
Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier Algorithm

Taqwa Hariguna ^{a,*}, Wiga Maulana Baihaqi ^a, Aulia Nurwanti^a

^aInformation System Study Program, Amikom Purwokerto University, Indonesia

* corresponding author

Abstract

In an e-commerce Shopee, the process of selling and buying continues to run every day, and the comments given by consumers will increase more and more. Comments given by consumers will be the reference/review of a product that has been purchased by consumers. Consumers freely provide a review containing positive comments and negative comments in the Comments field listed on the Shopee e-commerce website. With the above problems, researchers will do a research with the method of sentiment analysis to distinguish classes in product review comments that include positive comment class or negative comment class using a combination of K-means and naive Bayes classifier. K-means used to determine the grouping of classes; naive Bayes classifier used to get the value of accuracy. The results obtained based on clustering K-means include getting 116 negative comments on product reviews and 37 negative comments product reviews. Accuracy results obtained from product review comment data of 77.12%. Thus, the accuracy value using K-means and naive Bayes classifier without manual data get a higher accuracy value is compared using K-means, Naive Bayes classifier, and manual data get results lower accuracy of 56.86%. From the results above the most comments is a negative comment of 116 data review comments product, from the results of the study can be concluded that one of the products of Spatuaafa named high heels women know the Ribbon Ikat FX18 the condition of the product is not good enough due to the high negative comments compared to positive comments.

Keywords: naive bayes classifier; K-means; sentiment analysis

1. Introduction

In the development of the era, the needs of society are increasing with the ease of purchasing goods through the Internet. The purchase of goods can now be made from home, office or even a free natural place without having to find the seller. It's easy for customers to do online shopping process, customers not to pay attention to product reviews provided by other customers after purchasing products on shopee e-commerce website. Customers only pay attention to product ratings, prices of products that are too cheap, and the amount of discount that the store owner has provided on the e-commerce website without looking at the product reviews contained in the Comments field available on the e-commerce website[1][2]. Whereas if the customer is careful in looking at product reviews, customers will know the review of the services that the online store owner provided, knowing the review speed of shipping products, and the most important thing is to review products to be purchased customers[3].

The Review of items contained in each item in the online store is useful to see how the feedback from those who have previously purchased the goods. Feedback provided by users is usually divided into two kinds of comments, namely positive comments or negative comments[4]. Users can express their opinions about quality, price, service of a website, online seller, and speed of delivery. Online shopping users often use comments from other users when they will purchase goods. Therefore, the grouping of reviews of goods from consumers is influenced by the emotions (sentiments) grouped or classified to determine their polarizing positive or negative [5]. The definition of this review can be both positive and negative, according to the person who did the review. However, most of a review is based on the information obtained earlier from the product so that it can be made in the right way or bring negative things to the product.

Opinion Mining is one of the areas of computer science that study linguistic computing, natural language processing, and text mining aimed at analyzing emotions, judgments, attitudes, opinions, sentiments, evaluation of a person on a speaker or writer pleased with a product, service, organization, individual, public figure, topic, event, or particular activity [6][7].

In the application of data mining, There is a text mining process to conduct decision-making analysis. To analyze customer comments on positive and negative sentiments, several methods can be used in the analysis process. In sentiment analysis, the method used is a classification method using the Naïve Bayes algorithm. Data Mining is the mining or discovery of new information by looking for specific patterns or rules from a massive amount of data [8]. Text Mining is not a function, but a collection of various functions combined and referred to as Text mining functions. The main functions of text mining include Searching, information Extraction, categorization, summarization, prioritization, Clustering, information Monitor and Question & Answer [9].

The method used is the classification method using the Naïve Bayes algorithm. The process of this research uses the combination of K-means and the algorithm of naive Bayes classifier. K-means is used to distinguish the class review comments of a product into a positive comment class or a negative comment class. While the naive Bayes classifier algorithm is used to derive the accuracy value of positive comment data and negative product reviews. The Naive Bayes classifier method is considered a potentially useful method for classifying data from other classification methods in terms of accuracy and computation [10][11].

2. Method

This research method is used through 5 (five) data processing processes, namely data collection, pre-processing data, clustering K-means, data Classification, and data evaluation, and subsequent withdrawal conclusions, as in Figure 1.

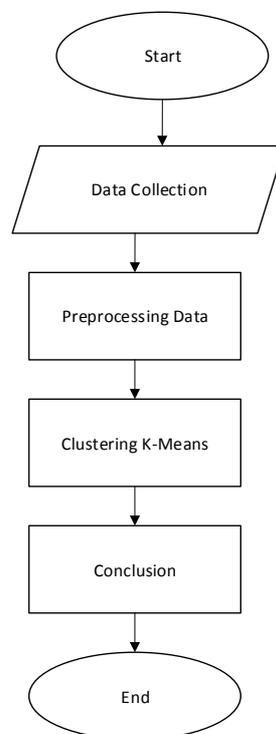


Fig. 1. Research Flowchart

Figure 1 shows a scheme of research concepts, and the following will be explained for each of them:

1. In this study, secondary data was taken based on the product review on a shopee e-commerce website. The way the research data is done by copying product comments based on the product review. The steps taken to retrieve the research data are, go to the Shopee e-commerce website, then find the product to be analyzed in the search field, look for a product that contains a lot of positive and negative comments, then look at the comments section, copy customer comments one by one then paste the comments into Notepad++, if you have obtained the data research, save the research data into .txt format.
2. In this process, the data obtained will be performed preprocessing stage to select usable data and remove unnecessary data for stages of analyzing data. There are several stages in preprocessing techniques, such as case folding, tokenization, filtering, and stemming.
 - a. Case Folding is the stage for converting all the letters in a document to lowercase (lowers case).

- b. Tokenizing is the cutting phase of the input string based on each word that follows it. Tokenizing is also used to remove some characters that are considered punctuation marks.
- c. Filtering is the stage of taking the crucial words of tokenizing results using recognized removal (removing the less important words). Stop word is an un-descriptive word that can be simplified.
- d. Stemming is a process of normalization in system Information retrieval used in finding root words in a document or term regarding specific rules [12][13].

3. Clustering of K-Means algorithm

In the process of clustering, the K-means is a process to categorize product review comments, including positive classes or negative classes. In this process, the review of the reviewed product is processed using the help of a specific value or membership degree by using the K-means algorithm to determine the positive and negative comments on the product review. The initial step of the K-Means algorithm is to determine the center of each cluster, which is almost a type, then called a centroid. Centroid can be determined randomly. Then calculate the distance between each cluster against the existing centroid, then group each cluster based on the closest distance from each Centroid object. Then recalculate the centroid, do this repeatedly until the centroid position does not move again[14].

4. Data Classification and Evaluation

In the classification process, using the Naïve Bayes classifier algorithm. Naïve Bayes Classifier makes a dominant assumption (naïf) to the independence of each class. Naïve Bayes Classifier is used to obtain the accuracy results of the comment data taken from the Shopee e-commerce website and is used to classify the positive and negative comments of the product review from the comments data that has been obtained.

The evaluation process conducted based on the category of positive comments and negative product reviews using the help of the confusion matrix to determine the accuracy value, the precision value, and the F-measure value, the recall value[15].

5. Withdrawal of conclusion

The next stage is the withdrawal of the final result conclusion. In this stage, the author undertook the results of the research. Withdrawing conclusions based on the outcome of the positive class and the negative class of the product review comment data obtained in the K-means process, as well as the final calculation using the confusion matrix, include accuracy value, precision value, F-measure value, and recall rate obtained using naïve Bayes classifier algorithm.

3. Results and Discussion

4.1. Data Collection

In this study, the data to be used is secondary data, which is the data obtained in this study based on positive comments and negative reviews of Indonesian-language products on the e-commerce website. The dataset used in the study amounted to 153 comment data.

4.2. Pre-processing Stage

- a. Case Normalization is the phase of converting sentences to lowercase. At this stage, the dataset will be checked per sentence to change the sentence to lowercase.

Table 1. Example of the Case normalization process (lowercase)

No.	Data prior to Case Normalization Process	Data after Case normalization process
1.	barang yang datang beda sekali sama yang di gambar.. customer amat sangat kecewa..	barang yang datang beda sekali sama yang di gambar.. customer amat sangat kecewa..

- b. Tokenization is a phase of solving text that can be a sentence, a paragraph, or a document, to be a token or a specific part based on each word that follows it.

Tabel 2. Examples of tokenization processes

No.	Data prior to tokenization process	Data after tokenization process
1.	barang yang datang beda sekali sama yang di gambar.. customer amat sangat kecewa..	“barang”, “yang”, “datang”, “beda”, “sekali”, “sama”, “yang”, “di”, “gambar”, “customer”, “amat”, “amat”, “sangat”, “kecewa”

- c. Filtering (Stopword) is the process of eliminating frequently appearing words but does not affect sentiment analysis. Examples of recognized are "dan", "di", "dari", "yang", "ke" and so on.

Tabel 3. Example of Filtering Process

No.	Data before the Stopword process	Data after the Stopword process
1.	barang yang datang beda sekali sama yang di gambar.. customer amat sangat kecewa..	barang datang beda sekali sama gambar..customer kecewa..

- d. Stemming is the process of eliminating all the words that exist in the word to be a basic word. For example, namely meng-, me-, kan-, di-, pe-, peng-, etc. Stemming aims to reduce variations of words that have the same base word.

Tabel 4. Example of Stemming Process

No.	Stemming Data before process	Data after Stemming process
1.	Goods that come different from the same thing in the picture. Very disappointed.	The goods come different once the picture. Customer disappointed.

4.3. Clustering of K-Means algorithm

In the next stage of the product review, comment data in Microsoft Excel is inserted in the rapid miner process. Process Clustering uses 2 clusters consisting of cluster 0 and cluster 1 that explains that cluster 0 is a negative comment and cluster 1 is a positive comment. The results obtained in the clustering process, among others, are cluster 0 amounting to 116 data comment product reviews, and cluster 1 amounted to 37 product review Comment data.

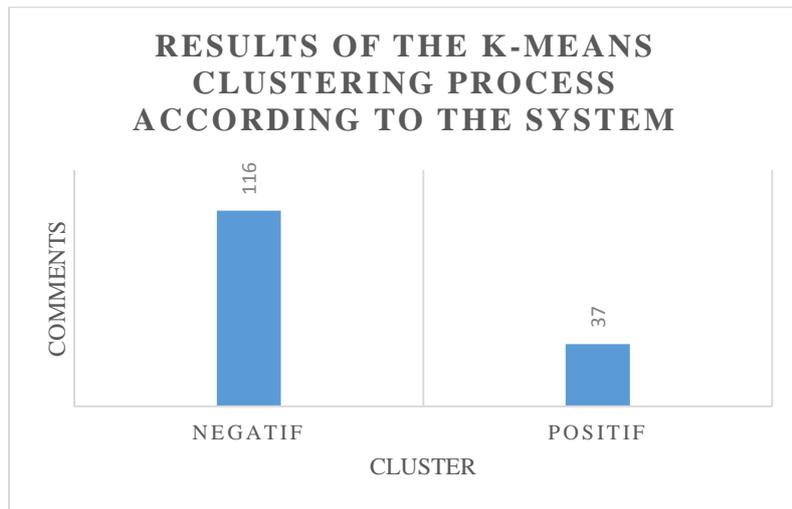


Fig. 2. Results Process Clustering K-means by System

The Data that has been obtained in clustering K-means is reprocessed manually to improve the product review, comment class. The results obtained in this process include cluster 0 amounting to 89 comment data, and cluster 1 amounted to 64 comment data.

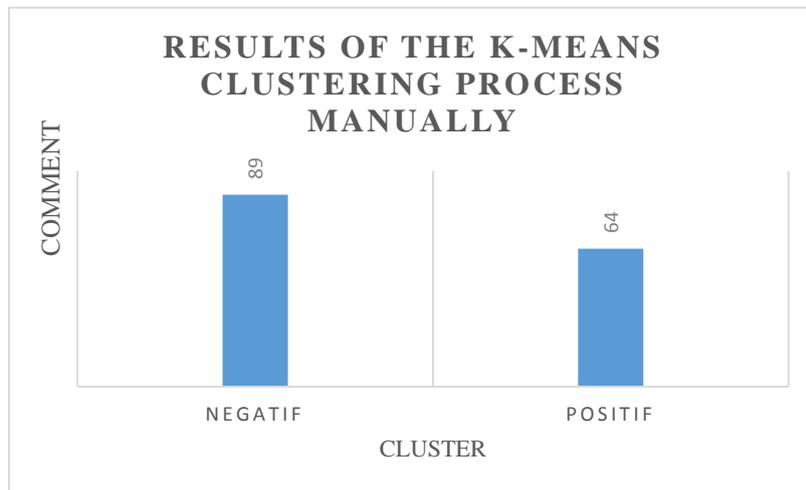


Fig. 3. Manual Process of Clustering K-means

The results of the clustering K-means process, indicating cluster 0 has the most product review comments compared to cluster 1 so that many product review sentiments contain negative comments.

The total word used from 153 comments Product Reviews with a total of 596 words consisting of positive and negative comments. The most important word in the clustering process is the most common word that appears. To determine the most common words the researcher uses in cloud Word as shown in Figures 4 and 5.



Fig. 4 Cloud Word Comments



Fig. 5 Cloud Word comments

Cloud word successfully performed and can be analyzed words such as "baik", "produk", "kualitas", is a high-frequency word in providing positive comments product reviews, and the word "baik", "saya", being a high-frequency word in commenting negative comments review.

4.4. Data Classification and Evaluation

Table 5. Evaluation Model

Model	Precision	Recall	F-Measure	Accuracy
K-means + NBC	76.82%	100%	86.89%	77.12%
K-means + manual + NBC	64.94%	56.18%	71.94%	56.86%

The results of table 5 were shown that after experiments were conducted there were differences between the K-means + NBC model compared to the K-means + manual + NBC model. Where the K-means + Manual + NBC model decreased accuracy value.

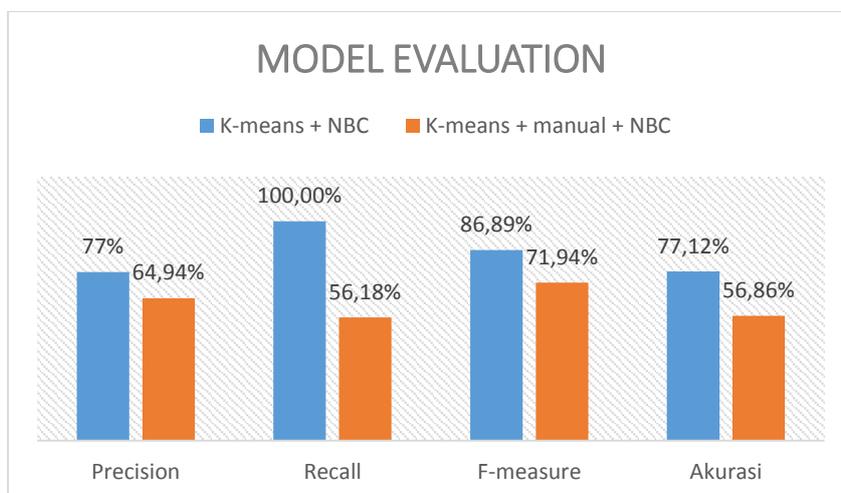


Fig. 6. Model Evaluation Chart

Figure 6. Shows the K-means + NBC model received the highest value of the precision value of 77%, recall value of 100%, F-measure 86.89%, and an accuracy value of 77.12%.

4. Conclusions

Based on the results of the research that has been done then it can be concluded that the combination of K-means and naive Bayes classifier (NBC) has successfully analyzed positive comments and negative comments product reviews, but the process of grouping using K-means is not yet fully maximized. Considered not maximum because at the time of class grouping, there is still an error in entering sentences into positive classes and negative classes. For that, it is necessary to have manual assistance to correct the grouping of positive classes and negative classes. The accuracy results obtained from the combination of K-means and naive Bayes classifier without manual amounted to 77.12% while the process is done manually get the accuracy result of 56.86%.

It can be concluded that the results of accuracy using a combination of K-means and naive Bayes classifier without manual get higher accuracy value compared to using a manual process, the highest comment result is on negative comments using K-means and naive Bayes classifier without a manual of 116 comment data product review. With the results of such research can be used as a consideration in the process of purchasing a product in the website of e-commerce Shopee, because the highest comment is a negative comment, it can be concluded that one of the products of Spatuafa named high heels women know the Ribbon Ikat FX18 the condition of the product is not good enough. For customers who want to purchase products to not race on the product rating only, but see a review of customers who have purchased products at Spatuafa. Then from the store Spatuafa to be used as a reference to improve the quality of products that will be marketed in the Shopee e-commerce website. For products that customers buy can satisfy customers.

References

- [1] Indriati, A. Ridok. "Sentiment Analysis For Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (NWKNN)". *Journal of Environmental Engineering & Sustainable Technology*. Malang. 2016.
- [2] B, Liu. 2012. "Sentiment Analysis and Opinion Mining". *Synthesis Lectures On Human Language Technologies*, 5 (1), 1-167.
- [3] B, Davies, "Database Systems 3rd Edition", Basingstoke, UK: Palgrave, 2004.
- [4] A. Mustafa, A. Akbar, A. Sultan. "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization". *International Journal of Multimedia and Ubiquitous Engineering*, 4 (2), 183-188. 2009.
- [5] L. S. Ting, A. H. Tsang. "Is Naive Bayes a Good Classifier for Document Classification?". *International Journal of Software Engineering and Its Applications*, 5 (3), 37-46. 2011.
- [6] M. Joshi, H. Vala. "Opinion Mining For Sentiment Data Classification". *International Journal of Research in Information Technology*, 3 (1), 1-13. 2014.
- [7] L. Agusta. "Comparison algorithm Stemming Porter with Algoritma Nazief & Adriani for Stemming Bahasa Indonesia text document". *National Conference of Systems and Informatics*. Kns & 109-036. 2009.
- [8] Balazs, J. A., Velasquez, J. D. (2016), "Opinion mining and information fusion: a survey", *Information Fusion*, 27, 95-110.
- [9] Boynukalin, Z., and Karagoz, P., (2013), "Emotion analysis on Turkish texts", *ISCIS 2013, Lecture Notes in Electrical Engineering*, Volume 264, pp. 159-168.
- [10] Catal, C., Sevim, U. and Diri, B., (2011), "Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm", *Expert Syst. Appl.*, 38, 3 (March 2011), 2347-2353.
- [11] Celik, H., (2013), "Sentiment analysis for Turkish language", MSc Thesis, Istanbul Kultur University.
- [12] Cetin, M., and Amasyali, F., (2013), "Active learning for Turkish sentiment analysis", *INISTA 2013*, Albana, Bulgaria, pp. 1-4.
- [13] Chalothorn, T., Ellman, J., (2015), "Simple approaches of sentiment analysis via ensemble learning", *Information Science and Applications, Lecture Notes in Electrical Engineering*, Volume 339, 631-639.
- [14] Demirtas, E., and Pechenizkiy, M., (2013), "Cross-lingual polarity detection with machine translation", *Int'l Workshop on Issues of Sentiment Discovery and Opinion Mining*, ACM, NY, 8 pages.

- [15] Erogul, U., (2009), "Sentiment analysis in Turkish", MSc Thesis, Middle East Technical University (METU), Department of Computer Engineering.