Jurnal Elektronik Ilmu Komputer Udayana
Volume 8, No 3. February 2020

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Sentiment Analysis of Snack Review Using the Naïve Bayes Method

I Gede Cahya Purnama Yasa[a1] Ngurah Agus Sanjaya ER[a2], Luh Arida Ayu Rahning Putri[a3]

[a]Informatics Department, Udayana University
Bali, Indonesia
[1]cahya.purnama456@gmail.com
[2]agus_sanjaya@unud.ac.id
[3]luh.arida@cs.unud.ac.id

Abstract

Fast food is a product that we often encounter in stores such as convenience stores. Ready-to-eat products can now be easily found by consumers. One of the reasons is due to the expansion of minimarkets in areas that are easily reached, such as housing complexes, school areas, and offices. Sentiment analysis is used to determine whether an opinion or comment on a product has a positive or negative interest and can be used as a reference in improving service, or improving product quality. In this research, we study the sentiments of consumers towards snack food products as a reference to improve the level of service and quality of these products. We classify the sentiment of a review on snack food products as positive and negative. To classify the sentiments we apply the Naïve Bayes and Multinomial Naïve Bayes methods. We compare the two methods to study the most effective and efficient method for classifying sentiments on reviews of snack food products.

Keywords: *Sentiment Analysis, TF-IDF, Naïve Bayes,Multinomial, Review, Snack, Preprocessing*

1. Introduction

There are various types of food and drinks on the Indonesian market, food and beverages are often found in minimarkets, for example, which provides various types of food and beverages in the form of fast food products. Through the Worldpanel Indonesia 2017 Study in urban communities shows that ready-to-eat products are three times more salable outside the home than consumed at home. Ready-to-eat products are becoming increasingly easy to find by consumers, one of them thanks to the expansion of minimarkets in easily accessible areas such as housing complexes, school areas, and offices [1]. Snack is one of the foods that are favorite and most in demand. Because the majority of snacks are dry and have a small size so it is very easy to take and put into the mouth. In addition, a quick snack makes people thirsty. After drinking, the person will consume again [2].

Not all fast food has good taste and is satisfying for consumers. Nowadays, internet users often use forums as a consideration to buy an item or food and drink. Generally, to provide a product review, users are asked to fill in a review sentence and rating results that are usually in the form of stars. Users are asked to write down their experiences using the product and assess whether the experience deserves a rating. However, users also often give results of reviews and ratings which are actually inversely proportional. For example, often the user writes down how satisfied he is with the product but instead gives a low rating, which means the review is negative.

Sentiment analysis studies one's perspective, behavior and feelings or emotions towards an individual, problem, activity, subject [3]. Sentiment analysis is done to determine whether an opinion or comment on a problem, has a positive or negative tendency and can be used as a reference in improving a service, or improving product quality

In previous studies relating to Sentiment Analysis conducted by [4] on opinions written in Roman-Urdu and English extracted from the blog. In this study a classification method or model which includes the Naïve Bayes method, decision tree, and K-Nearest Neighbor (KNN) is used.

The dataset used for the training process consisted of 150 positives and 150 negatives. Based on test results using precision, recall, and f-measure, the Naïve Bayes classification method produces better performance than the decision tree and KNN.

For the classification method itself many researchers use Naive Bayes where a text will be classified in machine learning based on probability [5]. Naïve Bayes classifiers are very simple and efficient [6]. In addition to its simplicity, the Naïve Bayes classifier is a popular machine learning technique for text classification, and has good performance in many domains [7].

Sentiment analysis helps a seller to evaluate the opinions and behavior of clients towards their products, so that the seller gets a review of their goods directly from the client. Therefore, sentiment analysis is needed to analyze a person's view of a product so that it can increase the usability and sale of the product by knowing the weaknesses of goods from the user's point of view. In this study, the authors conducted a study of snack or snack reviews to find out whether the sentiments were positive or negative. The final result that will be produced is the level of accuracy achieved in conducting sentiment analysis using the Naïve Bayes method.

2. Reseach Methods

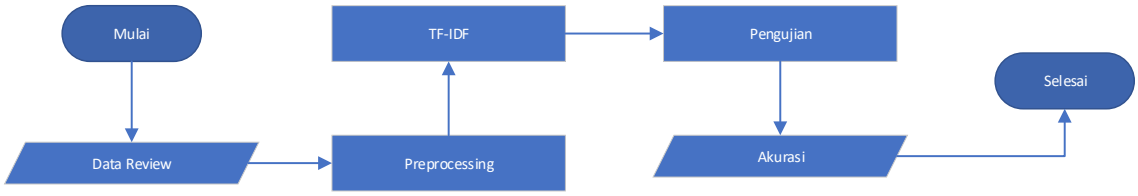

Figure 1. General System Flow

In the research method that the author uses regarding the stages that must be passed as shown in figure 1. The earliest stage that is passed is reading documents reviewing documents as a data system, after which a preprocessing process will be carried out, such as handling punctuation, deletion stopwords, stemming, tokenization and handling negative words. Then a feature that uses word weighting will use a Term Frequency - Inverse Document Frequency (TF-IDF) that will produce vector features. After completing the preprocessing and weighting stages, there will be a training and testing phase of the system.

2.1. Dataset

In this study the authors used snack food product review data obtained from https://www.hometesterclub.com which consisted of 50 positive reviews and 50 negative reviews.

2.2. Preprocessing

Preprocessing stage that the author uses for the dataset that is owned are:

a. Tokenization
   From each review sentence contained in the dataset will go through the tokenization process. In the process of tokenisation each sentence is broken down into words.
b. Stopwords Removal
   Namely the elimination of words that are not relevant, like this, is, that, only and so on.
c. Punctuation Management
   In this method, the punctuation review will be removed.
d. Stemming
   Stemming is a method used to convert words into root words by eliminating affixes and suffixes on words, such as using, using, and using where the basic words of all are use words.
e. Handling Negative Words

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Changing the words in front of it there are negative words like, 'no', 'ga', 'not', and others will be given treatment, i.e. the addition of 'no_' before the word. For example, the word 'does not match' then the words become 'not not suitable'. This treatment serves to distinguish the meaning of the word 'match' if it is preceded by a negative word and which is not preceded by a negative word.

Table 1. Preprocessing Results

Data	Rasanya hampir tidak ada lebih ke rasa kentang pada umumnya!
Tokenization	'rasanya','hampir',' tidak','ada','lebih','ke','rasa','kentang','pada','umumnya!'
Stopwords Removal	'rasanya',' tidak','rasa','kentang','umumnya!'
Penanganan Puctiation	'rasanya','tidak','rasa','kentang','umumnya'
Stemming	'rasa', 'tidak','rasa','kentang','umum'
Penanganan Negative Words	'rasa', 'tidak','tidak_rasa','kentang','umum'

2.3. Word Weighting

Word weighting (term) aims to give weight to each word (term) contained in the text document to be processed. The stages in word weighting are as follows:

a. Term Frequency (TF)
Term Frequency is the frequency of occurrence of words in a text document. Term Frequency (tft, d) is defined by the number of occurrences of the term t in the document d.

$$tf(t,d) = \frac{f(t,d)}{n} \quad (1)$$

b. Invers Document Frequency (IDF)
Inverse Document Frequency is the frequency with which the term appears in all text documents. Term that rarely appears in the whole text document has a value of Inverse Document Frequency greater than the term that often appears.

$$idf(t) = log\frac{n}{1+df(t)} \quad (2)$$

c. Term Frequency- Inverse Document Frequency
The tf-idf value of a word is a combination of the tf value and idf value in the weight calculation.

$$tf - idf(t,d) = tf(t,d) \times idf(t) \quad (3)$$

2.4. Naïve Bayes

Naïve Bayes is one of the probability statistical methods based on the application of the Bayes theorem with strong (naive) independent assumptions, to predict the class of a document based on its probabilities [8]. Naïve Bayes classifier assumes that the presence of certain features of a class is not related to the presence of other features (independent). Naïve Bayes is a very simple approach to classification text.

In the Bayes theorem, P (c | d) is determined where c is the class and d is the object (document) to be classified. P (c) is the prior probability of class c, P (d | c) is the probability of document (d) in a class (c). Bayes' theorem has the following formula

$$\mathrm{P(c|d)} = \frac{\mathrm{P(d|c)\ P(c)}}{\mathrm{P(d)}} \quad (4)$$

a. Multinomial Naïve Bayes
Multinomial Naïve Bayes is one of the specific methods of the Naïve Bayes method that uses conditional probability. Conditional probability can be done by using the frequency of occurrence of a word in a class (raw term frequency) [8]. Multinomial Naïve Bayes calculates the frequency of each word that appears in the document. For example there are documents d and class c. To calculate the class of document d, it can be calculated using the formula:

$$P(c|term\ dokumen\ d) = P(c)x\ P(t1|c)x\ P(t2|c)x\ P(t3|c)x \ldots x\ P(tn|c) \quad (5)$$

The probability of prior class c is determined by the formula:

$$\mathrm{P(c)} = \frac{\mathrm{Nc}}{\mathrm{N}} \quad (6)$$

Multinomial Naïve Bayes for the value of input x are shown in the following equation:

$$P(tn|c) = \frac{W(c,tn)+1}{(\sum \mathrm{W\prime \text{€} VW\prime(c,tn)} + \mathrm{B`})} \quad (7)$$

Where :
$W(c, tn)$: Nilai pembobotan tfidf atau W dari term t di kategori c
$\sum W' \text{€}\ VW'ct$: Jumlah total W dari keseluruhan term yang berada di kategori c.
B' : Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen

b. Experiment Scenarios
Calculation of the accuracy of the system to measure how well the system performance that we made is measured using the formula:

$$Akurasi = \frac{Db}{N} x100\% \quad (8)$$

From that formula we obtained from the percentage of the results of a properly classified review divided by the total of all review tested.

3. Result and Discussion

The proposed method uses 100 review data where 50 are positive reviews and 50 are negative reviews. For each dataset 20% is used as test data and the rest is used as training data.

Previously the data had to go through preprocessing as follows: (1) First, we separated each sentence into words, (2) Then eliminated irrelevant words, (3) After that it eliminated punctuation, (4) basics, and (5) lastly set the word negation. After going through the preprocessing stage, the results will be weighted using the tf-tdf formula as in equation (3). After the weights are obtained, the classification is done using the Naive Bayes and Multinomial Naive Bayes methods.

The results of the system evaluation collected with the calculations obtained are the amount of data classified correctly divided by all test data.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

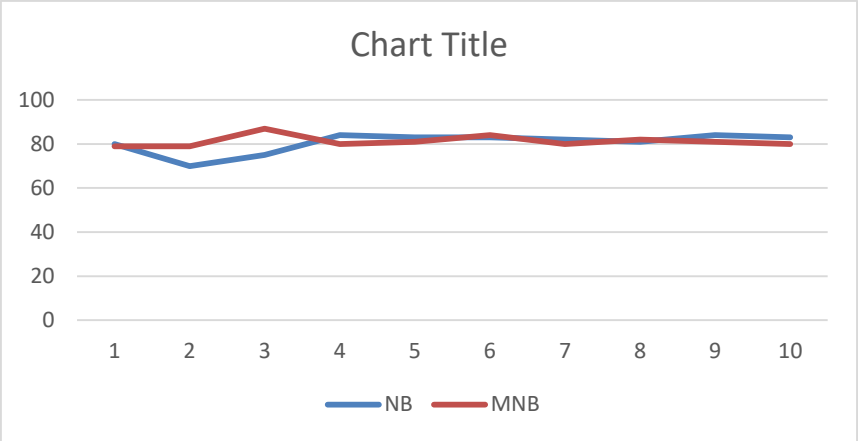

Figure 2. Accuracy

From Figure 2 we conducted a number of experiments that obtained varying accuracy, this is because the distribution of training data and testing data was randomized by a percentage as stated. From these results it was found that the highest accuracy obtained when using the Naïve Bayes method was 84% and when using the Multinomial Naive Bayes method was 87% with an average accuracy obtained from 10 trials was 80,5% using the Naïve Bayes method and 81,3% using the Multinomial Naïve Bayes method.

4. Conclusion

Naïve Bayes and Multinomial Naïve Bayes are some methods that used for classification. Where in this system the Naïve Bayes and Multinomial Naïve Bayes method is used to classify sentiments from the review of snacks, whether the review is included in positive sentiments or included in negative sentiments. Where the feature extraction method used is the weighting of words using the tf-idf method that provides statistics on the appearance of words and the level of importance of documents that support it.

The use of both methods to classify the system can already be implemented by applying feature extraction with TF-IDF, this is proven and also with the test results produced by the system which shows quite good results. This can be seen from the highest results obtained with Naïve Bayes is 84% and Multinomial Naïve Bayes is 87%, with an average accuracy 80,5% and 81,3%.

From the research conducted, both methods can classified review into a positive or negative sentiment quite well but it still cannot be grouped directly based on the product to determine the overall assessment of a product. In the future it might be possible to make improvements by directly evaluating based on the product.

References

[1] N. A. SAM, "Kompas," Ekonomi Kompas, 28 12 2017. [Online]. Available: https://ekonomi.kompas.com/read/2017/12/28/135401226/studi-kantar-makanan-dan-minuman-kemasan-lebih-laku-di-perkotaan. [Accessed 24 05 2019].

[2] T. Sipahutar, "Kompas," Lifestyle Kompas, 26 05 2011. [Online]. Available: https://lifestyle.kompas.com/read/2011/05/26/08592040/kerupuk.atau.keripik.favorit.orang.indonesia. [Accessed 24 05 2019].

[3] A. Basari, B. Hussin, . I. Ananta and J. Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization," *Procedia Engineering,* vol. 53, p. 453 – 462, 2013.

[4] M. Bilal, . H. Israr, . M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions Using Naı¨ve Bayesian, Decision Tree and KNN classification techniques.," *ing Saud University.,* 2015.

[5] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification," *Advanced in Control Engineering and Information Science,* vol. 15, p. 2160–2164, 2011.

[6] J. Chen, H. Huang, . S. Tian and . Y. Qu, "Feature selection for text classification with Naïve Bayes.," *Expert Systems with Applications,* vol. 36(3), p. 5432–5435, 2009.

[7] Q. Ye, . Z. Zhang and R. Law, "xpert Systems with Applications Sentiment classification of online reviews to travel destinations by supervised machine learning approaches.," *Expert Systems With Applications,* vol. 36(3), p. 6527–6535, 2009.

[8] A. Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford,* vol. 12, p. 1, 2009.

[9] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory," *CoRR,* vol. 14105329, pp. 1-20, 2014.