International Research Journal of Computer Science (IRJCS)
Issue 11, Volume 5 (December 2018)

ISSN: 2393-9842
www.irjcs.com

UTILIZING DATA MINING IN GETTING INFORMATION FOR VEHICULAR USERS IN INDONESIA

Dhanny Permatasari Putri*
Computer Science, Mercu Buana University, Indonesia
dhanny.permatasari@mercubuana.ac.id

Manuscript History
Number: **IRJCS/RS/Vol.05/Issue11/DCCS10082**
Received: 03, December 2018
Final Correction: 13, December 2018
Final Accepted: 24, December 2018
Published: December 2018
Citation: Dhanny (2018).Utilizing Data Mining in getting information for vehicular users in Indonesia, IRJCS:: International Research Journal of Computer Science, Volume V, 525-535. **doi://10.26562/IRJCS.2018.DCCS10082**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract— Today micro-blogging such as twitter has become very popular social media. Millions of tweets are posted to the internet every second. Those high volumes of uncertain data produced by social media can be collected or generated and become valuable information. This research tries to collect data from twitter using Twitter API and Python data science to know what kind the information needs and search by the vehicular user in their way head to their destination. The data taken from twitter user in Jabodetabek area (Indonesia). First by identify the twitter data streaming, testing prerequisite, collecting data, categorize it and analyze it. Then measure it with Precision ratio, Recall ratio and Accuracy ratio using Confusion Matrix Table. The result of all the measurements show that more than 90%, which mean the system method using data science to mine the data and analyze the data is valid to represent.

Keywords— Data Mining; Data Science; Phyton; Big Data;

I. INTRODUCTION

The rapid development of telecommunication has impact to services delivered by telecommunication operator, which one of the popular services is internet. The increasing of using internet by mobile devices make people can access all content provide in the internet [14]. Not only young people whom call millennials like to post anything what in their mind to social media, but everyone who can reach internet also like to post and searching something at their social media. According to statistics there are around 4.15 billion active internet users in the world, which is around 54% of the world population.

There are 2.62 billion active users in social media. [22]. Today the world of internet makes social media become popular. And Twitter is one of the popular social networking that used worldwide. Twitter also known as microblog service on the internet. The number of monthly active Twitter users worldwide reaches 336 million active users in 1st quarter in 2018. And the Number of Twitter users in Indonesia reach 20.9 million users in 2018 [23]. Most of the internet users in Indonesia using social media by their mobile devices. And with the number of active Twitter users mention by statistic number, there are an extremely large data sets of information posted by social media users in the internet every day. That big data set of social media posts from Twitter users can be collected or generated and become valuable information. Vehicular user behaviour in Indonesia, -whom using public transportation or own vehicle, usually like to search information and posts something to their social media with their mobile devices, while they are in the way head to some place or destination. Beside Facebook, Path, Instagram, one of the social medias they mostly used Twitter. And with twitter we can curate the information from user's post and can report/search news in real time. Based on the description above, the author will do research on data science in getting information for vehicular users using python data science.

The formulation of problems can be determined in this study as following:
1. What kind of specific words/ information can be collected in twitter for vehicular users?
2. The amount of data generated by social media is a huge of data to be analyze. The data itself is unstructured, so first that need to be processed to be meaningful information. How to process and analyze the information based on big data twitter?

The scopes of this research are limited to these areas:
1. The social media used for this research is Twitter.
2. The information of social media which is used are tweets in Indonesian language, and the area is near Jakarta area called Jabodetabek (Jakarta, Bogor, Depok, Tanggerang & Bekasi), Indonesia.

II. LITERATURE REVIEW

Research about Twitter was done by Jimmy Lin and Alek Kolcz in 2012 [1] , both of them work at Twitter, Inc., San Francisco, CA, USA, their research was about "Large-scale machine learning at twitter". It was presented in the proceeding ACM SIGMOD International Conference on Management of Data. Since then there many researches arise about social media, data science and machine learning.

In 2013 Ryan Compton, Craig Lee, Tsai-Ching Lu, Lalindra De Silva, Michael Macy made research about "Detecting future social unrest in unprocessed Twitter data: Emerging phenomena and big data" where public API and Geocoding then identify demographics recent posts for specific keywords where the place of the research was in Latin America [2].

In 2014 Li Bing and Keith C.C. Chan had research about "Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data". Where they proposed a novel matrix-based fuzzy algorithm to mine the defined multi-layered Twitter data [3].

In 2017 Muhammet Baykara and Ugur Gürtürk had research in "Classification of social media shares using sentiment analysis". They researched how objectively indicates whether a phrase is positive, neutral or negative of tweet to performed by sentiment analysis. Then used Bayes algorithm in the analysis phase. [4].

In the same year (2017) Alfredo Cuzzocrea [5] research about "Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions", where the author propose an innovative late validation methodology, also the challenges and directions to be considered in future, some of them such as research focus, accuracy vs privacy, Security Issues, Quality and Utility of Data, Performance and so on. In this paper research about how to get data from twitter, -which are unstructured data, to be processed it into valuable information. The idea is to processed the data as aim to know what kind of information can get on social media (twitter) about traffic information and vehicular user behaviour using Twitter API and Python data science.

III. MATERIAL AND METHODOLOGY

A. Big Data

The According to Oxford dictionaries online, "Big Data is extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Or sets of information that are too large or too complex to handle, analyze or use with standard methods".While according to Techopedia online, "Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines". Big data at first has three characteristics; they are volume, variety and velocity. As the develop of the technology and complexity big data, according to Bernard Marr (2014) explain that big data known by the characteristics as "five V which are Volume, Velocity, Variety, Veracity and Value" [13]. So, in term of technical or computer science, author concludes the definition of big data. Big data is a very large data set that cannot be covered to process by traditional technology like relational database, data mining and data processing. The data set usually a very complex data, unstructured data and grow rapidly that need tens, hundreds, or even thousands of servers to store. The data set gather and grow rapidly every second by numerous information over the Internet or by information-sensing Internet of things devices (IoT devices) such as mobile devices, software logs, camera CCTV, public microphones network for giving information, wireless sensor networks, aerial (remote sensing), radio-frequency identification (RFID) readers, and other devices that can connect online through the internet (usually has IP address).

B. Unstructured Data

The challenge of processing big data is data storage (usually extremely large), searching, capturing the data, transferring, sharing, updating, querying, to visualize the data, information privacy and the data source which usually unstructured data. According to Technopedia online "Unstructured data represents any data that does not have a recognizable structure. It is unorganized, raw and can be non-textual or textual.

Unstructured textual data include Word documents, PowerPoint presentations, instant messages, collaboration software, documents, books, social media posts and medical records. Non-textual unstructured data is generally created in media, such as MP3 audio files, JPEG images and flash video files". Then author conclude that Unstructured data consider not as relational database hierarchy or data model that usually known, and it cannot fit in traditional database as Relational Database Management System. Unstructured data is comprised of data that is usually not as easily searchable, it's including formats like audio, video, and social media postings. While structured data neatly fit to a standard and rigid format that has consistency in storing, updating, processing and analyzing it.

Characteristic of unstructured data:

- Unorganized, because differ from common database model hierarchy.
- The information is inconsistent and scattered.
- Containing objects or documents that not under control and free size, it can be combining elements such as containing text, images, audio, video, email, office documents, etc
- The data spread all over the internet.
- Difficult to access and query in traditional database, so it needs additional preprocessing tools..

C. Data Mining

Data mining is a study to analyze a large data digital collection to find unique, interesting and useful structure of relationship in a large data by combining some study from statistics, computer science and others area of study.

D. Data Science

Data science is interdisciplinary board areas cover of mathematics, statistics, information science, and computer science and other sub area such as artificial intelligence (like neural networks or machine learning) and others. It used field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured data. Data Science usually has a good knowledge in programming language (coding), SQL database/coding, knowing statistical theory and tools (SAS, R, etc), also understand to work with unstructured data such as social media, video feeds, or audio.

E. Data Streaming

Data Streaming refer to collect raw data of tweets from internet that are applied as input to produce the analytics goal. Data streaming is used to merge and access the real-time feeds in social media and archived the data for analytics.

F. Python

Python is an open source created by Guido van Rossum and first released in 1991; Python is interpreted high-level programming language for general-purpose programming. It is an easily readable language. Python does not use curly brackets to delimit blocks and semicolons but uses whitespace indentation. It also an Object-oriented programming and structured programming, also many of its features support functional programming and aspect-oriented programming. Python is one of the programming language uses for data science beside Java, Perl, C/C++. Python is a powerful language that provides many services and many libraries. Author will use some additional libraries that have to be installed first when running Python.

G. Proposed Methodology

This research is directed how to get the data and to analysis the twitter data using variety prospective to define the process and implementation part to get the result. This research author study how data is collected, mined, processed, classified data and algorithm used. To get expected result in this research, the proposed methodology and the activities are:

H. Methodology Validation

Step#1 – Observing Twitter Data
- Preparation to determine the twitter data will be used: observation, getting sample data collection
- Identify the problem which are the possible causes that effect the problem
- Data review from sample data collection for possible result

Step#2 – Twitter API
- Getting requirement for Twitter API
- Learn about the Twitter's object that can be used for research

Step#3 – Testing Prequisite

Step#4 – Data Mining, Collection and Processing
- Categorize the selection data

Step#5 – Analyze the data and Result

Step#6 – Conclusion

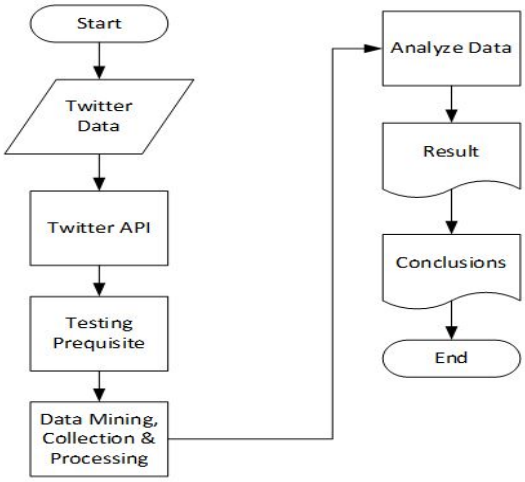

Fig. 1 Flowchart Proposed Methodology

Tools are used in this research are:

1. Twitter API.
2. Online internet, -used when mining and collecting data from Twitter.
3. Python programming language (using version 3.x).
4. Python modules and libraries.
5. Spyder; is the Scientific Python Development Environment. The Python coding can be written and run in Spyder. Spyder may also be used as a library providing powerful console-related widgets to Python. https://pythonhosted.org/spyder/
6. Operating System: Windows.

IV. RESULT AND DISCUSSION

Research result obtained from the implementation of the steps taken to validate the process of the proposed methodology and the validation, the results of research are:

A. Step#1 – Twitter Data

The result of pre-processing in observing and identification of the Twitter Data

1. Observe people behaviour using questionnaire to random people. The results are:
   i. What mostly people do when using internet? They use social media.

Fig. 2 What mostly people do when using internet

   ii. The social media used to get information about road or traffic: twitter and google map

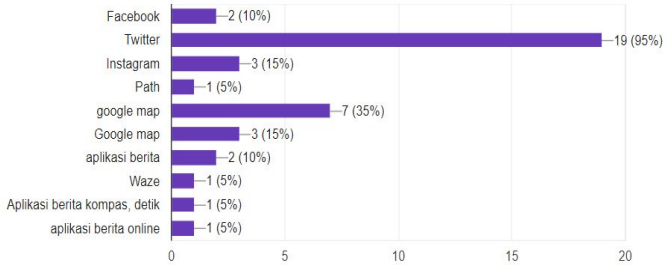

Fig. 3 What social media used to get traffic information

iii. People followed trusted twitter account for traffic report: @TMCPoldaMetro

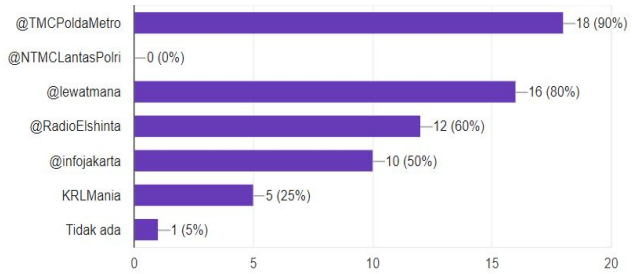

Fig. 4 What usually twitter account to follow to get traffic information

iv. Twitter user looking for their reference word related to road traffic in their way to destination. Indonesian words are: lalin/lalulintas (traffic) and macet (jammed)

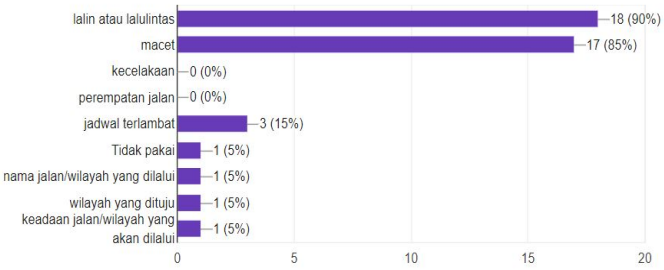

Fig. 5 What word using to get traffic information to the way of destination

v. Participation of twitter user in Jabodetabek area to share traffic information (scale 1 to 10)

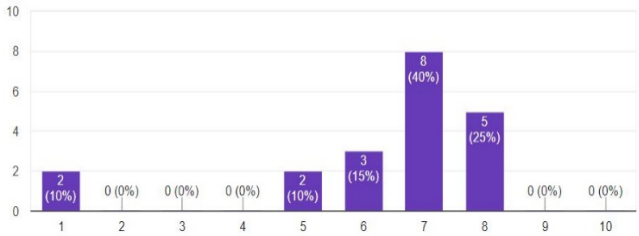

Fig. 6 Participation of twitter user to share traffic information

2. **Observing Twitter Streamline.**
   Observe the twitter trusted account timeline, people twitter streamline and everything related to traffic information for vehicular users or commuter which using public transportation or own car/motorcycle.

B. Step#2 – Twitter API

1. Setting Twitter API

To use Twitter API, first we already have Twitter account. There are the steps to get the API key and token access. The steps are:

1. Sign up at Twitter website and get user id and password for login purpose.

Fig. 7 . Twitter Access Token

2.Go to https://apps.twitter.com to get Twitter API access key.
3. Create basic information application for Twitter API
4. Go to Application Settings > Key and Access Tokens
5. Go to the bottom and click and generate the token

With Twitter API platform we can access data from twitter account and use it for the analysis. For additional information about twitter credentials we can access Twitter Dev website. API is created, user can know his/her customer key, customer secret key, access token key and access secret key. These keys will be needed to add to the coding letter to authenticate user when user want to access twitter data.

2. **Twitter's Objects**

To know more information about the objects, which describe in Twitter. The object will be used to get the data needed in the Python coding.

1. Tweet object — Twitter Developers: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/ tweet-object
2. User object — Twitter Developers: https://developer.twitter.com/en/docs/tweets/data-dictionary / overview /user-object

C. Step#3 – Testing Prequisite

1. **Installation Additional Libraries for Python**

After Python and Spyder already installed, before creating the script in Python environment others additional libraries will be needed. The library usually can be install using command "pip install" in Python. The libraries such as:

a) Tweepy

- Tweepy is Open Source Python library that can enable Python to communicate with twitter by using its API to collect data that needed to be analyzed. Example of command is: import tweepy
- In this script author use all the keys and secret keys which generated from Twitter API (which already describe in Step #2.

Testing Tweepy. The script to test is as follow:

```
# General:
import tweepy  #To consume Twitter's API
from tweepy import OAuthHandler

consumer_key = 'xxxxxxxx'
consumer_secret = 'xxxxxxx'
access_token = 'xxxxxxx'
access_secret = 'xxxxxxx'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth)

# Using the API object to get tweets from your timeline,
# and storing it in a variable called public tweets
public_tweets = api.home_timeline()
                # foreach through all tweets pulled
for tweet in public_tweets:
   # printing the text stored inside the tweet object
   print (tweet.text)
```

Fig. 8 . Python code testing twitter API token to get data using tweepy

When write the codding, first create listener class to load the data from the twitter. OAuth protocol used as standard protocol for authorization, to allow user to log-in at user timeline or any third-party user in the social twitter social network. OAuth provides security and authorization to user. Command "print(tweet.text)" is a command to print the object which is the tweet in a form of a text that the result will display in the console screen. The test the result is successful.

b) **Csv** module, used to import or export spreadsheets and databases for use in the Python interpreter, you must rely on the CSV module, or Comma Separated Values format. Example of command: import csv.

c) **Re**, this module provides regular expression matching operations similar to those found in Perl. A regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern. Example of command: import re

D. Step#4 – Data Mining, Collection and Processing

For getting the data, collecting the data and processing the data the step taken to implement the process are:

a) Some Trusted Twitter Account

After observing twitter streamline, there are some trusted twitter accounts have many followers.

TABLE I - SAMPLE OF TRUSTED TWITTER ACCOUNT

Twitter Account	Description
@TMCPoldaMetro	Official account of Indonesia Police, Jabodetabek city (Jakarta, Bogor, Depok, Tanggerang, Bekasi)
@NTMCLantasPolri	Official Corps Police who handle traffic in Indonesia (national)
@lewatmana	Live Traffic CCTV & Current Traffic Info

b) To get tweet data from another user

The script shows how to get raw data tweet from the specific user @TMCPoldaMetro

```
# The Twitter user who we want to get tweets from
name = "TMCPoldaMetro"
# Number of tweets to pull
                        tweetCount = 20

# Calling the user_timeline function with our parameters
results = api.user_timeline(id=name, count=tweetCount)
```

Fig. 9 . Python code to get TMCPoldaMetro post

c) Some useful twitter objects to test

The script shows how to get raw data tweet from other twitter object to be shown. As already explain in the previous sub chapter about twitter objects. Here some sample to get the detail information about twitter objects that will be used latter on coding.

```
   # printing the text stored inside the tweet object
            print (tweet.text)         #for tweet
print (tweet.created_at)  #for time created
print (tweet.user.screen_name)  #for user name show
print (tweet.user.location)     #for location
```

Fig.10 .Python code to get the twitter object

d) Identification for social-aware of twitter user in social media in searching or giving information to their destination. Terminology used in tweet information, the words are in Indonesia language: macet, lalulintas (lalin), kecelakaan, tabrakan. Usually the use format is like this: @twitter_account : Time + information area (including words macet, lalulintas/lalin, kecelakaan, tabrakan).

e) Print specific word need to be show

The script shows how to get raw data tweet from twitter object q for query specific word or hashtag. Variable **q** is the object read in twitter. For language object use ISO 639-1 standards, for the abbreviation. For Indonesia code language standard is "id". Query identify as lalin (short of traffic), lalulintas (traffic), macet (jammed).

f) Store the data

After getting data from Twitter API, then store the data. It will be used to be analyzed at next step. When ran the script and get the data collected with some classification related vehicular user in getting information, the result will be extracted into csv file.

E. Step#5 – Analyze the data and Result

The focus of this research to get data from big data twitter about what twitter user do related to give and search information about traffic on road as socially aware of human behavior for vehicular users. The data collected based of raw data twitter already extract using data science Python.

a) Shorting the data

To short the data, author using this code

b) Analyze

For this research, the word used only lalin and lalulintas, and account mention to TMCPoldaMetro

i. Data A. After getting the data collection using object twitter q for query for the word "lalin" (the short word for lalulintas Indonesia word for traffic) then saved it in katakunci2.csv then next steps:

- Short by deleting empty space to file#2.csv there are 1486 tweet, then
- Short it only for Jabodetabek area or the commuter area (Jakarta, Bogor, Depok, Tanggerang, Bekasi) become file#3.csv there are 586 data tweets.
- Then short again all the data and become 556 data (30 data false)

```
import re
                          import csv

# Make a case-insensitive regex to match the words "lalin" or
"macet"
pattern = re.compile(r'\blalin\b|\bmacet\b', re.I)

with open('file#1.csv', 'r', newline='') as csvFile,
open('file#2.csv', 'w', newline='') as newFile:
    reader = csv.reader(csvFile)
    writer = csv.writer(newFile)

    for row in reader:
        # Skip empty rows
        if not row:
            continue
        if pattern.search(row[2]):
            print(row)
            writer.writerow(row)
```

Fig. 11. Python code to short data by specific word

ii. Data B. By collecting from mention TMCPoldaMetro
- Fitering to word classification "lalin" and "macet" become 348 data
- After checking the data, the data all already for Jabodetabek area.

iii. To evaluate the result, author will use the measurement. The measurement use in this research are Precision and Recall. It used to determine the accuracy of a system in which simple computations of accuracy usually used in machine learning.

The goals of the measurement are:
- To check if the implementation already run success
- The checking is to test the feature of system.
- The checking is to determine the accuracy of system

After shorting the data then put it into the table of Confusion Metrics. Table Confusion Metrics is used to describe the performance of a classification model on a set of test data that shows the actual and predicted labels from a classification problem.

TABLE II - CONFUSION MATRIX TABLE. **DATA A**

	P' (Predicted)	n' (Predicted)
P (Actual)	TP – 556	FN – 20
n (Actual)	FP – 30	TN – 880

TABLE III - CONFUSION MATRIX TABLE. **DATA B**

	P' (Predicted)	n' (Predicted)
P (Actual)	TP - 348	FN - 0
n (Actual)	FP - 0	TN - 0

Where:
TP = True Positive = Relevant data found
FP = False Positive = Irrelevant data found
TN = True Negative = Relevant data were not found
FN = False Negative= Irrelevant data that cannot be found

c) Precision Measurement

Precision (also called positive predictive value). Precision is the number ratio of relevant data obtained by the system with total number of data that picked by the system, either relevant or irrelevant.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

d) Recall

Recall (also known as sensitivity). Recall is the number ratio of relevant data obtained by the system with the sum of all relevant data in the collection of data (wheatear it is drown or not drown by the system).

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

e) Accuracy

Accuracy is the performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

f) Result

The result measurement of Precision, Recall and Accuracy:

TABLE IV - MEASUREMENT RESULT

	Precision	Recall	Accuracy
Data A	94%	96%	97%
Data B	100%	100%	100%

The result after doing all the implementation and analysis,

1) The kind of information that vehicular user search in the twitter streamline is related to the information of the road is traffic report from official twitter account, such as @TMCPoldaMetro. The classification of the word for commonly used are lalin, lalu lintas or macet
2) To analyze the big data using Python data science.
3) To measure the data and to see the accuracy of the actual and predicted labels from a classification problem, author use Precision ratio, Recall ratio and Accuracy ratio.
4) Basically, Indonesia twitter users are socially-aware. They use the information given by the official twitter account as their reference to get to someplace or to their destination. They also ask the official twitter account about the traffic report to avoid traffic jammed or to share information about the traffic itself.

V. CONCLUSIONS

1) Twitter is a good tool for understanding the public opinion. Using twitter data author can do research and analysis about information needed by the vehicular user in Indonesia, specially Jakarta which well known as the city that has heavy traffic at rush hour.
2) Twitter API is a good way for creating automated tools to get insights related to the scope of work to maintain research.
3) The information usually get from twitter for vehicular user need is the road traffic information or accident in the road. The common words usually using such as lalin (lalulintas) and macet.
4) The result of the measurement show that the Precision ratio is 94% and 100%, Recall ratio is 96% and 100% also the Accuracy ratio is 97% and 100%. Which mean the system method using data science to mine the data and analyze the data is valid to represent.
5) Indonesia vehicular user in are mostly commuters and social-aware. They like to use twitter, specially who live in Jabodetabek area to update information about the situation of the road/ traffic from the trusted twitter account. They sometimes like to share information at their social media about the situation and about traffic report.
6) Indonesian vehicular user and commuters are also social media aware, they like to ask to the trusted twitter account admin (such as account @TMCPoldaMetro and @Lewatmana) about how is the traffic situation at the road, -in the way of their destination. The two twitter accounts also most trusted to follow by netizen which can be used as reference to get information about the road traffic.
7) The author suggest that the future work can be done using another method, programming language or algorithm to get data and process it to be meaningful information, such as using another language beside Indonesian language or other key-words to be searched.

REFERENCES

1. Jimmy Lin; Alek Kolcz, "Large-Scale Machine Learning at Twitter", Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data Pages 793-804. SIGMOD 2012, ACM, NY, USA ©2012, ISBN: 978-1-4503-1247-9 doi> 10.1145/2213836.2213958, CrossRef: https://dl.acm.org/citation.cfm?id=2213958
   Download PDF : http://users.umiacs.umd.edu/~jimmylin/publications/Lin_Kolcz_SIGMOD2012.pdf
2. R Compton ; C Lee ; T-C Lu ; L De Silva ; M Macy, "Detecting future social unrest in unprocessed Twitter data: Emerging phenomena and big data", IEEE International Conference on Intelligence and Security Informatics 2013, E-ISBN: 978-1-4673-6213-9. Print ISBN: 978-1-4673-6214-6. DOI: 10.1109/ISI.2013.6578786
   CrossRef:https://www.researchgate.net/publication/261457063_Detecting_future_social_unrest_in_unprocessed_Twitter_data_Emerging_phenomena_and_big_data and https://ieeexplore.ieee.org/document/6578786/

3. Li Bing; Keith C.C. Chan, "Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data", Proceeding UCC '14 Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing Pages 652-657. IEEE Computer Society Washington, DC, USA ©2014. ISBN: 978-1-4799-7881-6 doi>10.1109/UCC.2014.105 CrossRef : https://dl.acm.org/citation.cfm?id=2760017
4. Muhammet Baykara ; Ugur Gürtürk, "Classification of social media shares using sentiment analysis", International Conference on Computer Science and Engineering. Antalya, Turkey 2017. Electronic ISBN: 978-1-5386-0930-9. DOI: 10.1109/UBMK.2017.8093536 CrossRef: https://ieeexplore.ieee.org/document/8093536/
5. Alfredo Cuzzocrea, "Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions", 2017 IEEE International Conference on Data Mining Workshops (ICDMW) 18-21 Nov. 2017 , New Orleans, LA, USA ,DOI: 10.1109/ICDMW.2017.140 CrossRef : https://www.computer.org/csdl/proceedings/icdmw/2017/3800/00/3800a992-abs.html
6. Jian Ming ; Lingling Zhang ; Jinhai Sun ; Yi Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis", 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 20-22 April 2018,, Chengdu, China , DOI: 10.1109/ICCCBDA.2018.8386516
7. Amir Gandomi; Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management Volume 35, Issue 2, April 2015, Pages 137-144View Record in Scopus : https://www.sciencedirect.com/science/article/pii/S0268401214001066 https://doi.org/10.1016/j.ijinfomgt.2014.10.007
8. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research Volume 70, January 2017, Pages 263-286 View Record in Scopus : https://www.sciencedirect.com/science/article/pii/S014829631630488X https://doi.org/10.1016/j.jbusres.2016.08.001
9. Bradley Voytek, "Social Media, Open Science, and Data Science Are Inextricably Linked" Neuron Volume 96, Issue 6, 20 December 2017, Pages 1219-1222 View Record in Scopus : https://www.sciencedirect.com/science/article/pii/S0896627317310681 CrossRef : https://www.cell.com/neuron/fulltext/S0896-6273(17)31068-?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0896627317310681%3Fshowall%3Dtrue https://doi.org/10.1016/j.neuron.2017.11.015
10. Ibrahim Abaker Targio Hashem; Victor Chang; Nor Badrul Anuar; Kayode Adewole; Ibrar Yaqoob; Abdullah Gani; Ejaz Ahmed; Haruna Chiroma, "The role of big data in smart city" International Journal of Information Management Volume 36, Issue 5, October 2016, Pages 748-758 View Record in Scopus : https://www.sciencedirect.com/science/article/pii/S0268401216302778 https://doi.org/10.1016/j.ijinfomgt.2016.05.002
11. Marco Bonzanini, "Mastering Social Media Mining with Python", Packt Publishing, UK, 2016. Book ISBN 978-1-78355-201-6. [Online] Available book: https://books.google.co.id/books?id=SuvUDQAAQBAJ &printsec=frontcover&hl=id#v=onepage&q&f=false–Accessed on: June 10, 2018
12. Jake Vander Plas, "Python Data Science", Book O'Reilly Media Publishing Inc, USA 2017. [Online] Available book: https://books.google.co.id/books?id=xYmNDQAAQBAJ &printsec=frontcover#v=onepage&q&f=false – Accessed on June 16, 2018
13. Bernard Marr,"Big Data: The 5Vs Everyone Must Know" Published Article 2014 [Online]. Available: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/Accessed on: April. 17, 2018
14. DP Putri; RB Bahawares; M Alaydrus, "Analysis Provisioning Process in SIM Card Activation at Telecommunication Operator Using BPM and Balance Scorecard", International Seminar on Intelligent Technology and Its Applications http://isitia.its.ac.id/ (ISTIA), Surabaya 2014. Download PDF: https://www.researchgate.net/publication/315458867_Analysis_Provisioning_Process_in_SIM_Card_Activation_at_Telecommunication_Operator_using_BPM_and_Balance_Scorecard
15. Courses on data mining with machine learning techniques, The University of Waikoto, [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/courses.html - Accessed on: June. 11, 2018.
16. Lecture Notes Data Science 101, [Online]. Available: https://web.stanford.edu/class/stats101/#data-summaries - Accessed on: June 11, 2018.
17. Lecture COMPSCI Introduction to Data Science, [Online]. Available: https://canvas.harvard.edu/courses/29726/pages/lectures - Accessed on: June 11, 2018.
18. Video lectures from Foundations of Data Science, [Online]. Available: https://data.berkeley.edu/news/video-lectures-foundations-data-science-now-available-online - Accessed on: June 12, 2018.

19. "Insight: Big Data, Changing the way business compete and operate", April 2014, Ernst & Young Global Publication,Download PDF : https://www.ey.com/Publication/vwLUAssets/EY_-Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf [Online]. Available: https://goo.gl/C3kW2r - Accessed on: June. 12, 2018.
20. "Big data: The next frontier for innovation, competition, and productivity", May 2011, McKinsey Media, Download PDF : https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx [Online]. Available: https://goo.gl/kdVxLw - Accessed on: June. 12, 2018.
21. "The Age of Analytics: Competing In A Data-Driven World", Dec 2016, McKinsey Media, Download PDF : https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.ashx [Online]. Available: https://goo.gl/ZDqGG6 - Accessed on: June. 12 2018.
22. Data Statistics Report [Online]. Available: http:// www.internetworldstats.com- Accessed on: July. 1, 2018
23. Data Statistics Report [Online]. Available: http:// www.statista.com - Accessed on: July. 1, 2018.

