Statistika, Vol. 21 No. 2, 99 – 106
November 2021

Classification of Public Opinion on Social Media Twitter concerning the Education in Indonesia Using the K-Nearest Neighbors (K-NN) Algorithm and K-Fold Cross Validation

INTAN MONICA HANMASTIANA[1], BUDI WARSITO[2], RITA RAHMAWATI[3], HASBI YASIN[4], PUSPITA KARTIKASARI[5]

[1,2,3,4,5]Departement of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
e-mail: [1]inmonicaa@gmail.com

ABSTRACT

Developing country is a country that has perspective and idea which reflect its awareness of the importance of advancing the education sector. Assessment of the quality of education in Indonesia from the perspective of the community gets different responses. Therefore, it makes people respond differently. The community response is often found on social media, one of which is Twitter. Twitter is one of the application service that is popular due to its uses to interact and communicate with people in daily life. The sentiment analysis on Twitter can be a choice to see the community's responses to the condition of education in Indonesia. The responses are classified into positive sentiments and negative sentiments using the K-Nearest Neighbors (K-NN) algorithm with a 10-fold cross validation model evaluation. K-NN has several advantages, they are fast training, simple, easy to learn, resistance toward training data which has noise, and effective if the training data is large. In this study, the sentiment classification uses Cosine Similarity distance measurement and four k value parameters which are 3, 5, 7, and 9. Data labelling is done manually and done by scoring sentiment. Visualization of positive and negative sentiments use Word Cloud. The test results show that public sentiment about education tends to be positive on Twitter and the parameter k = 7 obtained the highest accuracy value in data labelling that was done manually and done by scoring sentiment. In labelling data manually, it obtained an accuracy of 76.93% whereas, in labelling the data with scoring sentiment, it obtained an accuracy of 77.87%. Sentiment analysis is made using the RStudio programming language as the support software.

Keywords: education, sentiment analysis, twitter, k-nearest neighbors.

1. INTRODUCTION

Online Social Networks (OSN) is a provider of popular application service which can be used to interact and communicate with people in daily life (Teljstedt, 2016). Indonesia is one of countries in the world which has the largest population considered social media as an important need to socialize between individuals and groups. One of internet based application that is currently in demand by many people is Twitter. The data and information published through Twitter are very diverse, one of topic which is often discussed by Twitter users is education system in Indonesia. Education is one of the keys of the improvement of a nation. Expectations and problems that lurk the education system in Indonesia create different responses from people both in positive comment and negative comment to the government and the public itself. According to the previous statement, the author wants to observe how the society's responses toward the topic of the education system in Indonesia by analyzing the tweets of society on social media Twitter. The sentiment analysis on social media Twitter becomes the choice of the writer to analyze how the society's responses related with education topic.

Text mining is the implementation of concept and data mining technique to observe a pattern in a text in order to obtain useful information. The key element of this process is combining information that is successfully extracted to make new hypothesis from various sources for further exploration (Hearst, 2003). Sentiment analysis or opinion mining is a process of understanding, extracting, and obtaining textual data automatically to obtain sentiment information that is consisted in an opinion sentence. Sentiment analysis focuses on the opinion that contain negative or positive message (Liu, 2012).

In the previous study, the performance of the K-Nearest Neighbors as classification of algorithm is adequate. K-Nearest Neighbors (K-NN) algorithm is a method that classifies the object based on the distance learning data that is closest to the object. The purpose of this algorithm is to classify object based on attribute and training sample using the most voted between the classifications of k object. K-Nearest Neighbors (K-NN) algorithm uses neighboring classification as the prediction value from the new query instance (Larose & Larose, 2014). The method commonly used to calculate the distance between vectors is cosine similarity. Some of advantages from this algorithm are fast, having high accuracy, and simple. Wu et al. (2008), stated that the K-NN algorithm as one of the best algorithm in Top10 algorithms in data mining. In the previous study by Putrianti et al. (2019), K-Nearest Neighbors to classify sentiment produces a high degree of accuracy when classifying public sentiment regarding restaurant reviews using the K-NN algorithm. Therefore, the issue that will be discussed in this study is how the result of the society sentiment classification concerning the education system in Indonesia on social media Twitter using the K-Nearest Neighbors method.

In this study used a considerable amount of tweet data compared to previous studies. This study also uses performance evaluation model, the k-Fold Cross Validation, from the result of data labelling by using sentiment scoring and by manual. Difference that existed with previous study, labeling methods with sentiment scoring and evaluation on this research was carried out in order to accelerate the time in the labeling process while obtaining the best accuracy. The data labelling is limited to two sentiment classes which are positive and negative. The classification of Algorithm uses the K-Nearest Neighbors with the parameter k value which are 3, 5, 7, and 9, also cosine similarity distance measurement. The performance evaluation model using the k-Fold Cross Validation as much as 10 folds. So, researchers chose to use the K-NN method because in addition to being simple in classifying, the accuracy of K-NN in previous studies to classify sentiment is quite good.

2. RESEARCH METHODS

The type of data used in this study is tweets data of Twitter users. The data used are 1,500 tweets which are the opinions of Indonesia's society about education on 12 to 19 February 2020. The data are extracted with keywords "pendidikan Indonesia" in the Indonesian language tweets category. The steps of the analysis are:

1. The extraction of tweets
2. Pre-processing data.
3. Data labelling into positive class or negative class by sentiment scoring and manually.
4. Feature selection: filtering stopwords and tokenization
5. Document weighting with TF-IDF, TF-IDF is calculated using equation as down below:

$$W_{j,i} = \frac{n_{j,i}}{\sum_k n_{k,i}} . \log_2 \frac{D}{d_j} \quad \text{... (1)}$$

$W_{j,i}$ is the weighting of TF-IDF for the term j in the document i, $n_{j,i}$ is the number of j term appearance in the document i, $\sum_k n_{k,i}$ is the number of all term appearance in the document i, D is the number of all documents in collection, and d_j is the number of document that contains term j.

6. Developing the K-Nearest Neighbors classification model by calculating the cosine similarity for the distance similarity measurement, selecting of parameter k value, and using of 10-Fold Cross Validation performance evaluation model. The formula used to calculate cosine similarity is as follow:

$$\mathrm{CosSim}(x,y) = \frac{\sum_{i=1}^{m} x_i y_j}{\sqrt{\sum_{i=1}^{m} x_i^2} \cdot \sqrt{\sum_{j=1}^{m} y_j^2}} \quad \text{... (2)}$$

7. Calculating entire accuracy based on confusion matrix. Furthermore, according to confusion matrix, the performance evaluation result from the classification in the term of accuracy value, recall, precision, and f-measure can be sought. The formula used to calculate the accuracy of classifier is:

$$Accuracy = \frac{TP + TN}{P + N} \quad \text{... (3)}$$

8. Determining the best k parameter.
9. Visualizing data with word cloud.

3. RESULT AND DISCUSSION

Tweets Extraction

The tweet extraction from Twitter uses API (Application Programming Interface) platform. In the process of tweet extraction, it needs four access codes, namely API Key, API Secret, Access Token, and Access Token Secret. Those codes are obtained after registering Twitter account as developer on https://developer.twitter.com/app/new.

Text Pre-Processing

The text data is processed using the text mining method with using the tm package in RStudio software. The steps done are:

1. Case folding

Case folding will change word into one form in text document. In this step, all texts are transformed into small words (lowercase)

2. Remove URL

Remove URL will remove the link of URL (Uniform Resource Locator) that appears in text document. URL link usually contain “http://” word

3. Remove mention

Mention is a term of Twitter user that mention other users’ username. Remove mention will remove words that contain “@” symbol user

4. Remove emoticon

Remove emoticon will remove emoticon symbols such as smiling, crying, and other symbols. These symbols are called emoji.

5. Remove punctuation

Remove punctuation will remove the punctuation or symbol in tweets data because this study only classifies text data. Therefore, besides alphabet character, it will be removed from the document.

6. Remove number

Remove number will remove number character into space because number does not show emotion.

7. Convert word

Convert word functions to change non-standard Indonesian words, typos, local languages, or slangs found in tweets data. The dictionary of convert word is created manually in Microsoft Excel with 1,254 words.

Sentiment Class Labelling: Manual and Sentiment Scoring

In this study, tweets data are classified into positive sentiment or negative sentiment. The tweets data labelling process is done with two techniques namely sentiment scoring and manually. Based on the result of crawling 1,500 tweets about education, it is obtained 1,005 tweets from manual positive sentiment labelling and 495 tweets from negative sentiment. Sentiment class labelling with sentiment scoring uses the support of three dictionaries that calculate every score of tweets in text document by summing them with the following provisions:

1. Every word in tweets that occurs in sentiment dictionary will obtain score in accordance with the sentiment dictionary. If the word does not occur in the sentiment dictionary, it will obtain 0 score.
2. Every word that has negation word in the previous word will obtain the opposite score in sentiment dictionary.
3. If a word in sentiment dictionary valued > 0 that is followed by boosterwords in the previous or in the next word, then the score of sentiment word is added with the score of boosterwords word.
4. If a word in sentiment dictionary valued < 0 that is followed by boosterwords in the previous or in the next word, then the score of sentiment word is reduced with the score of boosterwords word.

The labelling of sentiment class with sentiment scoring resulted in tweets with 1,039 positive labels, meanwhile, there are 461 tweets with negative labels. Sentiment scoring gives several differences in doing the data labelling. After labelling the data manually on 1,500 tweets, there are 68 tweets that are incorrect in the labelling. It means that 4.53% tweets obtain incorrect label.

Filtering Stopwords and Tokenizing

After going through the stages of labelling sentiment classes on each tweet data, the feature selection process will then be performed. The process of removing features through two stages, namely remove stopwords and tokenizing. This stage is useful for selecting features so that the data is ready to be used for the classification process. The choice of meaningful words by removing the unimportant words in building the model can improve the accuracy of system classification result. Stopwords used in this study are 592 words and 431 words for manual stopwords. Tokenizing is a process of separating words by words in a document into interdependent words. Space is used to separate those words.

The Weighting of Term Frequency Inverse Document Frequency (TF-IDF)

The weighting process with TF-IDF will weight every term based on the importance level of the term in a collection of feedback documents. The word weighting values with Term Frequency-Inverse Document Frequency (TF-IDF) can be seen in the Table 1. The word weighting will be used to build the classification model.

Table 1 Result of Weighting

tweet	need	indonesia	issue	moral	education	...	crack
221	0	0.07969	0.7227	0.97862	0.0807	...	0
286	0	0.11156	0	1.37006	0.11298	...	2.11015
785	0.70394	0	0	0	0.05135	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1500	0	0.13945	0	0	0.14123	...	0

Sentiment Classification

After weighting every term in tweet document, data is ready to be processed in sentiment classification. The data will be divided into training data and testing data that its sentiment class will be predicted. Training data is data used to train algorithm and testing data is data that have been identified specifically used in the testing. In this study, the testing uses 1,500 tweets document with 90% : 10% division of training data and testing data. The percentage division of training data and testing data refers to the 10-fold cross validation that will repeat the process of prediction 10 times so that every data can be testing data exactly one time division into 90% : 10%. It means that 1,350 data as training data 150 testing data. The class of 150 data will be predicted by classifier using K-NN algorithm.

This study uses *K-Nearest Neighbors* (K-NN) method that uses cosine similarity distance formula and four parameters, namely k = 3, k = 5, k = 7, and k = 9. This study uses *k-fold cross validation* to examine the performance of model built with certain data test. The k value used in this k-fold cross validation is 10 with the review of 150 data test as testing data and with using 10-fold cross validation so that the prediction will be repeated 10 times. The result of accuracy calculation in its each iteration uses K-NN algorithm and 10-fold cross validation. Manual data labelling can be seen in Table 2, while the result of labelling accuracy with sentiment scoring is shown in Table 3.

Table 2 The Accuracy Value of Labelling Result Manually

k-Fold	**Parameter k**			
	k = 3	**k = 5**	**k = 7**	**k = 9**
1	74.00	76.67	77.33	74.67
2	78.00	74.67	72.00	73.33
3	74.67	74.67	72.00	72.00
4	67.33	72.67	73.33	71.33
5	69.33	77.33	72.67	76.00
6	70.00	74.67	77.33	76.67
7	80.00	82.67	85.33	84.67
8	77.33	76.67	80.00	75.33
9	74.67	72.67	77.33	78.67
10	73.33	81.33	82.00	83.33
Max value	80.00	82.67	85.33	84.67
Average	73.87	76.40	76.93	76.60

Table 3 The Accuracy Value of Labelling Result with Sentiment Scoring

k-Fold	**Parameter k**			
	k = 3	**k = 5**	**k = 7**	**k = 9**
1	80.00	76.00	78.67	76.67
2	72.00	76.00	76.00	77.33
3	72.00	69.33	69.33	66.00
4	70.67	74.67	75.33	69.33
5	75.33	76.67	76.00	74.67
6	74.00	78.00	82.00	82.67
7	80.67	85.33	84.00	85.33
8	75.33	84.00	83.33	82.00
9	73.33	74.00	75.33	80.00
10	72.00	76.00	78.67	81.33
Max value	80.67	85.33	84.00	85.33
Average	74.53	77.00	77.87	77.53

According to Table 3 from the accuracy average of entire 10 fold, the optimum accuracy value of k parameter for manual data labelling is at k = 7 because it obtains the largest accuracy value of 76.93%, while on Table 4, the accuracy value of parameter k for data labelling with sentiment scoring shows that the optimum k value is also at k = 7 because it obtains the largest accuracy value of 77.87%. The process of validation model uses 10-fold cross validation in Table 3 and Table 4 shows that the most maximum accuracy value is in fold 7 for both labelling method. Taken the 7th fold in the k = 7 parameters to see the results of the confusion matrix which can be seen in Table 4 and Table 5.

Table 4 Table Confusion Matrix Fold 7 for k = 7 from the Results of Manual Data Labeling

Actual	Prediction	
	Negative	**Positive**
Negative	36	8
Positive	14	92

Table 5 Confusion Matrix Fold 7 for k = 7 from the Results of Labeling Data with Sentiment Scoring

Actual	Prediction	
	Negative	**Positive**
Negative	30	8
Positive	14	96

After analyzing using four k value parameters, namely 3, 5, 7, and 9, the best k parameter is obtained from the results of manual data labeling and sentiment scoring is the parameter k = 7. The performance of the classification results from manual data labeling and sentiment scoring can be seen in Table 6.

Table 6 Best k Parameters for Manual Data Labeling and Sentiment Scoring

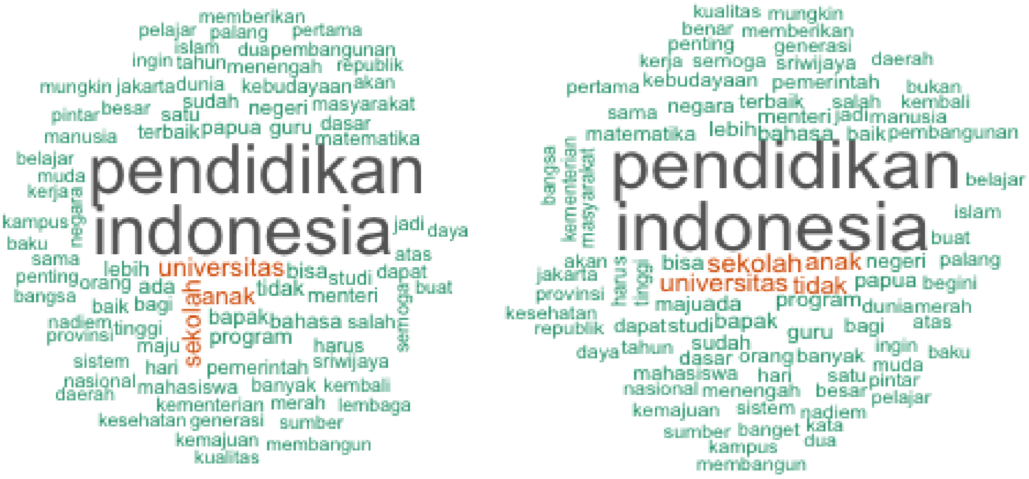
Model Evaluation	Manual (k = 7)	Sentiment Scoring (k =7)
Total Accuracy	76.93%	77.87%
Recall	71.88%	70.25%
Precision	74.12%	74.62%
F-Measure	73.00%	72.44%

Based on Table 6, the comparison between the two labeling methods based on the average total accuracy, recall, precision, and f-measure on the parameter k = 7 has the highest average total accuracy, namely 76.93% and 77.87%. The performance measurement results from data labeling with sentiment scoring have an accuracy value that is slightly greater than the accuracy value on manual labeling data. This is because the difference in the number of positive and negative classes between sentiment scoring data and manual data affects the success of the classifier in predicting the test data class. The test of two methods describes the same pattern. K-NN method has low accuracy when the value of k is small at k = 3. This is because in the small k, the data that enter in the nearest k neighbor is too small and cannot represent the class in testing data. The accuracy value of k = 5 and k = 7 indicate the increasing of accuracy value. However, in the k = 9, accuracy value decreases. Therefore, it is not always the higher the k parameter, the higher the accuracy. Through both of data labelling methods that are used and calculated its 10 fold average accuracy, it obtains the best k parameter in k = 7 that has compatible result in classifying the society's sentiment about education in Indonesia.

Based on the results and discussion can be known that the results of sentiment analysis using twitter data that has initially unstructured data managed to get the same accuracy as previous sentiment analysis research, and additional use of sentiment scoring method for data labeling proved to produce good accuracy and can be an option for other research that has a lot of data in order to save time in the process of labeling data.

The Data Visualization with Word Cloud

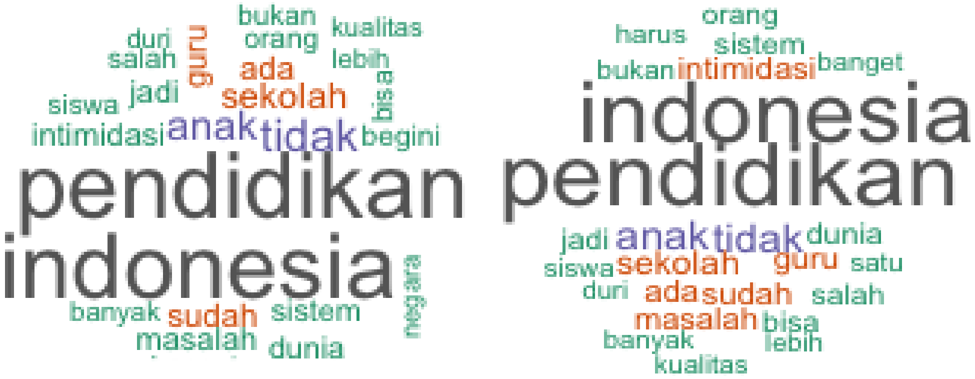
The visualization of word cloud is done using RStudio software with comparing positive sentiment tweets and negative sentiment tweets. Here are the data of word cloud concerning the topic of education in each labelling data by sentiment scoring and manual.



(a) (b)

Figure 1 (a) Word cloud positive sentiment by manual labelling, (b) Word cloud positive sentiment by sentiment scoring

Both data labelling methods of world cloud that are shown in Picture 1 (a) and 1 (b) show the words that frequently appear in the positive sentiment category. Overall, both data seem to have similar word feature that frequently appear. The most frequently words that appear in the both labelling methods are “pendidikan”, “indonesia”, “universitas”, “sekolah”, and “anak”. It shows that Twitter users mostly discuss about school and university as the important factor of child education.



(a) (b)

Figure 2 (a) Word cloud negative sentiment by manual labelling, (b) Word cloud negative sentiment by sentiment scoring

Then, both data labelling methods of world cloud that are shown in Picture 2 (a) and 2 (b) show the words that frequently appear in the negative sentiment category. Both labelling methods have the same pattern for words that frequently appear with little difference in the number of frequency of some words such as “masalah” and “intimidasi” that look bigger in the sentiment scoring labelling than in manual labelling. In the negative sentiment category, the world cloud of both labelling methods are dominated by word “anak”, “sekolah”, “guru”, “intimidasi”, and “masalah”. It indicates that Twitter users mostly discuss issues, mainly the intimidation case in the child’s school environment.

4. CONCLUSION

According to the result of the study of the responses of the society on 12 February 2020 to 19 February 2020, tends to has positive sentiment on Twitter. The result of sentiment classification from the manual and sentiment scoring data labelling result using K-Nearest Neighbors (K-NN) algorithm and 10-fold cross validation with cosine similarity distance measurements regarding education result the best accuracy level on the parameter k = 7. Through two labeling methods data can be seen that sentiment scoring is proven to be an option for labeling large amounts of data so that it is more efficient even though there are still errors in the labeling process. This can be minimized in further research by making modifications to the three labeling dictionaries. The two test methods describe the same pattern, where in this study the selection of the correct k parameter is an important factor to improve the accuracy of the classification results of the K-NN method. It is expected for further research to test more k value parameters in the K-NN model and if possible increase the number of folds for evaluation, so that the resulting sentiment analysis is likely to be better.

REFERENCES

Hearst, M. (2003). *What Is Text Mining?*

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data. In *Discovering Knowledge in Data*. https://doi.org/10.1002/9781118874059

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Http://Dx.Doi.Org/10.2200/S00416ED1V01Y201204HLT016*, *5*(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Putrianti, R. P., Kurniati, A., & Agustin, D. (2019). Implementasi Algoritma K - Nearest Neighbor Terhadap Analisis Sentimen Review Restoran Dengan Teks Bahasa Indonesia. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, *0*(0). https://journal.uii.ac.id/Snati/article/view/13397

Teljstedt, E. C. (2016). *Separating Tweets from Croaks (Detecting Automated Twitter Accounts with Supervised Learning and Synthetically Constructed Training Data)*. www.kth.se/csc

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. In *Knowledge and Information Systems* (Vol. 14, Issue 1). https://doi.org/10.1007/s10115-007-0114-2