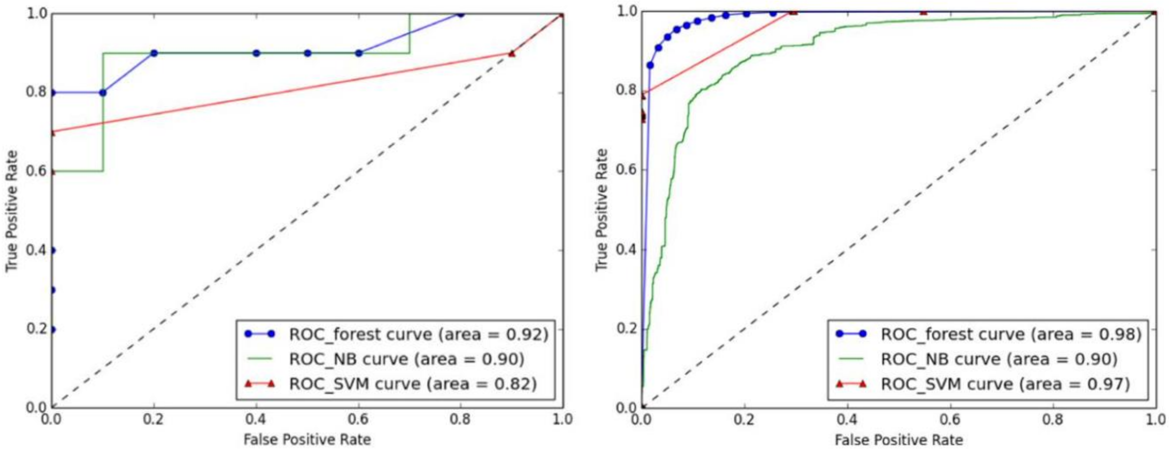
Table 3. Pre-processing Data Classification Test Results With NLP and Without NLP

Test results	No NLP Approach	With the NLP Approach
Accuracy	67.52%	76.98 %
Precision	80.00 %	80.00 %
Recall	64.23%	74.29%

Table 4. Comparison of ROC Algorithm Results Without and With SMOTE

Test results	Low Probability		High Probability	
	No SMOTE	With SMOTE	No SMOTE	With SMOTE
SVM	0.8552	0.8242	0.9802	1,000
Naive Bayes	0.8592	0.9059	0.9599	0.9649
Random Forest	0.8572	0.9202	0.9827	0.9982

To train the classifier, the training data is transformed into a vector containing tokens (feature vectors). The dataset is tested with low data of 200 feature vectors and high data of 20,000 feature vectors. As a result, the classification model shows the same level of performance based on the F1 score, where the three scores all have the same value of 0.85 for low probability and the value will increase on high probability data. In figure 5 part (a), the ROC curve shows that the three models have a fairly good performance on the lower probability data, the Naïve Bayesian classifier is superior to SVM, with a larger area under the curve. In general, the Random-Forest model performs best.



(a) Test DataLow Probability (b) Test DataHigh Probability

Figure 5. ROC Curve With SMOTE Method

In figure 5 part (b), the more the model gets more training data, the F1 scores all increase. The SVM model experienced the most significant increase from 0.61 to 0.94 as its training data increased from 200 to 20,000. The SVM model outperforms the Naïve Bayesian model and the Random-Forest model performs best for datasets across all scopes.

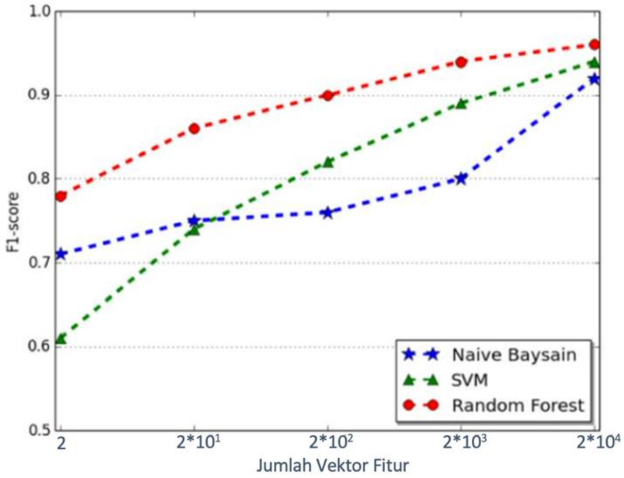


Figure 6. F1 Score Categorization Sentence Level

Three million feature vectors were created for star rating categorization. Vectors resulting from reviews that have at least a 4-star rating are labeled positive, while vectors labeled as negative are generated from 1-star and 2-star reviews. 3-star reviews are used to prepare class-neutral vectors. This complete set of vectors is uniformly labeled into three classes: positive, neutral, and negative. In addition, three subsets are obtained from the complete set, where the subset A contains 300 vectors, the subset B contains 3,000 vectors, the subset C contains 30,000 vectors, and the subset D contains 300,000 vectors each.

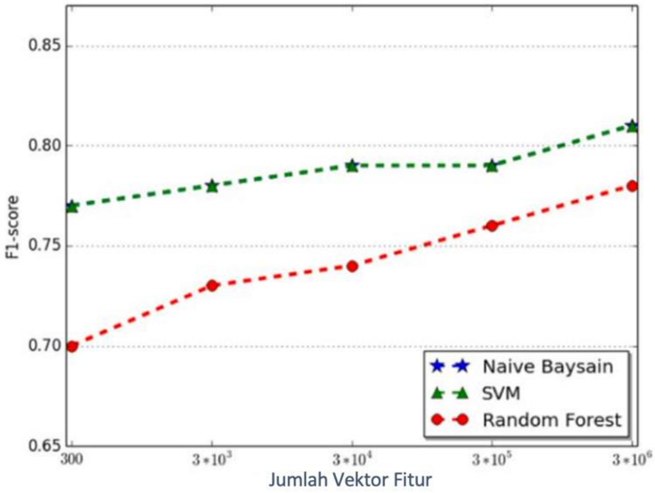


Figure 7. F1 Score Categorization Star Rating Level

Figure 7 shows F1 scores obtained based on star ratings only. It can be clearly observed that the SVM model and the Naïve Bayesian model are identical in terms of performance. Both models are generally superior to the Random-Forest model in all vector sets. However, neither of these models can achieve the same level of performance when used for sentence-level categorization, due to their relatively low performance in the neutral class. From the results of this study, it was observed that the average sentiment score was able to achieve an F1 score of more than 0.8 for the level of sentence categorization using the SMOTE method and for the star rating category the F1 score was above 0.73.

4. Conlcolusion

This study discusses the fundamental problem of sentiment analysis, namely the categorization of sentiment polarity. Online product reviews from Tokopedia.com were chosen as the data used for this study. Research for sentence level categorization and star rating level categorization has been carried out according to the detailed description of each step according to the flow of the sentiment polarity categorization process. It can be clearly observed that the SVM model and the Naïve Bayesian model are identical in terms of performance. Both models are generally superior to the Random-Forest model in star rating categorization. However, neither of these models can achieve the same level of performance as the Random-Forest model when used for review sentence level categorization, due to its relatively low performance in the neutral class.The limitations of this study are when the reviews contain implicit sentiments. Implicit sentiments usually contain a few neutral words, making it difficult to assess the polarity of the sentiment. For example, the sentences "Item as described" or "Item corresponds", which often appear in positive reviews only consist of neutral words. This will be a further research task that can be developed from this research. Including testing the sentiment polarity categorization scheme with other algorithm methods.

REFERENCE

[1] Pak A, Paroubek P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh conference on International Language Resources and Evaluation. European Languages Resources Association, Valletta, Malta

[2] Elik H. M, Kusrini, Emha T. L. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan NLP. Jurnal Eksplora Informatika, Vol. 10, No.1, pp. 32-42

[3] Mukherjee A, Liu B, Glance N. (2012). Spotting fake reviewer groups in consumer reviews. Proceedings of the 21st, International Conference on World Wide Web, WWW ’12. ACM, New York, NY, USA. pp 191–200

[4] Tan LK-W, Na J-C, Theng Y-L, Chang K. (2011). Sentence-level sentiment polarity classification using a linguistic approach. In: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation. Springer, Heidelberg, Germany. pp 77–87

[5] A.A.Lutfi, A.E.Permanasari, S.Fauziati. (2018). Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine. *J. Inf. Syst. Eng. Bus. Intell.*, vol. 4, no. 1, p. 57

[6] Muljono, D. P. Artanti, A. Syukur, A. Prihandono, D. R. I. M. Setiadi. (2018). Analisis Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes. in *Konferensi Nasional Sistem Informasi 2018*, pp. 8–9

[7] Hu M, Liu B. (2004). Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA. pp 168–177

[8] Pang B, Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL ’04. Association for Computational Linguistics, Stroudsburg, PA, USA

[9] Zhang Y, Xiang X, Yin C, Shang L. (2013). Parallel sentiment polarity classification method with substring feature reduction. In: Trends and Applications in Knowledge Discovery and Data Mining, volume 7867 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Heidelberg, Germany. pp 121–132

[10] Pang B, Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL ’04. Association for Computational Linguistics, Stroudsburg, PA, USA

[11] Marcus M. (accessed Apr. 28, 2022). Upenn part of speech tagger. http://www.cis.upenn.edu/~treebank/home.html

[12] Gan W-JK, Day J, Zhou S. (2014). Twitter analytics for insider trading fraud detection system. In: Proceedings of the sencond ASE international conference on Big Data. ASE

[13] Hermanto, Kuntoro, A. Y., Asra, T., Pratama, E. B., Effendi, L., & Ocanitra, R. (2020). Gojek and Grab User Sentiment Analysis on Google Play Using Naive Bayes Algorithm and Support Vector Machine Based Smote Technique. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012102