Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 1, No. 1, Mei 2023, pp. 39-46

K-Nearest Neighbor and Naive Bayes Classifier Methods for Expedition Service Comparison Analysis of User Sentiments

Nurul Hikmah [a,1], Muhammad Habibi [b,2,*]

[a,b]Departement of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia
[1] nurulhik2@gmail.com; [2] muhammadhabibi17@gmail.com*

* corresponding author

ABSTRACT

Depending on the service chosen, expedition is one of the freight forwarding companies that operate in the domestic market. The availability of expedition services can make it easier for traders to transfer items to purchasers who conduct online transactions, as well as encourage shipping businesses to collaborate with online dealers. The JNE, JNT, and Pos Indonesia excursions were utilized in this study. The goal of this project is to develop an analytical model that will make it simpler for online merchants to find collaborators for effectively and securely transporting their goods. Based on user sentiment on the social networking site Twitter, this study uses sentiment analysis. With the keywords "JNT, JNE, and Pos Indonesia," this study compares the accuracy results using the Naive Bayes Classifier (NBC) and K-Nearest Neighbor (KNN) algorithms. According to this study, testing accuracy for the NBC method was 80% and training accuracy was 83%. While the accuracy of the KNN approach is 68%. According to public opinion, the JNE expedition is the best one for distributing products, scoring 68.58% in favor of it and 30.64% against it.

ARTICLE INFO

Article history
Received: 2 Maret 2023
Revised: 5 April 2023
Accepted: 10 Mei 023

Keywords
Sentiment Analysis
K-Nearest Neighbor
Naïve Bayes Classifier
Expedition Service
Text Mining

1. Introduction

According to the Indonesia Digital Report for 2021, there are 170 persons in Indonesia who actively use social media. Because the rise of social media among Indonesians has made it easier to collect and share knowledge, social media has the potential to transform people's perspectives and become one of the media for digital literacy [1]. YouTube, WhatsApp, Instagram, Facebook, and Twitter are some popular social media networks. The level of public interest in the most recent and easily accessible news to obtain clear information via the Twitter network. Twitter has the most users when compared to other social media platforms. According to katadata, Indonesian Twitter users are the fifth most numerous in the world, with 24 million users [2].

The Covid-19 pandemic hit the world, particularly Indonesia, in 2020. The pandemic forced individuals to stay at home instead of working or going to school. This pandemic's influence has also resulted in an increase in people's internet buying habits. According to SIRCLO and the Katadata Insight Center (KIC) research, 74.5% of consumers prefer online buying over offline

shopping during a pandemic [3]. As the community's e-commerce activity grows, freight forwarders play a crucial role in distributing items from sellers to purchasers.

We refer the delivery of goods or freight forwarding businesses to as expedition. Expedition provides freight forwarding services throughout the United States. Depending on the service selected, the consumer will receive the products sooner or later. The expedition offers services such as express delivery, normal delivery, and others. The availability of expedition services can make it easier for traders to supply items to online buyers and encourage shipping businesses to collaborate with online traders. The sentiment analysis method is used in this study to examine public perceptions about freight forwarder reviews. Sentiment analysis is a technique for gathering data from social media networks. The goal of sentiment analysis is to collect both positive and negative conflicts or opinions. Identification of user's social media text to analyze information that leads to negative, good, or neutral attitude [4].

Based on these issues, the author analyzes public opinions about courier services on Twitter, because shipping companies must provide the finest service in transporting products bought by receivers. The shipping firm can increase service quality and customer satisfaction by doing sentiment analysis on public perception regarding the expedition's quality. The Nave Bayes classifier (NBC) and K-Nearest Neighbors (KNN) were utilized in this work. The NBC approach is a straightforward probability method with a good degree of accuracy [5], [6]. The NBC approach is also commonly utilized in sentiment analysis, with one example being a sentiment analysis study for evaluating online learning site services [7]. The KNN approach is used to classify scientific articles based on abstracts [8], but it is also utilized in the learning system to classify student comments [9].

2. Method

2.1. Data Collecting

Figure 1 displays the steps of research that were used. For this study, we used data from the Twitter streaming API between April 28, and May 24, 2022. Throughout that time, we were able to collect 13.811 tweets concerning Expedition Service in Indonesia by using the keywords JNT, JNE, and Pos Indonesia. The Twitter streaming API to retrieve the entire set of tweet properties provides an interface.

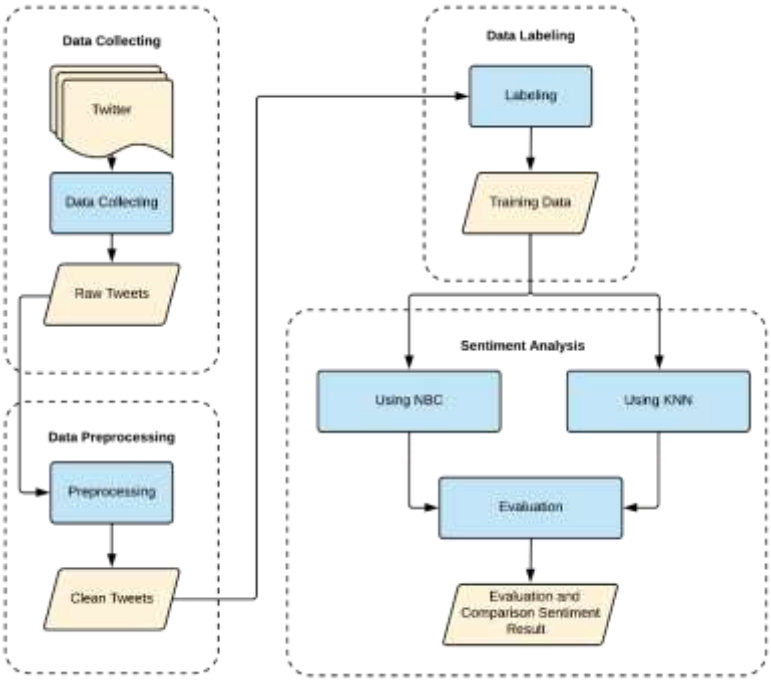

Fig. 1. Research Stages

2.2. Data Preprocessing

Following data acquisition, follows data preparation. Tweets with a lot of noise, typos, and bizarre sentences, as well as a range of acronyms and slang terminology, are common in raw form. These words typically contradict the generated tweet sentiment and degrade the efficiency of the categorization model. As a result, features extracted from tweets must first be pre-processed. On the tweet data that will be processed, we do the following preparation activities [10]: (1) Tokenizing, dividing, or separating text characters, whether such letters are regarded as word separators. (2) Cleaning, remove unnecessary features such as hashtags, numerals, emoji, and urls from received documents. (3) Case folding, which converts the original form (uppercase) to the usual form (lowercase). (4) Remove the stopword Texts should be rid of words that provide minimal information. (5) Stemming is the process of reducing all words to their most basic forms. Normalization is the process of converting shortened words into full words.

2.3. Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is a field of study that investigates how people feel about various aspects of products, organizational services, people, themes, events, situations, and traits [11]. Sentiment analysis, often referred to as subjective analysis, opinion generation, judgment extraction, and other names, is connected to emotional computing in several ways, including computer recognition and emotional expression [12].

2.4. Term Frequency-inverse document frequency

In this study, the Term Frequency-Inverse Document Frequency (TF-IDF) characteristic was used. In text categorization, the TF-IDF measure is commonly employed [13]. The TF-IDF approach is a well-known statistical technique for detecting the significance of a term in a corpus document [14]. The TF-IDF weighting system provides weight to term t in document d [15], as indicated in equation (1).

$$tf.idf_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

The value of $tf_{t,d}$ is the weight of a term t in document d, while idf_t is the inverse document frequency of term t. Equation (2) is an equation for finding the value of idf_t. The value of idf_t is obtained from the result of the logarithm of N divided by df_t. is the total number of documents where df_t is the number of documents containing term t.

$$idf_t = \log \frac{N}{df_t} \quad (2)$$

2.5. Naïve Bayes Classifier (NBC)

The Nave Bayes Classifier is a classifier based on the Bayes theorem. This classifier assumes that a feature's presence in a class is unrelated to other features. [16]. Equation (3) is the Bayes theorem equation.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (3)$$

Where P(X|Y) is the probability of occurring X if it is known Y. P(Y|X) is the chance of occurring Y if it is known X. P(X) is the probability of occurring X and P(Y) is the probability of occurring Y.

2.6. K-Nearest Neighbor

K-Nearest Neighbor is a supervised learning decision-making method in which it categorised the results of new input data based on the nearest in the value data [17]. The KNN approach employs a supervised algorithm from the instance-based learning group as well as a lazy learning strategy, searching for groups of k objects in the training data that are closest (similar) to the objects in the new data or testing data [18]. By giving query points, KNN will discover several k objects or (training points) closest to the query point [19]. There are numerous methods for calculating the distance between the closeness of the new data and the old data (training data), but the method used in this study is cosine similarity, which is a method for calculating the similarity between two objects expressed in two vectors with the keywords of a document as the size.

Cosine Similarity is frequently used to determine the similarity of text documents. The dot product is used to calculate the cosine similarity. The dot product is a straightforward calculation for each vector component. A vector is a representation of each document, with the number of terms in each document serving as the vector's dimensions. Cosine Similarity can be viewed as a document comparison because it considers not just the magnitude of each word count (weight) of each text, but also the angle between papers. The Cosine Similarity approach is denoted by Equations (4) and (5), where $|| \bar{a} ||$ is the Euclidean norm of vector a and $|| \bar{b} ||$ is the Euclidean norm of vector b [4].

$$\bar{a} \cdot \bar{b} = \left|\left|\bar{a}\right|\right| \left|\left|\bar{b}\right|\right| \cos\theta \quad (4)$$

$$\cos\theta = \frac{\bar{a} \cdot \bar{b}}{\left|\left|\bar{a}\right|\right| \left|\left|\bar{b}\right|\right|} \quad (5)$$

From equation (5) a mathematical equation can be formed which is shown by equation (6) [20]

$$Cos(a,b) = \frac{\sum_{i=1}^{n} x_{ai} y_{bi}}{\sqrt{\sum_{i=1}^{n} x_{ai}^2 \cdot \sum_{i=1}^{n} y_{bi}^2}} \quad (6)$$

Where:

x_{ai}: term i contained in document a

y_{bi}: the ith term contained in document b

2.7. Evaluation

We evaluate the built classifier model using the k-fold cross-validation method. This method separates the data into k equal-sized chunks. During the procedure, we select one division for testing and the rest for instruction. This technique uses one divisionque and is repeated k times to ensure that it use exactly once each partition for the test [21]. The total number of errors is obtained by adding the errors from all k processes. A Confusion matrix is also used to evaluate the accuracy of the developed classifier model. A confusion matrix is a useful tool in machine learning visualization, which typically comprises two or more categories. [22].

3. Results and Discussion

3.1. Naïve Bayes Classifier Evaluation Results

Before beginning the testing phase, the model used for classification must be evaluated using training data from 1120 tweets, each of which contains 560 positive and 560 negative data tagged as human. We conducted ten trials to determine the classification model's accuracy. Each computation has a distinct level of precision. Table 1 shows the results of the 10-fold calculation in cross validation.

Table 1. K-Fold Cross Validation

Fold	Accuracy	F1-Score
Fold 1	83.9%	83.93%
Fold 2	80.8%	80.75%
Fold 3	84.4%	84.32%
Fold 4	79.5%	79.46%
Fold 5	76.8%	76.79%
Fold 6	78.1%	78.09%
Fold 7	77.2%	77.04%
Fold 8	80.8%	80.78%

Fold	Accuracy	F1-Score
Fold 9	77.7%	77.54%
Fold 10	81.7%	81.7%

According to Table 1, the best accuracy was obtained at fold 4 with an accuracy of 84.4% and the lowest accuracy was obtained at fold 7 with an accuracy of 77.2%. The average accuracy gained is 80.09%, with an F-Score of 80.04%. After determining the classification model's accuracy value, conduct data testing on up to 400 tweets, we have designated 200 of which as positive and 200 as negative. The data utilized in testing is distinct from the data used in training. Table 2 displays the testing accuracy findings.

Table 2. NBC Accuracy Result

Information	Score
Acuraccy	80%
Precission	87 %
Re Call	76 %
F1 Score	81 %

The resulting accuracy value is 80%, the precision value is 87%, the recall value is 76%, and the F1 Score value is 81%, as shown in the table.

3.2. K-Nearest Neighbor Evaluation Results

The KNN was trained with 1120 tweets, whereas the testing data consisted of 400 tweets, 200 of which were positive and 200 of which were negative. The KNN algorithm test does not use training data to construct a model, but it does use the data to determine how far apart the testing and training data are. Table 3 shows the results of the testing method in terms of accuracy, precision, recall, and F1 score.

Table 3. KNN Accuracy Result

Information	Score
Acuraccy	68%
Precission	78%
Re Call	65%
F1 Score	71%

Table 3 demonstrates the KNN technique calculation utilizing cosine similarity, which yielded pretty good results, with an accuracy of 68%, recall of 65%, and F1 Score of 71%.

3.3. Comparison of NBC and KNN Evaluation Results

The accuracy levels produced by the NBC and KNN systems are considerably different. The following Table 4 shows the difference in accuracy results:

Table 4. Confusion Matrix Result

	Method	
	NBC	KNN
Akurasi	80%	68%

The NBC and KNN accuracy comparison numbers are obtained by comparing the Testing process results using testing data. The NBC and KNN Testing Methods make use of data from 400 tweets, 200 of which were categorized as good and 200 as negative. Even though the amount of NBC and KNN testing data is the same, the comparison accuracy with the two NBC and KNN methodologies is different. The NBC approach is more accurate than the KNN method, with an accuracy of 80% versus 68% for the KNN method.

3.4. Sentiment Analysis Results

The process of extracting labels from data is known as classification. The objective is to make decisions and forecast a situation. Based on the classification results obtained in this study utilizing the NBC approach. The data used is a collection of 13,812 tweets about JNT, JNE, and Pos Indonesia. According to public opinion on Twitter, we conducted the classification in this study out to discover the ideal expedition to distribute goods properly and safely. Figure 2 depicts the graph of the results of the tweet classification that addresses JNT. Figure 2 illustrates that there are more negative sentiments than positive sentiments, with 3,192 tweets for negative sentiments and 1,808 tweets for good sentiments. Negative opinion about JNT focuses on the evaluation of JNT adventures, which frequently encounter delays in sending items and a lack of responsibility from the expedition, resulting in products being lost or destroyed. Meanwhile, favorable mood focused on electronic products security and the faith of some customers who had utilized JNT's expedition services.

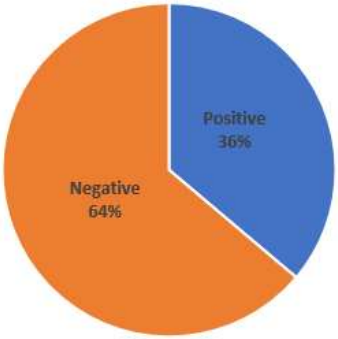

Fig. 2. JNT Sentiment Analysis Results

Aside from the JNT voyage, this study also covers the JNE expedition, with a total of 5,000 tweets. Figure 3 depicts a graph showing the JNE Expedition Classification results. According to Figure 3, positive sentiment outnumbered negative sentiment, with 3,429 positive sentiment tweets and 1,532 negative sentiment tweets. Negative feedback about JNE trips focuses on exorbitant pricing, lengthy delivery times, and delays in updating receipts. While positive emotions focus on security during the shipping process, the JNE admin responds fast to complaints from JNE users.

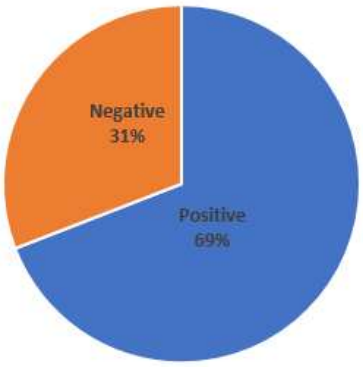

Fig. 3. JNE Sentiment Analysis Results

The Pos Indonesia voyage was the most recent expedition used in this investigation. The amount of data used was 3,812 tweets, and Figure 4 shows a graph of the Pos Indonesia expedition categorization findings. According to Figure 4, good emotion outnumbers negative sentiment. It expressed positive sentiment towards Pos Indonesia in 2,853 tweets, while negative attitude was expressed in 960 tweets. Positive attitude concerns the lack of clarity in the delivery payment mechanism and the excessively long delivery of items. While positive attitude highlights the simple process of delivering overseas, the expedition organized by this BUMN ensures the safe delivery of both items and documents.

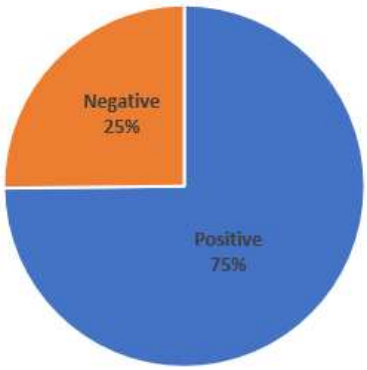

Fig. 4. Pos Indonesia Sentiment Analysis Results

4. Conclusion

We can take the following conclusions from the conversation and results of sentiment analysis on the use of courier services on Twitter social media: (1) The NBC approach has an accuracy of 80%, whereas the KNN method has an accuracy of 68%. As a result, the NBC approach outperforms the KNN method in terms of accuracy. (2) This work was successful in developing a freight service classification model with positive and negative labels. According to the classification findings, JNE expedition provided the finest expedition service based on public opinion on Twitter social media.

References

[1] M. A. Harahap and S. Adeni, "TREN PENGGUNAAN MEDIA SOSIAL SELAMA PANDEMI DI INDONESIA," *Prof. J. Komun. dan Adm. Publik*, vol. 7, no. 2, pp. 13–23, Dec. 2020.

[2] C. M. Annur, "Pengguna Twitter di Indonesia Capai 24 Juta hingga Awal 2023, Peringkat Berapa di Dunia?," *Katadata Media Networks*, 27-Feb-2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/02/27/pengguna-twitter-di-indonesia-capai-24-juta-hingga-awal-2023-peringkat-berapa-di-dunia. [Accessed: 08-Mar-2023].

[3] G. Nurcahyadi, "Riset : 74,5 % Konsumen Lebih Banyak Berbelanja Online Daripada Offline ," *Media Indonesia*, 22-Oct-2021. [Online]. Available: https://mediaindonesia.com/ekonomi/441793/riset-745-konsumen-lebih-banyak-berbelanja-online-daripada-offline. [Accessed: 27-Apr-2023].

[4] M. Habibi and E. Winarko, "Klasifikasi Komentar Mahasiswa Menggunakan Kombinasi KNN berbasis Cosine Similarity dan Supervised Model," no. x, pp. 1–11, 2017.

[5] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 5, no. 2, pp. 697–711, Sep. 2021.

[6] S. Rahmawati and M. Habibi, "Public Sentiments Analysis about Indonesian Social Insurance Administration Organization on Twitter," *IJID (International J. Informatics Dev.*, vol. 9, no. 2, pp. 87–93, Dec. 2020.

[7] M. Habibi, A. Priadana, and M. R. Ma'arif, "Sentiment Analysis and Topic Modeling of Indonesian Public Conversation about COVID-19 Epidemics on Twitter," *IJID (International J. Informatics*

Dev., vol. 10, no. 1, pp. 23–30, Jun. 2021.

[8] M. Habibi and P. W. Cahyo, "Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine," 2020.

[9] M. Habibi and S. Sumarsono, "Implementation of Cosine Similarity in an automatic classifier for comments Program," 2018.

[10] M. Habibi, A. Priadana, A. B. Saputra, and P. W. Cahyo, "Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA)," pp. 260–264, Jan. 2021.

[11] B. Liu, "Sentiment Analysis and Opinion Mining," *http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016* , vol. 5, no. 1, pp. 1–184, May 2012.

[12] M. Habibi, A. Priadana, and M. R. Ma'arif, "Hashtag Analysis of Indonesian COVID-19 Tweets Using Social Network Analysis," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, Jul. 2021.

[13] N. N. Amir Sjarif, N. F. Mohd Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS spam message detection using term frequency-inverse document frequency and random forest algorithm," in *Procedia Computer Science*, 2019, vol. 161, pp. 509–515.

[14] A. Thakkar and K. Chaudhari, "Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks," *Appl. Soft Comput. J.*, vol. 96, p. 106684, Nov. 2020.

[15] A. A. Putri Ratna, A. Kaltsum, L. Santiar, H. Khairunissa, I. Ibrahim, and P. D. Purnamasari, "Term Frequency-Inverse Document Frequency Answer Categorization with Support Vector Machine on Automatic Short Essay Grading System with Latent Semantic Analysis for Japanese Language," in *ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science, Proceeding*, 2019, pp. 293–298.

[16] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy, and G. R. Mohan Reddy, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, 2020, pp. 603–607.

[17] Z. Mundargi, S. Mulay, D. Navale, V. Talnikar, A. Nawale, and V. Sonkusale, "An Aviation Industry Recommender System(AIRS) using K-nearest Neighbour and Cosine Similarity," *2023 Int. Conf. Adv. Technol.*, pp. 1–6, Jan. 2023.

[18] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-January, pp. 294–298, Feb. 2018.

[19] Y. Sari, M. Maulida, E. Gunawan, and J. Wahyudi, "Artificial Intelligence Approach for BAZNAS Website Using K-Nearest Neighbor (KNN)," *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, 2021.

[20] M. Habibi and P. W. Cahyo, "Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 4, no. 3, pp. 48–55, 2020.

[21] M. Habibi and E. Winarko, "Analisis Sentimen dan Klasifikasi Komentar Mahasiswa pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi KNN Berbasis Cosine Similarity dan Supervised Model," Universitas Gadjah Mada, 2017.

[22] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.