The application of \( k \)-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality

R Damarta, A Hidayat, and A S Abdullah*  
Department of Computer Science, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia  
*atje.setiawan@unpad.ac.id

Abstract. Social media has an important role in human life. In its implementation social media is used as a media for opinion and self-expression. One of the social media that is often used in Indonesia is Twitter. PT PLN (Persero) as a State-Owned Enterprise that is engaged in providing electricity always tries to provide optimal services. The text mining method can be used to control PT PLN (Persero) service quality by classifying Twitter data with the \( k \)-Nearest Neighbors algorithm. Text mining is used to extract information from unstructured textual data to produce useful information. Data classification is a text mining application for information retrieval. In this study the data collected will pass the preprocessing stage, using the \( k \)-Nearest Neighbors algorithm to classify data into negative, neutral, or positive classes. The data used in this study was sourced from Twitter. Data is taken from 1 December 2019 to 1 February 2020 using the Twitter API with the keyword ‘@ pln_123’. We obtained 3,000 tweet data successfully. The results are implemented in web-based applications that are built using the Python programming language. Evaluation of the \( k \)-Nearest Neighbors model produces an accuracy value of 87.41%. The classification prediction results also show that there is a tendency of the positive sentiment of 35%, the neutral sentiment of 28%, and the negative sentiment of 37%.

1. Introduction  
In the current era of globalization, social media is very commonly used and has become a daily need. In its implementation, social media is widely used to convey information and as a medium for self-expression [1]. Social media Twitter is ranked 6th as the most frequently used social media in Indonesia [2]. Various information about daily life is conveyed via Twitter, including information about the needs of life. One example of primary needs is electricity needs. Related to this, the task of PT PLN (Persero) as the only State-Owned Enterprise that is engaged in providing electricity for a better life in Indonesia.

PT PLN (Persero) can use the information on Twitter to be able to find out what consumers think about the electricity services and provide the best service to the community to meet these primary needs. To be able to analyze sentiment on the service quality of the official PT PLN (Persero) Twitter account, we can use sentiment analysis to understand sentiment information related to a particular topic that involves grouping comments or opinions in a text into categories. Sentiment analysis in marketing terminology is also called customer voice. Sentiment analysis can be a very useful tool for checking affinity for a brand, product, or domain [3].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
This is necessary because based on one of the missions of PT PLN (Persero) in carrying out the electricity business and other related fields, it is oriented towards customer satisfaction, company members, and shareholders [4].

Customer satisfaction is the keyword and main orientation in carrying out PT PLN’s duties as the only provider of electricity in Indonesia. This study will use the classification method with the k-Nearest Neighbors algorithm which is already popular for pattern recognition because of its effective performance [5].

Based on the introduction, it is necessary to conduct a sentiment analysis of the tweet data using the classification method with the k-Nearest Neighbors algorithm, with the aim of being able to find out public opinion about the PT PLN (Persero) electricity services.

2. Method

![Figure 1. Research procedure](image)

Figure 1 is the stage of the research procedure which consists of the stages of data collection, data labelling, data preprocessing, Term Frequency (TF)-Inverse Document Frequency (IDF) classification, and k-Nearest neighbors classification. The data to be collected is tweet data with the keyword "@pln_123" from Twitter. The data collection stage was carried out by the crawling process using the Tweepy library. Data collection was carried out within a period of two months from 1 December 2019 to 1 February 2020.

The data that has been collected will then go through the data labelling stage. The data labelling stage is the process of marking data with meaningful and informative tags. The purpose of this stage is to provide a learning basis for further data processing. In this study, three labels were used, namely negative, neutral, and positive labels. Negative, neutral, and positive labels are converted to the numbers 0, 1, and 2. This is done using the label encoder. The process of labelling data at this stage is done manually by the author. The author determines the label of each data that has been obtained based on the words in each tweet data. Data that has gone through the data labelling stage will then go through the data preprocessing stage.

The preprocessing stage is carried out with the aim of obtaining more structured data. The preprocessing process in this study consists of four processes. Noise removal, case folding, stopword removal, and tokenization. The noise removal process is used to remove numbers, special characters, and URLs in tweets that can interfere with the tweet analysis process [6]. In the case folding process all the letters in the tweet will be converted to lowercase. The stopword deletion process is the process of removing words that are considered to have no effect on the classification process because they have no sentimental value. The tokenization process will be carried out by separating sentences into words. The next data will go through the data weighting stage.

The data weighting stage is carried out because the k-Nearest Neighbors algorithm cannot accept data in text form, so it needs to be changed to numeric form. One way to change text form to numeric form is to use TF-IDF weighting. Tweet data that has gone through the preprocessing stage will be converted into numeric data using TF-IDF weighting. The TF-IDF weighting was chosen because TF-IDF can reflect how relevant a term is in a document [7]. The document used here is a tweet text document from Twitter which can contain various text data according to the author's wishes. Due to this, a weighting is needed that can reflect how relevant a term is in a document.

TF-IDF is a weighting scheme used to determine how far a term is associated with a document by giving weight to each word. TF is the frequency at which a term appears in a document. The more
often a term appears in a document, the greater the weight of the document for that term, and vice versa. IDF is the Inverse document frequency which was created to reduce the effect of too high a frequency word in the document. The more words in a document, the lower the weight. TF-IDF can be calculated using the following equation [8]

\[ idf_t = \log \frac{N}{df_t} \]  

where:
\( N \) : number of documents
\( df_t \) : number of documents containing the term for which the IDF value is sought.

\[ tf.idf_{t,d} = tf_{t,d} \times idf_t \]  

where:
\( tf.idf_{t,d} \) : the weight of the word \( t \) against the document \( d \)
\( tf_{t,d} \) : the number of occurrences of the term searched for in the document.

The most common theme in analyzing complex data is classification, also known as data categorization. Data classification aims to classify data sets by dividing data into various predefined categories, the process of finding the correct category (or topic) for each document becomes the main task of text classification [9]. The classification stage uses the k-Nearest Neighbors algorithm to classify data into three classes or labels, namely negative, neutral, and positive. The k-Nearest Neighbors algorithm is used as a classification algorithm because it is easy to implement and understand at each stage. The k-Nearest Neighbors algorithm is an instance-based learning and memory-based approach. KNN classifiers can adapt immediately when we collect new training data. This makes the algorithm respond quickly to any input changes that occur during real-time use [10].

The data that has been collected will be divided into two, namely training data and test data. The k-Nearest Neighbors algorithm is an algorithm that can be used to classify objects based on the closest distance to objects in the training data set. The process will continue until all objects have a class. The k-Nearest Neighbors algorithm uses the closest calculation as the predictive value of the new instance value [9]. The k-Nearest Neighbor algorithm works by calculating the distance between the test object and the object in the training object set. In general, the Euclidean distance measurement theory is used as a distance matrix to calculate the distance between objects. The Euclidean distance can be calculated using the following equation [11].

\[ d_{ij} = \sqrt{\sum_{n=1}^{m}(x_{in} - x_{jn})^2} \]  

where:
\( d_{ij} \) : the Euclidean distance of data object and data object \( j \)
\( m \) : the number of dimensions of the total object
\( x_{in} \) : data object \( i \) of dimension \( n \)
\( x_{jn} \) : data object \( j \) in dimension \( n \).

The working path of the k-Nearest Neighbors algorithm is as follows [12]
a. Determine the value for \( k \) to be the number of closest neighbors.
b. Calculate the Euclidean distance between the test object and all training data objects.
c. Sort the Euclidean distance between the test object and the training object based on the shortest to the farthest distance.
d. Set the training data objects belonging to the \( k \) (nearest neighbors) based on the sorted object group.
e. Get a class of specimens based on the majority of your closest neighbors.
f. Repeat steps 4 through 5 until all test objects are classified.

To clarify the data classification stage using the k-Nearest Neighbors classification algorithm, a figure will be used.
As shown in Figure 2 it is known that the triangle shape represents one class, let's say that the class is negative. The star shape represents one other class, let's say that the class is positive. The box shape represents an object whose class is not yet known. The position of each object on the figure is determined based on the weight of the TF-IDF result. In the example image, a k value of 5 will be used, then the five closest objects to the box-shaped object will be used to represent objects whose class is unknown. To find out the five closest objects, we need the Euclidean distance value from the rectangular object that represents an object whose class is not yet known. It can be seen in Figure 2 that the triangle shaped object is the majority of the objects. By using the k-Nearest Neighbors classification algorithm, the box-shaped object will become a positive class object.

The determination of the value of k in the k-Nearest Neighbors classification has no standard rules. In general, a k-value that is too small can result in inconsistent results and a large k-value has a more consistent decision threshold which means lower variance of results but increased bias. [10]. In this study, the value of k that will be used for classification is determined by comparing the error rate value of k = 1 with the value of k = 40. This number was chosen by calculation if the value of k is too small it can result in inconsistent results and a large k value. has a more consistent decision threshold which implies lower variance of results but increased bias. The k value with the lowest MPE (Mean Percentage Error) value can be said to be the best value.

Figure 3 shows that the value of k = 12 produces the lowest error rate with a value of 0.125. Then the model created will use 12 as the k value. The k-Nearest Neighbors model was created using training data as input and labelling as an objective. The k-Nearest Neighbors model that has been created will be tested using test data. The prediction results of the test data will be displayed in a confusion matrix graph and the calculation of accuracy, precision, recall, and f-score as an evaluation of the performance of the k-Nearest Neighbors model.
Accuracy is the ratio of the correct prediction of every class to the overall data. Precision is the ratio of the correct prediction of each existing class compared to the overall predicted results of each existing class. Recall is the ratio of correct prediction of each class compared to the total correct data in each class. F Score is a weighted average comparison of precision and recall.

3. Result and Discussion
The data used is tweet data from Twitter taken from December 2019 to February 2020 using the keyword "@ pln_123" which is the official Twitter account of PT PLN (Persero). The data collected is 3,000 tweets. There are 3,000 tweet data, 1,230 data are labeled negative, 850 data are labeled neutral, and 920 data are labeled positive.

The k-Nearest Neighbors model that is made will classify the data into three classes, namely positive, negative, and neutral. We got 3,000 tweet data that has been collected and it will be divided into training data and test data using train_test_split from the Sklearn library with a ratio of 60 to 40 in order to obtain 1800 training data and 1200 test data.

Table 1. The results of the confusion matrix from the 1200 test data were used

| Actual Class | Predicted Class | | |
| | Negative | Neutral | Positive |
| Negative | 409 | 32 | 51 |
| Neutral | 29 | 292 | 19 |
| Positive | 12 | 8 | 348 |

Table 1 is the result of the confusion matrix which shows how many successful and failed data are classified according to class. By using a comparison of training data and test data of 60 to 40. The table shows that from 1,200 test data, there are 492 negative data, 340 neutral data, and 368 positive data. The k-Nearest Neighbor algorithm successfully predicted 409 negative data, 292 neutral data, and 348 positive data correctly.

Table 2. The results of the accuracy, recall, f-score of each label, and the accuracy value of the k-Nearest Neighbor classification

| Sentiment | Precision | Recall | F-score |
|-----------|-----------|--------|---------|
| Negative  | 90,88%    | 83,13% | 86,83% |
| Neutral   | 87,95%    | 85,88% | 86,90% |
| Positive  | 83,25%    | 94,56% | 88,54% |
| Accuracy  | 87,41%    |        |         |

Table 2 is a table that contains the presentation of the precision, recall, and F-score for each class. Based on the classification results using the k-Nearest Neighbors algorithm, it can be seen that the accuracy of the model in the test data is 87.41% with the number of correct classifications of 1,068 tweet data. It can also be seen that the highest average recall and precision score is in the positive class with a value of 88.54%, while the negative class has the lowest score with a value of 86.83%. These results indicate that the model can predict more accurately the positive rather than negative and neutral categories.

The results of the model that have been created are used to predict which files contain new tweet data that can be uploaded by users.
Figure 4. Comparison between error rate and k value

Figure 4 shows the contents of each tweet in the uploaded file and the class prediction results of each tweet in the tweet file uploaded by the user.

4. Conclusion

Based on the classification results of k-Nearest Neighbors from tweet data about PT PLN (Persero), the following conclusions can be drawn. The process of analyzing tweet data to determine public opinion about the official Twitter account service of PT PLN (Persero) can be done by analyzing the tweet data sentiment using the k-Nearest Neighbors calcification algorithm. The k-Nearest Neighbors algorithm is used to view public opinion about the service from the official Twitter account of PT PLN (Persero). The k-Nearest Neighbors algorithm is used to determine the quality of public opinion on services from the PT PLN (Persero) Twitter account. By using the comparison of training data and test data from 60 to 40, the accuracy of the data tested was 87.42%. The prediction results of 1200 test data show that there are 492 negative data, 340 neutral data, and 368 positive data. These results indicate that the tweets that mention the Twitter account of PT PLN (Persero) have almost the same positive and negative tendencies.

The results of making the k-Nearest Neighbors model which are trained and tested using tweet data related to the service quality of PT PLN (Persero) can be used or tested using a new tweet data file. The k-Nearest Neighbors model will predict test data based on the previously used training data. This model can be improved further by adding training data with tidier data quality.

Acknowledgments

The authors thank Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Academic Leadership Grant 2020 with contract number: 1427/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

References

[1] Szabo G, Polatkan G, Boykin O, and Chalkiopoulos A 2019 Social Media Data Mining and Analytics (John Wiley & Sons.)

[2] Kemp S 2019 Indonesia Digital Report BOC Indonesia Bali
[3] Chatterjee S, and Krystyanczuk Z 2017 *Python social media analytics: analyze and visualize data from Twitter, YouTube, GitHub, and more* (Birmingham, UK: Packt Publishing Ltd.)

[4] Zaini Z 2020 Profile Perusahaan https://www.pln.co.id/tentang-kami/profil-perusahaan. [Accessed 11 January 2020]

[5] Hota S, and Pathak S 2018 KNN classifier based approach for multi-class sentiment analysis of twitter data *International Journal of Engineering & Technology* VII 3 1372-1375

[6] Uysal A, and Gurnal S 2014 The Impact Preprocessing on Text *Journal Information Processing and Management* L 1 104-112

[7] Herwijayanti B, Ratnawanti D, and Mufliah L 2018 Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* II 1 306-312

[8] Manning, Christopher, Prabhakar R, and Hinrich S 2008 *Introduction to Information Retriveal* (New York: Cambridge University Press)

[9] Feldman R and Sanger J 2007 *The Text Mining Handbook* (New York: Cambridge University Press)

[10] Imandoust S, and Bolandraftar M 2013 Application of K-nearest neighbor approach for predicting economic events theoretical background *International journal of engineering research and applications* III 5 605-610

[11] Huang H and Tsai K 2007 *The Text Mining Handbook* 9 (New York: Cambridge University Press)

[12] Gorunescu and Florin 2011 *Data Mining: Concepts, Models, and Techniques* (Verlag Berlin Heidelberg)