ISSN: 04532198
Volume xx, Issue xx, January 2020

Classification of Political Data on Social Media Twitter using Naive Bayes Algorithm

Raisha Citra Chairani[1], Casi Setianingsih[2], Budhi Irawan[3]

School of Electrical Engineering, Telkom University-Bandung, Indonesia[1,2,3]

Abstract - Twitter is social media that can be used to exchange ideas and give opinions. Twitter users can write their opinions on the issue of President Joko Widodo's government. Tweet data or public opinion can be done sentiment analysis method to analyze public opinion. *The Naïve Bayes method* is used to classify Twitter data to determine sentiment and grouping into positive class and negative class. Furthermore, topic modeling is carried *out with the Latent Dirichlet Allocation* (LDA) method to find out the topic of discussion in each sentiment group. In the classification process, the value of accuracy depends on the *preprocessing stage and* depends on the amount of data. In train data 80% and test data 20% obtained accuracy 84.58%, *recall* 85%, *precision* 85% and *F1-Score* 85%. At the LDA stage, performance testing with *perplexity resulted* in *a perplexity value* of 7.1049 based on the number of iterations of 30 for the positive sentiment group. Furthermore *the perplexity value* is 7.3165 with the number of iterations is 60 for the negative sentiment group.

Keywords: Classification, Naïve Bayes, Latent Dirichlet Allocation, Tweet

1. INTRODUCTION

Developments in the field of technology and information bring a role to human life with a rapid exchange of information, especially in the form of text. Social media is an internet-connected medium that allows its users to access information, discuss and create content. One social media that allows its users to share opinions in the form of text is twitter. Twitter's daily user growth jumped 34% in the second quarter of 2020 making Twitter as a social media *platform* that can be used to retrieve data to further analyze the issues that exist in the community [1]. With Twitter, users can write their opinions into 280 characters called tweets. Tweets can contain opinions, criticisms and suggestions on issues that occurred in the government of President Joko Widodo. Tweets written by users will be analyzed to find out the sentiments and topics of conversation of outstanding issues and public views on issues related to the government of President Joko Widodo. Sentiment analysis and knowing the topic of public discussion on the issue of President Joko Widodo's government can be one of the efforts to improve and evaluate the issues discussed by the public. One method for conducting sentiment analysis is Naïve Bayes. The Naïve Bayes method is used because it has high speed and accuracy values when applied to large files and diverse databases [2]. The Naïve Bayes classifying method also has the advantage of fast, high accuracy and simple [3]. One method to find out the topic is by the *Latent Dirichlet Allocation* (LDA) method. LDA is a method used to summarize, group, connect and process data.

In this study, sentiment analysis will be conducted on the issue in president Joko Widodo's government using the classification model Naïve Bayes and modeling topics using the LDA to evaluate the government of Joko Widodo to be better based on public comments.

2. RELATED WORK

Billy Gunawan et al, conducted an analysis sentiment research on online product reviews in Indonesian. Review data is classified using Naïve Bayes and the sentiment analysis system is divided into 5 classes which are very negative, negative, neutral, positive and very positive. The research aims to help companies know the feedback on products owned from public assessments based on existing opinions and reviews. The results of the study obtained the best value in testing with 3 classes (positive, negative and neutral) for data sharing 90% training data and 10% test data with accuracy values of 77.78%, recall 93.33% and precision 77.78% [4].

Winda Estu Nurjanah et al, researched television based on opinions written by the public on social media Twitter. The study used the K-Nearest Neighbor method and used weighting with the number of retweets. Retweets are a feature on Twitter that has a function to share or share tweets. The weighting with the number of retweets depends on the number of retweets, the more retweets the more positive the sentiment of the tweet. The study obtained a maximum value with k = 3 with an accuracy rate of 80.83%, precision 72.28%, recall 100% and f-measure 83.91%. The study also obtained an accuracy rate of 82.50% when using textual weighting obtained 60% accuracy and when combining the two obtained accuracy of 83.33% with the values k = 3, $\alpha = 0.8$ and $\beta = 0.2$ [5].

Mesut Kaya et al conduct research on political news from various Turkish political news sites. This study classifying sentiment against Turkish political news and is divided into positive and negative classes. The study compared 4 algorithms namely Naïve Bayes, Maximum Entropy and SVM. In the study obtained the highest accuracy value Naïve Bayes 72.05%, while Maximum Entropy obtained an accuracy of 69.44% and in SVM obtained an accuracy of 66.81% [6].

3. RESEARCH METHOD

3.1 Text Mining

Text mining is defined as the process of collecting data to obtain valuable information from unstructured text [7]. The purpose of text mining is to obtain valuable information from a set of unstructured text data.

3.2 Preprocessing

The preprocessing process usually significantly impacts performance for machine learning algorithms [8]. Eliminating distractions is one of the stages that must be done to obtain the expected results. In this study, cleaning activities were carried out on @username, number, and http: //URL on tweets that have been obtained at the text mining stage. After the cleaning phase, the results will be obtained in the form of tweet data that is free of numbers, @username and http: //URL. Furthermore, the case folding stage is carried out. Case folding is a stage to convert capital letters on data into lowercase letters [9]. The process is continued with the stop removal stage. Stop removal is a list of words that are repeated frequently and appear in any text data. Common words such as and, or, it etc. need to be omitted because it has no meaning and when omitted does not have a significant effect on the classification process. The last stage of preprocessing in this study is stemming. Stemming is the process of eliminating the initial and final words to obtain the base word and to reduce the number of features to improve the performance of the classification. [10]

3.3 Feature Extraction

TF-IDF is the process of converting text data into numerical data [11]. Changing text data to numerical data is intended to give weight to each word or feature. TF-IDF is an evaluation process by analyzing statistical measures based on the importance of words in a document. TF or term frequency calculates the

frequency with which a word appears in each document. The more often a word appears the more important it is in the document. DF or document frequency indicates how common the word is. While the IDF is the inverse of the DF. To calculate the Term Frequency in the search for a term (t) in a document (d) use the following equation:

$$W_{t,d} = TF_{t,d} * IDF_t \quad (1)$$

Description :

$W_{t,d}$: weight of term (t) in a document

$TF_{t,d}$: frequency of occurrence of term (t) in document (d)

IDF_t : Inverse frequency of documents, where

$$IDF_t = \log \frac{N}{Nt} \quad (2)$$

Description :

N : number of all documents

N_t : number of documents (d) containing term (t)

3.4 Naïve Bayes Classifier

Naïve Bayes is a method of classification with probability and statistical approach on the assumption that the feature is an independent class [12]. The use of the Naïve Bayes method in this research is based on research that has been done before by obtaining high accuracy but simple and fast. There are two stages of classification using Naïve Bayes, namely learning or training and testing. The Naïve Bayes method can be done in the following ways :

1. Calculate the probability of the prior class

$$P(c) = \frac{Nc}{N} \quad (3)$$

Description:

P(c) : the probability of class occurrence
Nc : number of documents with class 'c'
N : number of documents

2. Calculate conditional probability

$$P(w\,|c) = \frac{ni+1}{|c|+v} \quad (4)$$

Description:
P(w | c) : value of the word 'w' in class 'c'
ni : weighting value of the word 'w' against class 'c'
|c| : total number of words of class 'c'
v : vocabulary count

3. Calculating posterior class

$$C_{MAP} = = argmax\ P(\ c)\ x\ P(Wi\ |C) \quad (5)$$

Description:
P(c) : the probability of class occurrence
P(Wi | C) : the probability of occurrence of the word 'w' against class 'c'

3.5 Latent Dirichlet Allocation

Latent Dirichlet Allocation or LDA is a method for detecting topics through the probability modeling of a data set [13]. In the context of text modeling, topic probability provides an explicit representation of a document. The LDA assumes that the word represents the topic contained in each document [14]. Or in other words, the document is a mixture of topics of a certain proportion [15].

3.6 Performance Evaluation

3.6.1 Confusion Matrix

Confusion Matrix is one of the methods used to calculate the performance of an algorithm. The performance of the classification system illustrates how well the system classifies data. The confusion matrix can be seen in the following table :

Table 1. Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Based on TP, FN, FP and TN values can be obtained accuracy, precision, recall and F1-Score values. Here is the formula for calculating the performance of the classification system :

$$Accuration : \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (5)$$

$$Precision : \frac{TP}{TP+FP} * 100\% \quad (6)$$

$$Recall : \frac{TP}{TP+FN} * 100\% \quad (7)$$

$$F1\ Score : 2\ x \frac{Precision*Recall}{Precision+Recall} * 100\% \quad (8)$$

3.6.2 Perplexity

One method for evaluating topic modeling performance is to calculate the value *of perplexity. Perplexity is* a metric *to evaluate* the accuracy of the information in the document to the topic obtained. Perplexity *calculations* are performed by specifying text logs from invisible documents. The smaller or lower the *perplexity value* the better or better the resulting model. Here is the formula for calculating *perplexity score* [13] :

$$Perplexity(D_{\text{test}}) = \exp\{\frac{-\sum_d \log p(w_d)}{\sum_d N_d}\} \quad (9)$$

Description :

$p(w_d)$: chance of total word count

N_d : Total number of words in document 'd'

.

4. PROPOSED METHOD

4.1 Data Collection

In this study, the dataset is tweet data obtained from Twitter social media in the Indonesian language based on #jokowi queries taken from 23 September 2018, to 15 December 2020.

Datasets are obtained by crawling by entering date queries and #jokowi which are then stored in csv form. Furthermore, the dataset has been validated by the West Java Language Hall office.

4.2 Preprocessing Data

After the tweet dataset is obtained, the dataset will be preprocessing intending to obtain data ready for the next process. Here are the preprocessing stages of the study:

Figure 1 Preprocessing Diagram

As in the diagram above that the first stage of preprocessing is clearing text, the table below is an overview of the process of clearing text by removing @username URLs, http://, numbers, characters, and punctuation marks. Here is an example of preprocessed text in Indonesian:

Table 2 *Cleaning text*

Input	Output
@aniesbaswedan Harapan semoga 2019 menjadi #presiden yang akan datang. https://t.co/QEFqyJ4kuf	Harapan semoga menjadi presiden yang akan datang

Furthermore, the case folding stage is carried out to convert capital letters into lowercase letters.

Table 3 *Case folding process*

Input	Output
Harapan semoga menjadi presiden yang akan datang	harapan semoga menjadi presiden yang akan datang

In the next step, remove the stop word to remove words that do not have meaningful meaning.

Table 4 *Stop removal process*

Input	Output
harapan semoga menjadi presiden yang akan datang	harapan semoga presiden datang

Then followed by the stemming stage. Stemming is done to obtain basic words and avoid words that are not standard in Indonesian to reduce the ambiguity that can affect system performance.

Table 5 *Stemming*

Input	Output
harapan semoga presiden datang	harap semoga presiden datang

Here is the *tokenizing process. Tokenizing is* done to convert a sentence into several words in the form of a token.

Table 6 *Tokenizing*

Input	Output
harap semoga presiden datang	'harap', 'semoga', 'presiden', 'datang'

4.3Feature Extraction

At this stage, a feature extraction will be performed using TF-IDF to calculate the weight of each word or feature. The less often a word is found in a document, the greater the weight, and the more often a word is found in the document, the smaller the weight. Next, the data will be labeled 0 and 1. Where label 0 is a negative sentence, while 1 is a text with a positive sentence.

Table 7 Example of Twitter tweets with Indonesia Language

Dokument	Text	Label
D1	kinerja presiden jokowi buruk	0
D2	presiden jokowi baik	1
D3	kinerja dpr buruk	0

An example of a tweet with Bahasa Indonesia in table 1 has been translated into English as shown in table 7 below.

Table 8. Example of Twitter tweets with English

Document	Tweet	Label
D1	president jokowi's poor performance	0
D2	president jokowi good	1
D3	poor parliament performance	0

The next step is the TF-IDF calculation to calculate the weight of each term.

Table 8. TF-IDF Score

	baik	buruk	dpr	jokowi	kinerja	presiden	Label
D1	0	1,17	0	1,17	1,17	1,17	0
D2	1,17	0	0	1,17	0	1,17	1
D3	0	1,17	1,17	0	1,17	0	0

4.4 Classification of Naïve Bayes

At this stage, classification is carried out using the Naïve Bayes method using the libraries available in python. The output of this stage is the confusion matrix which is then calculated accuracy, recall, precision and F1-Score.

4.5Topic Modeling using Latent Dirichlet Allocation

This stage aims to obtain topic modeling from the tweet dataset. At this stage, conduct experiments by changing the topic until the highest results are found. After finding the highest value, then the parameters

of the number of topics in this study were set with 2 topics, then only changed the iteration value from 10 to 100 until the best coherence value of each accuracy was found.

5. EXPERIMENTAL RESULTS

5.1Data Retrieval Results

The following is the result of collecting tweet data with the #jokowi query starting from 23 September 2018 to 15 December 2020. Obtained data as much as 1200 by determining polarity manually so that obtained 600 tweet data with negative polarity and 600 tweet data with positive polarity:

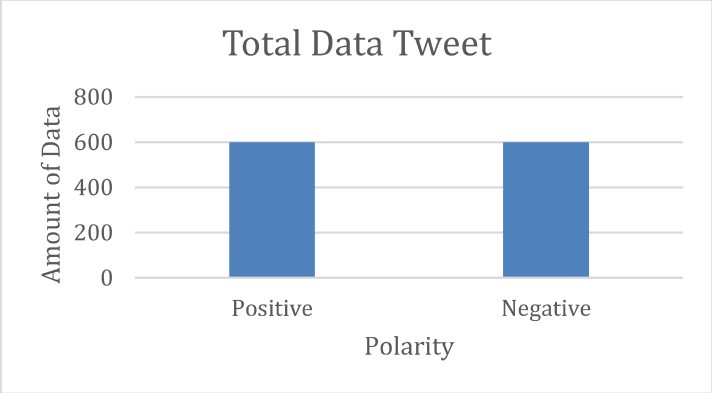

Figure 2 Total Data Tweet

5.2 Naïve Bayes Classifier Performance Results

This study was classified with 600 positive tweets and 600 negative tweets. Here are the results of classification performance using Naïve Bayes :

Table 9. Naive Bayes Performance

Testing	Test Data (%)	Train Data (%)	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	F1-*Score* (%)
1	10	90	84.17	84	84	84
2	20	80	84.58	85	85	85
3	30	70	83.33	83	84	83
4	40	60	83.96	84	84	84
5	50	50	84.50	84	85	84

5.3 Topic Modeling Results with LDA

On modeling topics using LDA performed an analysis of each class separately. In concluding the topic of the cluster is done manually to ensure the accuracy of the results by considering the opportunities of the largest as a word that better reflects the topic of each cluster.

The following are the results of the LDA experiment by changing the number of iterations for positive polarity:

Table 10. Perplexity score of LDA

Positive		Negative	
Number of Iterations	***Perplexity***	**Number of Iterations**	***Perplexity***
10	7,1211	10	7,3419
20	7,1067	20	7,3358
30	7,1049	30	7,3285
40	7,1081	40	7,3236
50	7,1055	50	7,3257
60	7,1064	60	7,3165
70	7,1061	70	7,3206
80	7,1052	80	7,3229
90	7,1052	90	7,3206
100	7,1076	100	7,3241

Here are the results of topic modeling for positive and negative tweet data consisting of 2 clusters and their conclusions have done manually :

Table 9 LDA Results

Positive				Negative			
Topic 0		**Topic 1**		**Topic 0**		**Topic 1**	
Words	**Opportunities**	**Words**	**Opportunities**	**Words**	**Opportunities**	**Words**	**Opportunities**
jokowi	0,059	jokowi	0,046	jokowi	0,043	jokowi	0,033
indonesia	0,024	indonesia	0,026	presiden	0,022	presiden	0,019
maju	0,017	presiden	0,012	indonesia	0,007	rakyat	0,007
presiden	0,013	maju	0,011	tidak	0,006	korupsi	0,006
widodo	0,005	joko	0,006	rakyat	0,006	indonesia	0,006
prabowo	0,005	widodo	0,006	dpr	0,005	tidak	0,005
joko	0,005	prabowo	0,006	korupsi	0,005	negara	0,004
ekonomi	0,004	corona	0,006	kpk	0,004	kpk	0,004
corona	0,004	edhy	0,004	menteri	0,004	orang	0,004
perintah	0,004	kerja	0,004	fpi	0,004	perintah	0,004
Economy		Corona		Corruption		Corruption	

6. CONCLUSION

Based on the result of the final project, can be drawn some conclusions are:

1. Based on classification results, obtained an accuracy value of 84.58%, recall 85%, precision 85% and F1-Score 85%.
2. Based on topic modeling *obtained perplexity value* of 7.1049 with the number of iterations of 30 for positive sentiment and perplexity value of 7.3165 with the number of iterations of 60 for negative sentiment.

7. REFERENCES

[1] S. J, "Dunia Terisolasi Pandemi Covid 19, Pengguna Twitter Meningkat," Pikiran Rakyat Com, 2020.

[2] T. L. D, Naive Bayes Estimation and Bayesian Networks, in Data Mining Methods and Models, USA: John Wiley & Sons, 2006.

[3] M. R, A Comparison of the Accuracy of Support Vector Machine and Naive Bayes Algorithms in Spam Classification, 2009.

[4] B. Gunawan, H. S. Pratiwi and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *Jurnal Edukasi dan Penelitian Informatika,* vol. 4, no. 2, pp. 113-118, 2018.

[5] W. E. Nurjanah, R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 1, no. 12, pp. 1750-1757, 2017.

[6] M. Kaya, G. Fidan and I. H. Toroslu, "Sentiment Analysis of Turkish Political News," in *International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012.

[7] S. Dang and P. H. Ahmad, "Text Mining: Techniques and its Application," *International Journal of Engineering & Technology Innovations,* vol. I, no. 4, pp. 22-25, 2014.

[8] S. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data Preprocessing for Supervised Learning," *International Journal of Computer Science,* vol. I, no. 1, pp. 111-117, 2006.

[9] A. Go, R. Bhayani and L. Huang, "Twitter Sentiment Classification using Distant Supervision," 2009.

[10] A. I. Kadhim, "An Evolution of Preprocessing Techniques for Text Classification," *International Journal of Computer Science and Information Security,* vol. 16, no. 6, pp. 22-32, 2018.

[11] J. A. Septian, T. M. Fahrudin and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *Journal of Intelligent Systems and Computation,* 2019.

[12] I. R, "An Empirical Study of The Naive Bayes Classifier," pp. 41-47, 2001.

[13] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3,* pp. 993-1022, 2003.

[14] R. Krestel, P. Fankhauser and W. Nejdl, "Latent Dirichlet Allocation for Tag Recommendation," in *ACM*, 2009.

[15] J. Mazarura and A. d. Waal, "A comparison of the Performance of Latent Dirichlet Allocation and the Dirichlet Multinomial Mixture Model on Short Text," in *PRASA-RobMech International Conference*, 2016.