SINTECHCOM: Science, Technology and Communication Journal
Vol. 3, No. 2, February 2023, pp. 35-40, DOI: 10.59190/stc.v3i2.229

Performance comparison of the Naive Bayes algorithm and the k-NN lexicon approach on Twitter media sentiment analysis

Azhar[1,*], Siti Ummi Masruroh[1], Luh Kesuma Wardhani[1], Okfalisa[2]
[1]Dept. Informatics Engineering, UIN Syarif Hidayatullah Jakarta, South Tangerang 15412, Indonesia
[2]Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

Sentiment analysis or opinion mining is a natural language that processes words to find out opinions, attitudes, or moods about certain things. Word processing in this study related to the process of classification in textual documents, which was classified into three classes, positive, negative, and neutral. Data obtained from social media Twitter were related to netizens' comments as many as 1000 comments. These data were crawled using keywords of the "Pilpres2019" and "Jokowi". This study compared the performance of the Naive Bayes and k-Nearest Neighbor (k-NN) algorithms with the lexicon approach in classification. The aim of this study was to compare the level of accuracy, precision, and recall of Naive Bayes and the k-NN algorithm with the lexicon approach. From the evaluation, we concluded that the combination of the k-NN algorithm and the lexicon approach could improve accuracy in this sentiment analysis case. Generally, the k-NN algorithm with lexicon approach in which the k value is k = 5 has better performance with a 77% of accuracy level, followed by Naive Bayes with an accuracy of 81% of accuracy level.

ARTICLE INFO

Article history:

Received Jan 10, 2023
Revised Feb 5, 2023
Accepted Feb 20, 2023

Keywords:

Classification
k-Nearest Neighbor
Lexicon
Naive Bayes Classifier
Sentiment Analysis

* **Corresponding Author**
E-mail address: azhar.roses14@mhs.uinjkt.ac.id

1. INTRODUCTION

Social media such as Facebook, Twitter, and You Tube have experienced an increase in active users. From 2015 data, the number of Twitter users is more than 285 million. The majority of users are aged 16 years to 65 years and over [1-3]. Twitter social media has also recently experienced quite drastic changes in function. When Twitter was first established, this media was used as a tool for replying to messages and a place to share, but due to its high popularity, the use of its functions has increased, such as searching for trends, product promotions, and giving ratings [4, 5]. Recently, Twitter has also often been used in the socio-political realm, for example in joint social movements, reporting information about traffic jams, politics, weather conditions, natural disasters, and certain information such as warnings about events [6-8]. As a result, Twitter's function, which started as a place for questions and answers and sharing information, has become a means used in various aspects, one of which is as a research medium in the form of text mining, sentiment analysis, and artificial intelligence [9-11].

Twitter has an essential role as an indicator of trends that occur, for example during the 2019 presidential candidate election. Since 2018, there have been many trending topics on Twitter regarding presidential candidates for 2019, although the truth of the news is not certain. The high interest of netizens in responding to popular news in Indonesia has invited a lot of research based on machine learning, text mining, and artificial intelligence to try to use data from Twitter media. One of the studies on sentiment analysis using machine learning was carried out in 2016 by Ghulam Asrofi with

the title Sentiment Analysis of Candidates for Governor of DKI Jakarta 2017 and Devika M. D. for comparative research on performance analysis in sentiment analysis [12, 13].

Naive Bayes and k-NN are several machine learning methods that are often used in text classification and provide quite good results. In this research, a comparison of the performance of the two methods was carried out in the case of sentiment analysis. k-NN which is used as a classification method is combined with the lexicon approach, where in several studies this combination provides better results in the performance of the classification method [14-18].

Research data was obtained from a dataset originating from Twitter media of 1000 documents. This document was obtained using crawling techniques with the keywords "Pilpres 2019" and "Jokowi". This paper is divided into several discussions, namely an introduction explaining this research's motivation. Section 2 reviews the literature review that underlies the research. Next, section 3 explains the research methods used in this research. The results and discussion of this research can be seen in section 4. The final part of this research is the conclusion of the study.

2. LITERATURE REVIEW

2.1. Pre-Processing

Pre-processing in the document classification process is used to build an index from the document collection. The index is a set of terms that indicate the content or topics contained in a document [19]. Term extraction usually involves several main operations, including:

a. Case folding is the stage of changing all letters into lowercase.
b. Filtering is the stage of removing punctuation marks such as "@,.!" To make the next stage easier.
c. Separation of a series of terms (tokenization). Tokenization is the task of separating a series of words in a sentence, paragraph, or page into tokens or pieces of single words or termed words.
d. Normalization is the stage of re-checking whether each token matches the spelling in the KBBI or not and changing non-standard word terms to words that match the KBBI.
e. Removal of stop words. Stop words are defined as terms that are not related (irrelevant) to the main subject of the database even though these words are often present in the document. Examples of stop words are there are, are, there are, as for, rather.
f. Stemming, words that appear in documents often have many variants. Therefore, every word that is not a stopword is reduced to a suitable stemmed word (term), that is, the word is stemmed to obtain its root form by removing the prefix or suffix.

2.2. Naive Bayes

The Naive Bayes method utilizes probability or possible values. The basic concept used by naïve Bayes is Bayes' theorem, namely carrying out classification by calculating the probability value p(c), namely the class probability if the document is known. Naive Bayes considers a document as a collection of words that make up the document and does not pay attention to the order in which the words appear in the document. The probability calculation can be considered as the product of the probabilities of the words appearing in the document [20].

The probability that a document is in class c can be calculated using the posterior probability formula as follows:

$$P(c_j|w_i) = P(c_j) \times P(w_1|c_j) \times \ldots \times P(w_i|c_j) \quad (1)$$

Information:

a. $P(c_j|w_i)$ = Posterior probability is the chance of the results of the prior and likelihood appearing to determine the category of a class (document).
b. $P(c_j)$ = Prior probability is the chance of each class appearing.
c. $P(w_i|c_j)$ = Conditional probability (likelihood) is the probability of words in a certain class (document).

2.3. K-Nearest Neighbor

k-NN is an algorithm for classifying new objects based on attributes and training samples (training data). Where the results of the new test samples are classified based on the majority of the categories in k-NN. The k-NN algorithm uses neighborhood classification as a prediction value for new test samples [21]. Training data will be built by paying attention to the balance of documents with each other. The k-NN algorithm can be explained with the following information [22]:

a. Calculate the distance between the sample data (test data) and the training data that has been built. One of the equations for calculating closeness distance can use the Cosine Similarity equation.
b. Determine the parameter value k = number of nearest neighbors.
c. Sort the smallest distance from sample data
d. Match categories according to suitability
e. Find the largest number of nearest neighbors. Then assign categories.

The distance used in this research is cosine similarity:

$$cos(i,k) = \frac{\sum k(di.dk)}{\sqrt{\sum k\ d_{ik}^2}\sqrt{\sum k d_{jk}^2}} \quad (2)$$

Information:

$\sum k\ (d1\ dk)$ = vector product of i and k

$\sqrt{\sum k\ d_{ik}^2}$ = vector length i

$\sqrt{\sum k\ d_{jk}^2}$ = vector length k

2.4. Simulation Method

Simulation is a methodology for carrying out experiments using a model of a real system. There are various types of lifecycles according to Zhang *et al.* (2014) which can be used for modeling and simulation studies [23]. There are basic steps that must be considered when conducting a simulation study. Lifecycle does not have to be interpreted as a strict sequence, it is iterative, and sometimes also transitions in the opposite direction [24, 25].

3. RESEARCH METHODS

In this research, the author collected data and information that can support the research process the data collection process is as follows.

3.1. Literature Review

A literature study was carried out by collecting theories related to the writing of this article as material to complete this research. Sources of theory come from reference books, research results (journals and theses), and related articles. Apart from that, researchers also visited sites related to natural language processing applications, text mining, lexicon approaches, and classification algorithms regarding Naive Bayes and k-NN.

3.2. Crawling Data

Twitter data was obtained by crawling Twitter media using the Twitter API. The data taken is about netizen comments on one of the 2019 presidential election candidates, Joko Widodo, in the 2019 presidential election. Crawling data starts from February 8 2018 to February 24, 2018. The second

crawl started from June 12 to August 8, 2018. Data collection utilized features from Twitter intended for developers on the website https://developers.twitter.com/.

4. RESULTS AND DISCUSSIONS

Several previous studies obtained information that sentiment analysis can be carried out by comparing the Naive Bayes and k-NN algorithms as classification [2-6] [10, 11, 15-17]. The Lexicon method is used to make it easier to label training data of 900 data and test data of 100 data.

4.1. Simulation Phase

At this simulation stage, several simulation experiments were carried out related to testing the accuracy level of the Naive Bayes and k-NN algorithms. The factors in the simulation process can be seen in Table 1 below.

Table 1. Simulation stages.

Simulation Variables/Parameters	Simulation Stage
Factor 1	Training data sentiment classification stage using the Lexicon approach
Factor 2	Data training stage on training data
Factor 3	Test data testing stage with the Naive Bayes algorithm and k-NN algorithm based on the k value in the k-NN algorithm is 1, 3, 5, 9, and 10
Factor 4	The accuracy testing stage uses the confusion matrix model

4.2. Verification, Validation and Experimentation

In the validation process, the correctness of the system is tested, namely by comparing the classification results of the Naive Bayes and k-NN algorithms using the lexicon approach which is calculated manually with the results of the sentiment orientation analysis application, thereby producing system accuracy.

After the experiment was carried out six times, the results of the accuracy performance analysis of the two algorithms can be seen in Figure 1 below.

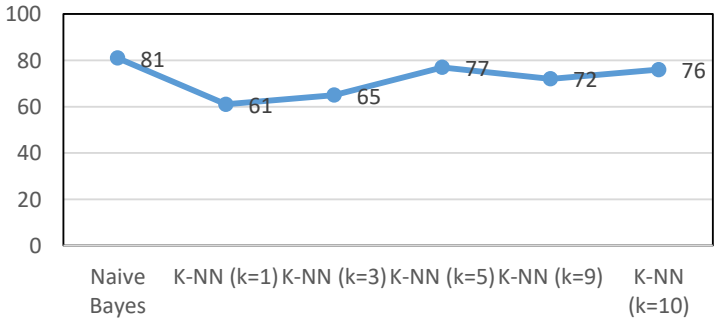

Figure 1. Accuracy performance analysis results.

The performance of Naive Bayes and k-NN is then compared with the best k value, to find out the difference, namely,

a. Results of the accuracy level of the Naive Bayes algorithm

Table 2. Table of Naive Bayes classification results.

Sentiment		Prediction result class		
		Positive	Negative	Neutral
Actual Class	Positive	a = 31	b = 1	c = 4
	Negative	d = 5	e = 18	f = 3
	Neutral	g = 2	h = 4	i = 32

Based on the test results from Table 2, the following accuracy values can be taken:

$$accuracy = \frac{31 + 18 + 32}{31 + 1 + 4 + 5 + 18 + 3 + 2 + 4 + 32} \times 100\% = 81\%$$

b. Results of the accuracy level of the k-NN algorithm with a value of k = 5.

Table 3. Results of k-NN classification with a value of k = 5.

Sentiment		Prediction result class		
		Positive	Negative	Neutral
Actual Class	Positive	a=32	b=3	c=1
	Negative	d=5	e=14	f=7
	Neutral	g=3	h=4	i=31

Based on the test results from Table 3 above, the following accuracy values can be taken:

$$accuracy = \frac{32 + 14 + 31}{32 + 3 + 1 + 5 + 14 + 7 + 3 + 4 + 31} \times 100\% = 77\%$$

5. CONCLUSION

The combination of a supervised learning classification algorithm with a lexicon approach can be applied to word sentiment analysis. The optimal k value in carrying out the k-NN algorithm classification process reaches an accuracy level of k = 5 with an accuracy level of 77% using a dataset from crawling Twitter data with the keywords "Pilpres2019" and "Jokowi". When comparing the Naive Bayes algorithm with k-NN, the difference in accuracy level reaches 4%. Based on tests carried out by researchers, the combination of the Naive Bayes and k-NN algorithms with the lexicon approach has been proven to increase accuracy in classifying sentiment. Accuracy at the k value in the k-NN algorithm experiences an increasing pattern before reaching the best k value and then decreases in accuracy level after the best k value. By using a small amount of training data, naive Bayes has a slight advantage over k-NN which may cause the accuracy of naive Bayes to be slightly above k-NN.

REFERENCES

[1] Patel, D., & Jermacane, D. (2015). Social media in travel medicine: A review. *Travel Medicine and Infectious Disease*, **13**(2), 135–142.

[2] Gezgin, D. M., Hamutoglu, N. B., Gemikonakli, O., & Raman, İ. (2017). Social networks users: Fear of missing out in preservice teachers. *Online Submission*, **8**(17), 156–168.

[3] Tsui, E. & Rao, R. C. (2019). Navigating social media in# ophthalmology. *Ophthalmology*, **126**(6), 779.

[4] Philander, K. & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, **55**, 16–24.

[5] Culotta, A. & Cutler, J. (2016). Mining brand perceptions from twitter social networks. *Marketing Science*, **35**(3), 343–362.

[6] Resnyansky, L. (2014). Social media, disaster studies, and human communication. *IEEE Technology and Society Magazine*, **33**(1), 54–65.

[7] Sangkhamanee, J. (2021). Bangkok precipitated: Cloudbursts, sentient urbanity, and emergent atmospheres. *East Asian Science, Technology and Society: An International Journal*, **15**(2), 153–172.

[8] Singh, N. J., & Bagchi, S. (2013). Applied ecology in India: scope of science and policy to meet contemporary environmental and socio-ecological challenges. *Journal of Applied Ecology*, **50**(1), 4–14.

[9] Buntoro, G. A. (2017). Analisis sentimen calon gubernur DKI Jakarta 2017 di Twitter. *Integer: Journal of Information Technology*, **2**(1).

[10] Hamdan, S. & Amiruddin, E. (2022). Schrödinger's equation as a Hamiltonian system. *Science, Technology and Communication Journal*, **3**(1), 19–24.
[11] Hamdan, S. & Amiruddin, E. (2022). Decomposition and estimation of gauge transformation for Chern-Simons-Antoniadis-Savvidy forms. *Science, Technology and Communication Journal*, **2**(3), 73–78.
[12] Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, **87**, 44–49.
[13] Indriani, A. (2014, June). Klasifikasi data forum dengan menggunakan metode naïve bayes classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*.
[14] Purnama, K. E. (2012). Classification of emotions in indonesian texts using K-NN method. *International Journal of Information and Electronics Engineering*, **2**(6), 899–903.
[15] Mohandes, M., Deriche, M., & Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, **6**, 19626–19639.
[16] Mustaqim, T., Umam, K., & Muslim, M. A. (2020). Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm. *Journal of Physics: Conference Series*, **1567**(3), 032024.
[17] Puspita, W., Defrianto, D., & Soerbakti, Y. (2021). Prediksi kadar particulate matter (PM10) menggunakan jaringan syaraf tiruan di Kota Pekanbaru. *Komunikasi Fisika Indonesia*, **18**(1), 1–4.
[18] Defrianto, D., Titrawani, T., Umar, L., & Asyana, V. (2022). Identifikasi hewan berdasarkan pola akustik dengan prinsip ekstraksi wavelet dan klasifikasi multi-label jaringan syaraf tiruan. *Indonesian Physics Communication*, **19**(1), 51–56.
[19] Yosnaningsih, Y. V. (2015). Klasifikasi dokumen bahasa Jawa menggunakan metode Naive Bayes. *Sanata Dharma University*.
[20] Krisandi, N. & Helmi, B. P. (2013). Algoritma K-Nearest Neighbor dalam klasifikasi data hasil produksi kelapa sawit pada PT. Minamas Kecamatan Parindu. *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, **2**(1).
[21] Oktinas, W. I. L. L. A. (2017). Analisis sentimen pada acara televisi menggunakan improved k-Nearest Neighbor. *Program Studi Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan*, **1**(2), 6–38.
[22] Madani, S. A., Kazmi, J., & Mahlknecht, S. (2010). Wireless sensor networks: modeling and simulation. *Discret. Event Simulations*, (2004), 1–16.
[23] Zhang, H., Gan, W., & Jiang, B. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. *2014 11th Web Information System and Application Conference*, 262–265.
[24] Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. *International Journal of Information Engineering and Electronic Business*, **8**, 54–62.
[25] Markard, J. (2020). The life cycle of technological innovation systems. *Technological Forecasting and Social Change*, **153**, 119407.