Journal of Engineering and Applied Sciences 14 (4): 1182-1188, 2019
ISSN: 1816-949X

Geovisualization Way for Exploiting Customer's Emotions on Twitter

[1]Mochamad Nizar Palefi Ma'ady, [2]Arif Djunaidy and [2]Renny Pradina Kusumawardani
[1]Department of Informatics Engineering, Universitas Nahdlatul Ulama Sunan Giri,
Ahmad Yani No. 10, Bojonegoro, Indonesia
[2]Institut Teknologi Sepuluh Nopember, Department of Information Systems,
Raya ITS No. 1, Surabaya, Indonesia

Abstract: Tight competition among companies has been making many companies to increase their efforts to collect information for the sake of their service analysis. Recorded data in Twitter can be used as a potential data which is becoming one of the most popular social media in the world. Naive Bayes is developed to observe positive and negative opinions of the customers on the services given by the company in the form of heat map. Data pre-processing was performed in order to determine attributes to choose data accompanying with their associated coordinates and to decide which words are considered as positive or negative opinion. Heat map was used to visualize density level gradation of opinions. The use of heat map is capable to observe customer's opinions density level based on their colour gradation visualization that are presented in certain radius from one coordinate observing point specified by the user.

Key words: Twitter, emotions, Naive Bayes, heat map, classification, opinion

INTRODUCTION

Tight competition requires companies to obtain information that is more valuable as an analysis of the needs of the service through various media. Now a days more and more people interact with each other through social media Twitter. This makes the recorded data in Twitter are very potential to be analysed.

In line with ease of the accessibility of the internet, Twitter users are increased quickly. In 2012, Indonesia became a country with fifth highest Twitter users in the world as shown in Fig. 1, there are two cities in Indonesia were among the top 20 cities with highest number of Tweets, Jakarta ranked first and Bandung ranked sixth. It indicates that the Twitter data in Indonesia are very potential to be processed to produce information that is more valuable to the company.

This study suggests a method of text classification using Naive Bayes along with data pre-processing to classify an opinion into positive or negative sentiment. In order to produce more valuable information, the data was displayed in the form of heat map to determine the density gradation of users who prefer to like or dislike into the services of a company as a case study is Telkomsel company, the largest telecommunication provider in Indonesia.

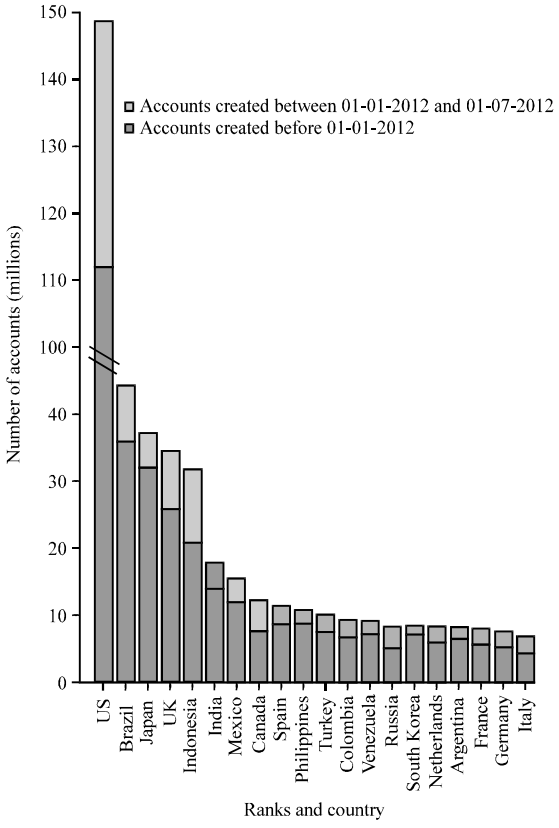

Fig. 1: Twitter accounts ranking (Semiocast, 2012)

Corresponding Author: Mochamad Nizar Palefi Ma'ady, Department of Informatics Engineering, Universitas Nahdlatul Ulama Sunan Giri, Ahmad Yani No. 10, Bojonegoro, Indonesia

Literature review: Twitter, a web services based social media, uses the API to allow other applications (3rd-party applications) to use or to interact with the data. Matthew Wagner stated that the API of Twitter acts as a bridge that facilitate interaction between Twitter with other programs including a program to retrieve data in real time.

There are three common text classification approaches for Twitter data objects, namely N-gram analysis, ontology based and Naive Bayes. Ghiassi *et al.* (2013) researched the N-gram analysis is the approach of text classification by extracting or taking the special characteristics of an object. While ontology-based by Kontopoulos *et al.* (2013) is focusing on modelling terms into a domain. Calvin and Setiawan (2014) used these two methods for text in English or any other Foreign languages and never been used for text in Indonesian. Therefore, in this matter the method to be used was Naive Bayes as Maharani who uses for Indonesian text.

The name of Naive Bayes method was taken from its inventor in 1702, Thomas Bayes. In its application, Naive Bayes method besides being used to classify text in the context of sentiment analysis, it is also used for email filtering, subject's categorization, researcher identification, gender or age identification and language identification.

According to Ikonomakis *et al.* (2005), Naive Bayes is the most frequently used in applying text classification due to its simplicity and effectiveness of the method. This method works well to classify text objects included for Indonesian language as Ma'ady *et al.* (2018) showed that Naive Bayes text classification can play suitable with textual data from social media like Twitter.

Naive Bayes method consist of two phases which training phase and testing phase. Training phase is the phase in which the results of pre-processed document through the process of learning to obtain training data. While the testing phase is the phase in which the new data will be classified using the training data which have been made previously, according to Lestari *et al.* (2013). Training phase is done after the document through the data pre-processing, this pre-processing is done manually. These training data will be used to process new documents in the testing phase that all as explained (Liu, 2012).

Twitter data results of the classification can be displayed in the form of heat map as the concept of Tan *et al.* (2006) about clustering which is a graphic that displays data from a table with colours to represent numerical values. Clustering algorithms relationships between rows or columns that have similarities. So that, the red colour graphics, means indicates that the density of data or the similarity of data, becoming higher. Conversely, if the less similarity among the data, it will be interpreted with a greenish yellow colour.

MATERIALS AND METHODS

Twitter ongoing process of data collection based on keyword 'telkomsel'. The Tweet data were drawn only Tweet with its keyword. This process took the Twitter data in real-time. While the data was already past or historical data cannot be retrieved. As seen in Fig. 2, the new Twitter data collection process would stop when the Twitter API stream was stopped.

Data pre-processing was done in three stages as follows: the selection of data attributes, elimination of data without coordinate and opinion data selection.

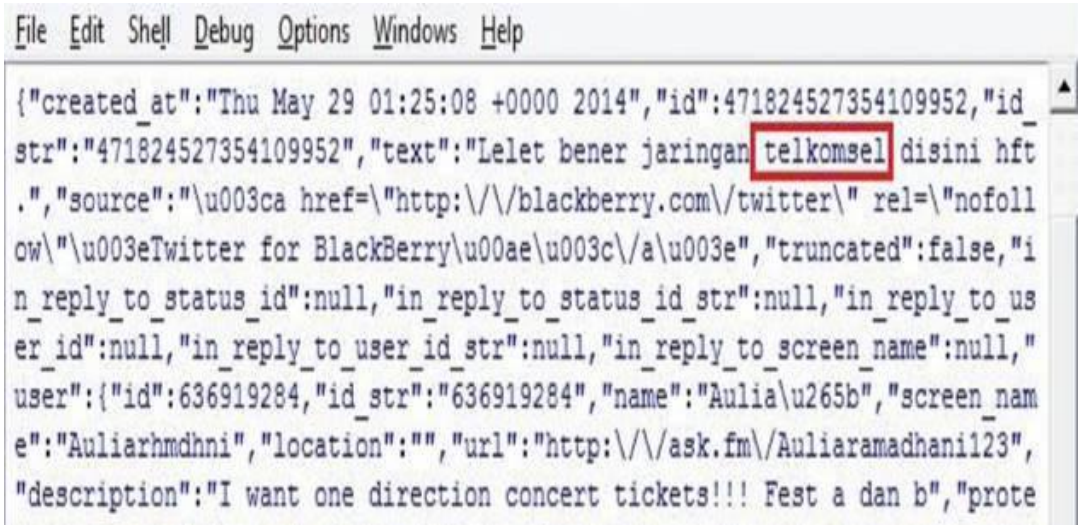

Fig. 2: Twitter crawling in Python

idteks	teks	Pscore	Nscore	idteks	teks	Pscore	Nscore
1460031	lelet	15	16	1460313	=))	2	1
1460032	makasih	7	6	1460314	:/	0	1
1460033	tdk	4	5	1460315	:(	19	20
1460034	ganteng	1	0	1460316	:D	2	1
1460035	emosi	2	3	1460317	x_x	2	3
1460036	cinta	1	0	1460318	-_____-	0	1
1460037	lancak	1	0	1460319	:)	5	4
1460038	aduh	1	2	1460320	:-(	0	1
1460039	sabar	0	1	1460321	:*	2	1

Fig. 3: Words and emoticon features in database

Initially selected raw data attributes that suit the needs of the mapping. Therefore, the raw data must have longitude and latitude coordinate. In addition, the data also must have a text attribute, date and unique number.

In mapping, the data attributes need to have latitude and longitude coordinates but in fact not all Tweets have coordinates. It was because some people when setting their account, they didn't activate the coordinates. Therefore, from the collected data select the data that has coordinates. The data must be eliminated if the data doesn't have coordinates.

Data that has coordinates can be opinion or non-opinion. Therefore, the data was then sorted out to ensure that the data was user opinion toward Telkomsel services. Because the concept of text classification is to classify the text data into a class. Therefore, the data of non-opinion must be eliminated from further processing. In this process also cleaned irrelevant items such as URLs (i.e., strings starting with 'http: //') and hashtags '#'.

The next step was to take the keywords (feature selection) to be categorized into positive or negative sentiment based on domain expert. The results were then verified by Telkomsel company. Once the company approved then the next process was performed text classification.

The process of Naive Bayes method can be computed using PHP. Whereas to accommodate unique words required database which in this case using MySQL. PHP program can be integrated with MySQL. As shown in Fig. 3, some examples of the features with a score that were obtained from the frequency of occurrence. Of the total features were taken, each feature given two classes, positive and negative. Each feature had a value of frequency for each class.

Categorizing these features also apply to features like emoticons. Feature selection that made not only in the form of words but also emoticons that indicated the class of positive or negative sentiment. Suppose there were emoticons such as sad ':(' or happy':)'. Both of these emoticons have different values for its class sentiment. For emoticons ':(' means negative sentiment class while emoticons':)' class means positive sentiment (Fig. 4).

Equation $P(V_j)$ was used to calculate the ratio of the number of classes existing document. Equation 1 respectively defined $|fd(V_j)|$, a class of documents and $|D|$, the number of classes of documents to be used as a model:

$$P(V_j) = \frac{|fd(V_j)|}{|D|} \quad (1)$$

Then, the equations used to calculate the index of the class with the involving keyword frequency documents written by $P(Wk|V_j)$ as shown in Eq. 2. Equation $P(Wk|V_j)$ was the result of $f(Wk|V_j)$, the amount of frequency of occurrence of unique words on a particular class with added one, divided by the number of N, the number of unique word frequency in the training data and $|W|$, the frequency number of unique words that appear whenever there was additional. For both equations, its required a database that accommodates the frequency of unique words in each document:

$$P(Wk/V_j) = \frac{f(Wk/V_j)+1}{N+W|} \quad (2)$$

The data that had been entered into the database were directly classified into training data. The results also can be known what class of the data were entered. By using the PHP programming language, the implementation of Naive Bayes application connected to a database in MySQL.

id	date	text	longitude	latitude	sentiment
14060001	Thu Apr 10 16:05:22 +0000 2014	masbudilo aku pakek 2 pin budl Kan pakek telkom aku tapi gang ku ini gapernah ada sinyal telkomsel-_-	3.2502055	98.5395362	negative
14060002	Thu Apr 10 16:37:30 +0000 2014	Telkomsel Sucks!!!	-4.04767	122.46685	negative
14060003	Thu Apr 10 16:54:28 +0000 2014	Aku pake axis. Lama bangat sumvah -_- RT shintaaun. widiainia pake telkomsel bendera merah burung garuda!!!	-6.2538583	106.8498934	positive

Fig. 4: Classification results in database

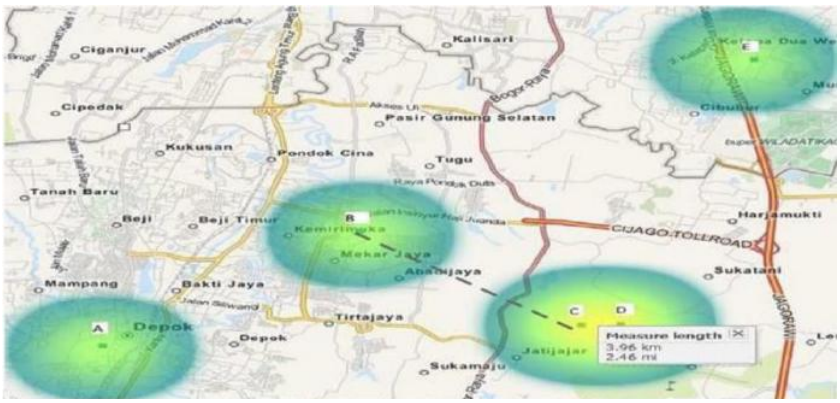

Fig. 5: Adjusted scale of heat map

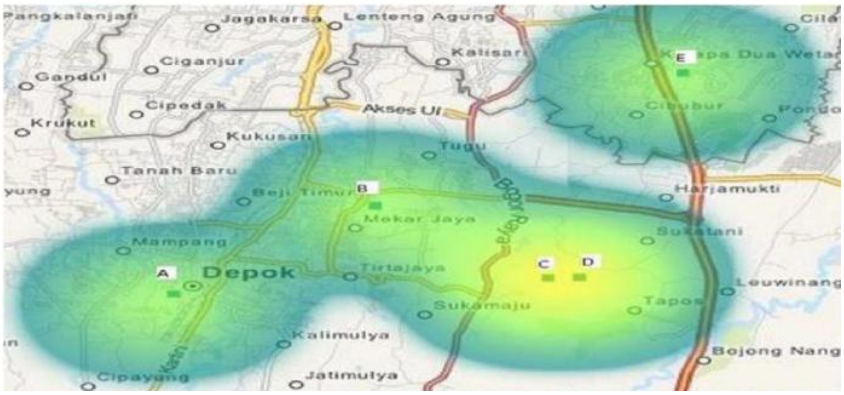

Fig. 6: Adjusted scale of heat map using location point

Heat map cluster which generated by the application OpenGeo suite was a hierarchical clustering with single linkage cluster similarity. Single linkage made by two resemblance closest centroids. Single linkage in the form of heat map can cope the difficulty making of non-elliptical shapes. There were 5 Twitter user location points or centroids, namely A-D and E which had a certain distance within a radius of 1,000 m. After going through the cluster similarity calculation it can be described the dendogram results of these calculations. The results of hierarchical clustering can be interpreted from the colour into a heat map. Heat map will be even greater when the radius of the map become wider as in Fig. 5-7.

Fig. 7: Adjusted scale of heat map within radius of l000 m

RESULTS AND DISCUSSION

Data used for training phase was the data that collected from the data collection process using Twitter API stream in the month of December 2013 and April 2014. While the testing data obtained in June 2014. Training data and testing data each one through the stages of the same data pre-processing. Raw data for training before going through the data pre-processing as much as 133.466. Then, after going through the data process stages, the amount of data become 623 data. While the testing data from the raw data as much as 22.733-150 data after going through data pre-processing stages.

All of 623 training data were processed using Naive Bayes method in order to select the feature. The features derived from the training data as many as 336 features. Each document data from the training data contains at least a feature of all total features. All of the features that have been analysed were used to perform testing on new data which obtained in June. After validation testing of text classifiers application, resulting contingency table with number of positive and negative opinions, respectively, 58 and 92. These data were then mapped in the form of heat map. The level of accuracy, precision and recall for the heat map.

With this contingency table, the resulting value of the accuracy test was 73%. While the precision and recall values are 34 and 91%. These results were likely due to two causes which are relatively many non-standard and ambiguous words used in the Twitter data in Indonesian language that causing incorrect in classifying. Text classifiers with Naive Bayes method is closely related to the selection of features from the training data. The high accuracy of this method depends on the accuracy in choosing the features that represent the training data. As explained earlier, the features obtained from 623 training data were 336 features.

Table 1: Training data and testing data

	Training data			Testing data
Stages	December	April	Total	June
Raw data	8.428	125.038	133.466	22.733
Preprocessing data:				
Attributes selection	8.428	125.038	133.466	22.733
Elimination data:				
without coordinates	570	2.463	3.033	763
Opinion data selection	161	462	623	150

The following example of test documents that has no features accommodated of the 336 features available. Because there were no features available in the database, the application of text classifiers automatically classifies the documents into a positive class but actually it should be negative class. Table 1 shows some of its testing data were not available in the database.

Besides the limitations of features derived from the training data, the other issue was the sarcasm sentences. Sarcasm or satire sentence is a sentence that consist of good words but it has bad intentions towards something. Some testing data which contain elements of sarcasm that has meaningful words, so that, by the application of text classifiers were classified into positive sentiment class.

The use of heat map is to map the coordinates position of the customers whose opinions have been classified will allow the user to observe the level of customer opinion density based on visualization of gradations of colour maps presented in a certain radius of the desired observation point coordinates. Yellowish green colour indicates a high density of negative opinion. However, because the maps produced by the heat map was based on a cluster method agglomerative based, then grouping the points that are incorporated in the mapping becomes sensitive to noise and outliers, so that, the cluster structure of data that formed may less precise. Figure 8 and 9 show examples of heat map area in Jakarta.

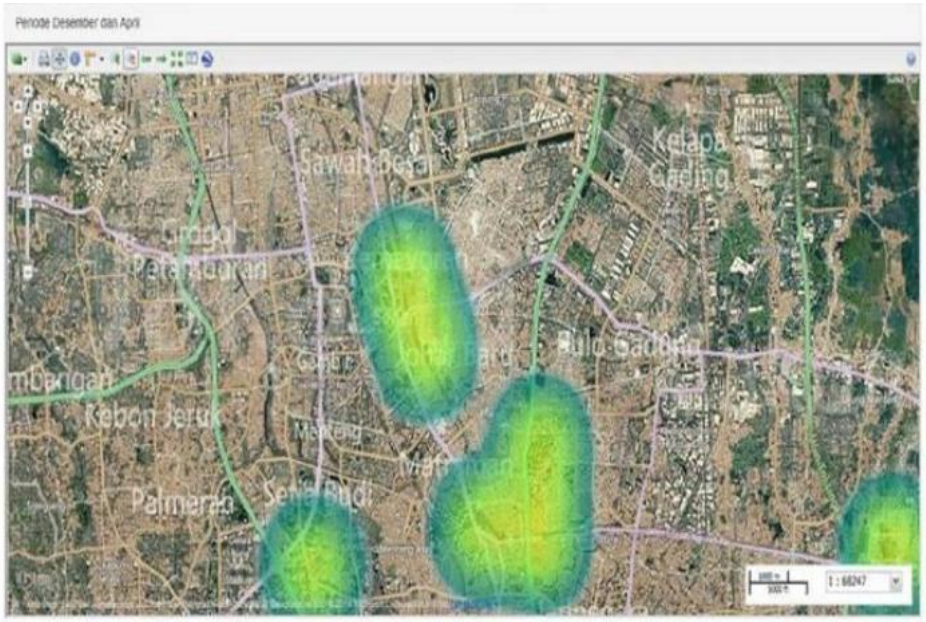

Fig. 8: Geovisualization presentation of positive sentiment

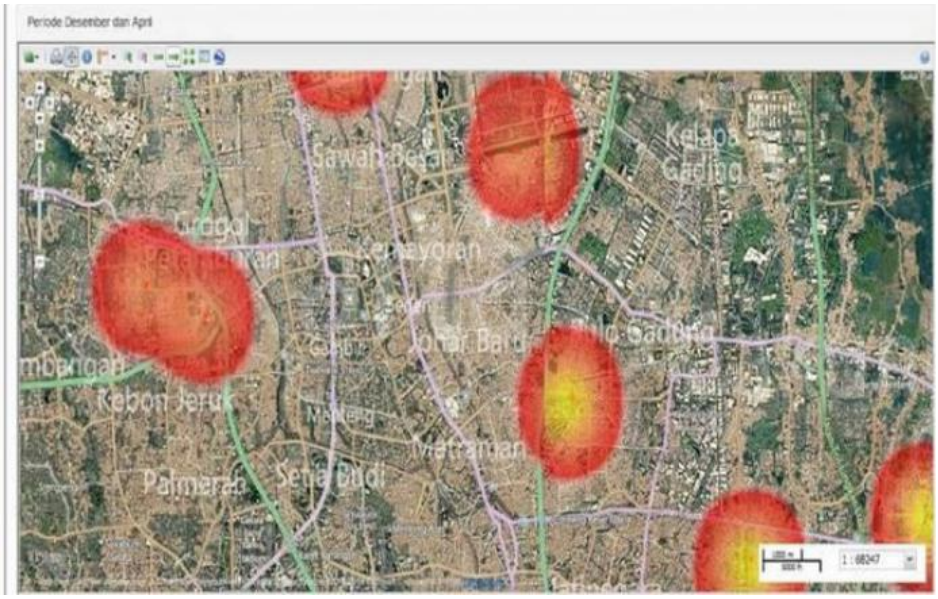

Fig. 9: Geovisualization presentation of positive sentiment in heat map

CONCLUSION

More valuable information that obtained from heat map depend on how accurate of text classifier application Naive Bayes based. It was built using the data of 623 Twitter subscribers of Telkomsel company in the period of December 2013 and April 2014 had been tested by using the data of 150 Twitter subscribers in the period of June 2014. Prior to develop the text classifier, data pre-processing was performed in order to determine attributes, to choose data accompanying with their associated coordinates and to decide which words are considered as positive or negative opinion. The results of the application test given results and sufficient recall with successive values of 73 and 91%. However, the results of the classification precision give less satisfactory results with a value of 34%. Relatively low precision results were most likely due to the number of words that were not

standard and ambiguity of Tweet used in building classifier, making it difficult to properly and accurately classified.

RECOMMENDATION

For that subsequent research should focus on improving the precision of text classification for data Twitter in Indonesian.

REFERENCES

Calvin and J. Setiawan, 2014. Using text mining to analyze mobile phone provider service quality (Case Study: Social Media Twitter). Intl. J. Mach. Learn. Comput., 4: 106-109.

Ghiassi, M., J. Skinner and D. Zimbra, 2013. Twitter brand sentiment analysis: A hybrid system using N-gram analysis and dynamic artificial neural network. Expert Syst. Appl., 40: 6266-6282.

Ikonomakis, M., S. Kotsiantis and V. Tampakas, 2005. Text classification using machine learning techniques. WSEAS. Trans. Comput., 4: 966-974.

Kontopoulos, E., C. Berberidis, T. Dergiades and N. Bassiliades, 2013. Ontology-based sentiment analysis of Twitter posts. Expert Syst. Appl., 40: 4065-4074.

Lestari, N.M., K.G. Putra and K.A. Chayawan, 2013. Personality types classification for Indonesian text in partners searching website using Naive Bayes methods. Intl. J. Comput. Sci. Issues, 2013: 1-8.

Liu, B., 2012. Sentiment Analysis: A Fascinating Problem. Morgan & Claypool Publishers, Chicago, Illinois,.

Ma'ady, M.N.P., C.K. Yang, R.P. Kusumawardani and H. Suryotrisongko, 2018. Temporal exploration in 2D visualization of emotions on Twitter stream. Telecommun. Comput. Electron. Control, 16: 376-384.

Tan, P.N., S. Michael and K. Vipin, 2006. Introduction to Data Mining. Pearson Education India Ltd., Boston, Massachusetts, ISBN:9780321420527, Pages: 769.