




Sentiment Analysis on Coffee Consumer Perceptions on Social Media Twitter Using Multinomial Naïve Bayes

Research Article

Nurul Qomariah 

Department of Statistics, Universitas Muhammadiyah Semarang, Semarang 50254, Indonesia

**nurulqomariah1809@gmail.com (coresponden author)*

Article history :

Received : 29Des 2019

Accepted : 29 Feb 2020

Available online : 10 April 2020

ABSTRACT

Coffee is an Indonesian plantation product that has high competitiveness in the international market. Demand for coffee can be influenced by people's perceptions of coffee in Indonesia. People easily provide opinions and opinions through social media. One of the most popular social media in all circles is Twitter. The types of coffee that are trending topics on Twitter today are civet coffee, black coffee, sick coffee and bitter coffee. In this study, we applied a multinomial naive bayes model to determine public perceptions that could influence coffee demand. The data used is sourced from Twitter sracpping data. The experimental results show that the level of accuracy, precision, and recall is 94%, 99%, 88%. The model used is successful in determining the positive sentiment of the community. It can be concluded that positive sentiment from the community has influenced the increase in coffee orders in Indonesia.

Keywords : Coffee, sentimen analysis, text mining, naïve bayes, machine learning.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. INTRODUCTION

In the era of globalization, the development of a country can be seen from the quality of the country's development. One of them is Indonesia, which has development programs such as consumption and production patterns. The agricultural sector is the main sector in sustainable consumption and a guaranteed production pattern. One of the potential agricultural subsectors is coffee plantations (Sitanggang & Sembiring, 2013).

Coffee is one of the plantation products which until now has become a national superior commodity and even has competitiveness in the international market. The demand for coffee products in Indonesia continues to increase every year. The demand for an item is seen from the public perception of the product. The public will find it easy to express an opinion using social media, one of which is Twitter. People will easily tweet about coffee products in Indonesia (Schwarz et al., 1994).

Twitter is a microblogging tool that can make it easier for people to exchange opinions on a variety of current topics. People can take advantage of opinions or responses by analyzing sentiments, such as expressing positive sentiment for a product or negative sentiment. Automatically extracting opinion information in text data from the Twitter platform is important. This approach is called Sentiment Analysis (Farha & Magdy, 2021).

Several previous studies such as (Nayak & Natarajan, 2016) compared the performance of five machine learning-based text classification algorithms such as Decision Rules, Decision Tree, K-Nearest Neighbor, Naïve Bayes and Support Vector Machine, with evaluations used are precision, recall, F-measure and accuracy. The results of the study reported that Naïve Bayes was superior to other comparative models.

Another study by (Artanti et al., 2018). They used Naïve Bayes to classify sentiment for online shopping website service ratings. The purpose of this study is to classify sentiments into negative, positive and neutral classes. The data used is 1200 data. The results obtained in this study using the Naïve Bayes Classifier method have a good value accuracy of 93.3%. Research by (Kalokasari et al., 2017) conducted research on the implementation of the Naïve Bayes Classifier multinomial algorithm to classify outgoing letters so that they can determine letter numbers automatically, obtained an accuracy level of 89.58% (Kalokasari et al., 2017).

Based on the foregoing, we applied the multinomial naïve bayes as a proven model for sentiment classification. Previous research above, can be applied text classification techniques using Multinomial Naïve Bayes. Text classification is the process of grouping documents into different categories. The advantage of naïve Bayes is that it is easy to implement and has high performance. Multinomial Naïve Bayes can also handle large vocabulary sizes and reduce high error rates (Suryanendra,

2018).

The aim of this research is to apply multinomial naïve bayes for sentiment classification in twitter data with case studies of coffee consumers in Indonesia.

This study is organized as follows. In section 2, there is a presentation of the proposed method. The experimental results of comparing the proposed method with others are given in section 3. Finally, the last section is devoted to concluding the work of this paper

2. RESEARCH METHOD

2.1. Coffee production

Indonesian coffee production has no effect on changes in coffee prices and substituted commodities in the domestic market, wage levels and area. The supply of coffee in Indonesia is influenced by the level of technology and the number of offers a year, while the effect of coffee and tea prices is not statistically significant. The negative coefficient of tea variable indicates that coffee and tea in Indonesia are competition products. Types of coffee that are often discussed on social media, especially on twitter, include bitter coffee, sick coffee, civet coffee and black coffee. From the topic of coffee has its own meaning. Bitter coffee is a coffee product that tastes bitter or without sugar. Then, sick coffee is a hashtag that has become a trending topic on social media, especially Twitter, sick coffee means that someone will get sick after consuming coffee. For civet coffee, it is a coffee product that comes from civet beans which until now has been popular with people in all circles. And black coffee is a coffee product that comes from the original black coffee beans which are usually brewed by adults [6].

2.2. Text mining

According to Feldman and Sanger (2007), text mining is the process of exploring and analyzing large amounts of unstructured text data aided by software that can identify concepts, patterns, topics, keywords, and other attributes in the data. This is also known as text analysis, although some people draw distinctions between the two terms; in that view, text analytics are applications that are enabled by the use of text mining techniques to sort data sets. Text mining is also a process of finding the latest information or trends that were previously unknown by processing and analyzing large amounts of data or big data.

Text mining has become more practical for data scientists and other users due to the development of big data platforms and deep learning algorithms that can massively analyze unstructured data sets.

2.3. Multinomial naïve bayes

Multinomial naïve bayes is the process of taking the number of words that appear in each document, where this

method assumes a document that has several occurrences in a word whose length does not depend on the class in the document. According to (Syaputri et al., 2020), the probability that a document d is in class c , this condition can be stated by the following formula:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq nd} P(t_k|c) \quad (1)$$

For the formula $P(t_k|c)$, which is the conditional probability of the word t_k contained in a document from class c . $P(c)$ is the prior probability of a document contained in class c . $(t_1, t_2, \dots, t_{nd})$ are tokens in document d which are part of the vocabulary used as classification and are the number of tokens in document d .

To estimate the prior probability $P(c)$, the following formula can be seen:

$$P(c) = \frac{N_c}{N} \quad (2)$$

where, N_c is number of training documents in class c , and N is the total number of training documents from all classes.

To pay attention to the conditional probability $P(t|c)$ it is stated by the following Eq. 3.

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (3)$$

where T_{ct} is the number of occurrences of the word t in a training document in class c . And the $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$ is the total number of words in the training document in class c t' is the total number of words in the training document

To eliminate zero values in a document. Laplace smoothing is used as the process of adding the value of 1 to each T_{ct} value in the calculation of conditional probability which is stated by the following Eq. 4.

$$P(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B'} \quad (4)$$

where B' is the total number of unique words in the entire class in the training document.

To get a high probability value for each word, the Laplace smoothing or add-one method is used. This method is used so that the value of the probability of each word meeting the conditions, which is not equal to zero. So, if a value on word probability is zero, then the data that is neither training nor testing data will never be sufficient to represent the frequency at which step occurs [7].

2.4. K-fold cross validation

K -fold cross validation is a statistical method used to partition data into data training and testing data. This method is often used by researchers because it can reduce bias that occurs in sampling. This method is used repeatedly to divide the data into two, namely training data and testing data, each data has the opportunity to become testing data.

The most optimal k value is 10. For this reason, 10-fold cross validation is one of the recommended K -fold cross validations for selecting the best model because it tends to provide less biased accuracy estimates compared to ordinary cross validation, leave-one-out cross. validation

and bootstrap. In 10-fold cross validation, the data is divided into 10-folds of approximately the same size, so that we have 10 subsets of data to evaluate the performance of the model or algorithm. For each of the 10 data subsets, cross validation will use 9-fold for training (training data) and 1-fold for testing (data testing).

2.5. Model evaluation

Evaluation of a calcification is generally carried out using a data set that is tested, not used in the classification training, at a certain measure. At this stage a number of measures can be used to reassess and evaluate the classification model, in our study using accuracy, error rate, recall, specificity and precision (Lim, et al. 2006). The evaluation used can be defined as follows:

$$Accuracy = \frac{TP+FN}{TP+FP+TN+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

2.6. Proposed method

In this study we used the secondary data. The data is taken from the Tweet database using the Twitter API (Application Programming Interface) with coffee product keywords, namely 'Kopi Luwak', 'Coffee Sick', 'Black Coffee' and 'Bitter Coffee'. The sample used is data taken as much as one month in November 2019.

In classification process, opinions conveyed via tweets contain many words which are considered noise and therefore need to be addressed. Preprocessing aims to prepare raw data to reduce words that are deemed unsuitable to be continued in the classification process. The preprocessing stages in this study were case folding, data cleaning, tokenizing, stemming, and stopword removal. Case folding will convert all tweet data to lowercase. Cleansing data consists of deleting url, username, character and short words. Followed by tokenizing which divides the data into word by word, then stemming or deleting the prefix and / or word endings with Indonesian rules using stemmer literature. Next is to remove stop words, or words that often appear in documents but are deemed useless for the classification process.

The next step is feature selection, in this phase we use the chi-square. The contingency used in the chi square feature selection process is the size $b \times k$, as shown in Table 2.

O_{ij} is the actual count of the two observed variables, in this case the feature as row (*line*), and column (*col*) class (negative, neutral, positive). Then find the expected count. The expected count table is shown in Table 2.

Finding the expected count of O_{ij} , or filling in cell E_{ij} following the Eq. 8 (Prabowo & Thelwall, 2009).

$$E_{ij} = \frac{b_i k_j}{N} \quad (8)$$

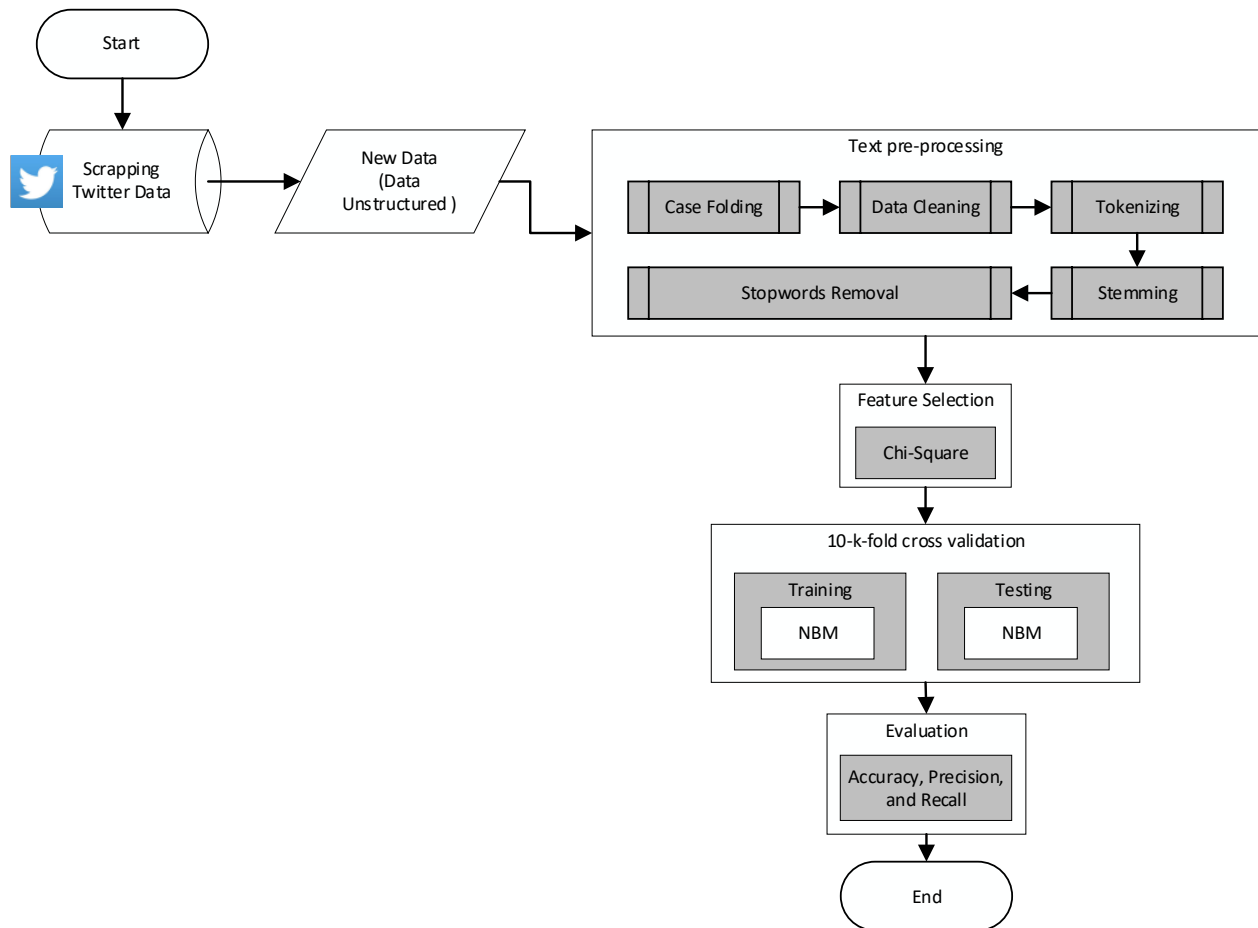


Fig 1. Flowdiagram of the proposed method.

Table 1. Labelling sentimen class.

	Col_1	Col_2	...	Col_j	Amount
$Line_1$	0_{11}	0_{12}	...	0_{1j}	b_1
$Line_2$	0_{21}	0_{22}	...	0_{2j}	b_2
...
$Line_i$	0_{i1}	0_{i2}	...	0_{ij}	b_i
Amount	k_1	k_2	...	k_j	N

Table 2. Contingency expected count.

	Col_1	Col_2	...	Col_j
$Line_1$	0_{11}	0_{12}	...	0_{1j}
$Line_2$	0_{21}	0_{22}	...	0_{2j}
...
$Line_i$	0_{i1}	0_{i2}	...	0_{ij}

After selecting the chi square feature, the next step is the classification process using naive bayes. The prediction results of the test data are displayed in the form of a configuration matrix that supports the calculation of the performance of the method used.

The classification results are displayed in the form of a discussion matrix which is then calculated to determine the percentage of accuracy, precision, and recall used to

see the performance of the method used.

3. RESULTS AND DISCUSSION

3.1. Scrapping data

Twitter data scrapping uses the Twitter API (Application Programming Interface) which accessed by

the Rstudio program as text data to be processed. The data used were trending tweets about coffee topics during November 2019, namely sick coffee, bitter coffee, black coffee and civet coffee, which were obtained as many as 3275 tweets.

3.2. Processing

In this stage, we prepare the text data that will be used for processing to the next stage. Tweet text data that is still unsupervised will be processed to become supervised data or have the same format. The following is the preprocessing result:

Table 1. Twitter data preprocessing results.

Hasil Case Folding dan Tokenisasi	
@PadantyaAbiyyu	minum luwak white coffee
@N_LalaJKT48	minum kopi nikmat tidak bikin lala
Luwak white coffee, kopi nikmat tidak bikin Lala kembang	kembang
Hasil Stemming dan Filtering	
luwak white coffee, kopi nikmat tidak bikin Lala kembang	minum kopi luwak nikmat tidak kembang
Hasil Data Preprocessing	
selesai minum kopi hitam aku jadi mules seperti nanggung mules 5 minggu	selesai minum kopi hitam aku jadi mules seperti nanggung mules 5 minggu

3.3. Descriptive analysis

After going through the preprocessing stage, the next step is followed by a descriptive analysis of the tweets on the words that have the most occurrence frequency. Descriptive analysis was conducted to find an overview of tweets about coffee on social media twitter. Information that can be retrieved is the number of tweets given by consumers per day with a frequency of 10 tweets. Descriptive analysis can be seen in the following graph in Fig 2.

3.4. Labeling words

We used two categories of sentiment classes namely positive and negative. Labeling is done automatically using a Colonial Lexicon dictionary. The results of the labeling for tweets about coffee can be shown in Tabel 2.

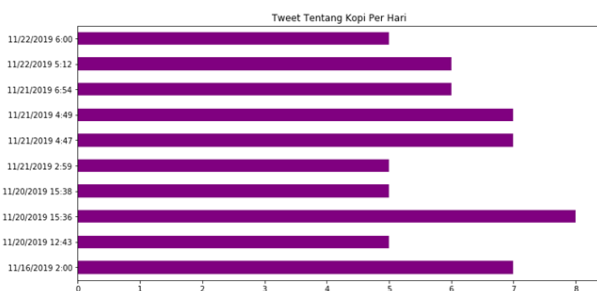


Fig. 2. The frequency of occurrence of the most words.

Table 2. Labelling sentiment class.

Sentimen	Jumlah <i>tweet</i>
Positif	2709
Negatif	566

Based on the results of automatic sentiment labeling, we received 2709 positive sentiments while less positive sentiment, namely 566 reviews. Tweets are classified as positive sentiment if they contain positive statements such as praise, positive tweets on coffee, happy, lovers, likes. A review is classified as a negative sentiment if it contains negative statements such as dislike, discomfort, drinking coffee becomes sick and so on.

3.6. Results

At this stage an evaluation is carried out on the classification using multinomial naïve bayes. The first result is obtained from the equation in the Classification Confusion Matrix in each class. Tabel 3, shown the results are obtained:

Table 3. Labelling sentiment class.

	Aktual			Jumlah
	Class	Positif	Negatif	
Prediksi	Positif	TP=259	FN=12	271
	Negatif	FP= 8	TN= 49	57
	Jumlah	267	61	328

Based on the results of calculations using the confusion matrix in the above tweets data classification, 271 tweets were predicted to be correct using the Multinomial Naïve Bayes. Meanwhile, 12 wrong positive predictions and 49 wrong negative predictions.

The next stage is the classification results using the Multinomial Naïve Bayes. The following results are obtained.

Table 4. Classification results using multinomial naïve bayes.

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
94%	99%	84%

Based on the classification results using the Multinomial Naïve Bayes in the table above for negative and positive tweets, the accuracy value is 94% which is high, this can be said that the accuracy in predicting a word is in accordance with the answer of a system. Whereas for precision of 99%, this means that the proportion of the number of relevant text documents recognized by all text documents is in accordance with the system, namely the positive class whose value is higher than the negative class. At the recall value of 84%.

It can be concluded that the success rate of the positive opinion data system in determining the results of an information is that it only has a few errors during the classification process.

5. CONCLUSION

The general picture is obtained from the results of descriptive analysis on tweets per day about coffee at most on November 22 at different times. Meanwhile, for tweets that often appear in negative sentiments are stomachaches, while positive sentiments are civet coffee and black coffee.

The results of the classification of sentiments on tweets about coffee in Indonesia using the Niave Bayes Multinomial Algorithm with data sharing using 10-fold Cross Validation, the result is that the accuracy value is 94% which is high, this can be said that the accuracy in predicting a word is in accordance with the responsibility of a system. Whereas for precession of 99%, this means that the proportion of the number of relevant text documents recognized by all text documents is in accordance with the system, namely the positive class whose value is higher than the negative class.

The recall value is 84%. It can be said that the success rate of the positive opinion data system in determining the results of an information is that it only has a few errors during the classification process

REFERENCES

- Artanti, D. P., Syukur, A., Prihandono, A., & Setiadi, D. R. I. M. (2018). Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes. *Konferensi Nasional Sistem Informatika 2018*, 8–9.
- Farha, I. A., & Magdy, W. (2021). A comparative study of effective approaches for Arabic sentiment analysis. *Information Processing and Management*, 58(October 2020), 102438. <https://doi.org/https://doi.org/10.1016/j.ipm.2020.102438>
- Kalokasari, D. H., Shofi, I. M., & Setyaningrum, A. H. (2017). Implementasi Algoritma Multinomial Naive Bayes Classifier pada Sistem Klasifikasi Surat Keluar (Studi Kasus : DISKOMINFO Kabupaten Tangerang). *Jurnal Teknik Informatika*, 10(2). <https://doi.org/10.15408/jti.v10i2.6822>
- Nayak, A., & Natarajan, D. S. (2016). Comparative study of Naïve Bayes , Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds. *International Journal of Advanced Studies in Computer Science and Engineering*, 5(1), 14–17.
- Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 143–157. <https://doi.org/10.1016/j.joi.2009.01.003>
- Schwarz, B., Bischof, H. P., & Kunze, M. (1994). Coffee, Tea, and Lifestyle. *Preventive Medecine*, 23, 377–384. <https://doi.org/10.1006/pmed.1994.1052>
- Sitanggang, J. T. N., & Sembiring, S. A. (2013). Pengembangan Potensi Kopi Sebagai Komoditas Unggulan Kawasan Agropolitan Kabupaten Dairi. *Jurnal Ekonomi Dan Keuangan*, 1(6), 33–48.
- Syaputri, A. W., Irwandi, E., & Mustakim. (2020). Naïve Bayes Algorithm for Classification of Student Major ' s Specialization. *Journal of Intelligent Computing and Health Informatics*, 1(1), 1–5. <https://doi.org/https://doi.org/10.26714/jichi.v1i1.5570>



Nurul Qomariyah is student graduate in Department of Statistics of Universitas Muhammadiyah Semarang in 2016. Her research interests include the statistics forecast, and data mining.