Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2018 13 (6): 277-284

Tweet Clustering in Indonesian Language Twitter Social Media using Naive Bayes Classifier Method

Rizal Tjut Adek, Sahlan Nasution

Received 14 September 2018 ▪ Revised 23 October 2018 ▪ Accepted 24 November 2018

Abstract: Twitter is one of the social media that has been widely used for various purposes, especially to facilitate the means of information, communication, entertainment and a means of expressing expression. We can find various kinds of information on twitter such as culture, sports, culinary, tourism, music, politics and others. The purpose of this research is to build an application that can group tweets from twitter into sports and non-sports categories using the Naive Bayes classifier method. Text mining is a technique used to handle classification, clustering, information extraction and information retrieval problems. To classify tweets from twitter automatically needed one of the mining Clustering text techniques. Learning outcomes in the form of probabilities will be used as material for processing tweet documents that are not yet known in the category. In the process, the tweet document will go through a text pre-processing process, and grouped into unigram (one word), bigram (two words), trigram (three words). For determining the category of a tweet document that is not yet known, the comparison is made between the results of the appearance of the categories of the three n-grams. From the results of testing the system using 100 to 2000 training data in each category, and 10 testing data in each category. The result is the accuracy of tweets that are categorized as 60% in training data as much as 100, accuracy of 65% in training data as much as 200, and accuracy of 90% in training data as much as 2000. The conclusion is that the more training data used as learning increases also the success rate of clusters to a tweet document.

Keywords: *Twitter, Clustering.*

INTRODUCTION

Social media networks have gained a lot of attention in recent years in terms of analysis for the use and detection of abnormal activities that occur on the internet (Kaur & Singh, 2016). Social media is an online media, with its users can easily participate, share and create content including blogs, social networks, wikis, forums and virtual worlds (Saif, He, Fernandez, & Alani, 2016). Blogs, social networks and wikis are the most common forms of social media used by people around the world (Zhang, Yu, & Meng, 2007).

Twitter is one of the social media that has been widely used for various purposes, especially to facilitate the means of information, communication, entertainment and a means of expressing expression (Ren, Wang, & Ji, 2016). We can find various kinds of information on twitter such as culture, sports, culinary, tourism, music, politics and others (Da Silva, Hruschka, & Hruschka, 2014; Ren et al., 2016). Data obtained through twitter are text, images and videos (T F Abidin, Hasanuddin, & Mutiawani, 2017).

The growth of Twitter usage causes the growth of digital data that exists in cyberspace (Yang & Ko, 2011). Everyone can now create an account on Twitter, which results in each user being able to upload news to twitter which will be consumed by their followers. This abundant data will make it difficult for Twitter users to classify or classify news that will be read through their twitter accounts (Bello-Orgaz,

Rizal Tjut Adek, Department of Information Technology, Malikussaleh University.
Sahlan Nasution, Department of Information Technology, Malikussaleh University.

Hernandez-Castro, & Camacho, 2017). Twitter has also been used to detect epidemics based on rumors developed in the Twitter network (Sicilia, Lo Giudice, Pei, Pechenizkiy, & Soda, 2018).

With the increasing trend of online social networks in different domains, social network analysis has recently become a research center (Taufik Fuadi Abidin, Ferdhiana, & Kamil, 2013). The Online Social Network (OSN) has attracted the interest of researchers to analyze its use and detect abnormal activity (Hasdina & Tjut Adek, 2016). Classification by (Kaur & Singh, 2016) an anomalous activities on social networks shows unusual and illegal activities that show behavior that is different from the others even though in the same structure. In (Dong et al., 2014), the Adaptive Recursive Neural Network approach is used in classifying twitter. An ensemble classifier has been proposed by (Ankit & Saleena, 2018) which combines the basic learning classifier to form a single classifier, with the aim of improving the performance and accuracy of sentiment classification techniques. The results show that the proposed ensemble grouping performs better than stand-alone standard classifiers.

In classifying a tweet, it can actually be determined by looking at the value of the tweet. but not all tweets give a hagg to the tweet (Zhou, He, & Wu, 2014). There is also a gift given, but the contents of the tweet do not exist at all with respect to the tag (T F Abidin et al., 2017). This is what makes us want to make a tweet classifier based on the content of the tweet.

Tweet data in this study were obtained by using the Application Programming Interface which will then be referred to as the API provided by Twitter. By utilizing the API an application is built to retrieve Tweet data. The data will be grouped based on the characteristics or characteristics that will be determined and will become information.

Based on the background description above, this paper will design and build a system that can classify (classify) tweet data in the form of writing into Sports or Non-Sports categories based on the characteristics of each category. The results of this study are in the form of Tweet information and the Tweet Category that has gone through the Cluster process.

METHODOLOGY

The object taken in this study is Twitter. Twitter is taken as an object because the tweet text that is posted is shorter and the purpose is more specific. The research material used in this research is tweet text posted on Twitter social media. Activities carried out in tweet data retrieval are by visiting the twitter website.

The Clustering Tweet application that will be built is web based using the Naive bayes classifier algorithm to classify tweet data in the form of text into Sports or Non-Sports categories based on the characteristics of each category. The first step that needs to be done in building this system is to get training data in the form of text tweets that will later go through a training process and generate probabilities in each word.

At the learning stage or training the processes carried out are as follows:

1. Input Tweet training based on the category of tweets stored in the database.
2. Then the system will do the text preprocessing process.
3. After doing the text preprocessing process the system does the filtering process, namely the stopword removal.
4. Then the system performs the stemming process.
5. The word stemming is then searched for by the N-gram, he said. Word n-grams are processed into Uni-gram, Bi-gram, and Tri-gram. Uni-gram, Bi-gram, and Tri-gram words appear compared to Uni-gram, Bi-gram, and Tri-gram words that are in the database.
6. If appropriate, then add the number of occurrence frequencies Uni-gram, Bi-gram, Tri-gram, word (nk). If it does not match, then the word is used as Uni-gram, Bi-gram, or Tri-gram new word and add the number of occurrences of the word (n_k).
7. Calculate the probability of each Uni-gram, Bi-gram, and Tri-gram word ($P(x_i|V_j)$) which is the probability x_i in the V_j category. To calculate, use

$$P(x_i|V_j) = \frac{n_k+1}{n+|vocabulary|} \quad (1)$$

where :

$P(x_i|V_j)$ = The n-word probability of the k in the category V_j

n_k = the number of times the occurrence of each word

n = the number of times the occurrence of words from each category

vocabulary = sum of all words from all categories

8. If the stem word list results from more than zero, the process will return to step number 5. Otherwise, the process will continue to the next step.
9. Add the number of documents.
10. Calculate the probability of Tweet documents in each category ($P(V_j)$).
11. The result is the probability value of each Uni-gram, Bi-gram, Tri-gram word and the probability value of Tweet for each Tweet category.
12. The training process is complete.

The process of training data tweets that will be used in this study can be described in Figure 1 below:

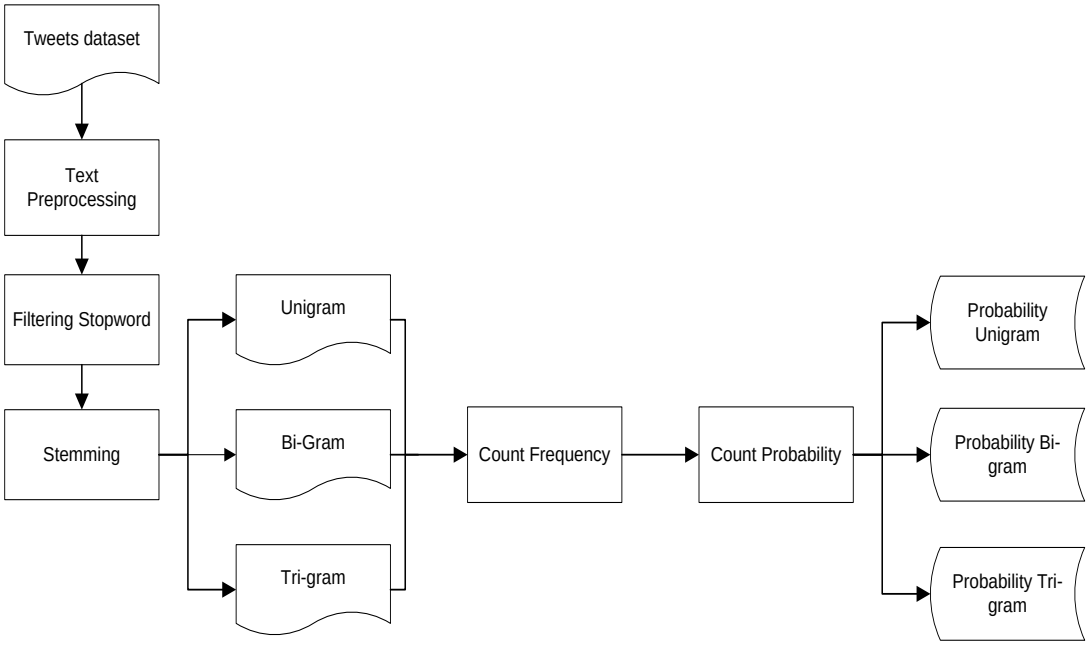

Fig.1: Data training process

Whereas in the training phase the processes carried out are as follows :

1. Input a tweet.
2. System will do *text preprocessing*.
3. After doing the text preprocessing process the system does the filtering process, namely the stop word removal.
4. Furthermore, the system carries out the stemming process, which is to change the word that has become a basic word.
5. The word stemming results from N-gram search for it. Word n-grams are processed into Uni-gram, Bi-gram, and Tri-gram. Uni-gram, Bi-gram, and Tri-gram words appear compared to Uni-gram, Bi-gram, and Tri-gram words that are in the database. If appropriate, the probability values Uni-gram, Bi-gram, and Tri-gram words from Uni-gram, Bi-gram, and Tri-gram words which are in the keyword database are used as word probability values ($(P(x_i \mid V_j))$ If not, the occurrence frequency of Uni-gram, Bi-gram, and Tri-gram word (nk) is zero and the probability value of Uni-gram, Bi-gram, and Tri-gram words ($(P(x_i \mid V_j))$ calculated.
6. If the stem word list results from more than zero then the process will return to step number 5, If not, the process will continue to the next step.
7. Calculate the VMAP value from each category of tweets on the keyword Uni-gram, Bi-gram, and Tri-gram.
8. Find for the highest VMAP value between Sports or Non-Sports categories in each keyword Uni-gram, Bi-gram, and Tri-gram.
9. Compare the results of each keyword Uni-gram, Bi-gram, and Tri-gram. The testing results are determined from the many occurrences of the categories of each keyword result.
10. Displays the Tweet category.
11. The testing process is complete.

Whereas in the testing phase the processes carried out are as follows:

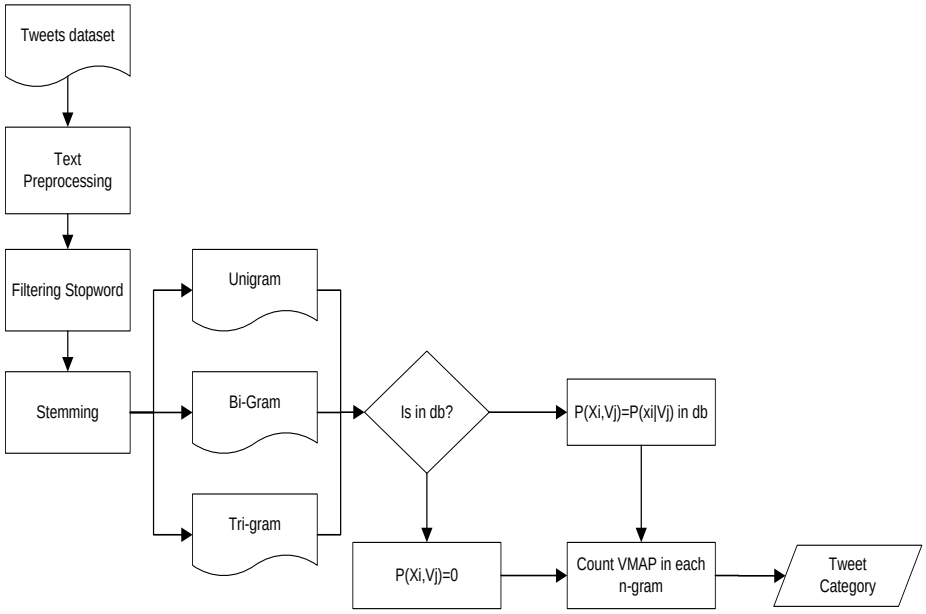

Fig.2: Testing Process

In this study Unified Modelling Language (UML) is used as a modelling language to design and design a tweet clustering system on Twitter social media. The UML model used is a use case diagram and activity diagram. Use Case Diagram in Figure 3 will explain what the system will do. Therefore, the use case diagram will present how the interaction between the user and the system. In the system that is built, it has an Active actor, User, the following is a picture of the use case diagram

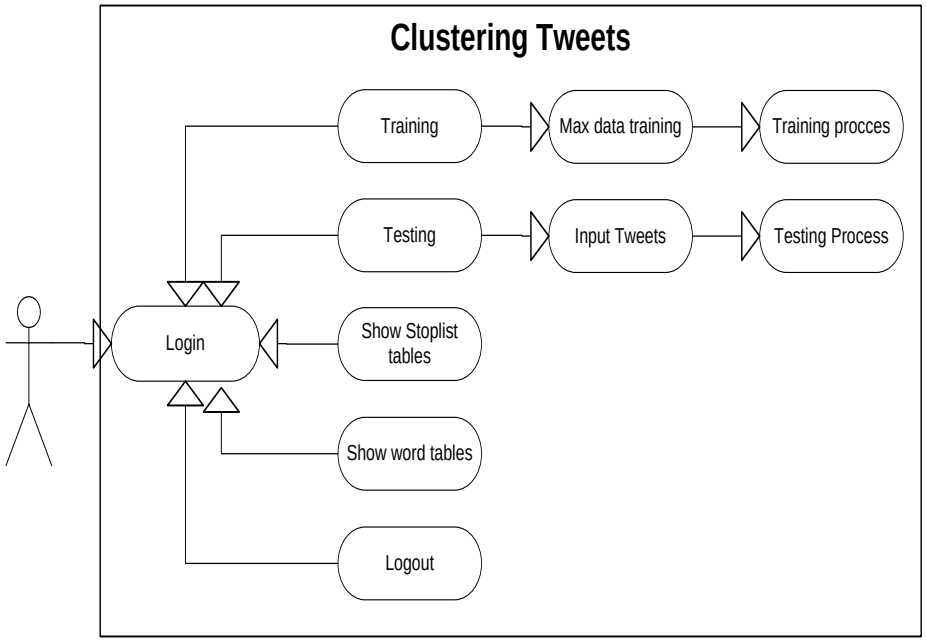

Fig.3: Use case diagram

The interactions that run on this tweet clustering system include Login sequence diagrams, Training Pages, Testing Pages. Test page sequence diagrams in figure 4. explain the testing process on the testing page. The test page menu is a page for clustering a tweet whose category is unknown. After the user selects the test page menu, the system will display a testing page. Then the user will input the testing data in the form of text. The system will start the testing process, by calling the data needed from the database, namely, stop word database, basic database, database, unigram, bigram database, and trigram database. The testing process includes text preprocessing, filtering, stemming and the naive bayes classifier process. After the testing process is complete, the system will send the testing results to the testing page and display the testing results on the testing page.

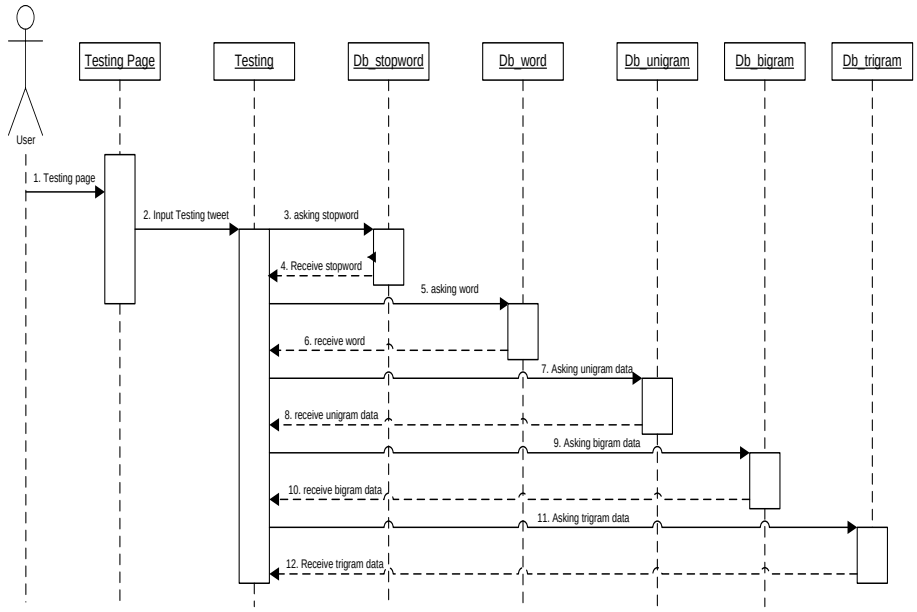

Fig.4: Sequence diagram testing

Class Diagrams describe the state (attributes / properties) of a system, while offering services to manipulate the situation (method / function). Class diagrams on this system can be seen in figure 5 below:

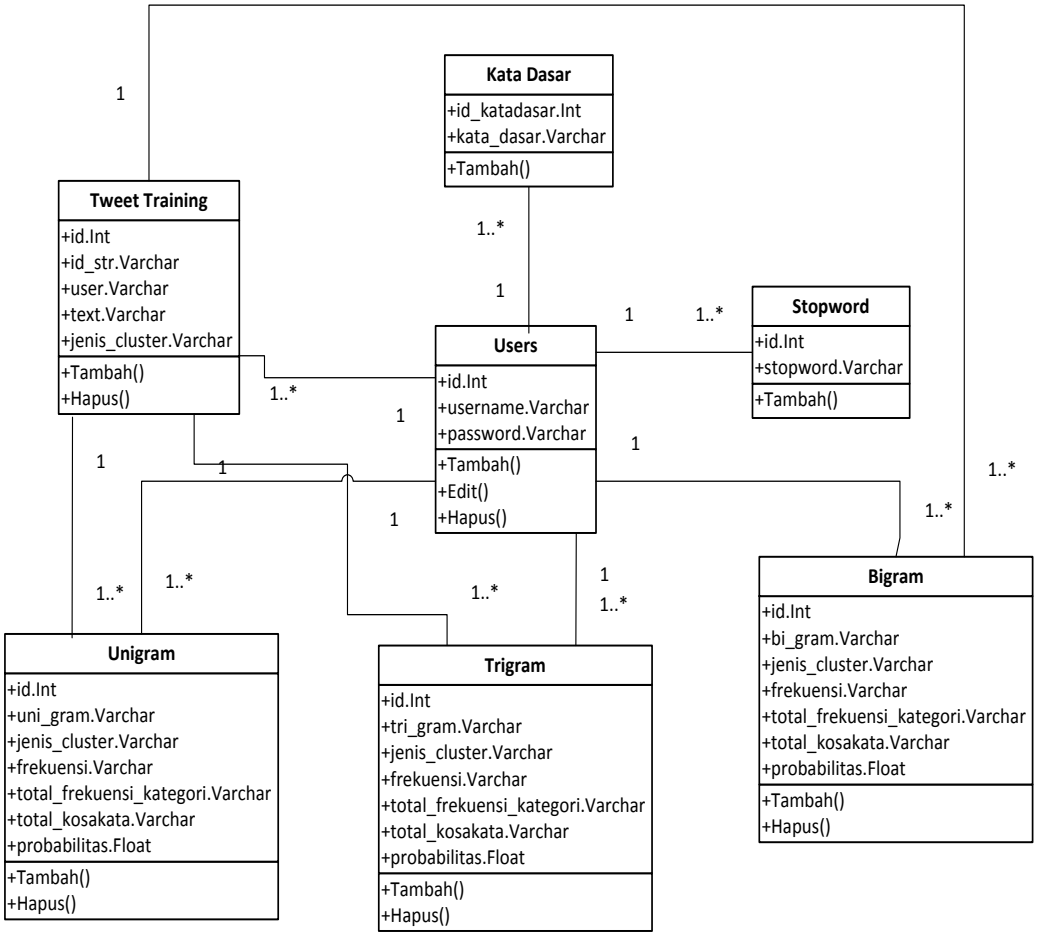

Fig.5: Class diagram

IMPLEMENTATION

In the training process, for each category there will be 3 types of training, namely unigram, bigram, trigram. Unigram is an n-gram consisting of one word, bigram (n-gram two words), trigram (n-gram three words). Later the probability data obtained will be stored in the database uni_gram, bi_gram, and tri_gram. Calculations to find probabilities in the training and testing process using the Naive Bayes classifier algorithm use the following equation:

$$P(x_i|V_j) = \frac{n_k + 1}{n + |vocabulary|}$$

Especially for |vocabulary| is the total number of words contained in the training process, namely the combination of the number of unigram words found in the sports category and non-sport categories. For the same word in different categories still counted 1.

From these values we can find the probability value of each unigram sport word in the training data by using the formula P (xi | Vj) which is like the example shown in table 1 below:

Table 1: Probability of Sports Category Unigram Words

No	Unigram	Frekuensi (n_k)	Probabilitas
1	selalu	1	0,0952
2	sepak	1	0,0952
3	bola	2	0,1428
4	teman	1	0,0952
5	fifa	1	0,0952
6	baru	1	0,0952

At the stage of the testing process begins by looking for the probability of each word in the document containing a tweet whose category is unknown. In this testing process, 3 testing will be carried out, which are Unigram, Bigram, and Trigram. Later the category in each n-gram will be searched. Then the results of each n-gram will be compared and obtained by the final category which will be the result of testing the document. Test results obtained in this test will be explained in the 2 below:

Table 2: Testing result

Training Data		***Testing*** Data		***Testing***				
Sport	**Non sport**	**Sport**	**Non sport**	**Sport**		**Non sport**		**Accuracy**
				T	**F**	**T**	**F**	
100	100	10	10	7	3	5	5	60%
200	200	10	10	7	3	6	4	65%
300	300	10	10	7	3	6	4	65%
400	400	10	10	8	2	6	4	70%
500	500	10	10	8	2	6	4	70%
600	600	10	10	8	2	7	3	75%
700	700	10	10	8	2	7	3	75%
800	800	10	10	8	2	8	2	80%
900	900	10	10	8	2	8	2	80%
1000	1000	10	10	9	1	8	2	85%
2000	2000	10	10	9	1	9	1	90%

The tableshows that first of all 100 training data are conducted in each category, then testing using 10 tweets in each category. The result, in the Sports category has 7 tweets that are True and 3 tweets are

False, while in the Non Sports category there are 5 tweets that are True and 5 tweets are False. So as to produce an accuracy value of 60%.

In the next step, the system was re-tested by using 10 test tweets in each category by adding training data every multiple of 100 to 2000 in each category. The result is the accuracy of the tweet that can be categorized as a system of 65% in the training data of 200 and accuracy of 90% in the training data of 2000

Clustering Tweet Application on Social Media Twitter uses the Naive Bayes classifier method which is built using web programming languages namely html, php and javascript. As for the database using the MySQL database.

The Training page serves to train the tweets that are already known in the category. The display of this training page can be seen in Figure 6 below:

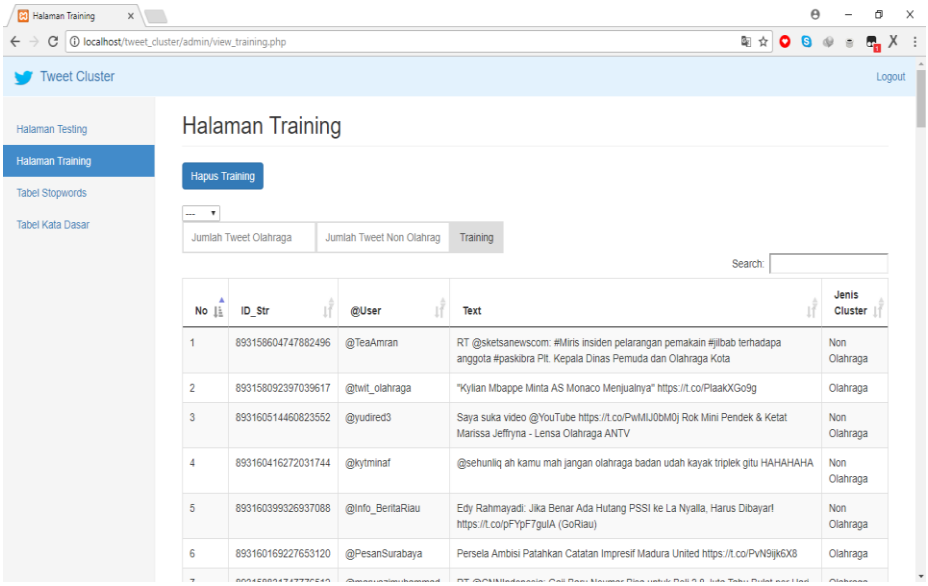

Fig.6: Training interface

On the training page above the user will do training on the tweet data that has been known to the category which will be used as learning in the testing process. Users will input how much data will be trained for each category, then choose the Training button. The system will process it.

The Testing page functions to test a tweet. The display of this testing page can be seen in the following figure 7:

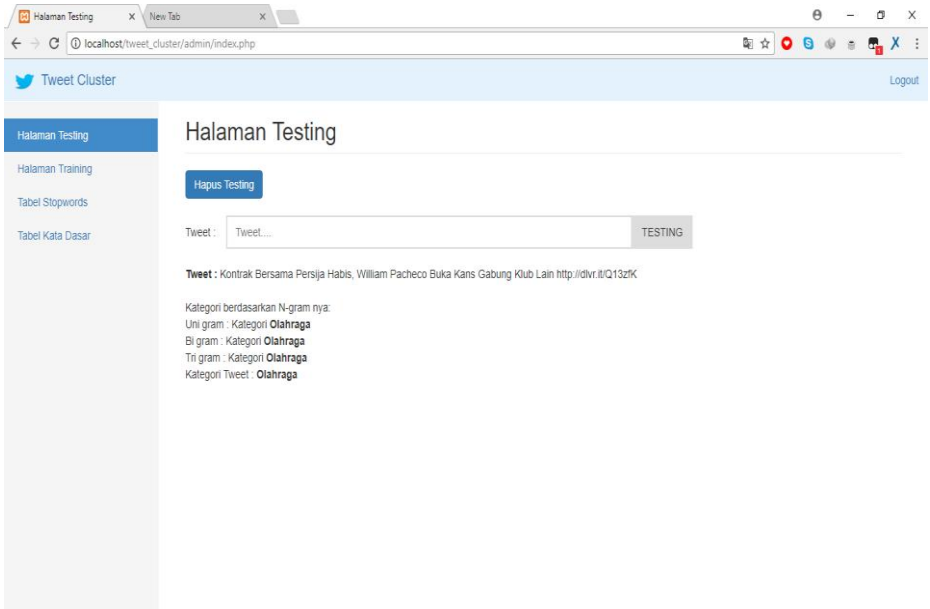

Fig.7: Testing interface

On the page above the user will enter a tweet whose category is unknown in the textbox, then choose the Testing button. The system will process the tweet and then display the results below.

CONCLUSION

Based on the analysis of the system and system testing as a whole that has been done in the previous chapters, then there are some things that can be concluded in this study, among others: The clustering process tweets on Twitter social media in general through 2 processes, namely the training and testing process. Before doing this process, tweet data in the form of text through the text preprocessing process. Then, the sentence is separated into n-gram words in the form of unigram, bigram, and trigram. The probability is searched using the Naive Bayes Classifier method. Based on the existing training probability value, a tweet was tested which was not known in the category. By seeing the emergence of more categories of tweets than the n-gram word unigram, bigram, trigram. The system was tested using 10 test tweets in each category and training data starting from 100 data and adding training data every multiple of 100 to 2000 in each category. The result is the accuracy of tweets that can be categorized as 60% in the training data of 100, accuracy of 65% in the training data of 200, and accuracy of 90% recorded in 2000 training. The classification process is more accurate if the training data used in learning is many However, it can also reduce accuracy if the words contained in the Tweet experience bias or double meaning.

REFERENCES

[1] Abidin, T. F., Ferdhiana, R., & Kamil, H. (2013). Automatic extraction of place entities and sentences containing the date and number of victims of tropical disease incidence from the web. *Journal of Emerging Technologies in Web Intelligence, 5*(3), 302–309.

[2] Abidin, T. F., Hasanuddin, M., & Mutiawani, V. (2017). N-grams based features for Indonesian tweets classification problems. *In International Conference on Electrical Engineering and Informatics (ICELTICs)*, 307–310.

[3] Ankit, & Saleena, N. (2018). An Ensemble Classification System for Twitter Sentiment Analysis. *Procedia Computer Science*, *132*, 937–946.

[4] Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems, 66*, 125–136.

[5] Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems, 66*, 170–179.

[6] Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. *In Proceedings of the 52nd Annual Meeting of Association for Computational Linguistics*, 49–54.

[7] Hasdina, N., & Tjut Adek, R. (2016). Implementasi Metode Cusum (Cummulative Summary) Untuk Menentukan Daerah Rawan Kecelakaan Berbasis Web Di Kota Lhokseumawe. *Techsi, 8*(1), 227–239.

[8] Kaur, R., & Singh, S. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal, 17*(2), 199–216.

[9] Ren, Y., Wang, R., & Ji, D. (2016). A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences, 369*, 188–198.

[10] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management, 52*(1), 5–19.

[11] Sicilia, R., Lo Giudice, S., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications, 110*, 33–40.

[12] Yang, S., & Ko, Y. (2011). Extracting Comparative Entities and Predicates from Texts Using Comparative Type Classification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1*, 1636–1644

[13] Zhang, W., Yu, C., & Meng, W. (2007). Opinion retrieval from blogs. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management - CIKM '07*.

[14] Zhou, D., He, W. H., & Wu, T. T. (2014). The Research on Tibetan Text Classification Based on N-Gram Model. *Applied Mechanics and Materials*, 543–547.