Inteligencia Artificial, 25(70) (2022), 1-12
doi: 10.4114/intartif.vol25iss70pp1-12

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Ontology-Based Traffic Accident Information Extraction on Twitter In Indonesia

Nur Aini Rakhmawati[1,A], Yasin Awwab[1], Ahmad Choirun Najib[1], Ahmad Irsyad [1]
[1]Information Systems Department, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
[A]nur.aini@is.its.ac.id

Abstract Traffic accidents become one of the events that often occur in Indonesia. From the three-monthly report by the Indonesian National Police Traffic Police, there are about 25,000 traffic accidents. Many social media users, especially Twitter, share information about traffic accidents. Twitter has various information regarding traffic accidents. Therefore, this study aims to process and map information about traffic accidents contained on Twitter in Indonesia language. We use the domain ontology and Named-Entity Recognition for the data extraction process. Named-Entity Recognition is used for obtaining keywords from a tweet based on class categories such as actor, time, location, and information on the cause of the accident. This research generates a Named Entity Recognition (NER) model that can provide a reasonably accurate level of accuracy. Also, we create an ontology that can categorize the causes of traffic accidents based on the Directorate General of the Land Transportation Office, Indonesia. We found that the traffic accidents are generally caused by inadequate vehicle conditions with the main problem in the vehicle caused by brake failure, while environmental factors rarely cause traffic accidents. Moreover, the vehicle is the subclass that mostly appears in the tweets, where car is the most popular actor, followed by truck and motorcycle.

Keywords: traffic accident, information extraction, ontology, Twitter, named entity recognition.

1 Introduction

In general, traffic accidents disrupt traffic flow. A significant accident may lead to injuries, damages, and sometimes fatality. Traffic accidents often occurred in Indonesia. From October to December 2018, the Ministry of Transportation Republic of Indonesia reported the number of traffic accident incidents in more than 25 thousand cases. In the report, 61% of the causes of traffic accidents were caused by driver negligence, 30% were due to facilities and environment, and the rest were vehicle factors [7]. In recent years, accident-related studies use data crowdsourcing to find new knowledge [3, 8, 5, 14]. Twitter, a microblogging platform that can create, consume, promote, distribute, and share information with limited characters. Nowadays, Twitter has been accepted as a direct user-contributed information source in event detection [20]. Hence, Twitter used as an information source for event detection, e.g., traffic accident [2, 19, 20] and incident response [21]. However, recent studies have different approaches for detecting the event and extracting information. Besides, most Twitter data present non-standard language and different linguistic styles. Therefore, information extraction from Twitter data is still a challenge. In this study, we explore Indonesian traffic accident tweets data from Twitter. We create a traffic accident ontology [10], extract the information, and visualize the data. We generate traffic accident ontology and extract the instances of traffic accidents based on the dataset. Then, we perform natural language

ISSN: 1137-3601 (print), 1988-3064 (on-line)
©IBERAMIA and the authors

Table 1: List of research on Traffic event using Twitter data

Paper	Ontology	NER	Labels	Classification	Place
Salas et al., 2018 [12]	No	No	Traffic, non-traffic	SVM	UK
Qian, 2016 [11]	No	No	accident, roadwork, hazards weather, events, obstacles vehicles	LDA	Pittsburgh and Philadelphia
AlFarasani, 2019 [4]	No	No	safe, needs attention, dangerous, and neutral	Lexicon based	Riyadh
Wongcharoen, 2016 [18]	No	No	Congestion traffic	J48	Thailand
Al-qaness et al., 2019 [1]	No	No	Traffic, non-traffic	SVM, TensorFlow Neural Network	Los Angeles
Bio et al., 2015 [2]	Yes	Yes	traffic-events, actors, locations and timestamps	SVM	Brazil

processing to obtain four aspects, such as actor, location, time, and the causes of the traffic accident. Finally, we visualize the data based on these aspects. Our main contributions are described as follows:

- Create Bahasa Indonesia traffic accident ontology
- Extract traffic accidents information and visualize the data

The rest of this paper structured as follows: Section 2 describes the related works. Section 3 describes the materials and notion of this work. Section 4 explains the methodology. Section 5 shows results and discussion. The rest brings to the conclusion.

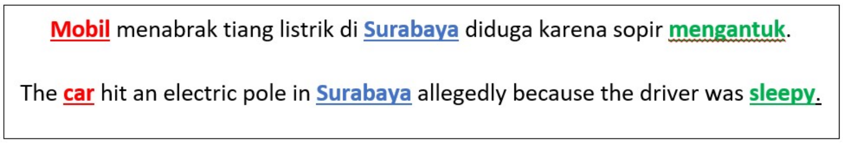
2 Related Works

Tabel 1 shows several works that extract information from Twitter for analysing traffic event. In the majority, those papers classify the Twitter data into several categories using machine learning and deep learning [2, 12, 11, 1, 4, 18]. Those papers generally predict whether the Twitter data are related to traffic event or not. Salas et al. [12] used the Support Vector Machine (SVM) for detecting traffic-related tweets in the UK. Meanwhile, Qian [11] applied the Supervised Latent Dirichlet allocation (LDA) to determine which category a traffic incident belongs to in Pittsburgh and Philadelphia. There are five incidents categories: accident, roadwork, hazards & weather, events, obstacles vehicles. Wongcharoen [18] also used one of the Machine Learning algorithms called J48 to classify traffic congestion in Thailand. Al-Qaness et al. [1] compared the performance of Machine Learning (SVM) and Deep Learning (TensorFlow Neural Network). TensorFlow performance is better than the SVM classifier. Classification of traffic-related events based on lexicon methods is done by Al Farasani et al. [4] for Twitter in Riyadh, Saudi Arabia. They divided data into five categories, namely safe, needs attention, dangerous, and neutral. Bio et al. [2] designed ontology for traffic event in Brazilian Twitter and detected entities in Twitter data. They used Named Entity Recognition (NER) based on SVM. In a similar thread, we also detected entities and classify the Twitter data for Indonesia traffic event.

3 Background Theory

3.1 Traffic Accident Cause Factors

Factors causing traffic accidents according to the Directorate General of the Land Transportation Office in Indonesia can be categorized based on four categories [17]. These categories are used in this research since we use Indonesia Twitter. The four categories can be described as follows:



Mobil menabrak tiang listrik di Surabaya diduga karena sopir mengantuk.

The car hit an electric pole in Surabaya allegedly because the driver was sleepy.

Figure 1: Entity Detection with NER.

1. Driver
   Traffic accidents caused by drivers can be caused by several factors, which include carelessness, drowsy, unskilled driver, drunk, high speed, pedestrian mistake, and animal disturbances.

2. Vehicle
   Vehicle factors consist of broken tires, brake failure, steering failure, axles/loose couplings and vehicle light problem.

3. Street
   The causes of traffic accidents caused by roads include intersections, narrow roads, open or controlled access, unclear road markings, no speed limit signs, slippery road surfaces.

4. Environment
   Accidents caused by the environment can occur due to mixed traffic between fast vehicles and slow vehicles, with pedestrians, supervision, and lack of law enforcement. Besides that, the weather is also an indicator of traffic accidents. The weather includes dark, rainy, smog.

3.2 Named Entity Recognition

Named Entity Recognition (NER) or Named Entity Recognition and Classification (NERC) is one of the main components of information extraction that has the purpose of detecting and classifying a named entity on a string (text) [9]. Objects detected in a tweet are called entities. An example for entity detection using NER depicted in Fig. 1. The red colour denotes the actor entity, the blue denotes the location entity, and the green denotes the causes of the traffic accident.

3.3 Ontology

Ontology is a specification of a classification [10]. Ontology is divided into four components, including:

1. Concept
   Concept (Concept) is a collection of various objects. The concept means a fundamental form of elements of a domain that usually represents a collection of groups that have something in common.

2. Instance
   An instance or entity is known as an individual is a ”ground-level” component that represents a specific object of a concept.

3. Relation
   A relation is a relationship between two concepts in a domain.

4. Axioms
   Axioms impose constraints or problems on the value of a class.

Ontology plays a vital role in playing role in the semantic web to support the exchange of information spread from various environments. The Semantic Web represents data in a machine-processable way, where it becomes a reason to be an extension of an existing web [15].

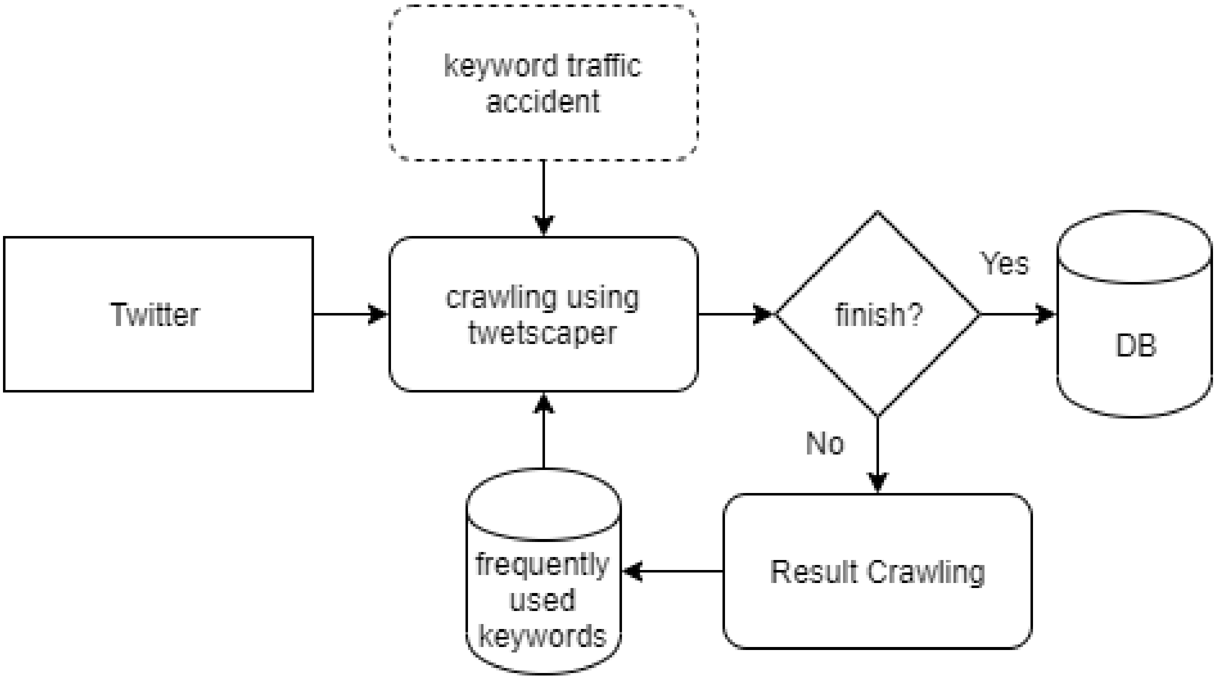


Figure 2: Flowchart for determining keywords from crawling results.

4 Methodology

4.1 Crawling data

Data crawling is a process of retrieving data from Twitter. The crawling uses the Tweetscraper[6]. There are two steps in the crawling process which illustrated in Figure 2. First, we filter the data using predefined keywords for traffic accidents. The second step is crawling used the most frequently used keywords based on the first crawling. In total, we retrieve 846,311 tweets. Table 2 shows the details of keywords and the number of tweets.

4.2 Pre-processing

At the pre-processing stage, we initially convert the text data to lowercase, delete redundant data, and clean data from Twitter symbols. The posting time is stored in a different column from the contents of the tweet. It is necessary to merge the contents of the tweet with the information posting time (timestamp) into the same column. In our assumption, the tweet must have four components: subject (S), verb (V), object (O), and context (C). Therefore, we remove a tweet that has less than four words. Moreover, the goal of our research is to investigate four components of accidents. After this process, the number of tweets has decreased significantly. The number of tweets after pre-processing data can be seen in Table 2.

All data collected are from 2009 to 2019. Due to the data requirements of this study only in 2019 and 2018, the total data obtained in the range 2018-2019 is 25,451 data. The total amount of data per year can be seen in Table 3.

4.3 Ontology design

We calculate the words that have the highest frequency of data. Those words are selected as instances and subclasses of the ontology class. Traffic Accident Ontology is designed by using the Protege software[13].

The ontology consists of two classes, namely the Actor Class and the Kecelakaan Lalu Lintas (Traffic Accident) Class, which can be seen in Figure 3.

1. Actor

   The Actor Class stores actors from the extraction results using the NER method. These actors are in the form of instances in the ontology. Instances can be in the form or abbreviations of subclasses in Indonesian that are commonly used on Twitter. Table 4 show subclasses and instances of Actors. Table 5

2. Traffic accident
   The Traffic Accident class is divided into several subclasses, namely as street, vehicle, Environment, driver. The instance of the subclass is the cause of the accident (Table 5). Table 6 Subclasses and instances of classes Traffic accident

4.4 Twitter Tagging

We tag the Twitter data by giving the name Span START: and END to the specified entity, namely Actor, Location, Information, and Time. The NER Tagging of each entity can be seen in Table 6. We label 15000 rows for each entity.

4.5 Tagging Token

Token Tagging is a labelling process that will be used as a reference in the tokenization process. The dataset used, for each token, word, the phrase you want to token separated by giving white space or special syntax like SPLIT. With provisions, the syntax is used at the end of each sentence marked with a period or a comma.

4.6 Model Training

The training process is done using the OpenNLP library and TokenizerTrainer library from Apache Open NLP[16]. The testing procedure will divide the data into two parts: 80% parts for training and 20% parts for testing. The statistical information of each testing model and training model can be seen in Table 7.

4.7 Classification Based on Ontology

To classify tweets, the extracted information is loaded into ontology. The only tweet that has at least one entity will be classified. The flow in this classification process can be seen in Figure 4. First, we detect the cleaned tweet whether contains actor, location, time and information using the NER model. If we find at least one entity in the tweet, then we classify the tweet based on the ontology. Lastly, we store the classified tweets in a database.

For example, in Figure 1; the picture shows that the entity detected in a tweet is a *Car* as an Actor, *Surabaya* as a description of the location, and sleepy as a description of the cause.

Table 2: Number of tweets after pre-processing.

Keyword	Number of Tweets
Kecelakaan Lalu Lintas (traffics accident)	274,925
Lakalantas (traffics accident)	112,298
Kecelakaan tunggal (single accident)	162,814
Kecelakaan beruntun (consecutive accident)	296,274
Total	846,311

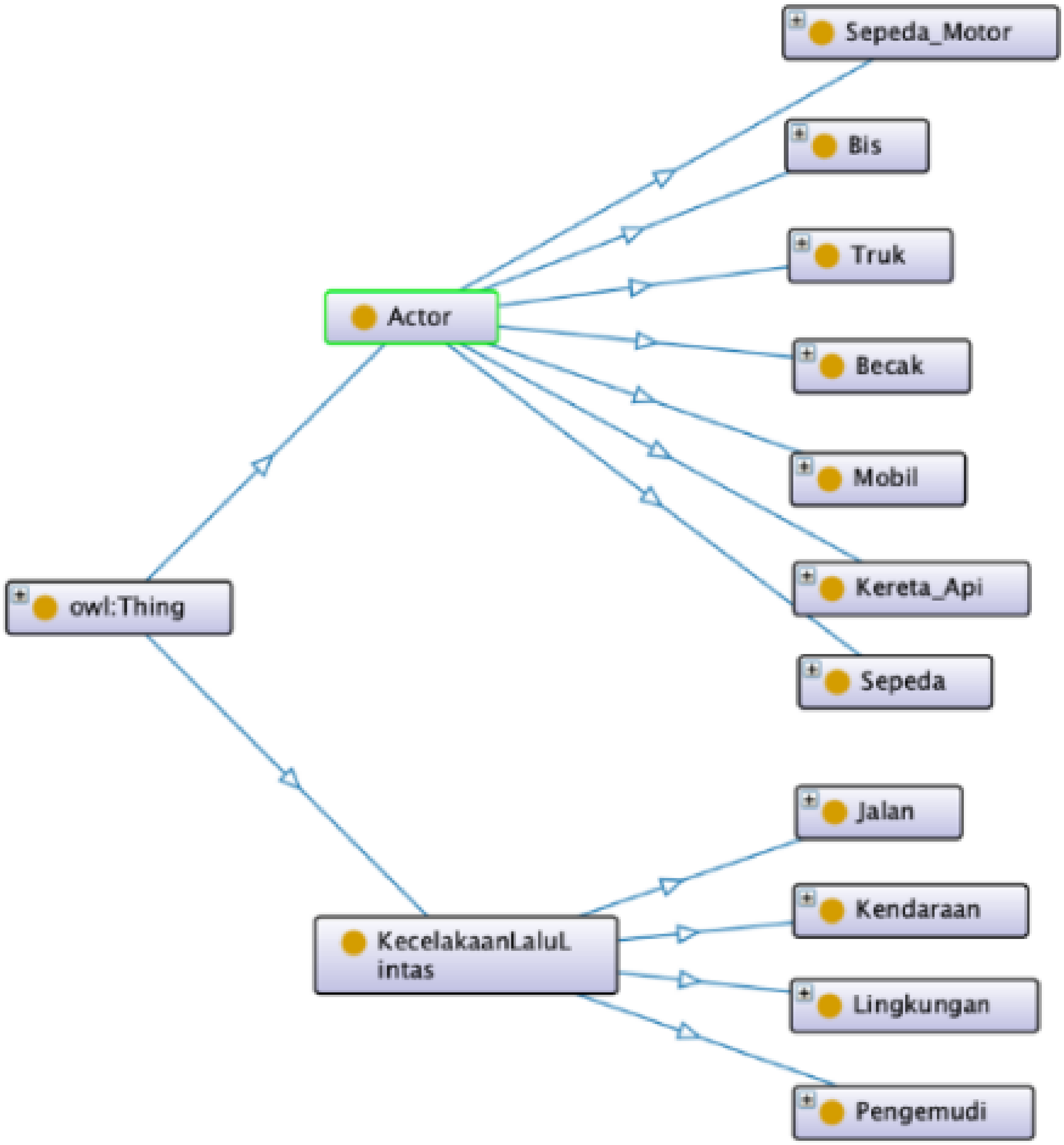


Figure 3: Ontology design.

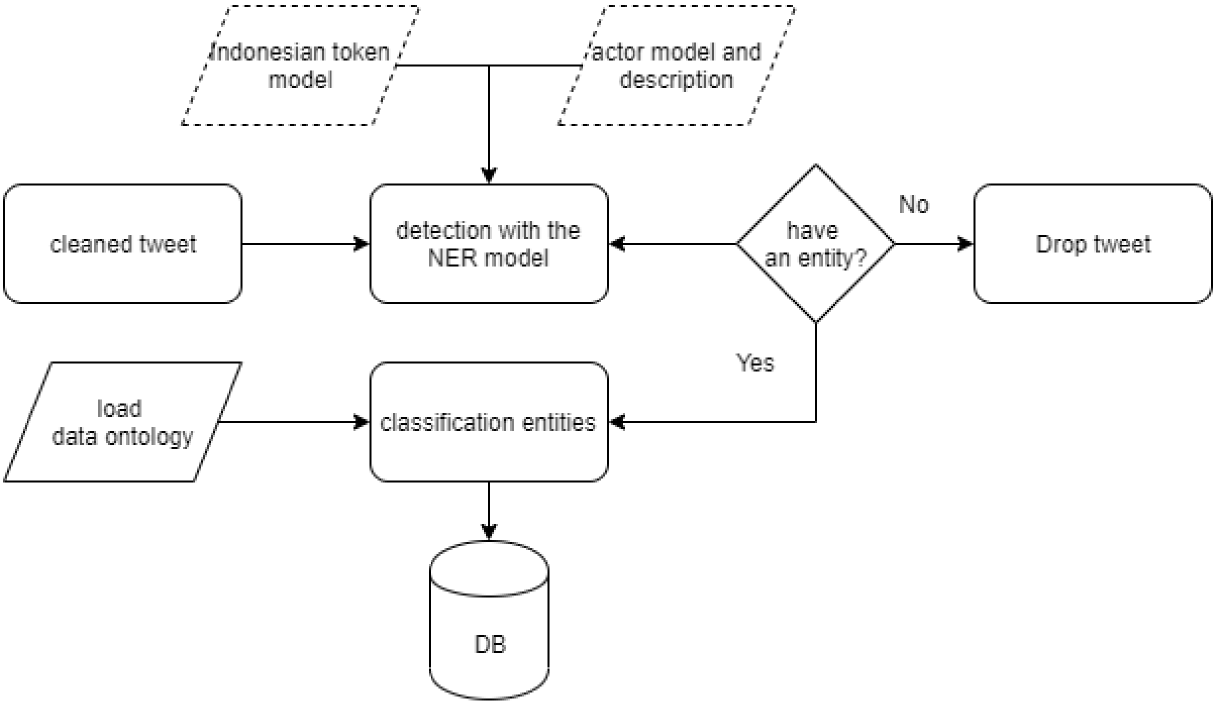


Figure 4: Classification of causes of traffic accidents and Actors.

5 Results

5.1 Model Testing Results

Model testing is performed on the entire named entity model. The results of model testing generated by OpenNLP are shown for the Actor, Location, and Remarks models. For Actor models, the precision is 99.31%, recall is 98.56%, and the F1 score is 98.93%. For Location models, the precision is 99.54%, recall is 98.37%, and the F1 score is 98.95%. The Remarks model has a precision of 99.83%, recall by 90.58%, and an F1 score of 94.98%.

Table 3: Number of tweets obtained by posting year.

Year	Number of Tweets
2019	8000
2018	17451
2017	19762
2016	22377
2015	22874
2014	22697
2013	44357
2012	34605
2011	17695
2010	6612
2009	553
Total	216983

Table 4: Subclasses and instances of the Actor class.

Subclass	Instances (in Indonesia)
Mobil(car)	Bajaj; mbl; mbl_elf; mobil; mobil_box; mobil_elf; mobil_travel; pick_up; pikap; sedan; taksi; taxi; travel
Sepeda_Motor (motorcycle)	Motor; Pemotor; sepeda_motor; spd_motor; ojek; ojol; ojek_online
Sepeda (bicycle)	Sepeda; sepedah; spd
Bis (bus)	Bis; bus; Kopaja
Kereta_api (train)	Kereta; kereta_api
Truk (truck)	Trek; truck; truk
Becak (pedicab)	Becak; becak_motor

Table 5: Subclasses and instances of the Traffic accident class.

Subclass	Instances (in Indonesia)
Jalan (street)	jalan_bergelombang (bumpy road); jalan_berlubang(potholes); jalan_licin (slippery road) ; jalan_rusak(damaged roads) ; jalan_sempit(narrow street) ; persimpangan(crossing) ; tergelincir(slippery)
Kendaraan (Vehicle)	ban_pecah(flat tire) ; lampu_mati(off) ; mogok(breaking down) ; pecah_ban(flat tire) ; rem_blong(brake failure)
Lingkungan (Environment)	banjir(flood) ; berkabut(foogy) ; genangan_air (puddle) ; hujan_deras (heavy rain) ; penyempitan_lajur (lane narrowing)
Pengemudi (Driver)	kecepatan_tinggi(high speed) kehilangan_kendali(lost control) ; kelelahan (fatigue) ; lalai (neglectful) ; mabuk (drunk) ; melamun (daydreaming) ; melanggar_lampu (breaking traffic lights) ; melawan_arus (against the current) ; menerobos_lampu (break through the traffic lights) ; mengantuk (sleepy) ; mengebut (speeding) ; ngantuk(sleepy) ; ngebut(speeding) ; rem_mendadak (sudden break) ; tertidur (fell sleep; tidak_terampil (unskilled)

Table 6: Tagging of NER entities.

Class	Tag	Description
Actor	START:actor entity_name END	The data needed is a car, motorcycle, bicycle, truck, bus, rickshaw, and train.
Location	START:location entity_name END	The data needed is data on the name of the province, district, and city.
Time	START:time entity_name END	The data needed is the date the tweet was posted.
Remark	START:keterangan entity_name END	The data needed is the cause of the accident.

Table 7: Number of model testing and training.

Model	Number of Testing	Number of Training
Actor	3397	13422
Location	15752	62635
Time	15511	62055
Information	14073	56077

Table 8: classification results based on ontology.

Class Actor	Number of Tweets	Class Traffic Accident	Number of Tweets
Car	4064	Street	171
Bus	1084	Vehicle	373
Truck	2642	Environment	9
Motorcycle	2122	Driver	252
Bike	99	-	-
Rickshaw	28	-	-
Total	10039	Total	805

5.2 Results of Ontology Classification

The results obtained after the classification process can be seen in Table 8. The results are tweets with a period of posting time in January 2018 - June 2019. The vehicle is the subclass that mostly appears in the tweets, where car is the most popular actor, followed by truck and motorcycle.

5.3 Dashboard Visualization

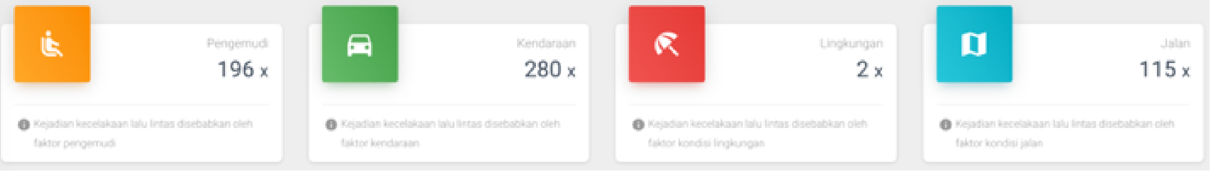
We display a visualization result by using Code Igniter on the location chosen or the actor selected.

Figure 5. presents the summary of the number of categories of causes of traffic accidents that occurred during 2019 from January to June. While Figure 6 displays the number of causes of traffic accidents during 2018. As seen in Figure 5 and 6, traffic accidents are generally caused by inadequate vehicle conditions, while environmental factors rarely cause traffic accidents.

Figure 7 shows statistics of traffic accidents every month during 2019 and Figure 8 in 2018. From the statistics, there is no specific pattern, so that shows that time does not affect the level of traffic accident events.

6 Conclusion

The purpose of this research is to extract information about Indonesia traffic accidents from Twitter. Also, we can find out the distribution of traffic accident data per the specified entities. The use of Named Entity Recognition with the OpenNLP API as a method of making models for the extraction of information on tweets can provide quite a high accuracy. The use of ontology as a keyword classification framework obtained from the NER process helps determine the categories of information extraction collected using the Jena library. The use of ontology as a knowledge model makes it possible to share the knowledge



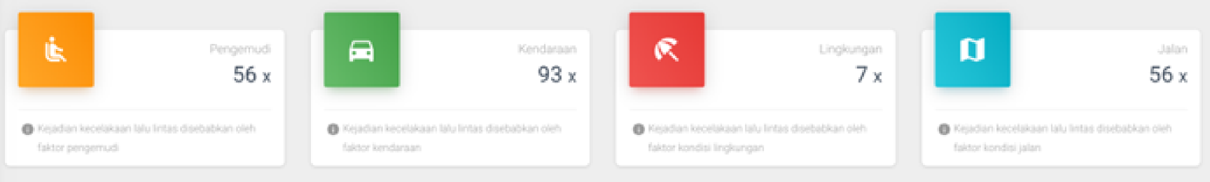
Pengemudi
196 x
Kejadian kecelakaan lalu lintas disebabkan oleh faktor pengemudi

Kendaraan
280 x
Kejadian kecelakaan lalu lintas disebabkan oleh faktor kendaraan

Lingkungan
2 x
Kejadian kecelakaan lalu lintas disebabkan oleh faktor kondisi lingkungan

Jalan
115 x
Kejadian kecelakaan lalu lintas disebabkan oleh faktor kondisi jalan

Figure 5: Number of traffic accident categories in 2019.



Pengemudi
56 x

Kejadian kecelakaan lalu lintas disebabkan oleh faktor pengemudi

Kendaraan
93 x

Kejadian kecelakaan lalu lintas disebabkan oleh faktor kendaraan

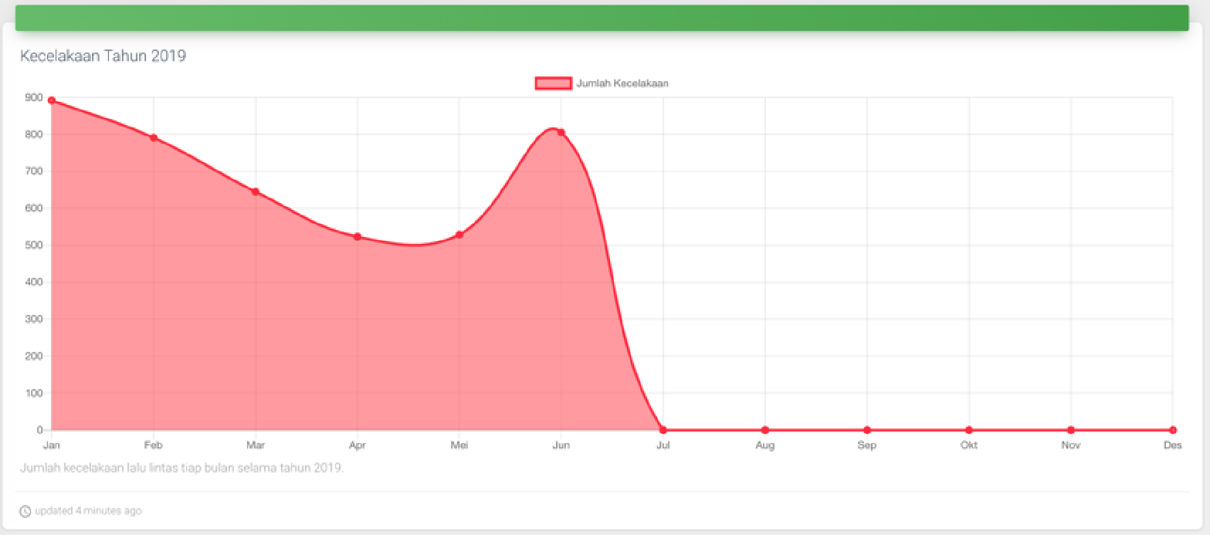
Lingkungan
7 x

Kejadian kecelakaan lalu lintas disebabkan oleh faktor kondisi lingkungan

Jalan
56 x

Kejadian kecelakaan lalu lintas disebabkan oleh faktor kondisi jalan

Figure 6: Number of traffic accident categories in 2018.

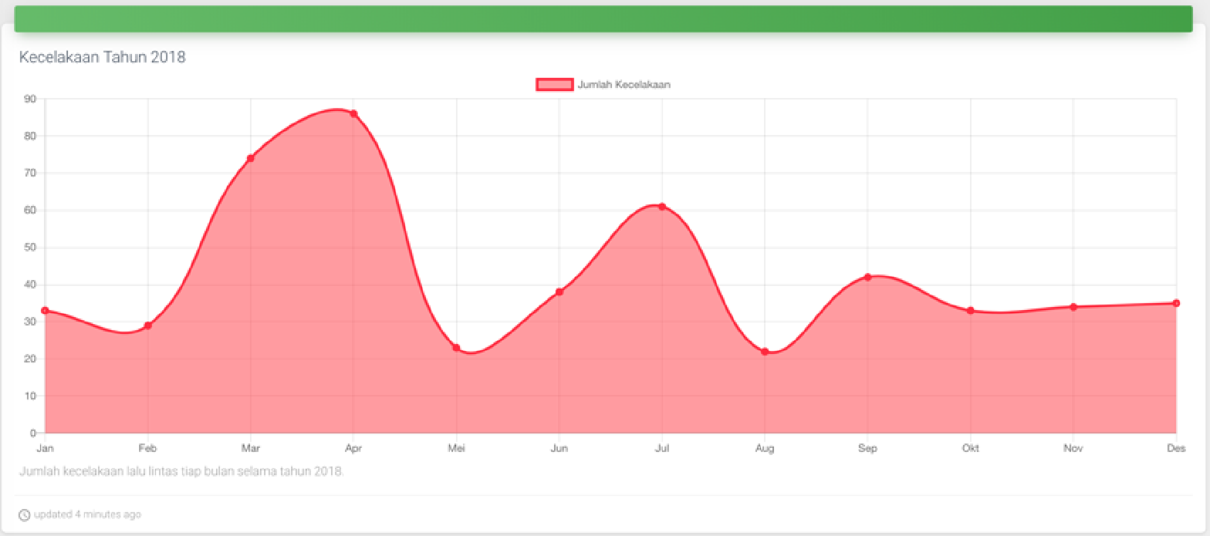


Kecelakaan Tahun 2019

Jumlah kecelakaan lalu lintas tiap bulan selama tahun 2019.

updated 4 minutes ago

Figure 7: Number of traffic accident each month in 2019 (*Jumlah kecelakaan* = Number of traffic accidents).



Kecelakaan Tahun 2018

Jumlah kecelakaan lalu lintas tiap bulan selama tahun 2018.

updated 4 minutes ago

Figure 8: Number of traffic accident each month in 2018 (*Jumlah kecelakaan* = Number of traffic accidents).

that has been formed in this study to be used and developed in further research. From the processing of the data generated, it shows that the most significant factor causing traffic accidents is the vehicle factor with the main problem in the vehicle caused by brake failure. This research can still be further developed by filtering tweets that do not specifically address traffic accident events. In the next steps, data retrieval and the process can be applied in real-time so that data can follow the development of information on Twitter. Thus it can become one of the more useful sources of information.

Acknowledgement

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

References

[1] Mohammed A. A. Al-qaness, Mohamed Abd Elaziz, Ammar Hawbani, Aaqif Afzaal Abbasi, Liang Zhao, and Sunghwan Kim. Real-time traffic congestion analysis based on collected tweets. In *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*. IEEE, oct 2019.

[2] Fábio C. Albuquerque, Marco A. Casanova, Hélio Lopes, Luciana R. Redlich, José Antonio F. de Macedo, Melissa Lemos, Marcelo Tilio M. de Carvalho, and Chiara Renso. A methodology for traffic-related twitter messages interpretation. *Computers in Industry*, 78:57–69, may 2016.

[3] Sam Aleyadeh, Sharief M.A. Oteafy, and Hossam S. Hassanein. Scalable transportation monitoring using the smartphone road monitoring (SRoM) system. In *Proceedings of the 5th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications*. ACM, nov 2015.

[4] Amani AlFarasani, Tahani AlHarthi, and Sarah AlHumoud. ATAM: Arabic traffic analysis model for twitter. *International Journal of Advanced Computer Science and Applications*, 10(3), 2019.

[5] Hasna El Alaoui El Abdallaoui, Abdelaziz El Fazziki, Fatima Zohra Ennaji, and Mohamed Sadgal. Decision Support System for the Analysis of Traffic Accident Big Data. *Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018*, pages 514–521, jul 2018.

[6] jonbakerfish. Tweetscrapper. https://pypi.org/project/tweetscraper/.

[7] Kementerian Komunikasi dan Informatika. Rata-rata Tiga Orang Meninggal Setiap Jam Akibat Kecelakaan Jalan, aug 2017.

[8] Lei Lin. *Data science application in intelligent transportation systems: An integrative approach for border delay prediction and traffic accident analysis.* PhD thesis, State University of New York at Buffalo, 2015.

[9] Amin Shahraki Moghaddam, Javad Hosseinkhani, Suriayati Chuprat, Hamed Taherdoost, and Hadi Barani Baravati. Proposing a framework for exploration of crime data using web structure and content mining. *Research Journal of Applied Sciences, Engineering and Technology*, 6(19):3617–3624, oct 2013.

[10] Daniel E. O'Leary. Is knowledge management dead (or dying)? *Journal of Decision Systems*, 25(sup1):512–526, jun 2016.

[11] Zhen (Sean) Qian and Carnegie-Mellon University. Real-time incident detection using social media data. may 2016.

[12] Angelica Salas, Panagiotis Georgakis, and Yannis Petalas. Incident detection using data from social media. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, oct 2017.

[13] Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. Protege. https://protege.stanford.edu/.

[14] Mátyás Szántó and László Vajta. Towards an intelligent traffic control system using crowdsourcing, based on combined evaluation of weather information and accident statistics. *Időjárás*, 123(3):295–312, 2019.

[15] Mohammad Mustafa Taye. Understanding Semantic Web and Ontologies: Theory and Applications. *Journal of Computing*, 2(6), jun 2010.

[16] The Apache Software Foundation. Open nlp library. https://opennlp.apache.org/.

[17] Suwardjoko Probonagoro Warpani. *Pengelolaan lalu lintas dan angkutan jalan.* ITB Bandung, Bandung, 2002.

[18] Sakkachin Wongcharoen and Twittie Senivongse. Twitter analysis of road traffic congestion severity estimation. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, jul 2016.

[19] M. Anil Yazici, Sandeep Mudigonda, and Camille Kamga. Incident detection through twitter. *Transportation Research Record: Journal of the Transportation Research Board*, 2643(1):121–128, jan 2017.

[20] Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni. A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies*, 86:580–596, jan 2018.

[21] Fan Zuo, Abdullah Kurkcu, Kaan Ozbay, and Jingqin Gao. Crowdsourcing incident information for emergency response using open data sources in smart cities. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(1):198–208, oct 2018.