Classification of Twitter Comments About the Image of the People's Representative Council (DPR) Using the K-Nearest Neighbor (K-NN) Method and Naïve Bayes

Tri Ginanjar Laksana[1], Putri Rizqiyah[2], Rima Dias Ramadhani[3], Novanda Alim S.N[4]
{anjarlaksana@ittelkom-pwt.ac.id[1], 15102031@ittelkom-pwt.ac.id[2], rima@ittelkom-pwt.ac.id[3], novanda@ittelkom-pwt.ac.id[4]}

Telkom Institute of Technology Purwokerto,The Faculty of Industrial Technology and Informatics[1,2,3,4]

Abstrak. Currently, Twitter is not only a place to write microblogging messages, but has become a place where people express their aspirations. In 2018 the DPR received a lot of criticism from the public, especially through the twitter platform, therefore the data used in this study is the image of the community towards the DPR, which will be labeled positive and negative. The data used were 600 data consisting of 500 training data and 100 test data. The classification algorithm used in this study is K-NN and Naive Bayes, where K-NN is an algorithm that adheres to the concept of many neighborhoods while Naive Bayes is an algorithm that adheres to the concepts of probability and statistics. The final result of this study is to compare the accuracy of the two algorithms used, and after the data normalization processes to produce accuracy have been obtained results that K-NN get an accuracy of 80% at k = 19 and 20 while Naive Bayes get an accuracy of 77%. In this case the K-NN algorithm performs better than Naive Bayes because accuracy calculations can be performed repeatedly with different k until the best accuracy is achieved while the accuracy of Naive Bayes can only be done once. But even though K-NN has a higher accuracy than Naive Bayes, Naive Bayes still has a good performance in classification.

Keywords: Accuracy, DPR, K-NN, classification, Naive bayes, twitter

1 Introduction

Twitter is widely used because of its simple use, users simply register using e-mail and can enjoy services that are on Twitter, one of the most widely used services is the tweet service that is commonly used to provide opinions, critics, suggestions on things. The number of tweets that appear every day, will be useless if not utilized, so we need a technique that can process these tweets to produce valuable information.Twitter users in Indonesia are always increasing every year, even Indonesia is the fifth largest country that provides the most tweets every day where the center of this tweet comes from the city of Jakarta. Based on the sources of Techinasia, the city provides a figure of 2.4% of the 10.6 million twitter counts from January to March 2019. In addition, Indonesia ranked fifth in the number of Twitter users after the USA, Brazil, Japan and the UK [1]. Sentiment analysis is a field of study that analyzes one's opinions, sentiments, evaluations, judgments, behaviors and emotions through entities such as products, services, organizations, individuals, issues, events, topics and their attributes

ICOGESS 2019, March 14, Medan, Indonesia

DOI 10.4108/eai.14-3-2019.2292042

[2]. Sentiment analysis is widely used for business, education or government interests in analyzing public opinion on an event so that the policies made will be in accordance with the needs of the community. Sentiment analysis in the business world is usually used to analyze market needs, or community needs, which are expected to develop marketing strategies that can increase their company's income.

In this study, the Parliament is a state institution that holds legislative power which has the following duties and authority as follows:

1. Absorbing, gathering, accommodating and following up on people's aspirations.
2. Approval to the president to: (1) declare war or make peace with other countries; (2) appointments and members of the Judicial Commission.
3. Give consideration to the president in terms of: (1) amnesty and abolition granting; (2) appoint ambassadors and accept the placement of other ambassadors.
4. Selecting BPK members taking into account the considerations of the DPD.
5. Give approval to the Judicial Commission regarding Supreme Court Justices who will be appointed as justices by the president.
6. Select the votes of constitutional justices for further submission to the President [3].

Based on some of these incidents, many people gave opinions on the image of the DPR on twitter social media, this was interesting to do research because it would describe the image of the DPR in the eyes of the public. The existence of this phenomenon requires an algorithm to classify user comments both positive and negative. This is needed to find out how much the Indonesian people provide positive and negative responses to the image of the DPR, this number is needed because it can be an evaluation of the DPR's performance going forward so that it will be even better.

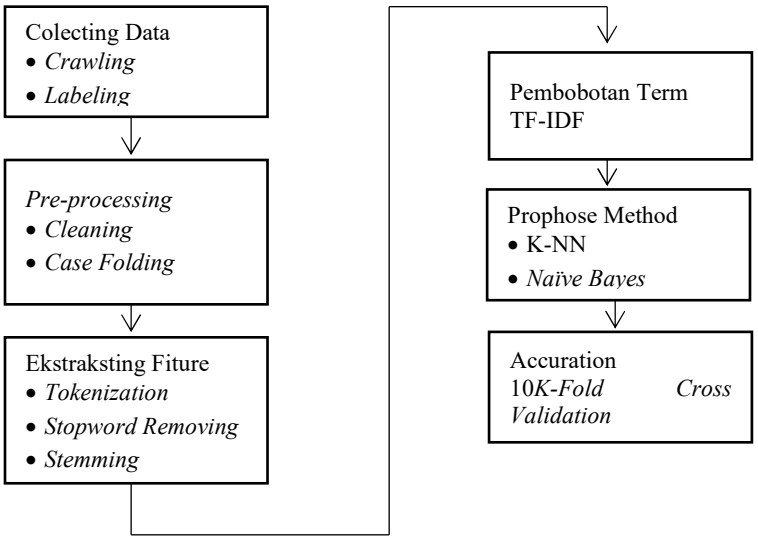
Naïve Bayes is one of the classification algorithms with probability and statistical methods proposed by British scientist Thomas Bayes, namely predicting future opportunities based on previous experience so that it is known as the Bayes Theorem. The advantages of this algorithm are simple and only require a lot of calculations. This algorithm is fast and easy and its performance is quite good but this algorithm has its disadvantages, the main disadvantage is that this algorithm cannot study interactions between features [5].

The last algorithm is the K-Nearest Neighbor (K-NN) algorithm. The K-NN algorithm is a simple algorithm other than the Naïve Bayes algorithm. K-NN is an algorithm that has no assumptions on the distribution of data that shelves it. K-NN also does not need training data so that the phase during training becomes fast and flexible. But the disadvantage of this algorithm is that the predictions given can contain a lot of noise when the data has a lot of noise [6].

Based on the results of the comparison, K-NN and Naïve Bayes are more suitable for use in text classification because the algorithms are simple, flexible and the average accuracy obtained is good. So, in this study, there will be a comparison between K-NN and Naïve Bayes in the classification of public comments on Twitter's social media regarding endorsement of UUMD3. The aim to be achieved in this study is to analyze the public response to the image of the DPR based on the performance results of the two algorithms.

2 Method

The following is the process of sentiment analysis used in this study as follows :



3 Result and Discussion

The classification of twitter comments about the image of the DPR (People's Representative Council) using the k-nearest neighbor (K-NN) and naïve bayes method is a supervised learning concept, where data consists of training data and testing data, supervised learning has its own characteristics of unsupervised learning, supervised learning has a label on training data used by computers to study data to make class predictions on testing data. Unlike clustering which is included in the concept of unsupervised learning, the concept of clustering allows computers to study data without any human guidance or labeling on training data and let the computer fully predict the existing data.

The preprocessing stage is a step that must be passed when doing text mining, because at this stage, the text that was originally in the form of the .json format will be normalized as normal as possible to facilitate the next stage. The preprocessing stage used in this study is a folding case to convert all data into lowercase letters, then remove text in the form of e-mail, URL, username, hastag, and non-alphabet using regular expression. Next, there are steps to move stopwords or words that have no meaning like and, with, who, this, that. The library used to move the stopword in this study is a literary library. This process will make it easier for further processing of data so that the data to be used later only data that has characteristics.

The next preprocessing stage is stemming, which is to change the verb into its basic word form, stemming is used to erase the suffix of a suffix prefix (confix), insertion (infix), suffix (suffix) and a combination of prefix suffix (confiks). The algorithm used in this study is the Nazief and Andriani algorithm that is available in the literary library. The use of Nazief and Andriani algorithms is known to have better accuracy than porter algorithms, but the time needed for execution is longer than the porter algorithm. The last stage is tokenization, or

dividing data into several tokens, tokenisation is done using the spacy library for Indonesian, this stage is needed to calculate the number of occurrences of the term term in a document by extracting the TF-IDF feature which will convert the word into a number based on the appearance of the term and based on the calculation of the IDF value.

Naive Bayes algorithm is a classification method that adheres to the concept of probability and statistics, Naive Bayes predicts opportunities in the future based on previous experience. The main characteristic of Naive Bayes classification is a strong assumption of independence from each condition / event, Naive Bayes works very well compared to other classification models, the advantage of this classification model is that it only requires a small amount of training data to determine parameter estimates which is needed in the classification process. Besides that, in the process of calculating naive bayes there is a condition of laplace correction or additive smoothing, which is a way to handle zero probability values, by adding a value of 1 (one) to each calculation of data in the training set, and will not make a significant difference in probability estimation so that it can avoid the case of a 0 (zero) probability value.

K-Nearest Neighbor Algorithm is a classification algorithm where the results of the new query instance are classified based on the majority of the proximity of the categories in K-NN, in this study the proximity of distance uses the formula cosine similiarity to calculate the similarity between documents, the highest number of neighbors in a neighbor will determine the position of the document located.

This study uses public image tweets to the DPR as data for processing, tweets taken as many as 600 tweets with the amount of training data as much as 500 and testing data as many as 100, data labeling is done using the questionnaire distribution method to respondents to label the data provided, then the results of respondents will put into SPSS software for reliability testing. The results to be achieved from this study is a comparison of predictions and accuracy of new data or test data using two classification algorithms, namely the Naive Bayes and K-NN algorithms, the algorithm with the best accuracy means that the algorithm is suitable for this study. After the stages before heading to classification, such as crawling, labeling, preprocessing, and weighting have been obtained accuracy for the two algorithms used, from the running process results that the Naive Bayes algorithm has an accuracy of 77%, while the K-NN has an accuracy of 80 % in conditions k = 19 and 20, with these results indicating that the K-NN algorithm has better accuracy than Naive Bayes, with a difference of 3%.

K-NN is able to have higher accuracy because the calculation of accuracy can be produced more than once according to the number of k or neighbor, so that during the process, researchers can try as many k values as possible until they get higher accuracy than Naive Bayes. But the accuracy obtained by K-NN is very conditional with the meaning that when the K-NN calculation process k value stops at 10 or 15 Naive Bayes has a higher accuracy than the K-NN value, but when the k value continues until it gets more accuracy both from Naive Bayes then that's where K-NN will get better accuracy than Naive Bayes. These things show that the K-NN algorithm can get better accuracy than Naive Bayes on the condition that several things, namely the k value and the condition of the training data, the conditions of the training data greatly influence the results of the K-NN algorithm prediction. In experiments carried out manually in stages 4.4.1 and 4.4.2, using the same dataset the Naive Bayes and K-NN algorithms produce different predictions, and if label matching is done, the predictions given by Naive Bayes are more appropriate than predictions given K-NN. This happens because the training data condition consists of 3 data that have a positive label and 1 data that has a negative label, because Naive Bayes adheres to the concept of probability and statistics so Naive Bayes does not depend on the number of positive and negative label comparisons

because the results with the highest probability are close a class, that's what is used as a result of predictions. However, the prediction of the K-NN algorithm is not appropriate, because the concept used in K-NN is the neighbor concept, where if k = 4 and in k = 4, more data is labeled positive than negative (example 3 positive data and 1 negative data) then the testing data is included in the positive label. So the use of the K-NN algorithm must pay attention to the conditions of the training data, for example training data consists of 100 data, and have positive and negative classes, the ratio of the number of positive and negative class data should not be too far like 10 positive data and 90 negative data, because predictions what will be given will not be good, so at least there are 45 positive data and 55 negative data, or better yet 50 positive data and 50 negative data, with this condition the prediction given will be good. But so far the two algorithms used are still relatively good because the resulting accuracy is more than 72%.

K-SVNN is a combination algorithm of SVM and K-NN, this combination is used because the problem faced by K-NN is the length of time predicted so that the K-SVNN method appears to shorten the prediction time but still maintain accuracy and reduce training data sets, and important parameters that have an effect on the results of reduction are K values. Based on the research that has been done the author agrees that the K value is very influential on the results of accuracy, it can be proven in table 10 that the higher the K value the accuracy obtained tends to rise for the accuracy range 1-20 , from here it can be concluded that the ideal K value used in this study is 1-20. However, in this study, the assumption is that the K-NN has a long time to make predictions, because in this study the predictions given by K-NN do not take a long time, this can occur due to several factors such as number of datasets, distribution of training data and test data, or data class. But in general, K-NN is included in an algorithm that has a good performance on classification.

Furthermore, the research conducted by Syahfitri et al. [8] entitled Sentiment Analysis in Indonesian Text Using Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN), where the problem is to compare the performance of K-NN and SVM, and determine the number of K for cross valiation , from the results obtained shows that K-NN provides good performance for independent data, that is if the test data is smaller than training data, the accuracy obtained will be good but if the test data is greater than the training data, the accuracy obtained is small. Based on the research that has been conducted, it is agreed that the distribution of training data and test data can affect the accuracy of K-NN. Then regarding the best K value for cross validation, Syahfitri et al. suggested that the best K value for cross validation was K = 10 because the highest accuracy was obtained when K = 10, and based on the research that had been done this study agreed that K = 10 was the best value for K-fold cross validation. Based on the research that has been carried out this study agrees that stemming has an effect on increasing the accuracy of the algorithm because when stemming all verbs that were previously entered into different bag of words, they will enter into the same bag of word. But in this study the accuracy obtained by Naive Bayes and K-NN before and after stemming remains the same, this can be caused because the stemming algorithm used is not good, so further research is needed to determine a good stemming algorithm.

4 Summary and Future Work

4.1 Summary

In this study it can be concluded that a number of things are as follows:

1. Predictions made by the Naive Bayes and K-NN algorithms are classified as well proven with more than 72% accuracy achieved
2. Accuracy obtained in the study The classification of twitter comments about the image of the DPR using the Naive Bayes and K-NN algorithms shows that the accuracy obtained by K-NN is greater than Naive Bayes, namely K-NN at 80% at k = 19 and 20 and Naive bayes by 77%. Based on the results of the research the level of accuracy obtained by K-NN and Naive Bayes for the classification of Twitter comments about the image of the DPR is due to differences in the second concept of the algorithm, namely Naive Bayes has the concept of probability and statistics and K-Nearest Neighbor has the closest concept of the document to its neighbors.

4.2 Future Work

Based on the results of this study there are several suggestions given for further research, namely:

1. Maximizing preprocessing stages, such as at the tokenization stage can be tried using the bigram or n-gram feature, and looking for a better library to use at the stage of removing stopword or stemming.
2. In the use of the K-NN algorithm the number of comparisons in the training data is not too far away and is recommended to be comparable
3. A combination of K-NN and Naive Bayes algorithms is carried out to determine the development of the accuracy obtained or the combination of K-NN with other algorithms and Naive Bayes with other algorithms
4. Try other classification algorithms such as the neural network, support vector machine or decision tree.

References

[1] E. Lukman, “Indonesia is Social: 2.4% of World’s Twitter Posts Come From Jakarta (INFOGRAPHIC),” 2013. .

[2] L. Zhang and B. Liu, “Sentiment Analysis and Opinion Mining,” Encycl. Mach. Learn. Data Min., no. May, pp. 1–10, 2016.

[3] S. J. D. RI, “Dewan Perwakilan Rakyat Republik Indonesia,” 2016. [Online]. Available: http://www.dpr.go.id/tentang/tugas-wewenang. [Accessed: 02-Jan-2019].

[4] A. S. Rahayu, “Polemik UUMD3,” Malang Post, Malang, p. 6, 2018.

[5] E. Chen, “Choosing a Machine Learning Classifier.” [Online]. Available: http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/.

[6] E. Alpaydın, “Introduction to machine learning,” Methods Mol. Biol., vol. 1107, pp. 105–128, 2014.

[7] E. Prasetyo, R. A. D. Rahajoe, and A. Arizal, “Perbandingan K-Support Vector Nearest Neighbour Terhadap Decision Tree dan Naive Bayes,” pp. 1–6.

[8] D. A. N. K. N. K-NN, “Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm),” vol. 2015, no. Sentika, pp. 1–8, 2015.

[9] E. Indrayuni, M. Wahyudi, S. Informasi, J. Selatan, I. Komputer, and J. Selatan, “penerapan character n-gram untuk sentiment analysis review hotel menggunakan algoritma naive bayes,” pp. 88–93, 2015.
[10] I. Surjandari, M. S. Naffisah, and M. I. Prawiradinata, “Text Mining of Twitter Data for Public Sentiment Analysis of Staple Foods Price Changes,” J. Ind. Intell. Inf., vol. 3, no. 3, pp. 253–257, 2014.
[11] M. S. Naffisah and I. Surjandari, “Penggunaan Text Mining pada Analisis Sentimen Masyarakat terhadap Perubahan Harga Bahan Pokok melalui Twitter,” pp. 1–20, 2014.
[12] P. Gil, “What Is Twitter & How Does It Work?,” Lifewire, 2018. [Online]. Available: https://www.lifewire.com/what-exactly-is-twitter-2483331. [Accessed: 05-Apr-2018].
[13] J. Strickland, “How Twitter Works.” [Online]. Available: https://computer.howstuffworks.com/internet/social-networking/networks/twitter2.htm. [Accessed: 05-Apr-2018].
[14] G. Balinggan, “Difference between Clustering and Classification,” 2018. [Online]. Available: http://www.differencebetween.net/technology/difference-between-clustering-and-classification/. [Accessed: 05-Apr-2018].
[15] L. Sofiyana, Z. Abidin, and H. Nurhayati, “Klasifikasi Emosi Untuk Teks Berbahasa Indonesia Dengan Menggunakan K-Nearest Neighbor,.” 2014.
[16] S. Raschka, “Introduction and Theory,” pp. 1–20, 2014.
[17] S. Vijayarani, M. J. Ilamathi, and M. Nithya, “Preprocessing Techniques for Text Mining - An Overview,” vol. 5, no. 1, pp. 7–16.
[18] D. Wahyudi, T. Susyanto, and D. Nugroho, “Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia,” J. Ilm. Sinus, vol. 15, no. 2, pp. 49–56, 2017.
[19] B. Góralewicz, “The TF*IDF Algorithm Explained,” 2018. [Online]. Available: https://www.elephate.com/blog/what-is-tf-idf/. [Accessed: 04-May-2018].
[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, Springer Texts in Statistics An Introduction to Statistical Learning. .
[21] “Python - Tutorial,” Tutorials Point. [Online]. Available: https://www.tutorialspoint.com/python/index.htm. [Accessed: 04-May-2018].