Visualization of Road Accident Data In Jabodetabek Area Year 2014-2017

Johan Setiawan¹ and Sudecanto²
¹Lecturer Paper Advisor, UMN University, Banten, Indonesia
²Student Paper Creator, UMN University, Banten, Indonesia

johan@umn.ac.id, sudecanto@student.umn.ac.id

Abstract. Road accidents are a problem for all countries across the world because it causes material losses and life losses. The aim of this paper is to share findings in a road accident in visualization data in Jabodetabek area. Road accidents data obtained by doing text mining from social media Twitter using Tweepy API. Tweets taken is a tweet that has the word “accident” during the period from January 2014 to December 2017. Data is processed and labelled automatically by using a PHP program which has been created with an algorithm that reads a text pattern and finds the keyword. The keyword information then stored to a new column for desired information. Data that has been labelled in structured table form is processed with PowerBI tools. Labelled data will be processed using Power BI tools to provide visualization data output that provides detailed information about road accidents occurred in Jabodetabek during the year 2014-2017. Cars and trucks are the types of vehicle that most often has accidents, accidents are more common in the rainy season, and accidents prone time is T2 (04.00 a.m to 08.00 a.m).

1. Introduction

Road accidents happen because of many factors. The main factors of road accidents are human or driver error, vehicle condition, road infrastructure, road maintenance, and traffic regulations. Drivers who have a bad attitude in driving, using drugs or alcohol, having sex, not using seat belts and underage drivers are the factors that most often cause road accidents.

According to Kompas.com news, 17/03/2016 04:20 P.M accessed on January 31, 2018, from Traffic Corps (Korlantas) Police data, the number of accidents in Indonesia in 2014 reached 85,756 incidents, with the death number reached 26,623 people, and the material loss reached Rp 224.2 billion.

The difficulty to obtain road accident data in Jabodetabek area makes this research useful to provide road accident information. This study aims to generate data visualization of road accidents in Jabodetabek area. All data are taken from Twitter with tweets containing the word “accident” from username “TMC Polda Metro Jaya”.

Findings from the visualization of road accident data in Jabodetabek area may help in viewing some location or road that is prone to an accident so that drivers can be more careful before passing that location. Further research can know the pattern and factors causing the occurrence of road accidents in each road so we could implement preventive action to reduce the number of traffic accidents, especially in Jabodetabek area, to minimize the losses due to the number of road accidents. Road accident data in Jabodetabek area results shown in a dashboard can provide information about traffic accidents in Jabodetabek area, considering every day there are many road accidents happens in Jabodetabek area.
2. Literature Review

2.1 Text Mining and Big Data.
Text Mining is the process of obtaining new unknown information using a computer by extracting data automatically from various types of written sources. Data is valuable and up-to-date information that has information to explain the events that occur, the longer the data progresses as well as the visual data. Big data is defined as Vs that is Volume related to the amount of data and disk space usage, Velocity is the ever-increasing data flow and the information contained in it, Variety which means the data come from many different linked sources and become merged.

2.2 Visual Analysis and Visual Data
The main purpose of the visual analysis is to develop science, methods, technologies, and practices that explain the combination of human capabilities and electronic data processing. Visualization is the result of humans and computers working together with their ability to produce effective results. Visual data is a process of representing data in the form of tables, graphics from simple data into complex and interactive information.

2.3 Relevance Literature

| Number | Researcher | Year | Title                                                          | Problem                                                                 | Result                                                                 | Implementation                                |
|--------|------------|------|----------------------------------------------------------------|------------------------------------------------------------------------|------------------------------------------------------------------------|-----------------------------------------------|
| 1      | Sachin Kumar dan Durga Toshniwa | 2015 | “A data mining framework to analyse road accident data” | The main problem accident data is to find and identify the main causes that cause traffic accidents. | The result is information about traffic accident data and main causes of traffic accidents. | This research used as a guideline in making cluster analysis. |
| 2      | Luis Ramos | 2015 | “Detection of road accident accumulation zones with a visual analytic approach” | At this era, traffic accidents are a major problem of death. There is a need for predictions of traffic accidents. Traffic accidents in the world cause 1,200,000 people die worldwide per year. | 1. The location accident occurred. 2. Visual data to give information about the concentration of accidents / gray spots | This research is used as a guideline in making the visual mapping of accident-prone location visual map |
| 3      | Rafał S. Jurecki dan Jaśkiewicz | 2012 | “Analysis of road accidents in Poland over the last ten years” | The need for analysis of the number of traffic accidents, injuries, casualties in recent years in Poland in various seasons including the types of vehicles used by victims and the causes of traffic accidents. | 1. Visual data on the number of traffic accidents and injuries in Poland 2. Analysis of the trend of car driver accidents 3. Visual data on the number of accidents per month. | This research is used as a guideline in making visual dashboard data on traffic accidents in Jabodetabek in 2014-2017 |
3. Methodology

This research uses CRISP-DM (Cross-Industry Standard Process for Data Mining) method to perform data mining process for road accident data in Jabodetabek area. The method consists of six (6) main processes:

1. Business Understanding
   Conduct objective analysis and the needs in terms of business perspective and convert the analysis into how to process the data and create a plan how we get the objective.

2. Data Understanding
   Analyse what data is needed to identify the main problems that need to be resolved, find solutions to process the data, create how to create a data hypothesis, and where the data can be retrieved.

3. Data Preparation
   Focus on selection and preparation of data up to the final set. At this stage, there are many changes in records, tables, and attributes as well as data cleaning and transformation.

4. Modelling
   The selection and application of Modelling techniques used, the parameters used to adjust to the problems of data mining.

5. Evaluation
   Evaluation and determines how to use data results. The results obtained depend on the algorithm, the model used, and after the review evaluate whether it has been according to objectives or not.

6. Deployment
   The final stage focus to organize report, socialize and implement visual data that has been processed.

K-means clustering is one of the non-hierarchical data clustering methods that classify data in the form of one or more clusters/groups. Data that have the same characteristics are grouped in one cluster or more group, each data have different characteristics clustered with other clusters/groups so that the data in one cluster/group has small variation level.

The steps of clustering with K-means method are as follows:

1. Select the number of clusters k.
2. Initialization the cluster center can be done in various ways. However, the most often done is by random. Allocate all data/objects to the nearest cluster.
3. The proximity of two objects is determined by the distance of the two objects. The proximity of a data to a particular cluster is determined by the distance between the data with the cluster center. In this stage, it is necessary to calculate the distance of each data to each cluster center. The most distance between one data and one particular cluster specifies a data entry in which
cluster. To distance all data to each cluster center can use Euclidean distance theory formulated as follows:

\[ D(i,j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots + (X_{il} - X_{jl})^2} \]  

**Formula 1. K-means**

4. Recalculate the cluster center with the current cluster membership. The cluster center is the average of all data/objects in a particular cluster. If desired it can also use the median of the cluster. So the mean is not the only size that can be used.

5. Reassign each object using the new cluster center. If the cluster center does not change again then the clustering process is complete Or, go back to step number three.

4. Results and discussion

4.1 Business Understanding

The needs for traffic accident information in Indonesia could increase public awareness of road accidents. Low road accidents can reduce material and life losses from road accidents in Indonesia.

The data obtained from Twitter during the period January 2014 - December 2017, the data scope is Jabodetabek area. The data is expected to provide information to the reader and increase the awareness of the danger of road accidents to citizen awareness at a certain time which is one of the results of road accident data in Jabodetabek in this research.

4.2 Data Understanding

Data Sample. Data for sample derived from Twitter tweets containing the word “accident” during the period of December 1, 2016, until December 31, 2016. This data will be used to choose user name who frequently tweets accident news or facts and suitable for this research criteria. The result is shown below:

| Username          | Location Scope | Total tweets containing word accident | Total tweets which are “true” accident news/facts |
|-------------------|----------------|----------------------------------------|--------------------------------------------------|
| TMC Polda Metro Jaya | Jabodetabek    | 252                                    | 168                                              |
| Radio Elshinta    | Jabodetabek    | 216                                    | 128                                              |
| NTMLantasPolri    | Jabodetabek    | 127                                    | 58                                               |
| sonoraFM92        | Jabodetabek    | 88                                     | -                                                |
| RadioDakta        | Jabodetabek    | 28                                     | -                                                |
| LiputanSekitar     | Jabodetabek    | 20                                     | -                                                |
| Lalulantas (breakingnews.id) | Jawa | 317                                    | -                                                |
| PTJASAMARGA       | Indonesia      | 650                                    | -                                                |
| TmcrestroBekasi   | Bekasi         | 17                                     | -                                                |
| e100ss (radio suara surabaya) | Surabaya | 105                                    | -                                                |
| Harian_Jogja      | Jogia          | 93                                     | -                                                |
| RadioANDIKA       | Jawa Timur     | 157                                    | -                                                |
| RPKD926FM         | Bali           | 18                                     | -                                                |
| Detikcom          | Indonesia      | 39                                     | -                                                |

Table 2 shows user page who frequently tweet accident news. By looking at the table and screening the user page relevant to this research aim within Jabodetabek area and the highest number
of tweets “true” accident, username “TMCPoldaMetroJaya is chosen as a representative of population data that will be collected.

Data scope for the population was taken within a period of 4 years from January 2014 until December 2017. The data are taken from tweets containing the word "accident" and username "TMCPoldaMetroJaya". Information obtained from the tweet are accident date, accident time, number of retweets, number of favorite tweets, and tweet text. Besides those mention before there are additional information such as vehicle type, accident location, and the city of the accident. This additional data obtained by using program that has been made to read text tweet automatically and generate new columns such as a column of vehicle type, accident location columns and city columns by input old CSV file and the program will generate new excel CSV files with additional column information.

4. 3. Data Preparation

Data obtained using Tweeppy (python language). The modules needed for text mining Twitter are Tweeppy-Master, GetOldTweets-Python-Master, Pyquery-Master. The Python version used in this research is Python version 2.7.13.

Retrieving data tweets from social media Twitter is quite easy if you already make the settings and modifications configuration appropriate to the modules. The next step is simply running the command, for example, to retrieve data from January 1, 2017, until December 31, 2017, with a maximum tweets 10,000

```
python Exporter.py --username "TMCPoldaMetro" --since 2016-12-31 --until 2017-01-01 --query search 'accident' --maxtweets10000
```

The command needs to run to retrieve the data for 4 years 2014 - 2017. The Results Data obtained for 4 years is:

**Table 3. Data obtained per year**

| Year | Total Rows |
|------|------------|
| 2014 | 8,462      |
| 2015 | 6,732      |
| 2016 | 5,254      |
| 2017 | 3,216      |
| **Total** | **23,664** |

**Table 4. Data after cleansing**

| Year | Total Rows | Rows after Cleansing | Rows after Delete Redundant Data |
|------|------------|----------------------|---------------------------------|
| 2014 | 8,462      | 5,812                | 4,878                           |
| 2015 | 6,732      | 3,899                | 3,254                           |
| 2016 | 5,254      | 2,473                | 2,189                           |
| 2017 | 3,216      | 1,667                | 1,494                           |

Data cleansing is done by eliminating non-accident data (tweets contain warning and suggestion about accidents), accidents that happened outside Jabodetabek area, and eliminating the same accident data. The process of labeling the tweets is done by using PHP program that has been created to produce new
additional column such as location, time, vehicle type, city, brand, and seasons according to information from tweets sentence.

4.4 Modeling

The results of the modeling stage are several dashboards with various graphical information that allows readers and testers understand the importance of the information. Tools used in this research is Microsoft PowerBI (later called only PowerBI) to make model and visualization.

PowerBI is selected because PowerBI has many features for cleansing, visualization, and analysis. The features in PowerBI application are also easier to use compared with other visualization tools. Besides PowerBI has many custom visual modules and PowerBI support R-Scripting.

R-Scripting is often used by programmers/data analysts to do data modeling with certain algorithms and to predict data

![Figure 1. Dashboard 1](https://example.com/dashboard1.png)

The dashboard at Figure 1 gives important information that represents road accident information in Jabodetabek Area within the year 2014 - 2017. The detail information about dashboard, graph, table and cluster analysis could be accessed using this link: [https://tinyurl.com/roadaccident2017](https://tinyurl.com/roadaccident2017).

The dashboard indicates Jakarta (7,447) as a city with the highest number of accidents followed by Bekasi (1,760) and Tangerang (1,753). Car 45% and Truck 35% is the most vehicle type with accidents. T2 (04.00 a.m to 08.00 a.m) is the highest accident interval time 26% followed by T3(08.00 am to 12.00 a.m) 20% and T4 (12.00 p.m to 04.00 p.m) 19%.

The pattern accident in 4 years (2014-2017) indicates March usually has an increasing number of road accidents than in February and January. Accidents in December also typically have higher accidents than in November. Rainy Seasons has higher accidents than dry seasons but the difference only 128 accident (0.1%). Vehicle brands with high accidents are Toyota (677), Honda (500), and Daihatsu (149).

K-means clustering algorithm is used in this study which results in some information about accident factors: car, truck, motorcycle, and unknown type vehicle have a higher accident in the rainy season.
16/25 variable has a higher accident in rainy season which indicates rainy season have higher accidents (66% True). The time during T2 interval applied to all type of vehicle (100%) which means all vehicle type have the highest accident rate during T2 (04.00 a.m to 08.00 a.m).

Figure 2 shows the location accident number in Jabodetabek area for 4 years (2014 - 2017). The location with the highest accident number is Cikampek and Cikarang Toll which has 493 accidents, the second is Cipularang with 348 accidents and the third is Soekarno Hatta Airport with 343 accidents. Following the big 3, consecutively Karawang (245), Cikunir (239), Jagorawi (230), Merak (221), Priok (212), and Semanggi (207).

4.5 Evaluation

The data produce visual data and models that have good quality information before performing deployment, this model needs to be evaluated and reviewed whether it has been reached objectives or not especially users, mentors, and examiners. The process of preparing this report has been evaluated by the supervisor and accepted by TMC Polda Metro Jaya.

The model evaluated by lecturer, supervisor, and user which is TMC Polda Metro Jaya, visual data result has been reviewed and accepted for use in TMC Polda Metro Jaya Jakarta for their use. The visual result evaluated by lecturer and user directly. There are many processes including screening, interview, user presentation to TMC Polda Metro Jaya for this data visual result. There are supporting document which already authorized and has been signed by TMC Polda Metro Jaya.
4.6 Deployment

The data that have been processed and the results have been informative. The results need to be shared with the citizen and the people who use it. The visual data could also be used for decision making as well as solutions to solve problems related to road accidents.

The visual data is uploaded to the PowerBI student account: sudecanto@student.umn.ac.id and Google Drive. The purpose of the upload is the ease to access the visual data that has been made for TMC Polda Metro Jaya.

5. Conclusion

By looking at the visual data that have been made, it could be summarized that the number of traffic accidents from 2014 to 2017 always decline. The car is a type of vehicles that the most often have accidents. Vehicle accidents based on the big three brands are Toyota, Honda, and Daihatsu.

The number of accidents occurring during Ramadhan, Christmas and New Year does not have a higher accident rate than any other month. In March there is usually a higher accident rate than February and January. Accidents in December also typically experience higher accidents than November.

The rainy and dry season in Jabodetabek does not significantly affect road accidents. The number of accidents during the rainy season is higher than the dry season, the city most affected by the rainy season is the city of Bekasi, which has the difference in rainy season by more than 6% accident from the dry season. The accident-prone time most often occurs at T2 (04.00 a.m s.d 08.00 a.m), accident-prone time on T2 applies to all vehicle types. Three locations with the highest accident rates are Cikampek, Cikarang, and Soekarno Hatta Airport.

References

[1] Coleman, A. 2014 Road Traffic Accidents in Ghana: A Public
[2] Kompas. 2018, 1 31. Kompas.com. Retrieved from kompas.com: http://sains.kompas.com/read/2016/03/17/162000830/Gawat.Angka.Kecelakaan.di.Indonesia.Makin.Mengkhawatirkan.
[3] Gupta, V., & Lehal, G. S. 2009. A Survey of Text Mining Techniques and.
[4] Tonidandel, S., King, E. B., & Cortina, J. M. 2016. Big Data Methods Leveraging Modern Data Analytic Techniques to Build Organizational Science.
[5] Andrienko, N., & Andrienko, G. 2012 Visual analytics of movement: An overview of methods, tools, and procedures.
[6] Kumar, S. (2015). A Data Mining Framework to Analyze.
[7] Ramos, L. (2015). Detection of road accident accumulation zones with a visual analytics approach.
[8] Jurecki, R. S., & Jaśkiewicz, M. (2012). Analysis of road accidents in Poland over the last ten years.
[9] Hill, R. L., Kennedy, H., & Gerrard, Y. 2016. Visualizations and the Need for Data Studies.
[10] Umair, S. 2014. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)