IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-9, Sept 2019 E-ISSN: 2347-2693

Comparison of Text Classification Algorithms of People Sentiments on Twitter (Case: Transjakarta)

Rexzy Tarnando[1*], Yuli Karyanti[2]

[1]Dept. of Computer Science and Information Technology, Gunadarma University, Depok, Indonesia
[2]Information Technology, Gunadarma University, Depok, Indonesia

[*]*Corresponding Author: rexzyt@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i9.812 | Available online at: www.ijcseonline.org

Accepted: 13/Sept/2019, Published: 30/Sept/2019

***Abstract*—** Nowadays social media is one to express things that are thought and felt by the community. One of the things that's much talked about is responses from the consumer of products or services. This is very useful for companies to find out the level of satisfaction of their products or services. Twitter is one of the most widely used social media by users. With this fact, it's really interesting for companies to use the data on Twitter for the company's progress generally in customer relations. In this study an analysis of public sentiments towards the use of Transjakarta. This study divides community sentiments into three classes, positive, neutral and negative. For data taken from Twitter with the results of research from June to July 2019 by dividing the data into training data and testing data. The amount of training data is 144 tweets and testing data are 36 tweets. Then for the text classification uses 3 algorithms, namely naïve bayes, k-nearest neighbor and logistic regression. Then after the results are obtained, next is to compare the performance levels of three methods by finding the highest f-measurement value using micro average formula. Micro average is chosen because it's the best for calculating imbalanced datasets. The results show the naïve bayes method has the best f-measurement with 0.861 value. For the next largest f-measurement value is the logistic regression method with an f-measurement value of 0.833, and the last is the k-nearest neighbor method with an f-measurement value of 0.806.

***Keywords*—** Naïve Bayes, K-Nearest Neighbor, Logistic Regression, F-measurement, Sentiment Analysis, Transjakarta

I. INTRODUCTION

The development of social media at this time has a considerable influence on social life, especially in Indonesia. Many people prefer social media to voice their thoughts. What can also be done and found on social media is people's responses to the products and services they have used. Film reviews, restaurant reviews, and even responses after using the mode of transportation. Twitter is one of the social media or microblog that is widely used today, especially users can argue there, so people can find out things that are the current trend, people's opinions and so forth. Data shows that as of September 2019 Twitter total monthly active users reached 330 million people with the number of tweets reaching 500 million tweets per day (omnicoreagency.com).

By using technology, it can be known that tweets written by Twitter users whether the tweets are positive, neutral or maybe negative responses. The technology that can classify responses on social media is sentiment analysis. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language[1]. There are many methods for classification, and for this study, the naïve bayes, k-nearest neighbor and logistic regression methods are used.

Naïve Bayes, k-nearest neighbor and logistic regression are methods that are often used in classification techniques. For this study, the three methods the researchers used in sentiment analysis were to determine community sentiment towards the use of Transjakarta. Transjakarta itself is one of the main modes of transportation in Jakarta. According to data taken from Kompas media, in 2018, Tranjakarta users reached 189.77 million people, this number rose by 31 percent compared to the previous year (Harian Kompas, 2019). Of course from this large number, it can be estimated that many users provide opinions, criticisms and suggestions for Transjakarta services, both directly given to Transjakarta or through social media, such as Twitter. This can be seen from the many people tweet to @transjakarta twitter account. From the results of the sentiment analysis, we will look for the best classification method from the three methods. Researchers hope to provide a little information for researchers and students who will later conduct research, especially on sentiment analysis which is the algorithm

which is better for classification techniques, even though this research does not necessarily provide definitive results that one method is better than the method the other, because in this study only used one object, the Transjakarta tweet.

II. RELATED WORK

Previous research that discusses the text classification and comparison of algorithms used, summarized as discussed below.

- Research conducted by M. Trivedi, N Soni, S. Sharma and S. Nair discusses the comparison of methods used in text classification. The method used for this research is the Naïve Bayes algorithm, C4.5 and Support Vector Machine. For data used regarding diabetes and calorie issues. The results show that the SVM algorithm has the best level of accuracy when the number of datasets or attributes are used a lot, while the SVM algorithm also has a smaller accuracy than the other two algorithms when the number of datasets or attributes is small [2].
- Research conducted by Kirti Couksey and Amit Ranjan has succeeded in making a technology of community sentiment analysis of elections in India. This study aims to determine the number of negative and positive sentiments from the general election in India. The algorithm used in this study consists of Naïve Bayes, Decision Tree, KNN, Logistic Regression, SVM, Adaboosted and Proposed algorithms[3].
- Research conducted by Bhawna Sharma, Sheetal Gandotra, Utkarsh Sharma et al discusses the comparison of machine learning algorithms to predict Chronic Kidney Disease. The algorithms used in this study include the Logistic Regression, SVM, KNN, Naïve Bayes, Stochastic Gradient Descent Classifier, Decision Tree and Random Forest algorithms. This study provides the results that the Logistic Regression algorithm, SGD Classifier, and Random Forest provide the best accuracy value compared to other algorithms [4].
- In this research, a web text classification application was made using the Naive Bayes algorithm. In this study divides categories into 10 categories such as sports news, games and so on. In this study, an accuracy rate of 96% was obtained so that it can be concluded that the application made was quite good[5].
- This study makes a sentiment analysis of movie reviews. The algorithms used in this research are Naive Bayes and SVM algorithms. In this study, to get good accuracy results is done by increasing the pre-processing features. The filter used in this study is a String to Word Vector with an accuracy for Naive Bayes and SVM of 81% and 82.1%. Then by using the Custom String to Word Vector filter get an accuracy value of 82.2% and 84.9% for Naive Bayes and SVM[6].

III. METHODOLOGY

The data taken is data from Twitter that is shown to the @transjakarta account. Transjakarta is one of the first Bus Rapid Transit (BRT) transportation modes in Southeast and South Asia which has been operating since 2004 in Jakarta, Indonesia. Data pulled from Twitter API. After the data is pulled, then the data is preprocessing. The preprocessing stage used is case folding, text cleaning, stopword removal, tokenizing and normalization.

Case folding is useful for shrinking letters to make it more easily recognized by computers for later calculations. Stopword removal works to get rid of words like "and", "you" etc. Tokenizing is useful for breaking sentences into words (tokens). Normalizatio to correct wrong words (typo)
After the preprocessing stage is completed, to proceed to the next stage is to create training data. The training data was taken from June-July 2019 with a total of 144 data tweets. Then the data is labeled, 2 for positive tweets, 1 for neutral tweets and 0 for negative tweets.

The final stage is to create a text classification program using the Naïve Bayes, K-Nearest Neighbor and Logistic Regression method. After the application is made, the next step is to compare the performance level of each method.

IV. RESULTS AND DISCUSSION

After all the initial steps to make the text classification machine learning run, the next thing is to make an application using the python programming language with the Naïve Bayes, K-Nearest Neighbor and Logistic Regression method. In the K-Nearest Neighbor algorithm using the value k = 15, because with these values get the optimal accuracy value for this method. It should be noted that the number of training data is 144 tweets and testing data is 36 tweets. Below is an image of the output program that has been made.

	Actual	Prediction
0	pos	pos
1	pos	pos
2	pos	pos
3	pos	pos
4	pos	net
5	pos	pos
6	pos	pos
7	pos	pos
8	pos	pos
9	pos	pos
10	pos	pos
11	pos	pos
12	net	net
13	net	net
14	net	net
15	net	net
16	net	net
17	net	net
18	net	net
19	neg	neg
20	neg	neg
21	neg	pos
22	neg	neg
23	neg	net
24	neg	neg
25	neg	pos
26	neg	neg
27	neg	neg
28	neg	neg
29	neg	neg
30	neg	neg
31	neg	neg
32	neg	pos
33	neg	neg
34	neg	neg
35	neg	neg

Figure 1. Naïve Bayes Program Output

	Actual	Prediction
0	pos	pos
1	pos	pos
2	pos	pos
3	pos	pos
4	pos	net
5	pos	pos
6	pos	pos
7	pos	pos
8	pos	pos
9	pos	pos
10	pos	pos
11	pos	pos
12	net	net
13	net	pos
14	net	net
15	net	net
16	net	neg
17	net	net
18	net	net
19	neg	neg
20	neg	neg
21	neg	net
22	neg	net
23	neg	neg
24	neg	neg
25	neg	net
26	neg	neg
27	neg	neg
28	neg	neg
29	neg	neg
30	neg	neg
31	neg	neg
32	neg	neg
33	neg	neg
34	neg	neg
35	neg	net

Figure 2. K-Nearest Neighbor Program Output

	Actual	Prediction
0	pos	net
1	pos	pos
2	pos	pos
3	pos	pos
4	pos	net
5	pos	pos
6	pos	pos
7	pos	pos
8	pos	pos
9	pos	pos
10	pos	pos
11	pos	pos
12	net	net
13	net	net
14	net	net
15	net	net
16	net	net
17	net	net
18	net	net
19	neg	neg
20	neg	neg
21	neg	pos
22	neg	net
23	neg	neg
24	neg	neg
25	neg	net
26	neg	neg
27	neg	neg
28	neg	neg
29	neg	neg
30	neg	neg
31	neg	neg
32	neg	neg
33	neg	neg
34	neg	neg
35	neg	net

Figure 3. Logistic Regression Program Output

Based on the 3 pictures above, a confusion matrix table can be made as follows:

Table 1. Confusion Matrix of Naïve Bayes Algorithm

		Prediction		
		Pos	**Net**	**Neg**
Actual	**Pos**	11	1	0
	Net	0	7	0
	Neg	3	1	13

Table 2. Confusion Matrix of K-Nearest Neighbor Algorithm

		Prediction		
		Pos	**Net**	**Neg**
Actual	**Pos**	11	1	0
	Net	1	5	1
	Neg	0	4	13

Table 3. Confusion Matrix of Logistic Regression Algorithm

		Prediction		
		Pos	**Net**	**Neg**
Actual	**Pos**	10	2	0
	Net	0	7	0
	Neg	1	3	13

To compare the three methods is to calculate the level of performance of each method. According to research conducted by titled "Evaluation Measures for Models Assessment over Imbalanced Data Sets" and "Imbalanced text classification: A term weighting approach" was told that the best way to measure the performance of imbalanced datasets is to use f-measurements. And to find the value of f-measurement for imbalanced datasets used micro average calculations to give better results [7,8]. The f-measurement calculations for each method are as follows.

A. Naïve Bayes

$$\text{Precision} = \frac{\sum \text{TPi}}{\sum (\text{TPi} + \text{FPi})}$$

$$\text{Precision} = \frac{11 + 7 + 13}{11 + 7 + 13 + 3 + 2 + 0} = 0.861$$

$$\text{Recall} = \frac{\sum \text{TPi}}{\sum (\text{TPi} + \text{FNi})}$$

$$\text{Recall} = \frac{11 + 7 + 13}{11 + 7 + 13 + 1 + 4 + 0} = 0.861$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1-Score} = 2 \times \frac{0.861 \times 0.861}{0.861 + 0.861} = 0.861$$

B. K-Nearest Neighbor

$$\text{Precision}=\frac{\sum TPi}{\sum (TPi+FPi)}$$

$$\text{Precision}=\frac{11+5+13}{11+5+13+1+5+1}=0.806$$

$$\text{Recall}=\frac{\sum TPi}{\sum (TPi+FNi)}$$

$$\text{Recall}=\frac{11+5+13}{11+5+13+1+2+4}=0.806$$

$$\text{F1-Score}=2\ x\ \frac{\text{Precision x Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{F1-Score}=2\ x\ \frac{0.806\ x\ 0.806}{0.806+0.806}=0.806$$

C. Logistic Regression

$$\text{Precision}=\frac{\sum TPi}{\sum (TPi+FPi)}$$

$$\text{Precision}=\frac{10+7+13}{10+7+13+1+5+1}=0.833$$

$$\text{Recall}=\frac{\sum TPi}{\sum (TPi+FNi)}$$

$$\text{Recall}=\frac{10+7+13}{10+7+13+2+0+4}=0.833$$

$$\text{F1-Score}=2\ x\ \frac{\text{Precision x Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{F1-Score}=2\ x\ \frac{0.833\ x\ 0.833}{0.833+0.833}=0.833$$

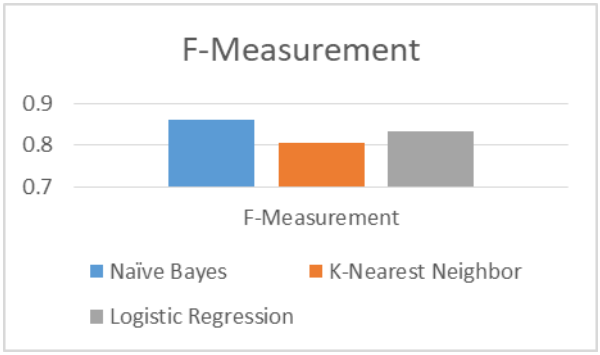

Figure 4. F-Measurement Value of Each Algorithms

The calculation shows that the naïve bayes method gets the highest value with an f-measurement value of 0.861, then logistic regression in the second position with an f-measurement value of 0.833, and the last is the k-nearest neighbor method with an f-measurement value of 0.806.
From the above results it can be explained that the Naive Bayes algorithm is the best algorithm in this research.

V. CONCLUSION AND FUTURE SCOPE

In this study, an application was made to analyze public sentiment towards the use of TransJakarta. The data used for this application is taken from twitter with the keyword or mention to the @transjakarta account between June - July 2019 with a total of 144 training data and 36 test data. Then the results of the comparison of the performance level of the algorithm can be seen that the naïve Bayes algorithm gets the highest f-measurement value with a value of 0.861. Furthermore, there is a logistic regression algorithm with a f-measurement value of 0.833 and the last is the k-nearest neighbor algorithm with an f-measurement value of 0.806.

From this research can be seen that Naïve Bayes is the best algorithm, but it can still be concluded that these three methods are very good to be used in text classification because the differences in the three values are not too far apart.

This research can still be developed by subsequent researchers using larger data, so that machines can learn better with large amounts of data and machines will provide better accuracy. This research is still difficult to determine ambiguous words, so the next researcher can combine supervised and unsupervised learning to give maximum results in ambiguous sentences.

REFERENCES

[1] B. Liu, "*Sentiment Analysis and Opinion Mining*", Synthesis Lectures on Human Language Technologies, Vol.**5,** No.**1**, pp.**1-167**, **2012**.

[2] M. Trivedi, N. Soni, S. Sharma. S. Nair, "*Comparison of Text Classification Alghorithms*", International Journal of Engineering Research & Technology (IJERT), Vol.**4,** Issue.0**2**, pp.**334-336**, **2015**.

[3] K. Chouksey, A. Ranjan, "*Analysis of Indian Election using Random Forest Algorithm*", International Journal of Computer Sciences and Engineering, Vol.**7,** Issue.**10**, pp.**50-57**, **2019**.

[4] B. Sharma, S. Gandotra, U. Sharma, R. Thakur, A. Mahajan, "*A Comparative Analysis of Different Machine Learning Classification Algorithms for Predicting Chronic Kidney Disease*", International Journal of Computer Sciences and Engineering, Vol.**7,** Issue.**6**, pp.**8-13**, **2019**.

[5] S.S. Bhadoria, R.K. Patel, "*Web Text Content Extraction and Classification using Naïve Bayes Classifier Algorithm*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**2,** Issue.**5**, pp.**1-4**, **2014**.

[6] K. Sarvakar, U.K. Kuchara, "*Sentiment Analysis of movie reviews: A new feature-based sentiment classification*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**6,** Issue.**3**, pp.**8-12**, **2018**.

[7] M. Bekkar, Dr.H.K Djemaa, Dr.T.A. Alitouche, *"Evaluation Measures for Models Assessment over Imbalanced Data Sets"*, Journal of Information Engineering and Applications, Vol.**3**, No.**10**, pp.**27-38**, **2013**.

[8] Y. Liu, H.T. Loh, A. Sun, *"Imbalanced text classification: A term weighting approach"*, Expert Systems with Applications, Vol.**36**, No.**1**, pp.**690-701**, **2009.**

Authors Profile

Rexzy Tarnando got his bachelor's degree in Information System at Gunadarma University in 2017. He is currently studying a master's degree in Gunadarma University with major Business Information System and working as a Business Analyst in IT Consultant company.

Yuli karyanti got her bachelor's, master's and doctoral degree in Information Technology at Gunadarma University. She has joined Gunadarma University in the UPT department since 2000 and has been a lecturer since 2005. Her research interest is image retrieval based on texture with dynamic segmentation.

