Congestion Correlation And Classification from Twitter and Waze Map Using Artificial Neural Network

Acihmah Sidauruk
Information System
Universitas AMIKOM Yogyakarta
Sleman, Indonesia
acihmah@amikom.ac.id

Ikmah
Information System
Universitas AMIKOM Yogyakarta
Sleman, Indonesia
ikhmadarwan01@amikom.ac.id

***Abstract*— Traffic congestion has become a big problem in cities around the world, especially in big cities. This causes information about traffic conditions very important to be known by the riders. Such information can be obtained quickly and easily through social media, but not yet known. Previous research has largely focused on classifying congestion data and traffic speed velocity analysis, while the correlation between congestion information from social media and actual traffic flow velocity has not been studied. In this study, we combine data from social media and traffic data collected for 1 week and focus on some major roads in Yogyakarta, Indonesia to investigate the correlation between congestion information in cyberspace through social media and actual traffic speed in Waze applications. The results in this study indicate that the highest precision value of all experiments is 84.01%, while the lowest precision value of all experiments is 0.37%.**

Keywords—congestion; traffic jam; data mining; classification; social media

I. INTRODUCTION

Traffic congestion problems are still often experienced by urban communities around the world. This can trigger a reaction from the public to write traffic jams and traffic conditions by sharing them through social media like Twitter. Indonesia is ranked as the 5th largest Twitter user in the world. Indonesia's position is only under the United States, Brazil, Japan and Britain. (source: www.kominfo.go.id) With active users reaching up to 100 million per day, Twitter tweets or submissions can reach 500 million submissions per day. This large amount of information can make Twitter a source of data for data mining research.

In this study, data and information from Twitter about congestion will be extracted and compared to real-time traffic flow data from Waze application. It aims to find out whether there is a relationship between congestion information that is shared through social media with the speed of actual traffic flow from Waze, in an attempt to determine the cause of the congestion that occurs so that the solution of the congestion problem can be found. To achieve this goal, we use data mining.
which is a classification by applying Artificial Neural Network machine learning algorithms to create training data and make predictions on the final results of the study. While to measure the level of accuracy, cross validation method and to determine the relationship or correlation of congestion with the number of posts used the Least Squares method.

II. RESEARC METHOD

This research started from extracting information from twitter account in Yogyakarta, twitter account that we use is @ jogjaupdate, @infojogja, @kota_jogja and @ jogja24jam. The account informs the traffic flow in Yogyakarta. We monitor real-time in both accounts, after which the existing tweets are extracted to get the results of data visualization on Waze. Then, several samples from a certain period we used to create a model system using the Artificial Neural Network method and measuring correlations to monitor congestion, observe accuracy and compare it with the reality of traffic flow from the Waze application.

A. Data Mining

Data mining is a process of processing data that is limited with the possibility of an unlimited model and aims to produce the model that best describes the existing data, by applying the data analysis algorithm and data discovery [1]. Data mining can also be said to be a process of extracting knowledge from a lot of data. This knowledge is in the form of order, patterns, and relationships in large and previously unknown data sets. There are several methods for processing in data mining and can be divided into 2 types in general, namely predictive method and descriptive method [2]. The Predictive method draws conclusions from existing data to make predictions on subsequent data, classification, regression and deviation are some examples of techniques in predictive methods. The Descriptive

978-1-5386-7082-8/18/$31.00 ©2018 IEEE

method generalizes the data characteristics contained in the database, clustering, association, and sequential mining are some examples of techniques in the descriptive method.

B. *Classification*

One of the most important of the myriad activities of data analysis is to classify or categorize the data into a set of categories or groups. Data objects that exist in the same group should display properties that are similar to some criteria. In this paper will be used one method of data mining is the method of classification. Classification is one of the most popular data mining techniques, which has extensively used in the analysis of big data [3]. Typically, the classification model is trained first on the historical dataset (training set). The purpose of the classification is to find the model of a training set that distinguishes attributes into appropriate categories or classes, the model is then used to classify attributes whose class is not previously known [1].

C. *Artifical Neural Network*

Artificial neural networks are information processing systems that have characteristics similar to biological neural networks. [4] ANN was formed as a generalization of the mathematical model of neural network biology, assuming:

a. Information processing occurs in many simple elements (neurons)
b. The signals are sent between the neurons through the connectors
c. Liaison between the neurons has a weight that will strengthen or weaken the signal
d. To determine the output, each neuron uses an activation function (usually not a linear function) imposed on the sum of inputs received. The magnitude of this output is then compared to a threshold.

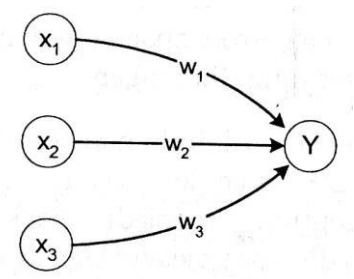

Fig 1 Simple Model of Artificial Neural Network

Y receives input from neurons x1, x2 and x3 with their respective weights are w1, w2 and w3. The three impulses of the existing neurons are summed up

net = x1w1 + x2w2 + x3w3 …….(1)

The amount of impulse received by Y follows the activation function y = f (net). If the activation function value is strong enough, the signal will continue. The activation function value (network model output) can also be used as a basis for changing the weight.

D. *Correlation*

Correlation is a number that indicates the direction and the strong relationship between variables or more. It is expressed in the form of positive or negative relationships, while the strength of relationships is expressed in the magnitude of the correlation coefficient [5]. Positive and negative correlations are shown in figures 2 and 3 as follows:

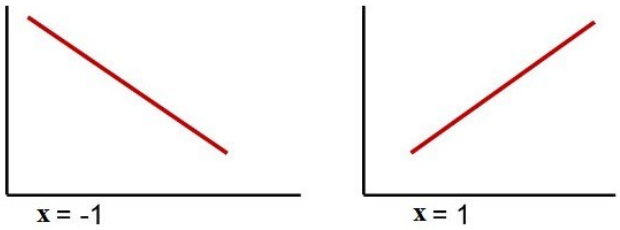

Fig 2 Positive And Negative Correlation

E. *Least Square*

The Method of Least Squares is a procedure, requiring just some calculus and linear algebra, to determine what the "best fit" line is to the data [6]. Least Square Method is used to obtain the linear regression coefficient estimator. A simple linear regression model is expressed by the equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon_1 \text{ general model} \quad \text{…….(2)}$$

The alleged model is expressed by:

$$\hat{Y} = \beta_0 + \beta_1 \hat{X} \text{ or } \hat{Y} = b_0 + b_1 X \quad \text{…….(3)}$$

Obtained an error, ε or ε_{1i}, as follows:

$$\varepsilon = Y_1 - \hat{Y}_1 = Y_1 - b_0 - b_1 X \quad \text{…….(4)}$$

III. SYSTEM DESIGN & OVERVIEW

What we get from this research:

1. Getting Visualization results of the city of Yogyakarta uses tokens via Waze
2. Observe directly for traffic congestion traffic at specific hours using the Artificial Neural Network.
3. Knowing the relationship between traffic information from twitter with actual traffic flow conditions

Due to some information about twitter that has low accuracy. For that we only use information from account @jogjaupdate @infojogja and @ jogja24jam.

A. System Overview

An overview of system design for displaying traffic flow information in Waze Map, it can be seen as follows:

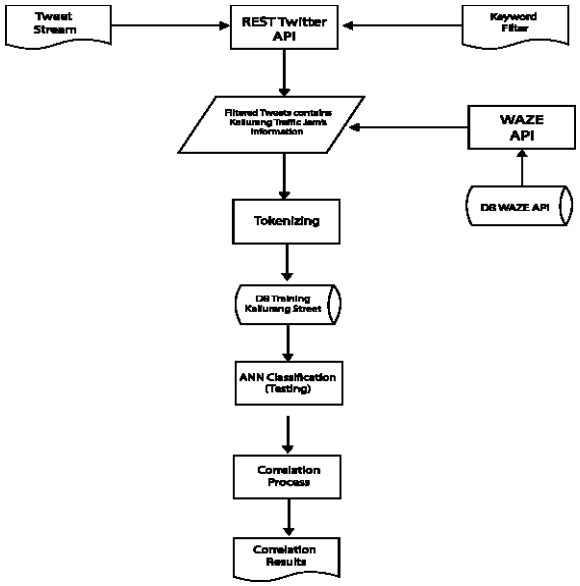

Fig 3 The Architecture of the Traffic Mapping System

B. Getting the Tweets

The most important first thing in this study was to collect tweets from Twitter and stored in the database. The twitter allow people to post information (tweets) reflect what they are looking, hearing, feeling. In other words, the people using such social media services can be regarded as a human sensor of physical world [7]. This motivated us to retrieve the relevant information from the human sensors to describe the traffic condition.

- The Twitter REST API & Engagement Library is used to help the data retrieval process. The Twitter Rest API is selected through Streaming to get targeted data from the three selected accounts. [8][9] Request to the APIs contain parameters that can include keywords, regions, hashtag and Twitter user ID's. Responses from APIs is in JavaScript Object Notation (JSON) format [10].
- Engagement Library is to help unlock the required data, developed by Adam Green. This framework leverages and makes it possible to collect tweets from targeted accounts and store them in the database.

C. Filterring

Not all tweets on @jogjaupdate @jogja24jam and @infojogja contain traffic congestion information. There fore we need to determine the keywords that represent the congestion situation. [11] in Yogyakarta city especially Kaliurang road. From the observation it was found that the keyword is:

a. Padat (Crowded)
b. Macet (Traffic jam)
c. Padat Merayap (Mostly traffic jam)

The next step is to check each incoming tweets, whether or not they have one. If there is at least one keyword then it will be stored in the database, but otherwise the tweet will be ignored.

D. Pre-processing

To display congestion information in the Waze Map, only Tokenization is required in the preprocessing step. But for ANN, more advanced steps are needed including cleaning, transformation and tokenizing as shown in Table 1 below:

Table 1: Illustartion Preprocessing

Process	Data
Data Tweets	17.00 Situation from Pandega Sakti street to Pandega Bhakti street keep slow speed on Juni 23 2018
Case Folding Process	17.00 Situation from Pandega Sakti street to Pandega Bhakti street keep slow speed on Juni 23 2018
Cleaning Process	17.00 Pandega Sakti street to pandega bhakti street slow speed juni 23 2018
Transformation Process	17.00-20.00 Pandega sakti pandega bhakti street slow speed on Saturday.
Tokenizing Process	[1] 17.00-22.00 [2] Pandega Sakti street - Pandega bhakti street [3] Slow speed [4] Saturday

Right after the last process, the data will be inserted into the traffic congestion dataset. The first array will be added to Column "Period" (period), 2nd Array Street Name ", 3rd Array" Class "and 4th" Day ", so the final result of preprocessing data is as follows in Table 2:

Table 2: Ilustration Data set

No	Day	Time Period	Street Name	Class
1	Sabtu (Saturday)	17.00-22.00	Pandega Sakti street-Pandega Bhakti street	Slow speed

E. Finding the traffic jam Location & Direction

The process of locating the location is done from the starting point and the destination point. Due to the characteristics of tweets from @ jogja24jam @jogjaupdate and @infojojga, it can be seen that the tweet tells where the traffic jam occurred.

Fig 4 sample tweets from @jogjaupdate

As seen in the example, there are two keywords typically used by all three accounts to state the location and direction of the traffic jam. These keywords (in Bahasa Indonesia): "Arah" and "Menuju". How it works by checking the words before and after the keyword and see if the keyword matches the street name already loaded in the database. If found then the system will provide information to the user. The process of finding the starting point and the end point of a standstill can be seen as follows:

F. Waze Map

Waze is a navigation application that provides information to users in the form of real-time traffic information and road information, such as traffic volume, road hazards, or accidents that can affect traffic. The information contained in Waze comes from other users or commonly referred to as crowds ourcing, so the information provided to the user is the latest information collected based on information obtained from previous users. The information that came from crowdsourcing was used by Waze to provide some of the shortest alternative routes that the user wanted to go to. The shortest route in this case is the route that requires the shortest time to get to the destination [12]. The beginning of the dot will be displayed with the flag location and the destination point will be displayed with the flag:

Fig 5 waze icon starting point and destination point

G. The auto Proces

As the traffic situation keeps changing and the flow of tweets always flows, the system is expected to automatically process the situation so that everything can always be done in real time and people can get accurate information. Cronjob, automated caller scripts running on unix-based servers are selected to help implement automated system processes. Cronjob will call the script every 5 minutes then collect the most recent tweets, update the database and will be deleted in every hour.

IV. THE RESULT

A. Showing Congsetion on Waz App

The data source of the traffic flow condition in this research is obtained from Live Map feature in Waze application. The feature can display a map of route search results with reports on traffic conditions coming from its users. On the Waze maps there are various icons and colors on the road that show the speed of traffic flow in real-time. The red color indicates a severe traffic jam, orange for traffic jams, and yellow indicates light jams. This color comes from the speed of each user when driving is tracked by the Waze system. In addition to color, on the map there are also various icons, especially car icon lined up. The icon is a live report entered by the user while on the go.

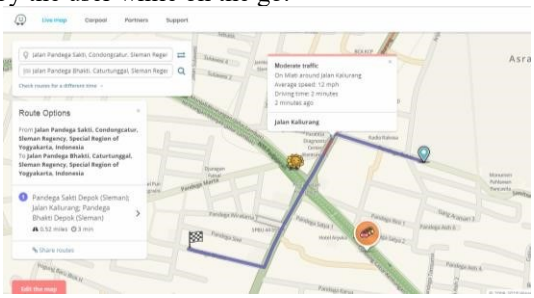

Fig 6 Display traffic congestion information on Waze app.

B. Classification use Artificial Neural Network

1. Data Preparation
   This study uses 245 data derived from Twitter posts in the preprocessing process for use in the next stage of performance testing phase.
2. Performance Testing
   At this stage, the data collected will be divided into two parts: training data and test data. A probability table will be created in training data and then will be test based on the probability table that has been made. Performance will be obtained by assigning a value to the confusion matrix to calculate the accuracy and error values of the test results. This study will perform five tests with different training sections of the data in each test.

Table 3: Summary of each test

Test	Data Partition		Confusion Matrtix Table		Predicted Class	
	Training data	Testing data			Fast	Slow
1st	80	60	Actual Class	Fast=23	16	7
				Slow=37	20	27
2nd	120	100	Actual Class	Fast=39	27	12
				Slow= 61	15	46
3rd	150	120	Actual Class	Fast=45	36	14
				Slow=75	16	59
4th	200	180	Actual Class	Fast=80	64	16
				Slow=84	84	16
5th	245	220	Actual Class	Fast=90	73	17
				Slow=130	22	108

Table 3 describes the distribution of the number of data into 5 subsets to measure the accuracy value, of which 245 data are divided into 5 data subnumbers in which 80 datasets with 60 data for testing, 120 datasets with 100 data for testing, 150 data sets with 120 data for testing, 200 data sets with 180 data for testing, and 245 datasets with 220 data for testing.

Table 4: summary of performance system

Test	Performance System	
	Accuracy (%)	Error (%)
1st	71,013%	0,0037
2nd	75,033%	0,0033
3thrd	79,028%	0,0027
4th	82,05%	0,0020
5th	84,01%	0,0019

Table 4 shows the results of accuracy and error values generated through testing in Table 3, where the first test resulted in an accuracy of 71.013% and an error with a value of 0.0037, the second test produced an accuracy of 75.033% and an error with a value of 0.0033, in the third test produced accuracy of 79.028% and error with a value of 0.0027, the fourth test resulted in an accuracy of 82.050% and an error with a value of 0.0020, in the fifth test resulted in an accuracy of 74.010% and an error with a value of 0.0019. Based on these data, the results obtained with the lowest value in the first test with a configuration of 80 datasets and 60 data testing, while the highest value in the fifth test with a configuration of 245 datasets and 220 data tests.

3. Result Analysis

From the test results using cross validation, the accuracy and error values for each test were found. The highest precision value of all experiments is 84.01%, while the lowest precision value of all experiments is 0.37%. For the highest accuracy value of all experiments is 84.01%, while the lowest accuracy value of all experiments is 71.013. For the highest error rate of all experiments is 0.0019%, while the lowest error value of all experiments is 0.037. These values result in comparisons between tweets and speeds shown in graphical form as in **Fig. 7**.

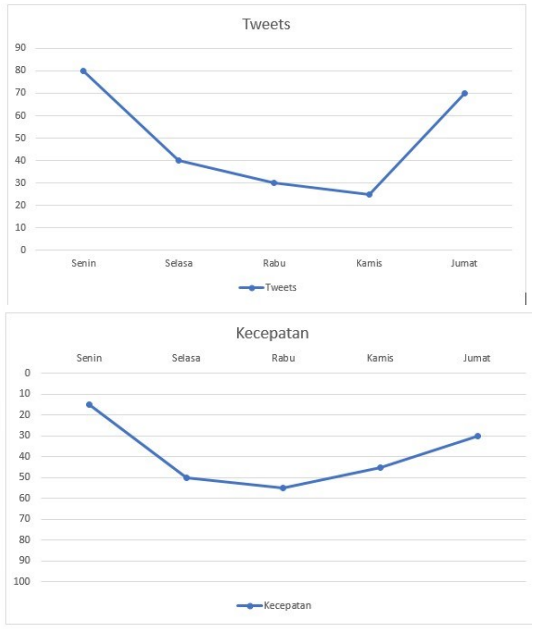

Fig 7 comparisons between tweets and speeds

C. *Corelation Analysis Result*

Correlation coefficient is used to determine whether there is a relationship between congestion information shared through social media with the actual speed of traffic flow from Waze. And this is the Least Square Method, if the result is mostly 1 means it has a strong correlation, and if it is below 1 or 0 it means weak correlation. The type of correlation between Tweet and Speed includes the type of positive correlation with a strong level of closeness with a value of 0.997.

V. CONCLUSION

Based on the classification and testing process that has been done above, the conclusions obtained are as follows:

1. The conditions of traffic congestion reported by Twitter do not always show accurate results in accordance with the real traffic conditions.
2. Based on test the highest precision value of all experiments is 84.01%, while the lowest precision value of all experiments is 0.37%.
3. Between complaints on Twitter with the speed of traffic flow has a correlation with a strong level of accuracy with a value of 0.997.
4. The results of this study in the future can be use to predict and provide notifications for traffic jams, so that drivers can find alternative roads.

REFERENCES

[1] Han, Jiawei., Kamber, Micheline.2006. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

[2] Witten, Ian H. et al. 2017. Data Mining: Practical Machine Learning Tools and Techniques Fourth Edition. United States: Todd Green Publishers.

[3] Liu, Shen et al. 2016. Computational and Statistical Methods for Analysing Big Data with Applications.

[4] Hertz, John., et al. 2018. The Theory of Neural Network Computation: CRC Press Publisher.

[5] Li Wei, Ni Zhigang, Li Shuhua. 2016. Cluster-inmolecule local correlation method for post-Hartfree-Fock calculations of large system. An International Journal at the Interface Between Chemistry and Physics. ISSN: 0026-8976.

[6] Miller, S.J. The method of least squares. Mathematics Department Brown University Providence, RI 02912.

[7] Pan, Bei et al. 2013. Crowd Sensing of Traffic Anomalies based on Human Mobility and Social Media. Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ISBN: 978-1-4503-2521-9. Pages 344-353.

[8] Montreal, Quebec. 2016. BotOrNot: A System to Evaluate Social Boots. Proceedings of the 25th Internation Conference Companion on World Wide Web. ISBN: 978-1-4503-4144-8. Pages 273-274. Yu Yang, Wang Xiao. 2015. World Cup 2014 in the Twitter World: A big Data analysis of sentiments in

[9] U.S. sports fans' tweets. Rochester Institute of Technology, USA. Vol. 48.

[10] Steiger Enrico, et al. 2015. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Transactions in GIS.

[11] Kumar, Samanth et al. 2013. Twitter Data Analytics.Spinger New York Heidelberg Dordrecht London.

[12] Fitri Riri, et al. 2017. Location-Based Mobile Application Software Development: Review of Waze and Other Apps. American Scientific Publisher. Vol. 23.