Application of the Naïve Bayes Classifier Algorithm using N-Gram and Information Gain to Improve the Accuracy of Restaurant Review Sentiment Analysis

Apriani Solikhatun [1*], Endang Sugiharti [1]

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: solikhatunapri@students.unnes.ac.id

ARTICLE INFO

Article history
Received 4 August 2020
Revised 10 September 2020
Accepted 12 October 2020

Keywords
Text Mining
Sentiment Analysis
Naïve Bayes Classifier
N-Gram
Information Gain

ABSTRACT

A consumer's review is an essential aspect for influencing others in determining decisions. The process of identifying positive or negative reviews can be conducted through sentiment analysis. One of the popular techniques in the sentiment analysis is the Naïve Bayes Classifier (NBC) algorithm, which has optimal performance. The purpose of this study was to improve the accuracy of the classifier in the analysis of restaurant review sentiments by applying N-Gram as feature extraction and Information Gain as a feature selection. N-Gram is used to produce new features that are more varied, while information gain functions to select relevant features with high weights. The dataset used in this study is the sentiment labeled dataset from UCI machine learning. The results of applying the NBC have an accuracy of 82.5%. The research results revealed that the Naïve Bayes Classifier's accuracy by using N-Gram and information gain of 86%. The application of N-Gram and information gain in the NBC algorithm can be concluded that it has succeeded in improving the classification accuracy of the restaurant review sentiment analysis with an increase in accuracy of 3.5%.

1 Introduction

With social media advancements, more and more people write their opinions about a product or service (Rana & Singh, 2017). Provided reviews are in the form of in-text thoughts, including food/restaurant reviews (Zhang, Ye, Zhang, & Li, 2011), books, and films (Moraes, Valiati, & Neto, 2013). In identifying positive or negative opinions, a method called sentiment analysis is needed.

Sentiment analysis or opinion mining analyzes opinions, sentiments, evaluations, assessments, attitudes, and public emotions towards an entity such as products, services, organizations, individuals, problems, events, topics, and their attributes. Sentiment analysis is popular because of its efficiency. Thousands of documents can be processed for sentiment analysis (Baid, Gupta, & Chaplot, 2019).

In conducting sentiment analysis, a classification technique is required, the classification techniques used in the sentiment review analysis include Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) (Dehkharghani, Mercan, Javeed, & Saygin, 2014). The Naïve Bayes Classifier (NBC) is a popular machine learning technique in text classification and performs well in many domains. However, the NBC has a very sensitive weakness in selecting features (Fauzi, Arifin, Gosaria, & Prabowo, 2016). In increasing accuracy, it is necessary to add a method to extract features to make it more varied and identify opinions more optimal, namely by using N-Gram (Zhang *et al.*, 2011).

According to Zhang *et al.* (2011), N-Gram's use affects the resulting classification accuracy of how N-Gram works by combining words to see the difference in sentiment from each word combination. However, N-Gram's use requires additional methods that can select important words because N-Gram produces many less relevant features. The feature selection method is considered to be used to obtain better results (Pujadayanti, Fauzi, & Sari, 2018).

There are two types of main feature selection methods in machine learning, namely wrappers and filters (Muthia, 2014). Filter types include information gain, chi-square, and log-likelihood ratio, while wrapper types comprise forward selection and backward elimination (Chandani, Wahono, & Purwanto, 2015). According to Uğuz (2011), information gain is a method that is often used in feature selection and has a good performance for English text classification. In the information gain method, we can see each feature to predict the correct class label because it chooses the highest value and is more effective in optimizing the classification results.

In their research, Chandani *et al.* (2015) compared the five most suitable feature selection algorithms for sentiment analysis on the film review dataset. The five algorithms are information gain, chi-square, forward selection, and backward selection. The information gain algorithm shows the best performance by obtaining the highest average classification accuracy with a percentage of 84.57%.

Meanwhile, Pranckevičius & Marcinkevičius (2017), in their research, compared several classification methods, namely Naïve Bayes, random forest, decision tree, SVM, and logistic regression with classification values based on the size of the training data, and the number of N-Gram. The N-Gram used were unigram, bigram, trigram, bigram unigram, and a combination of the three. The dataset uses product review data from Amazon. Combining unigram, bigram, and trigram results in a much higher added value for each classification method.

Based on the description above, the purpose of this study is to improve the classification accuracy of the NBC by applying N-Gram and Information Gain to restaurant review sentiment analysis so that it can better classify restaurant reviews based on positive or negative sentiments.

2 Methods

This research consists of several stages, including text pre-processing, sharing training data and test data, feature extraction, feature selection, and classification using the NBC algorithm. The process of classification of sentiment analysis in this study is carried out in Figure 1.

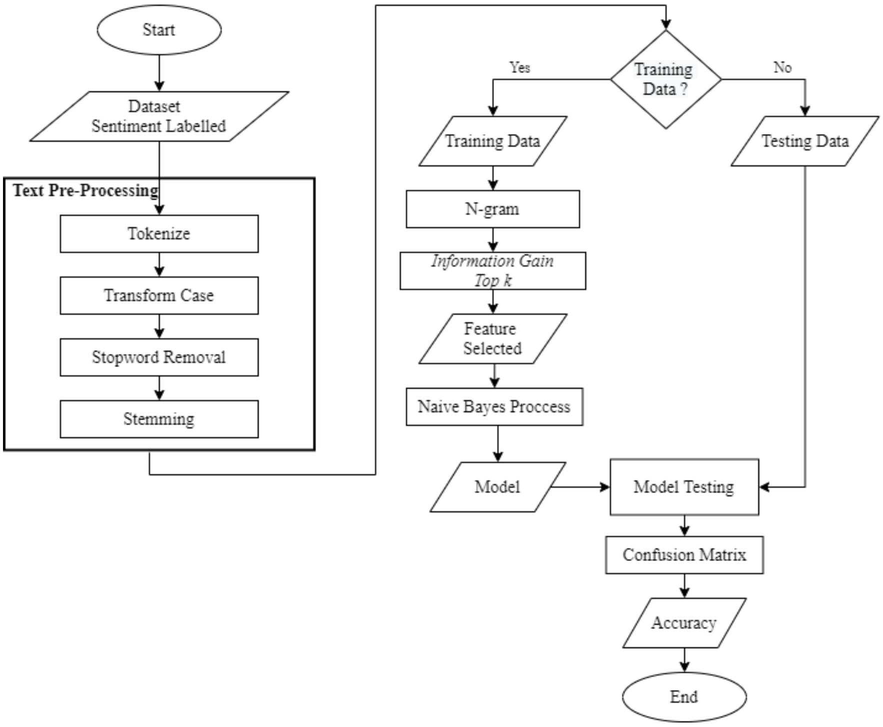

Figure 1. Flowchart of the NBC algorithm with N-Gram and information gain

2.1 Dataset

The data used is a sentiment labeled dataset with 500 documents, namely documents labeled positive and negative in the yelp_labelled field obtained from the UCI machine learning repository. The data obtained is in the form of a file with a .txt extension. The file is then converted into a file with the extension * .xlsx in the form of a table with two columns; namely, the first column contains text, and the second column includes labels defined as negative (0) and positive (1) as in Table 1.

Table 1. Sample data yelp_labelled.xlsx

Review	Sentiment
Not tasty, and the texture was just nasty.	0
The selection on the menu was great, and so were the prices.	1
The fries were great too.	1
A great touch.	1

2.2 Text Preprocessing

Text pre-processing is an important part of the text mining process system because the characters, words, and sentences identified at this stage are the basic units that are passed on to all further processing stages (Kannan & Gurusamy, 2014). The text pre-processing process stages include:

1. Tokenize is the stage of cutting the input string based on each word that contains it.
2. Transform case is the process of changing the entire text in a document to be lowercase or lower case.
3. Stopword removal is the process of removing words that often appear but have no effect on text extraction.
4. Stemming is the process of transforming a word into a root word by removing all word affixes.

2.3 N-Gram

In the sentiment analysis, the N-Gram method is used to analyze the sentiment of a text or document. N-Gram is distinguished by the number of character pieces of n. N-Gram with several chunks of one character is called unigram, while two characters are called bigram, and three characters are called trigrams (Maulana & Karomi, 2015). Meanwhile, N-Gram cuts off n-words taken from a sentence. In this study, using a combination of N-Gram unigram, bigram, and trigram. The N-gram results are shown in Table 2.

Table 2. The results of the N-Gram process

Token/Term	N-Gram Result
'tasti' 'textur' 'nasti'	'tasti' 'textur' 'nasti' 'tasti textur' 'textur nasti' 'tasti texture nasti'
'select' 'menu' 'great' 'price'	'select' 'menu' 'great' 'price' 'select menu' 'menu great' 'great price' 'select meu great' 'menu great price'
'fri' 'great'	'fri' 'great' 'fri great'
'great' 'touch'	'great' 'touch' 'great touch'

2.4 Information Gain

Information gain is a feature selection method that is often used to determine an attribute based on the limit of the importance of an attribute. In measuring the information gain value, it will determine the attributes that will be used or removed later. Attributes that meet the weighting criteria are later used in the classification process of an algorithm (Maulana *et al*., 2015). The following is the process of calculating information gain (Astuti, Muslim, & Sugiharti, 2019).

1. Find the entropy value before splitting with Equation 1.

$$Entropy\ (S) = -\sum_{i-1}^{k}(P_i) log2(P_i) \tag{1}$$

Whereas S : Case set
k : The number of partitions S
P_i: The proportion of S_i to k

2. Find the entropy value after splitting with Equation 2.

$$Entropy\ (S, A) = \sum_{i-1}^{k}\left(\frac{|Sv|}{|s|} * Entropy\ (Sv)\right) \tag{2}$$

Whereas S : Case set
A : Attribute
k : The number of partitions attribute
$|Sv|$: Number of cases on the partition to v
$|s|$: Number of cases in S
$Entropy\ (Sv)$: Total entropy in the partition

3. Find the information gain value with Equation 3.

$$Gain\ (S, A) = Entropy(S) - Entropy\ (S, A) \tag{3}$$

Whereas $Gain\ (S, A)$: Information Atribut (S,A)
$Entropy(S)$: Total entropy

$Entropy\ (S, A)$: Entropy (S,A)

2.5 Naïve Bayes Classifier (NBC)

NBC is a simple method but has high accuracy and text classification performance (Suhendra & Ranggadara, 2017). This method is often used to solve problems in machine learning because this method is known to have a high degree of accuracy with simple calculations (Aggrawal & Zhai, 2012). The following is the Bayes theorem used in Equation 4.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{4}$$

With the NBC approach, which assumes that every word in each category is independent of one another, the calculation can be simplified and can be written as Equation 5.

$$\mathrm{p}(\mathrm{X}\ |\mathrm{c_j}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{w_i}|\mathrm{c_j}) \tag{5}$$

By using Equation 5, Equation 6 can be written as.

$$\mathrm{C_{MAP}argmax}\ \ \mathrm{p}(\mathrm{c_j}) \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{w_i}|\mathrm{c_j}) \tag{6}$$

2.6 Confusion Matrix

A confusion matrix is a performance measurement technique that is widely employed for model classification. The confusion matrix can also be called an error matrix. There are two types of classification: "yes" and "no". If the presence of an incident, then "yes" means that the event has occurred, and "no" means no event (Avinash, Yasaswi, & Malleswari, 2019). Confusion matrix in the performance of each classifier that uses the help of its components, namely True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) (Patro & Patra, 2015). The table of a confusion matrix is shown in Table 3.

Table 3. Confusion matrix

Classification		**Predicted Class**	
		Class = YES	Class = NO
Observed class	Class = YES	(true positive-TP)	(false negative-FN)
	Class = NO	(false positive-FP)	(true negative-TN)

To gauge the accuracy, Equation 7 is applied.

$$\text{Accuracy (\%)} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{7}$$

3 Results and Discussion

3.1 Research Results

The results of research that has been carried out by the NBC algorithm using N-Gram and Information Gain are:

3.1.1 Text Preprocessing Process

Before the classification process, the restaurant review data goes through the text pre-processing stage. In the text pre-processing stage using tokenize, transform case, stopword removal, and stemming. This stage generates a token/term that will be used in the next step of the process. The results of the text pre-processing process can be seen in Table 4.

Table 4. The results of the term frequency process with the highest DF value

Features	DF
place	96
food	92
service	69
good	68
great	54
back	50
go	50
time	48
never	42
like	39

3.1.2 Feature Extraction Process

Before being used in the classification process, the dataset used is processed first at the text pre-processing stage to produce a token/term. The token/term is then carried out with the feature extraction process using N-Gram. N-Gram used a combination of unigram, bigram, and trigram, which produces 7812 features consisting of various words, two words, and three words. After the N-Gram process is carried out, the next process is to find the frequency/number of words/terms that appear in all documents (TF). Meanwhile, Document Frequency (DF) is the number of tokens/terms occurrences throughout the document. The following shows the features of the results of the Term Frequency (TF) process with the highest DF value can be seen in Table 5.

Table 5. The results of the term frequency process with the highest gain value

Features	Gain Value
great	0.045277
good	0.026134
bad	0.024444
never	0.021933
delici	0.017971
love	0.013646
dont	0.010181
friendli	0.0104465
fantast	0.009802
amaz	0.009665

3.1.3 Feature Selection Process

The process of selecting features in this study used Information Gain. The results of this stage, the features were selected based on the number of $top\ k$ specified, meaning that the feature with the highest gain was then ranked as many as the number needed, and the number of features that produce the best accuracy was sought. Table 6 shows the features with the highest gain value due to the feature selection process using Information Gain.

3.1.4 Classification Process

After the feature selection process was done, the classification process can then be carried using the NBC algorithm. The proposed model then went through the accuracy testing stage using a confusion matrix.

3.1.5 Model Evaluation

After building a model from the proposed method, the model was evaluated to determine its accuracy based on the calculation of test data using the confusion matrix method. Based on this research, the results of applying N-Gram and Information Gain to the NBC algorithm obtain an accuracy of 84.5%. The results of the confusion matrix calculation can be seen in Table 7.

Table 6. The results of the confusion matrix

	Predict Neg	*Predict Pos*
Actual Neg	87	8
Actual Pos	20	85

The calculation of classification accuracy uses the results of the confusion matrix as in Equation 8.

$$\text{Accuracy } (\%) = \frac{87+85}{87+85+8+20} = 0{,}86 \tag{8}$$

3.2 Discussion

This study applies N-Gram and information gain to improve accuracy in the classification process of restaurant sentiment analysis. The classification algorithm used was the NBC algorithm. The testing process uses the python programming language. The data used was a sentiment labeled dataset (yelp_labelled field), which contained 1000 texts with 500 texts labeled positive and 500 texts labeled negative.

The dataset used was processed in the first stage, namely, text pre-processing, which will produce data with a new format in terms/words. Then proceed with dividing the data into training data and test data using the splitter method in the Sklearn library with the proportion of 80% training data and 20% test data. Then obtained training data with a total of 800 data and test data with 200 data.

After splitting the data into training and test data, the classification process was carried out using the NBC algorithm. The result of the NBC algorithm is 82,5%.

The next step was the application of feature extraction using N-Gram. In this study, using a combination of unigram, bigram, and trigram with the total features produced amounted to 7812. The accuracy results of the NBC algorithm after being combined with N-Gram and using were 85.5%. The increase in accuracy obtained by the NBC after being combined with N-Gram is 3%.

Furthermore, by implementing the information gain feature selection to reduce the number of features that are too large and less relevant and increasing classification accuracy. The application of information gain as a selection feature uses top k = 5700 features on the NBC algorithm with N-Gram. Based on the experiments that had been carried out using the Information Gain feature selection with top k = 5700, it produces the best accuracy, and fewer features are used. The experimental results based on top k can be seen in Figure 2.

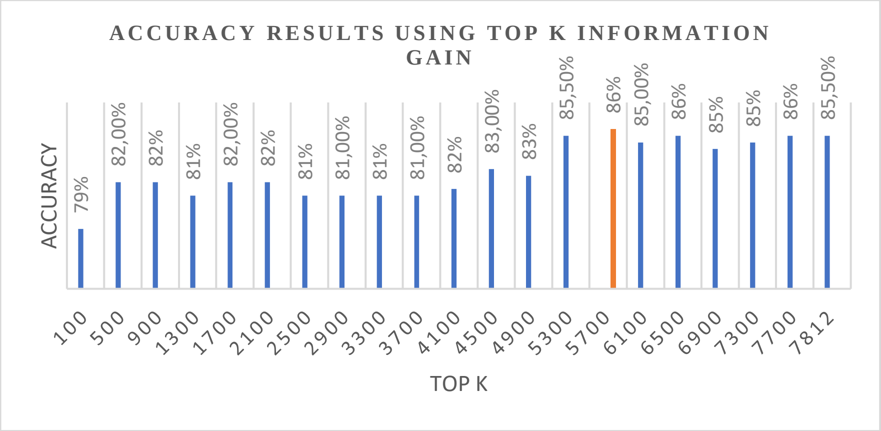

Figure 2. Accuracy results using top K information gain value

The increasing accuracy results of the NBC with N-Gram and Information Gain is 86%.

Comparison of the accuracy obtained from the proposed method with the method using the NBC algorithm without the combination of N-Gram and Information Gain and the NBC algorithm with N-Gram can be seen in Table 7.

Table 7. The comparison of accuracy results

Classification Algorithm	Accuracy Results
Naïve Bayes Classifier Algorithm	82.5%
Naïve Bayes Classifier Algorithm with N-Gram	85.5%
Naïve Bayes Classifier Algorithm with N-Gram and Information Gain	86%

Based on the experiments' results, the accuracy value for the classification of restaurant reviews' sentiment analysis using the NBC is 82.5%. After being combined with N-Gram, the NBC algorithm is 85.5%, while the NBC algorithm after being combined with N-Gram and Information Gain is 86%. So that the accuracy gain 3.5% increases.

With a given level of accuracy, this model can analyze the sentiment of different review data properly. A comparison of the previous sentiment analysis classification results was conducted to find out that this method is better than the existing methods. The comparison results are shown in Table 8.

Table 8. The comparison of accuracy with previous research

Author	Method	Accuracy
Pranckevičius & Marcinkevičius (2017)	Naïve Bayes + unigram, bigram dan trigram	45.22 %
Laksono *et al.*, (2019)	Naïve Bayes	72.06%
	TextBlob	69.12%
Muthia (2014)	Naïve Bayes	78.5%
Utami & Wahono (2015)	Naïve Bayes + Information Gain + Genetic Algorithm	83%
	Naïve Bayes	73%
	Naïve Bayes + Information Gain + AdaBoost	81.5%
Proposed method	Naïve Bayes +N-Gram+ AdaBoost	86%

Based on the results of the comparison above, it can be concluded that the NBC algorithm using N-Gram and Information Gain higher accuracy compared to previous related studies.

Furthermore, this study's advantages can improve the accuracy in the classification of restaurant review sentiment analysis so that further researchers can use it to research text classification.

4 Conclusion

From this research, it is known that applying the NBC algorithm combined with N-Gram as feature extraction and Information Gain in the feature selection process can improve the classification accuracy of restaurant review sentiment analysis. The accuracy results obtained are much better than using only the NBC algorithm without being combined with N-Gram and Information Gain to classify restaurant review sentiment analysis.

References

Aggarwal, C. C., & Zhai, C. X. (2012). A Survey of Text Classification Algorithms. *Mining Text Data*, 163-222. doi:10.1007/978-1-4614-3223-4_6

Astuti, W. T., Muslim, M. A., & Sugiharti, E. (2019). The Implementation of The Fuzzy Neuro Method using Information Gain for Improving Accuracy in Determination of Landslide Prone Areas. *Scientific Journal of Informatics, 6*(1), 95-105. doi:10.15294/sji.v6i1.16648

Avinash, K., Yasaswi, B., & Malleswari, D. N. (2019). Risk Assessment Strategy Performance Measure using Confusion matrix. *International Journal of Recent Technology and Engineering,* 7(6), 635-637. Retrieved from https://www.ijrte.org/wp-content/uploads/papers/v7i6s/F03250376S19.pdf

Baid, P., Gupta, A., & Chaplot, N. (2019). Sentiment Analysis of Movie Reviews using Machine Learning Classifiers. *International Journal of Computer Applications, 182*(50), 25-28. doi:10.5120/ijca2017916005

Chandani, V., Wahono, R. S., & Purwanto. (2015). Komparasi Algoritma Klasifikasi Machine Learning dan Feature Selection pada Analisis Sentimen Review Film [Machine Learning Classification Algorithm Comparison and Feature Selection in Film Review Sentiment Analysis]. *Journal of Intelligent Systems,* 1(1), 56-60. Retrieved from http://www.journal.ilmukomputer.org/index.php?journal=jis&page=article&op=view&path%5B%5D=10&path%5B%5D=21

Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y. (2014). Sentimental Causal Rule Discovery from Twitter. *Expert Systems with Applications, 41*(10), 4950–4958. Retrieved from http://research.sabanciuniv.edu/26264/1/final_version_9_09.pdf

Fauzi, M. A., Arifin, A. Z., Gosaria, S. C., & Prabowo, I. S. (2016). Indonesian News Classification Using Naïve Bayes and Two-phase Feature Selection Model. *Indonesian Journal of Electrical Engineering and Computer Science, 2*(3), 401-408. doi:10.11591/ijeecs.v8.i3

Kannan, S., & Gurusamy, V. (2014). Pre-processing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16. Retrieved from https://www.researchgate.net/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019). Sentiment Analysis of Restaurant Customer Reviews on Tripadvisor using Naïve Bayes. *International Conference on Information and Communication Technology and Systems (ICTS)*, 49-54. doi:10.1109/ICTS.2019.8850982

Maulana, M. R., & Karomi, M. A. (2015). Information Gain untuk mengetahui Pengaruh Atribut [Information Gain to Determine the Effect of Attributes]. *Jurnal Litbang Kota Pekalongan, 9*, 113-123. Retrieved from https://jurnal.pekalongankota.go.id/index.php/litbang/article/view/28

Moraes, R., Valiati, J. F., & Neto, W. P. (2013). Document Level Sentiment Classification: An Empirical Comparison Between SVM and ANN. *Expert Systems with Applications, 40*(2), 621–633. doi:10.1016/j.eswa.2012.07.059

Muthia, D. A. (2014). Sentiment Analysis of Hotel Review Using Naïve Bayes Algorithm and Integration of Information Gain and Genetic Algorithm as Feature Selection. *International Seminar on Scientific Issues and Trends (ISSIT)*, 25-30. Retrieved from http://issit.bsi.ac.id/proceedings/index.php/issit2014/article/view/30/30

Patro, V. M., & Patra, M. R. (2015). A Novel Approach to Compute Confusion matrix for Classification of n-Class Attributes with Feature Selection. *Transactions on Machine Learning and Artificial Intelligence, 3*(2), 52-64. doi:10.14738/tmlai.32.1108

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing, 5*(2), 221-232. doi:10.22364/bjmc.2017.5.2.05

Pujadayanti, I., Fauzi, M. A., & Sari, Y. A. (2018). Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Naïve Bayes dan N-gram [Automatic Ratings Prediction on Beauty Product Reviews with the Naïve Bayes Method and N-gram]. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK), 2*(1), 4421–4427. Retrieved from http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2921

Rana, S., & Singh, A. (2017). Comparative Analysis of Sentiment Orientation using SVM and Naïve Bayes Techniques. *International Conference on Next Generation Computing Technologies (NGCT)*, 106-111. doi:10.1109/NGCT.2016.7877399

Suhendra & Ranggadara, I. (2017). Naïve Bayes Algorithm with Chi Square and NGram. *International Research Journal of Computer Science (IRJCS), 4*(12), 28–33. doi:10.26562/IRJCS.2017.DCCS10087

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of Sentiment Reviews using N-gram Machine Learning Approach. *Expert Systems with Applications, 57*, 117–126. doi:10.1016/j.eswa.2016.03.028

Uğuz, H. (2011). A Two-Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorithm. *Knowledge-Based Systems, 24*(7), 1024-1032. doi:10.1016/j.knosys.2011.04.014

Utami, D. L., & Wahono, R. S. (2015). Integrasi Metode Information Gain untuk Seleksi Fitur dan AdaBoost untuk Mengurangi Bias pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes [Information Gain Methods Integration for Feature Selection and AdaBoost to Reduce Bias in Restaurant Review Sentiment Analysis Using the Naïve Bayes Algorithm]. *Journal of Intelligent Systems, 1*(2), 120-126. Retrieved from https://www.neliti.com/publications/243716/integrasi-metode-information-gain-untuk-seleksi-fitur-dan-adaboost-untuk-mengura

Zhang, Z., Ye, Q., Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment Classification of Internet Restaurant Reviews Written in Cantonese. *Expert Systems with Applications, 38*(6), 7674-7682. doi:10.1016/j.eswa.2010.12.147