Journal of Soft Computing Exploration

Optimize naïve bayes classifier using chi square and term frequency inverse document frequency for amazon review sentiment analysis

Anisa Falasari[1], Much Aziz Muslim[2]
[1,2]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[2]Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Feb 10, 2022
Revised Feb 20, 2022
Accepted Mar 25, 2022

Keywords:

Sentiment Analyst,
Naïve Bayes Classifier,
Chi Square,
TF-IDF

ABSTRACT

The rapid development of the internet has made information flow rapidly which has an impact on the world of commerce. Some people who have bought a product will write their opinion on social media or other online site. Long-text buyer reviews need a machine to recognize opinions. Sentiment analysis applies the text mining method. One of the methods applied in sentiment analysis is classification. One of the classification algorithms is the naïve bayes classifier. Naïve bayes classifier is a classification method with good efficiency and performance. However, it is very sensitive with too many features, which makes the accuracy low. To improve the accuracy of the naïve bayes classifier algorithm it can be done by selecting features. One of the feature selections is chi square. The selection of features with chi square calculation based on the top-K value that has been determined, namely 450. In addition, weighting features can also improve the accuracy of the naïve bayes classifier algorithm. One of the feature weighting techniques is term frequency inverse document frequency (TF-IDF). In this study, using sentiment labelled dataset (field amazon_labelled) obtained from UCI Machine Learning. This dataset has 500 positive reviews and 500 negative reviews. The accuracy of the naïve bayes classifier in the amazon review sentiment analysis was 82%. Meanwhile, the accuracy of the naïve bayes classifier by applying chi square and TF-IDF is 83%.

Corresponding Author:

Anisa Falasari,
Department of Computer Science,
Universitas Negeri Semarang.
Email: afalasari@students.unnes.ac.id

1. INTRODUCTION

The development of internet is increasingly rapid, making information flow faster wich has an impact on the world of commerce. Some people who have bought a product will write their opinion through social media or other online sites. One of the online sites for buying and selling is Amazon.com.

An opinion engine is needed to process large text data. Sentiment analysis is process that applies the text mining method. Text mining is the discovery of new information or trends that were not previously revealed by processing and analyzing large amounts of data [1]. According to them, text mining can provide solutions such as processing, organizing, or grouping and analyzing large amounts of unstructured text. One mining method is classification. There are several classification methods including the Bayessian classification.

Naïve bayes classifier or Bayessian classification is a statistical classification method based on the Bayes theorem wich can be used to predict the probability of membership of a class [2], [3]. This method is a

classification method with good computational efficiency and predictive performance [4]. However, it is very sensitive to too many features that is makes the accuracy low [5]. One of the problems in text classification is too many features or attributes [6]. Feature selection is used to reduce a number of irrelevant attributes [7], one of wich is the chi square.

In addition to feature selection, feature weighting can affect the accuracy of a method. One of the feature weighting methods is Terms Frequency Inverse Document Frequency (TTF-IDF). TF-IDF is a method of text mining that is fast and efficient [8].

The purpose of this study is to improve accuracy wich focuses on the sentiment analysis of amazon reviews using the naïve bayes classifier method with chi square and TF-IDF.

2. METHOD

The steps taken in this research are pre-processing, feature weighting, feature selection, and text classification. this research begins by entering the Amazon review dataset, then continues with pre-processing including transform case, tokenize, stopword removal, stemming, and punctuation removal. Then the feature weighting is carried out using TF-IDF. After obtaining the result from the feature weighting, feature selection is carried out using chi square. Based on the selected features, the classification is carried out using the naïve bayes classification method. Then the classification is tested using test data and evaluated using a configuration matrix to obtain an accuracy value. The research method flowchart can be seen in Figure 1.

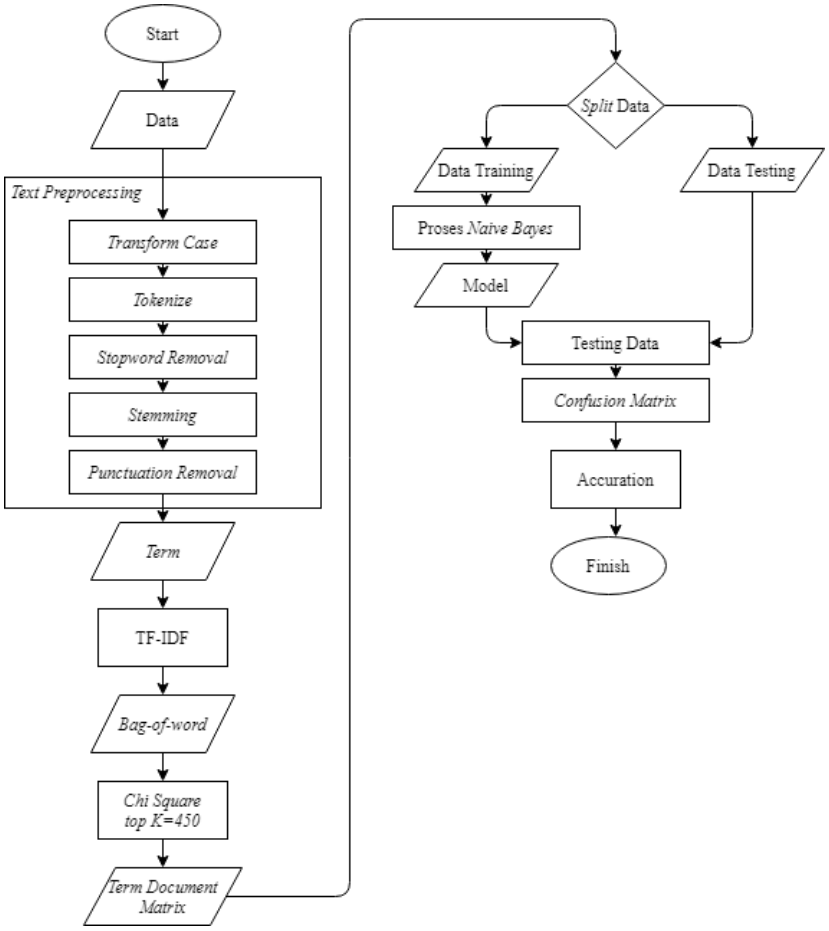

Figure 1. Flowchart of the naïve bayes classification method using chi square and TF-IDF

2.1. Dataset

This research uses a dataset obtained from the UCI machine learning repository, namely sentiment labelled dataset (fieldamazon_labelled) in English. This dataset has 1000 reviews consisting of 500 positive

reviews and 500 negative reviews. This dataset with a .txt extension is then converted into a table with the .xlxs format which has two column. Column label consists of "0" which means negative and "1" which means positive. The dataset with the .xlxs format can be seen in Table 1.

Table 1. The amazon_labelled dataset in .xlxs format

Review	Label
So, there is no way for me to plug it in here in the US unless I go by a converter.	0
Good case, Excellent value.	1

2.2. Text Pre-Processing

Text pre-processing is the stage for preparing textual data that will be used at a lates stage. The text pre-processing stage in this research is transform case, tokenize, stopword removal, stemming and punctuation removal.

2.2.1. Transform case

This stage aims to change all characters in the data to lowercase. The result of the case transformation process can be seen in Table 2.

Table 2. Result of the transform case

Review	*Transform case*
So, there is no way for me to plug it in here in the US unless I go by a converter.	so, there is no way for me to plug it in here in the us unless i go by a converter.

2.2.2. Tokenize

Tokenize is the stage for converting a document into a token/term. Based on the result of the previous process, the result of the tokenize process can be seen in Table 3.

Table 3. Tokenize result

Review	Token
so there is no way for me to plug it in here in the us unless i go by a converter.	"so" "there" "is" "no" "way" "for" "me" "to" "plug" "it" "in" "here" "in" "the" "us" "unless" "I" "go" "by" "a" "converter" "."

2.2.3. Stopword Removal

Stopword removal works by filtering all the tokens in the document then removing the tokens contained in the stopword list. This study uses an English stopword list as a stopword filter. Based on the result of the previous process, the result of the stopword removal process can be seen in Table 4.

Table 4. Stopword removal result

Review	*Stopword Removal*
so, there is no way for me to plug it in here in the us unless i go by a converter.	"way" "plug" "us" "unless" "go" "converter" "."

2.2.4. Stemming

Stemming is the process of converting words into their basic form. If there is more tha one identified word with the same name, it will be defined as one word. Based on the result of the previous process, the results of the stemming process can be seen in Table 5.

Table 5. Stemming result

Review	*Stemming*
so, there is no way for me to plug it in here in the us unless i go by a converter.	"way" "plug" "us" "unless" "go" "convert" "."

2.2.5. Punctuation Removal

Punctuation removal is a process to remove punctuation in a document. Based on the result of the previous process, the result of the punctuation removal process can be seen in Table 6.

Table 6. Result of punctuation removal

Review	*Stemming*
so, there is no way for me to plug it in here in the us unless i go by a converter.	“way” “plug” “us” “unless” “go” “convert”

2.3. Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF is one of the most well-known algorithms in text mining. Term frequency (TF) is used to measure the number of times a term appears in a document [9]. Meanwhile, the inverse document frequency (IDF) is the inverse probability log calculation of a term that often appears in a document [10]. So, TF-IDF gives inverse weight to a term based on the frequency in the document [11]. TF value is defined as $TF = t_{ij}$, wich is the term i in j. DF value is the number of documents containing the term i, used to calculate the IDF value. IDF value can be calculated by equation 1.

$$IDF = \log\left(\frac{N}{DF}\right) \quad (1)$$

Where the value of N is the number of all documents and DF is the number of documents that contain a term. So that TF-IDF can be calculated by multiplying TF by IDF as in equation 2.

$$TF - IDF = TF.\log\left(\frac{N}{DF}\right) \quad (2)$$

2.4. Chi Square Feature Selection

The selection of the chi square feature is used to test the independence of two events, namely term (t_i) and class emergence (c_k) [12]. The formula used in the calculation of chi square is shown in equation 3.

$$X^2(t_i, C_k) = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (3)$$

The variables related to the chi square calculation can be seen in Table 7.

Table 7. Related variables in the chi square calculation

	Include in C_k Category	Exclude in C_k Category
Number of documents containing the *term* t_i	A	b
Number of documents that do not contain the *term* t_i	c	d

2.5. Naïve Bayes Classification

Naïve bayes classification is a statistical classification method based on the Bayes theorem which can be used to predict the probability of class membership [2]. Naïve bayes classification works well for feature-level text categorization [13]. The steps in the naïve bayes classification process are as follows:

1. Prepare the dataset.
2. Divide the data into training data and testing data.
3. Naïve bayes classification process.

At this stage, training is carried out to obtain a classification model with training data. The stages of modeling naïve bayes classification are:

   a. Calculates the total probability of each class/label. The trick is to divide the amount of class data by the amount of training data.
   b. Calculates the probability of a variable detail in the class. This is done by counting the number of cases for each class.

4. Model testing is done by multiplying all class variables then comparing the result between classes.
5. The result is obtained from the calculation of the configuration matrix to determine the level of accuracy.

3. RESULTS AND DISCUSSIONS

The steps taken in this research are pre-processing, feature weighting, feature selection, and text classification. this research begins by entering the Amazon review dataset, then continues with pre-processing including transform case, tokenize, stopword removal, stemming, and punctuation removal. Then the feature weighting is carried out using TF-IDF. After obtaining the result from the feature weighting, feature selection is carried out using chi square. Based on the selected features, the classification is carried out using the naïve bayes classification method. Then the classification is tested using test data and evaluated using a configuration matrix to obtain an accuracy value. The research method flowchart can be seen in the research that was conducted was the application of chi square and TF-IDF to optimize naïve bayes classification in the Amazon review sentiment analysis. The level of accuracy generated by the naïve bayes classification algorithm without applying chi square and TF-IDF for sentiment analysis in Amazon review is 82%. In research, optimization will be carried out using the stages of pre-processing, weighting features, and feature selection.

The pre-processing stage is carried out by implementing transform case, tokenize, stopword removal, stemming, and punctuation removal. Followed by weighting the features using TF-IDF to produce 1375 features. The result of the pre-processing and feature weighting stages can be seen in Table 8.

Table 8. The results of pre-processing and feature weighting

Token	TF-IDF
way	0.3689381440614246
plug	0.22691093981246754
us	0.44363633321554674
unless	0.44363633321554674
go	0.317858952096447
convert	0.4704643608427174

After obtaining the results of the feature weighting, the data is used to calculate the chi square value of each word. The chi square value can be seen in Table 9.

Table 9. The result of the chi square calculation

Feature	Token	*Chi Square*
315	way	1.2330314635
66	plug	4.4306625991
1348	us	0.0001580372
257	unless	1.9647577093
1260	go	0.0007901859
857	convert	0.9823788546

After obtaining the result of the chi square calculation, feature selection is carried out based on the top-K value of the chi square value. Then proceed with the data splitting stage to divide the training data and test data on the naïve bayes classification algorithm with a comparation of training data and test data of 90:10. The accuracy value of the naïve bayes classification algorithm based on the top-K value can be seen in Figure 2.

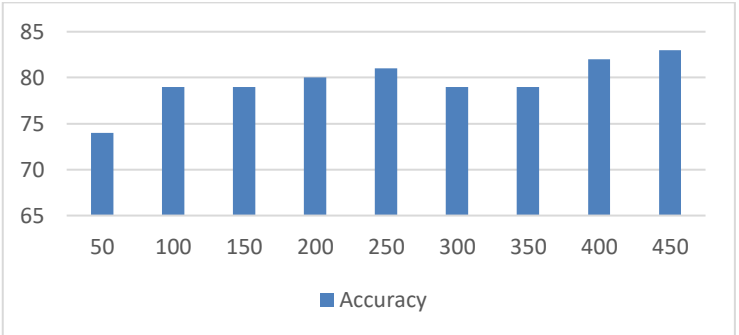

Figure 2. The accuracy of the naïve bayes classification algorithm based on the top-K value

The highest accuracy results are shown at the top-K value of 450 at 83%. Based on the result of the accuracy, it is proven that the application of chi square and TF-IDF can improve the accuracy of the naïve bayes classification algorithm by 1%. The selection of top-K is very influential in increasing the accuracy in this research.

4. CONCLUSION

In this research, the application of chi square and TF-IDF can improve the accuracy of naïve bayes classification algorithm in the Amazon review sentiment analysis. In the pre-processing and feature weighting stages using TF-IDF produces 1375 features. Then performed the feature selection using chi square to determine the optimization level of accuracy in the naïve bayes classification algorithm. The highest accuracy rate is shown on the top-K 450 at 83%. Meanwhile, the level of accuracy of the naïve bayes classification algorithm before applying chi square and TF-IDF was 82%.

REFERENCES

[1] A. Nurzahputra and M. A. Muslim, "Analisis sentimen pada opini mahasiswa menggunakan natural language processing," in *Seminar Nasional Ilmu Komputer (SNIK 2016)*, 2016, pp. 114–118.

[2] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "Optimasi naïve bayes classifier dengan menggunakan particle swarm optimization pada data iris," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), p-ISSN*, pp. 2355–7699, 2017.

[3] A. R. Safitri and M. A. Muslim, "Improved accuracy of naive bayes classifier for determination of customer churn uses smote and genetic algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020.

[4] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, 2009.

[5] L. D. Utami and R. S. Wahono, "Integrasi metode information gain untuk seleksi fitur dan adaboost untuk mengurangi bias pada analisis sentimen review restoran menggunakan algoritma naive bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[6] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8696–8702, 2011.

[7] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis," *Sci. J. Inform.*, vol. 6, no. 1, pp. 138–149, 2019.

[8] K. Oh, C.-G. Lim, S. S. Kim, and H.-J. Choi, "Research trend analysis using word similarities and clusters," *Int. J. Multimed. Ubiquitous Eng.*, vol. 8, no. 1, pp. 185–196, 2013.

[9] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018.

[10] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *2014 6th int. conf. inf. technol. electr. eng. (ICITEE)*, 2014, pp. 1–4.

[11] M. Liu and J. Yang, "An improvement of TFIDF weighting in text categorization," *Int. proc. comput. sci. inf. technol.*, vol. 47, pp. 44–47, 2012.

[12] A. Moh'd A Mesleh, "Chi square feature extraction based svms arabic language text categorization system," *J. Comput. Sci.*, vol. 3, no. 6, pp. 430–435, 2007.

[13] M. Govindarajan, "Sentiment analysis of restaurant reviews using hybrid classification method," *Int. J. Soft Comput. Artif. Intell. (IJSCAI)*, vol. 2, no. 1, pp. 17–23, 2014.