JURNAL RISET INFORMATIKA
Vol. 4, No. 3. June 2022

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.160

Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

SENTIMENT ANALYSIS OF THREE-PERIOD POLEMICS USING K-NEAREST NEIGHBOR WITH TF-IDF WEIGHTING

Siti Ernawati[1*)], Risa Wati[-2]

Program Studi Sistem Informasi
Universitas Nusa Mandiri
www.nusamandiri.ac.id
siti.ste@nusamandiri.ac.id

Program Studi Sistem Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
risawati.rwx@bsi.ac.id

(*) Corresponding Author

Abstrak

Berhembusnya isu wacana perubahan masa jabatan presiden yang awalnya 2 periode pemerintahan menjadi 3 periode menimbulkan pro-kontra pada masyarakat. Banyak tagar 3 periode bermunculan pada media sosial twitter. Sehingga dilakukan penelitian mengenai sentimen analisis terhadap polemik pemilihan presiden 3 periode. Tujuan dari penelitian adalah untuk menghasilkan nilai klasifikasi mengenai isu wacana perubahan pemilihan presiden menjadi 3 periode menggunakan metode K-NN dan apakah metode K-NN terbukti baik digunakan untuk pengklasifikasian teks pada review polemik pemilihan presiden 3 periode. Dataset berjumlah 1152 data, data diolah menggunakan Python dan Jupyter Notebook sebagai text editor. Data diklasifikasikan kedalam review positif dan review negatif selanjutnya data dibagi kedalam data latih dan data uji dengan perbandingan 90:10. Dilakukan pembobotan kata menggunakan TF-IDF dan dilakukan klasifikasi sentimen menggunakan metode K-NN. Dari hasil pengklasifikasian menggunakan metode K-NN diperoleh nilai akurasi tertinggi yaitu saat nilai k=17 dan k=18 dengan nilai akurasi sebesar 85.3%. Hasil analisis sentimen masyarakat terhadap review isu wacana perubahan masa jabatan presiden menjadi 3 periode cenderung negatif dengan jumlah persentase sebesar 21.26% sentimen positif dan 78.74% sentimen negatif.

Kata kunci: Sentimen Analisis; 3 Periode; K-Nearest Neighbor; TF-IDF

Abstract

The issue of changing the presidential term which was originally 2 periods of government into 3 periods raises pros and cons in the community. Many 3-period hashtags have sprung up on social media twitter. So that conducted research on sentiment analysis of presidential election polemics 3 period. The purpose of the study was to produce the value of classification on the issue of presidential election change discourse into 3 periods using the K-NN method and whether the k-NN method proved to be well used for classifying text in the review of presidential election polemics 3 periods. Dataset totaling 1152 data, data is processed using Python and Jupyter Notebook as a text editor. The data is classified into positive reviews and negative reviews, then the data is divided into training data and test data with a ratio of 90:10. Weighting words using TF-IDF and sentiment classification using K-NN method. From the results of classification using the K-NN method obtained the highest accuracy when the value of k=17 and k = 18 with an accuracy of 85.3%. The results of the analysis of public sentiment to review the issue of discourse on the change of presidential term into 3 periods tend to be negative with a percentage of 21.26% positive sentiment and 78.74% negative sentiment.

Keywords: Sentiment Analysis; 3 Periods; K-Nearest Neighbor; TF-IDF

INTRODUCTION

Currently, the issue of discourse regarding the 5th Amendment of the 1945 Constitution to change the term of office of the president who initially could be re-elected for 2 periods of government into 3 periods (Rauf & Rado, 2022). Discourse on the change of presidential term into 3

P-ISSN: 2656-1743 | E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.160
JURNAL RISET INFORMATIKA
Vol. 4, No. 3 June 2022
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

periods raises the pros and cons in society (Pin et al., 2021). A lot of 3periode hashtags are popping up on social media twitter. Twitter is a microblogging site used to write about topics and issues that are hotly discussed (Isnain et al., 2021).

Research conducted on sentiment analysis to classify reviews on twitter regarding the 3-term presidential election. Sentiment analysis is a method used to understand, extract opinion data and process textual data automatically to obtain sentiment in an opinion (Tri Romadloni et al., 2019). Several methods were developed and applied for text classification, one of which is K-Nearest Neighbor (k-NN) (Indriati & Ridok, 2016). K-NN is an already popular algorithm (Puspita & Widodo, 2021).

In this study, data was taken from social media twitter. Data is processed using Python and Jupyter Notebook as text editor. Review data is classified into two, namely positive reviews and negative reviews, then weighting is done using TF-IDF (Term Frequency-Inverse document Frequency). Weighting is the process of changing the term which is qualitative data into quantitative data so that it can be processed by a computer (Indriati & Ridok, 2016). Furthermore, modeling is done using the K-NN method. K-NN is a method for classifying objects based on learning data that are closest to the object (Tri Romadloni et al., 2019).

Some studies of text classification using the K-Nearest Neighbor (K-NN) method, among others, the application of sentiment analysis on Twitter users using the K-Nearest Neighbor method obtained an accuracy value of 67.2%, a precision value of 56.94% and a recall value of 78.24% (Deviyanto & Wahyudi, 2018). The application of the K-Nearest Neighbors algorithm on travel agent sentiment analysis with an accuracy of 87% with a value of k=8 (Ernawati & Wati, 2018). As well as research entitled Implementation of K-Nearest Neighbor (K-NN) Algorithm for Public Sentiment Analysis of Online Learning and get the results of accuracy of 84.93% with a value of k=10 (isnain et al., 2021).

RESEARCH METHODS

Several stages in research sentiment analysis on the issue of the 5th amendment to the 1945 constitution change of presidential term into 3 periods are described in the diagram below:

1. Identify the problem and Research objectives

The first stage in research is to identify research problems. The problem in this study is to assess public sentiment on the issue of discourse menganai presidential term change which was originally 2 periods of government into 3 periods. While the purpose of this study is to produce the value of classification on the issue of presidential election change discourse into 3 periods using the K-NN method and whether the k-NN method is proven to be well used for classifying text on the review of presidential election polemics 3 periods.

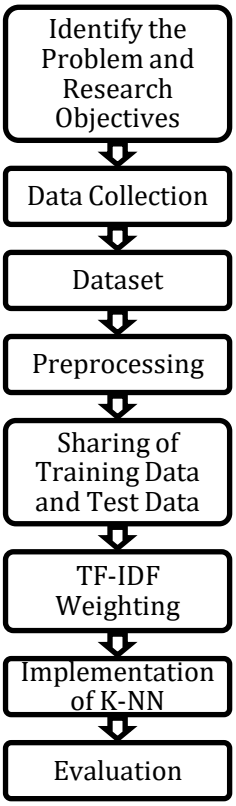

Figure 1. Stages Of Research

2. Data Collection

Data taken from the page https://netlytic.org/index.php. twitter data contains public opinion on the issue of presidential election change discourse into 3 periods. Data is stored in the form .CVS (Comma Separated Values).

3. Dataset

In this study the dataset taken amounted to 1152 data. Furthermore, the labeling of positive and negative manually, where the dataset contains elements of positive reviews as much as 241 data and negative reviews as much as 910 data. Data is processed using python and jupyter notebook as a text editor.

4. Prepocessing

Once the Dataset is collected and labeled positive and negative, the next step is preprocessing. Prepocessing is a stage for cleaning data (Duei Putri

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.160
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

et al., 2022). The need for preprocessing stages because the data taken from twitter is not structured and needs to be cleaned so that the data obtained is qualified so that data analysts can do it. Here are some stages of preprocessing:

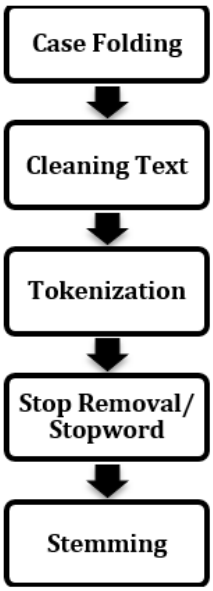

Figure 2. Stages Of Preprocessing

a. Case Folding
The process of changing uppercase to lowercase (Dervish et al., 2020).
b. Cleaning Text
The process of eliminating unneeded words such as HTML characters, emoticons, hastag (#), @ and url (Dervish et al., 2020).
c. Tokenization
The process of splitting words in a document into terms based on spaces (Dervish et al., 2020).
d. Stop Removal/ Stopword
The process of omitting words unrelated to sentiment analysis (Dervish et al., 2020).
e. Steaming
The process of removing affixes in words (Dervishes et al., 2020).

5. Sharing of training Data and test Data
The next process is to divide the data into training data and test data with a comparison of 90:10 data with the amount of training data as much as 1035 data and test data as much as 116 data.

6. TF-IDF weighting
In doing its job, machine learning and deep learning will work well by using data in numerical form. Sentiment analysis is a part of natural language that consists of a text. Before this sentiment analysis can be processed, the text must be converted into numerical form. The process of converting text into numerical form is called word weighting. This word weighting process is very important because it can represent each word into a numerical form. Some algorithms can be used in Word weighting such as TF-IDF, N-grams, Bag of Words and word2vecTerm. The Fruquency Inverse document Frequency (TF IDF) algorithm is often applied to text for sentiment analysis because it can be used to evaluate the importance of words in a corpus. The more often the word appears, the more useful it will be for the classification process (Chiny et al., 2021). Here's the formula in TF-IDF weighting (Liu & Yang , 2012).

$$a_{i,j} = tf_{i,j} * log\left(\frac{N}{n_j}\right) \quad \text{.....(1)}$$

$tf_{i,j}$ is the term frequency of term j in document i. N is the total number of documents in a dataset. n_j is the number of document appearances in term I.

7. K-NN implementation
K-NN is a supervised learning classification algorithm or so-called distance-based method (Septian et al., 2019). K-NN uses the same feature where it assigns data points based on how close they are to their neighbors. In the classification using k-NN unknown pattern will be determined as the most dominant class among the nearest neighboring classes (Hota & Pathak, 2018).

8. Evaluation
The final stage of this study is to evaluate to prove whether the K-NN method is a good classifier and has high accuracy in sentiment analysis on the issue of presidential election change discourse into 3 periods.

RESULTS AND DISCUSSION

In this study, data on public sentiment on the issue of discourse on the change of presidential term into 3 periods are taken from twetter through a link https://netlytic.org/index.php, the data is stored in the format .CVS. The Dataset amounted to 1152 data consisting of 241 positive reviews and 910 negative reviews. Next, preprocessing data is done with several stages, namely case Folding, Cleaning Text, Tokenization, Stop Removal/ Stopword, and Steaming. After going through the stages of preprocessing data is divided into training data and test data as many as 1035 training data and 116 Test data. Then weighting words using TF-IDF. After weighting the word, classification is carried out using the K-NN method. Testing on the value of

P-ISSN: 2656-1743 | E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.160
JURNAL RISET INFORMATIKA
Vol. 4, No. 3 June 2022
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

k to obtain the best accuracy results. Table 1 is the result of testing by entering the value of k from the range k=3 to k=20. Obtained the best accuracy value is k=17 and k=17 with an accuracy value of 85.3%.

Table 1. Test Results

K	Accuracy
3	81.9
4	81.9
5	74.1
6	84.5
7	84.5
8	82.2
9	84.5
10	82.2
11	82.2
12	81.9
13	83.6
14	83.6
15	84.5
16	84.5
17	**85.3**
18	**85.3**
19	82.8
20	83.6

The length of the tweets in the training data and test data are shown in the matplotlib diagram.

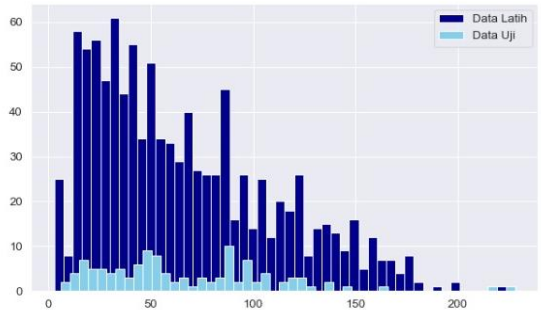

Figure 3. Matplotlib Diagram

Figure 4. Shows the confusion matrix test results with a percentage value of True negative 79.31%, False negative 12.93%, False positive 2.59% and True Positive 5.17%.

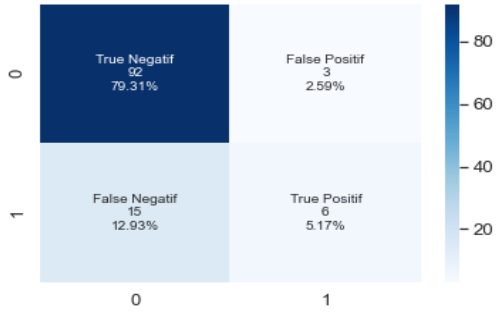

Figure 4. Confusion Matrix Test Results

Figure 5 is a diagram to calculate the error value of k between 1 to 20

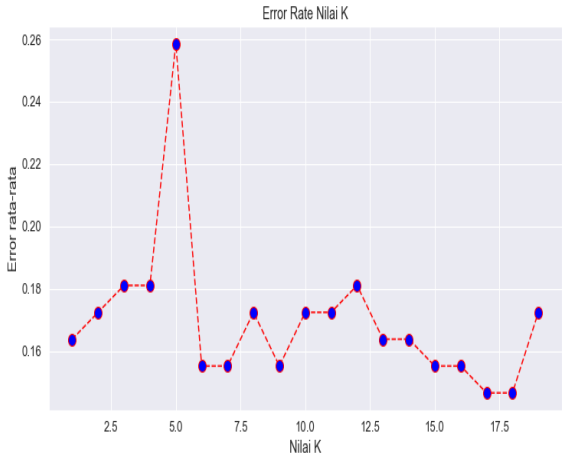

Figure 5. Error rate value k

Based on Figure 5, it can be seen that the error value is K=5. These results can be used as a reference when going to use the value of K with high accuracy results, so as to minimize the error of K value prediction.

Figure 6 is a wordcloud to count the number of words that often appear, the larger the word in the text, the greater the frequency of the word appears on the sentiment. Here are some words that often appear in sentiment, namely period, reject, jokowi, people, Students, president, support.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.160
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

Figure 6. Wordcloud

Seen in Figure 7. Pie chart results of sentiment analysis can be concluded that public sentiment to review the issue of discourse changes presidential term into 3 periods tend to be negative with a percentage of 21.26% positive sentiment and 78.74% negative sentiment.

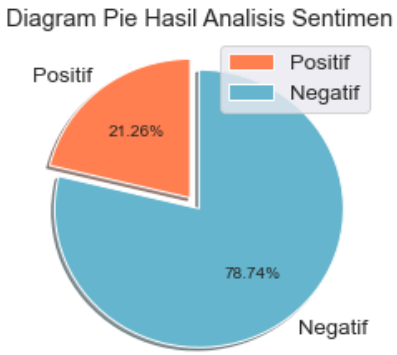

Figure 7. Positive and negative sentiment Pie chart

CONCLUSIONS AND SUGGESTIONS

Conclusion

From the results of this study it can be concluded that the k-NN method with TF-IDF weighting proved to be a good classifier method for sentiment analysis review polemic presidential election 3 period by obtaining high accuracy when the value of k= 17 and k=18 with an accuracy of 85.3%.

Suggestion

Expected in the next research can use the word weighting is N-Gram in order to compare the results of weighting between TF-IDF with N-Gram.

REFERENCES

Chiny, M., Chihab, M., Chihab, Y., & Bencharef, O. (2021). LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model. *International Journal of Advanced Computer Science and Applications*, *12*(7), 265–275. https://doi.org/10.14569/IJACSA.2021.0120730

Darwis, D., Pratiwi, E. S., & Pasaribu, a F. O. (2020). Penerapan Algoritma SVM Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Edutic - Scientific Journal of Informatics Education*, *7*(1), 1–11. https://doi.org/10.21107/edutic.v7i1.8779

Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(1), 1. https://doi.org/10.14421/jiska.2018.31-01

Duei Putri, D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Informatika Dan Teknik Elektro Terapan*, *10*(1), 34–40. https://doi.org/10.23960/jitet.v10i1.2262

Ernawati, S., & Wati, R. (2018). Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel. *Jurnal Khatulistiwa Informatika*, *6*(1), 64–69.

Hota, S., & Pathak, S. (2018). KNN classifier based approach for multi-class sentiment analysis of twitter data. *International Journal of Engineering and Technology(UAE)*, *7*(3), 1372–1375. https://doi.org/10.14419/ijet.v7i3.12656

Indriati, I., & Ridok, A. (2016). Sentiment Analysis for Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn). *Journal of Enviromental Engineering and Sustainable Technology*, *3*(1), 23–32. https://doi.org/10.21776/ub.jeest.2016.003.01.4

Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(2), 121. https://doi.org/10.22146/ijccs.65176

Liu, M., & Yang, J. (2012). An improvement of TFIDF weighting in text categorization. *International Conference on Computer Technology and Science (ICCTS)*, *47*(Iccts), 44–47. https://doi.org/10.7763/IPCSIT.2012.V47.9

Pin, P., Siahaan, J. T. H., Nellya, B., & Bangun, M. (2021). Presiden Indonesia Tiga Periode. *Jurnal Darma Agung*, *29*(2), 267–272.

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Rauf, M. A. A., & Rado, R. H. (2022). Menakar Peluang Masa Jabatan Presiden 3 Periode Dalam Konfigurasi Politik Hukum. *Al-Adalah : Jurnal Hukum Dan Politik Islam*, *7*, 30–47. https://mail.jurnal.iain-bone.ac.id/index.php/aladalah/article/view/2054

Septian, J. A., Fahrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter TerhadapPolemik PersepakbolaanIndonesiaMenggunakanPembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent Systems and Computation*, 43–49. https://t.co/9WloaWpfD5

Tri Romadloni, N., Santoso, I., & Budilaksono, S. (2019). Perbandingan Metode Naive Bayes, KNN, dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line. *Jurnal IKRA-ITH Informatika*, *3*(2), 1–9.