**MATICS** : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 12, No. 1 (2020), pp 90-93 DOI : 10.18860/mat.v12i1.8973
**Received** : January25th 2020; **Accepted** : March 9th 2020 ; **Avalaible Online** : April 5th 2020

# Sentiment Analysis of Perpustakaan Nasional Republik Indonesia Through Social Media Twitter

Fakhris Khusnu Reza Mahfud, Nita Siti Mudawamah, Wahyu Hariyanto

**civilization. Indonesia already has a Perpustakaan Nasional consisted of 27 floors and is equipped with facilities that are adequate for user needs. Apart from that, we need to see opinions from the community as users. Public opinion about the library is critical for library managers to evaluate services and facilities from the library. One way to find out the views of the community is by using social media twitter. Twitter social media is often used in channelling opinions or expressing opinions about specific topics; besides social media, twitter is commonly used for digital campaign movements. Submission of views and even digital campaigns on Twitter social media greatly influence the opinions and even behaviour of society in various ways. This study analyzes tweets about national libraries by classifying, positive opinions, negative opinions and neutral opinions. In this study, twitter data will go through the preprocessing, weighting, and classification stages. TF-IDF and TF binary are used in weighting in this study. The classification used in this study is Naive Bayes and KNN. Accuracy, precision, and recall values were also used in this study to evaluate classification performance. The highest classification performance using KNN classification with TF-IDF weighting resulted in the value of accuracy, precision, and recall of 83.33%, 79.2%, and 83.3% respectively.**

**Index Terms—Sentiment Analysis, Perpustakaan Nasional, Twitter, Classification**

## I. INTRODUCTION

Perpustakaan Nasional Republik Indonesia is a public facility owned by the Indonesian state provides several facilities and services including an art installation exhibition room, information center, bag storage lockers, borrowing hardcopy books or accessing books through online (e-books), rare book collections, public book collections, executive lounges, children's service rooms, elderly and people with disabilities, prayer rooms and canteens. There is also an audiovisual collection, photos, maps and paintings [1]. Perpustakaan Nasional is very concerned with the end satisfaction by improving facilities and services because the Perpustakaan Nasional is the mother of all libraries in Indonesia.

One way to find out the satisfaction of visitors to the facilities and services of the Perpustakaan Nasional is to do sentiment analysis on Twitter. Sentiment analysis known as opinion mining and subjective analysis is analyzing people's opinions, and attitudes form written reviews to rate products or services [2]. This sentiment analysis is used to find out whether the sentence is included in negative or positive opinion. From this sentiment analysis can be obtained whether the overall library facilities have met user satisfaction or not.

Twitter data is used in this study because there are quite many active Twitter users in Indonesia. In early 2017, Twitter has reached 330m per month of active users. In Indonesia, active Twitter users have reached 19m in 2017 [3]. Users prefer Twitter because of its simplicity in expressing thoughts. In previous studies, sentiment analysis data was used in the fields of e-government, e-commerce, education, agriculture.

Text classification is also applied to other fields as in the review of Thai restaurants [4], documents of medicinal plants and horticulture [5], twitter social network [6][7], news texts and academic abstracts [8], Facebook [9], a website “Lapor!” [10]. Previous research on E-government in Surabaya City Government is only about sentiment analysis on government performance [6][9].

The method used in this study is TF-IDF and TF binary for weighting. At the classification stage, the methods used are Naive Bayes and KNN.

This explanation consists of four parts; first introduction, second research method, third analysis and results, and fourth conclusion.

## II. RESEARCH METHOD

Some of the steps undertaken to conduct research on sentiment analysis are data preparation, data retrieval

Manuscript received March 2, 2020. This work was supported in part by Jurusan Perpusatakaan dan Ilmu Informasi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Fakhris Khusnu Reza Mahfud is with the Jurusan Perpusatakaan dan Ilmu Informasi Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (email fakhriskrm7@uin-malang.ac.id )

Nita Siti Mudawamah, was with Jurusan Perpusatakaan dan Ilmu Informasi Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail : nitastmudawamah@uin-malang.ac.id ).

Wahyu Hariyanto is the Jurusan Perpusatakaan dan Ilmu Informasi Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (email wahyuhariyanto@uin-malang.ac.id )

from Twitter, preprocessing, weighting, classification and evaluation of classification performance.

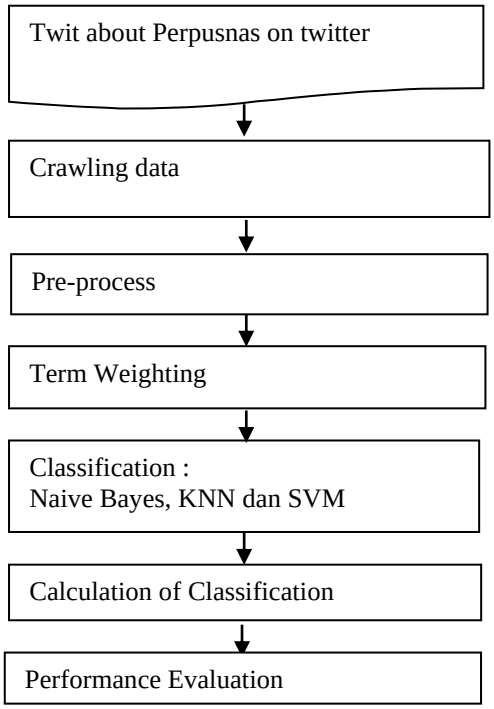


Figure 1 Research Method

In Figure 1 it can be seen that getting Twitter's fire is the first step of this research. The second step is crawling from Twitter about the Perpustakaan Nasional. The third step involves the four pre-processing stages. The fourth step is weighting with TF-IDF and TF-Binary. The fifth step is to classification with KNN and Naive Bayes. Evaluation of classification performance was carried out at the final stage of this study.

*A. Data Preparation*

**This study using data in the form of text obtained from Twitter. The tweet needed is about Perpustakaan Nasional that speak Indonesian. Data collected from January 2019 to August 2019. Collection using the software "R Studio" with the package and library "TwittER". The collected data is stored in CSV format.**

Table 1 Example of Tweet

| No | Tweet | Class |
|---|---|---|
| 1 | @dimahDonat malam....di perpusnas buku bukuny tidak bisa dibawa pulang, hanya bisa baca di tempat, karena sistem la… https://t.co/hnJv9gYXr6 | Negative |
| 2 | Kenapa perpusnas tidak buka 24 jammm | Negative |
| 3 | @clydealwyn @clowneryluvv Bukan bego. Lemot deh pas ke perpusnas | Negative |
| 4 | @evoIusi @hangyuliebe dngj kalo sejabodetabek mau ngajak main ke perpustakaan nasional ue ue | Neutral |
| 5 | Pernah nyari buku2 di senen kwitang atau lt bawah blokm square? Kyk gitu deh buku2nya. ☺ Duduk2 kerja ga cocok sih, ga senyaman perpusnas soalnya. https://t.co/EovS0zeaFn | Positive |

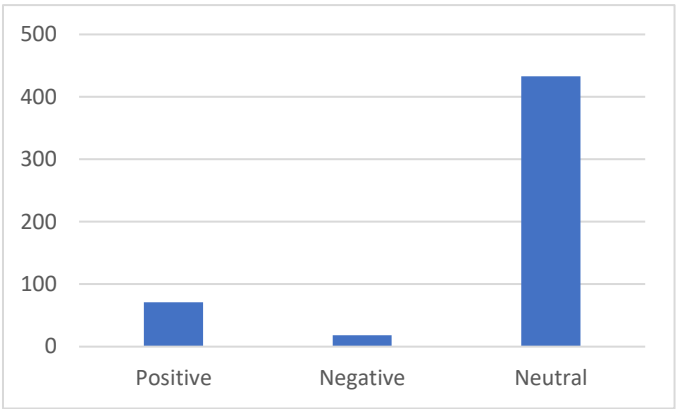


Figure 2 Graphical Image of a Tweet

The data obtained amounted to 522, consisting of 71 positive tweets, 18 negative tweets, and 433 neutral tweets. Positive data contains words that indicate good meaning, for example comfortable, good, delicious. Conversely, negative data contains bad meaning. Neutral data if the tweet contains words that do not have positive and negative meaning such as question words.

*B. Pre-process*

The pre-processing stage consists of four stages, case folding, tokenizing, filtering, and stemming. Data that has been done through preparation then carried out pre-preparation. The process of transforming data from unstructured data into structured data is called pre-processing. Structured data can be processed and analyzed by text mining.

.

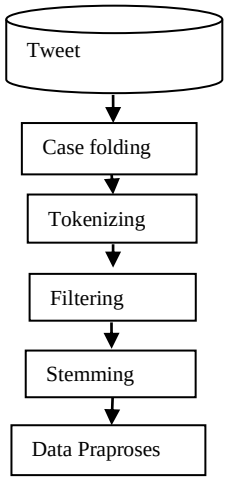


Figure 3 Phase of Data Pre-process

Case folding is a step to replace uppercase letters into lowercase letters and only involves the letters "a to z". Tokenizing is the stage of separating sentences into words. Filtering is used to eliminate stopwords or words that have no meaning. Perform stemming by changing

words that have an influence and begin to become basic words.

*C. Term Weighting*

TF-Binary and TF-IDF carried out at the weighting stage. By doing a weighting, the text containing the letters is transformed into a number that shows as weight.

*1) TF-IDF weighting*

Efficient, easy and accurate results are the advantages of TF-IDF [10]. TF-IDF is a method of weighting by combining local and global aspects of each term. The local aspect is calculated based on the term frequency (TF) of each document in the document collection multiplied by the global inverse document frequency (IDF) at each term [10].

$$tfidf(w) = tf \times log \frac{N}{df(w)} \quad (1)$$

The number of words that appear in a document called TF. Total documents containing words in the corpus are called Df (w). The total number of documents in a corpus is called N [11].

The term that often appears on each document has the lowest value because it cannot represent the document in the classification process, whereas the term that appears on a particular document has a high IDF value because it can represent the document in the classification process [11].

*2) TF Binary*

The TF-binary weighting method changes the words in the corpus to values 1 and 0. Value 1 if there are words in the document, and vice versa 0 if there are no words in the document.

*D. Data Classification*

This study uses Naive Bayes and KNN for the classification method. A simple classifier based on the Bayes theorem can be called the Naive Bayes. The assumptions of the Naive Bayes classification do not depend on other attribute values. This was done to simplify the calculation, therefore assuming the method is called naive. This method can be very efficient and accurate with simple designs and naive assumptions and a high number of variable compilations [10]. Another advantage of this algorithm is that it can work with little training data and easy implementation [12].

K-Nearest Neighbor (KNN) is an algorithm that works by comparing the distance of test data input with the closest number of training data [13]. Objects classified based on training attributes and samples are the objectives of this algorithm. The classification is given a query point. It will find several k objects or training points that are closest to the query point. The most votes among object classifications k are used in this classification. The adjacent classification is used as the predicted value of the new query instance in the KNN algorithm [14]. Noise training data can be overcome using this algorithm [12].

*E. Performance Calculation*

Accuracy, precision, and recall were used to evaluate the classification performance in this study. A comparison of correctly classified data with all data is shown at accuracy. A comparison of correctly classified data with all data in the same class is shown in precision. Also, a comparison between the amount of data in a class that is correctly classified and the correct data in the same class is shown in a recall.

## III. RESULT AND ANALYSIS

This section will describe the results of the classification analysis of tweets about perpustakaan nasional. The method used to do the analysis is accuracy, precision, and recall.

*A. Classification Results*

Data obtained from twitter contains the topic of the Perpustakaan Nasional Republik Indonesia (Perpusnas). Data consisted of 522 tweets collected from January 2019 to August 2019. Categories of tweets were positive, negative and neutral tweets. Before the classification is performed, the elimination of duplicate steps is performed on the tweet. Then the preprocessing stage is carried out, then weighted with TF-IDF to produce a total of 522 x 2008 features. TF-Binary was used as a comparison in this study because of its ease of application compared to TF-IDF. The classifications used are Naive Bayes and KNN. To measure the performance of the classification, this study uses accuracy, precision, and recall.

This research uses the Weka tool. In Weka, the KNN classification can be applied using the IBk library. As for the Naive Bayes classification, the library used in Weka is called the Naive Bayes classifier. Measurement of classification performance uses the ten fold cross-validation method.

**Table 2 Comparison of Classification Performance**

| Weighting | Classifier | Accuracy | Precision | Recall |
|---|---|---|---|---|
| TF-IDF | KNN | **83.33 %** | **79.2 %** | **83.3 %** |
| | Naive Bayes | 75.67 % | 77.7 % | 75.7 % |
| TF-Binary | KNN | 81.418 | 75.9 | 81.4 |
| | Naive Bayes | 78.736 | 79.2 | 78.7 |

Based on Table 2, the performance of the TF-IDF weighting and KNN classification has the highest values with accuracy, precision and recall values of 83.33%, 79.2%, and 83.3%, respectively, while the lowest performance is the combination of the TF-IDF weighting with the Naive Bayes classification with the values of accuracy, precision, and recall of 75.67%, 77.7% and 75.7%, respectively.

*B. Analysis of Results from Classification Method*

From table 2, it can be concluded that the classification performance by TF-IDF weighting and KNN classification has the highest value than the

others. This proves that TF-IDF is an efficient, easy weighting and has accurate results [10]. TF-IDF has maximum results when used in the classification with a few classes and low skew as in this study. In this study, TF-IDF can produce higher classification performance than others because with TF-IDF the weighting is more describing documents that contain text compared to TF-binary. KNN classification produces accuracy, precision, and recall of 83.33%, 79.2%, and 83.3%, respectively. KNN also supports the TF-IDF weighting of Naive Bayes when viewed from the results of the resulting performance.

The theoretical contribution of this study is that TF-IDF is more effectively used than TF-Binary. The KNN algorithm is also superior to Naive Bayes. The practical contribution of this research is that the Perpustakaan Nasional has been able to satisfy its users. It can be seen from the number of positive tweets more than the number of negative tweets. However, it is still necessary to improve the quality of services and facilities of the Perpustakaan Nasional because there are still negative tweets.

## IV. CONCLUSION AND FUTURE RESEARCH

The results of this study concluded that TF-IDF is still an effective weighting in terms of text mining. Especially with text data with few classes and low skew. Besides that the combination of TF-IDF with KNN classification can achieve maximum performance compared to other classification methods. The combination of TF-IDF with KNN can produce a maximum performance on the value of accuracy, precision, and recall.

To develop future research, we need both training data and test data with even more numbers. The more data, the more robust the research method is. The limitation of this research is the small amount of tweet data about the Perpustakaan Nasional, too many neutral tweets so that it does not appear to be as significant as the services or facilities of the Perpustakaan Nasional. A step is also needed to change non-standard words into standard words in Indonesian, also needs steps to change the negation sentence, but the sentence is positive.

## References

[1] S. Agmasari, “Melihat Fasilitas di Perpustakaan Nasional RI,” 07-Jan-2018.

[2] Y. Choi, “Finding ‘just right’ books for children: analyzing sentiments in online book reviews,” *Electron. Libr.*, Jun. 2019.

[3] H. Murfi, F. L. Siagian, and Y. Satria, “Topic features for machine learning-based sentiment analysis in Indonesian tweets,” *Int. J. Intell. Comput. Cybern.*, vol. 12, no. 1, pp. 70–81, Feb. 2019.

[4] N. Claypo and S. Jaiyen, “Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection,” in *2014 International Computer Science and Engineering Conference (ICSEC)*, 2014, pp. 394–397.

[5] A. D. Putri, “Klasifikasi Dokumen Teks Menggunakan Metode Support Vector Machine dengan Pemilihan Fitur Chi-Square,” 2013.

[6] N. Y. Faradhillah, R. P. Kusumawardani, and I. Hafidz, “Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin,” *SESINDO 2016*, vol. 2016, 2016.

[7] A. A. Arifiyanti, “EKSTRAKSI FITUR PADA KONTEN JEJARING SOSIAL TWITTER BERBAHASA INDONESIA DALAM PENINGKATAN KINERJA KLASIFIKASI,” 2015.

[8] A. Hamzah, “Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis,” in *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, 2012, pp. 269–277.

[9] P. Nomleni, M. Hariadi, and I. K. E. Purnama, “Sentiment Analysis Berbasis Big Data Sentiment Analysis Based Big Data,” *ReTII*, 2014.

[10] C. Megawati, “Analisis Aspirasi dan Pengaduan di Situs LAPOR! Dengan Menggunakan Text Mining,” *Depok Univ. Indones.*, 2015.

[11] F. K. R. Mahfud and A. Tjahyanto, “Improving classification performance of public complaints with TF-IGM weighting: Case study: Media center E-wadul surabaya,” in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2017, pp. 220–225.

[12] F. P. Shah and V. Patel, “A review on feature selection and feature extraction for text classification,” in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 2264–2268.

[13] A. D. Arifin, I. Arieshanti, and A. Z. Arifin, “Implementasi algoritma k-nearest neighbor yang berdasarkan one pass clustering untuk kategorisasi teks,” *ITS Surabaya*, 2012.

[14] S. K. Lidya, O. S. Sitompul, and S. Efendi, “Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) Dan K-Nearest Neighbor (K-NN),” in *Seminar Nasional Teknologi Informasi dan Komunikasi*, 2015.