Learning Drivers' Behavior Using Social Networking Service

Yueqing Li[1]✉, Acyut Kaneria[1], Xiang Zhao[2], Vinaya Manchaiah[3]

[1] Department of Industrial Engineering, Lamar University, Beaumont, TX 77710 USA
[2] Shanghai Chengtou Environment Group Co., Ltd, Shanghai 20060 China
[3] Department of Speech and Hearing Sciences, Lamar University, Beaumont, TX 77710 USA
{yueqing.li, akaneria}@lamar.edu, zhaoxiang@hjshy.com, vinaya.manchaiah@lamar.edu

Abstract. This study analyzed the driving behavior and accidents related to traffic accidents using twitter tweets as a tool for text mining. Sharing short real time messages on twitter is becoming a popular and powerful microblogging tool which conveys more than 400 million messages per day. Active users when encounter any traffic incidents, posts instant messages on twitter. Tweets with key word "traffic accident" were collected using Google Sheets with twitter's legal API keys obtained for research purpose. Various analyses were on 40,000 collected tweets performed on these tweets and was represented graphically using tableau analysis software and Rstudio. This method proved to be an effective and inexpensive method to study peoples' real time approach on traffic accident throughout the world. It proved to be a strong approach towards learning traffic accident behaviors.

Keywords: Text Mining · Twitter · Traffic Accident · Injuries · Keyword · Data Analysis.

1 Introduction

Road traffic accidents which disturbs ongoing traffic operation and can cause serious problems in the society all around the world. A minor human/weather/machine error causing major traffic accidents which mostly leads to fatalities, non-repairable damages and injuries. The direct participants of the accident and their families are the ones primarily affected by the consequences of road traffic accidents. According to statistical estimations provided by World Health Organization in 2018, yearly over 1.35 million people dies and between 20 to 50 million people are left disabled or injured due to road accidents. Unless immediate action is taken on this issue, this number will increase up to 2.4 million deaths by 2030 [1]. These large number of fatalities will consequently increase its effect on social sphere around us. National Highway Traffic Safety Association (NHTSA: a branch of US Department of Transportation) published most recent detailed traffic safety facts in 2015, which reports nearly 6.3 million Police-Reported Motor Vehicle Traffic Crashes [2]. About 28% of these 6.3

✉ Corresponding author: Yueqing Li, yueqing.li@lamar.edu

million reported crashes lead to fatalities/injuries, and the rest 72% lead to property damage only.

Twitter, a microblogging tool provides a strong platform and the ability for the users to chat, convey, share information and news reports, and discuss their ideas. Short messages containing text, links to websites, videos or pictures as well as hashtags (words beginning with #, eases in searching terms related to it) are called tweets. The treasure of information available from twitter where active users posts over 400 million tweets per day has prompted an extensive and diverse body of research in almost every area related to our daily life [3]. Active users when encounter any incidents, post instant messages on twitter including posts related to traffic incidents. As transportation plays an important role for a person's daily activities, the driving behavior and accidents related to traffic accidents was studied and analyzed using twitter tweets as a tool for text mining.

2 Literature Review

Researchers in 2015 presented real-time monitoring system for traffic event detection using Twitter stream analysis method [4]. They by solving multiclass classification problems obtained 88.89% accuracy in discriminating if traffic is caused by external event or not. Yiming Gu in 2016 applied the process of adaptive data acquisition methodology to show that mining tweets holds great potentials to complement to existing traffic incident data in a very cheap way [5]. The tweets were geocoded to determine their locations and found that tweets are more likely to report incidents near the center of a city and volume decays outwards from the center. A deep learning approach was studied and validated by comparing accident-related tweets and traffic accident logs to prove nearly 66% of accident-related tweets can be located by the accident log [6].

Technique of analyzing road accidents was proven as a reliable technique in 2016 by Sachin Kumar using *k*-means algorithm to divide the accident count based on high and low frequency accident locations [7]. A text mining-based prediction model using fault tree analysis (FTA) and Bayesian Network (BN) was used to find basic events concerning occupational accidents in steel industry sector [8]. Tibebe Beshah applied data mining technologies to link recorded road characteristics to accident severity in Ethiopia and developed a set of rules that could be used by the Ethiopian Traffic Agency to improve traffic safety [9]. A concept of data mining was used by a group of researchers in 2012 to automatically estimate the severity of traffic accidents [10]. The results proved that data mining classification algorithms and relevant features based on impact direction can be used to predict the severity of new accidents.

Researchers worked on to emphasize the significance of data mining classification algorithms in predicting the factors influencing road traffic accidents (dataset from-Fatality Analysis Reporting System) specific to injury severity [11]. Potential hidden data in the collected road accident statistics with large number of users and huge volume of data, are explored using Automatic and Visual Data Mining (VDM) methods [12]. Lawrence O. Gostin in 2010 discussed the risks and fatalities occurred as a result of distraction while driving and also educates about improvements in US Department

of Transportation rules and regulations towards incorporating advanced safety equipment for auto manufacturers [13].

Researchers have been conducting research on learning various means to measure traffic accidents, its location, time and other related information. High costs are associated with such research as expensive driving simulators and equipment are used. Despite its wide use in other domains, there is currently very less research performed on studying and analyzing traffic accidents using twitter data.

3 Methodology

This study uses tweets from twitter as a medium to source data from. Twitter developer platform assigns API keys which is used only for this research. The purpose of this study was to get an effective and efficient method to extract traffic accident data using twitter. TAGS (Twitter Archiving Google Sheet) a Google Sheet template allows us to setup and run automated collection of search results from Twitter. TAGS on Google Sheet was used to collect tweets related to keyword "traffic accident" using API keys acquired for research purpose only. Thus, we did not consider similar words like construction works, traffic congestion, etc. because these keywords may not always indicate traffic accidents and would hurt the accuracy of our results.

3.1 Data Acquisition:

Twitter data for eighteen consecutive days is extracted using "traffic accident" as the key word. Each tweet contains the following data along with it: user information, date/time posted, screen name, retweet information, user followers count, user friends count, longitude and latitude. The data collection was easy and accurate using TAGS template on Google Sheet as it allows us to automatically update results with fresh tweets every hour of the day. Preprocessing of raw tweet database was performed which is then used for further analysis. This collected data were then saved every day separately in Microsoft Excel files.

3.2 Data Analysis:

The tweets from eighteen individual data sets (one per day) were then combined all together and were saved in a single file. Combined data set consists of 41440 tweets including original tweet, retweet and @mention. Tableau Desktop Professional Edition was used as a data analysis tool to obtain meaningful results from the tweets. Also, word frequency chart was obtained using Rstudio to learn and understand the nature of accident related words in the tweets. The results section below represents and explains the graphical analysis performed on the tweets.

4 Results

In Fig. 1, distinct count of id Str (number of tweets represented by individual Id) vs. Hour of Time (Jan, 23 through Feb, 9) chart, three types of tweets i.e. original tweet,

retweet and @mention are represented using orange, red and blue colored lines respectively.

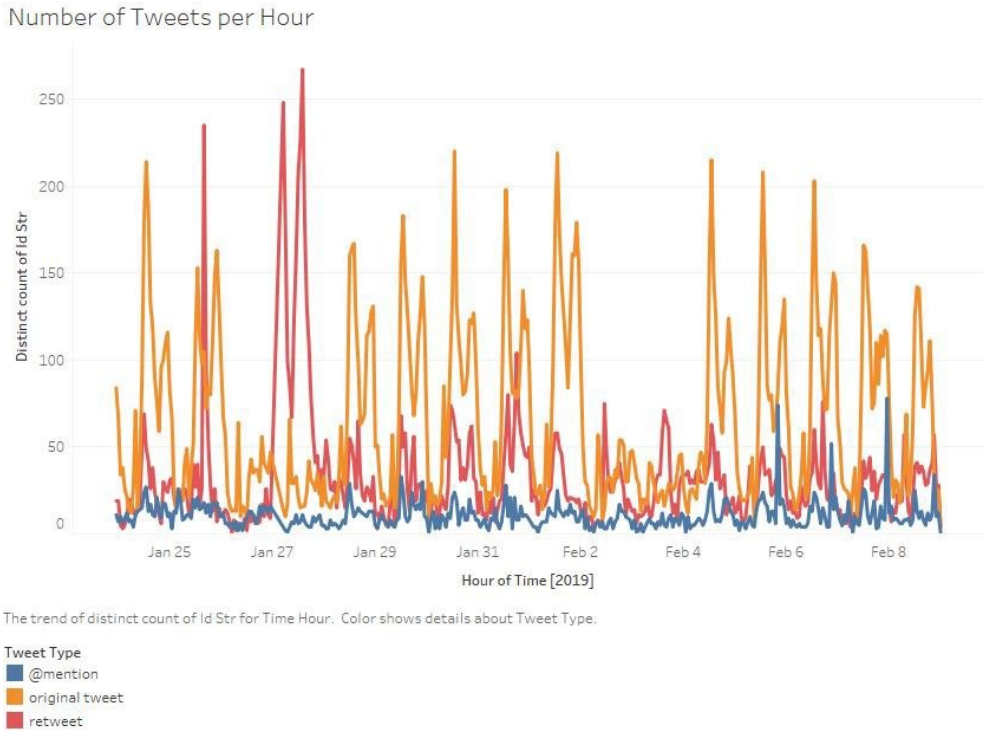

Fig. 1. Number of tweets per hour from Jan 23 through Feb 9.

Similarly, Fig. 2 shows the number of tweets in terms of days. It can be observed from the graph that original tweets are posted the most during weekdays and, on the other hand people retweets the most during weekends. This is because tweets related to accidents are usually posted by people and organizations related to news where they use twitter as a news sharing and communicating tool. Very few of them are usually active during the weekends and so the flow of tweets decreases during that period. The target of these news companies would be to make their news reach to as many common people as possible. Whereas, most of these common people seems to be more active to twitter during weekends who encounter these posts and retweets the most during that period. And @mention which is the post being referred to people or any other post related to similar incident, are comparatively very less and are rather consistent throughout the period of data collection.

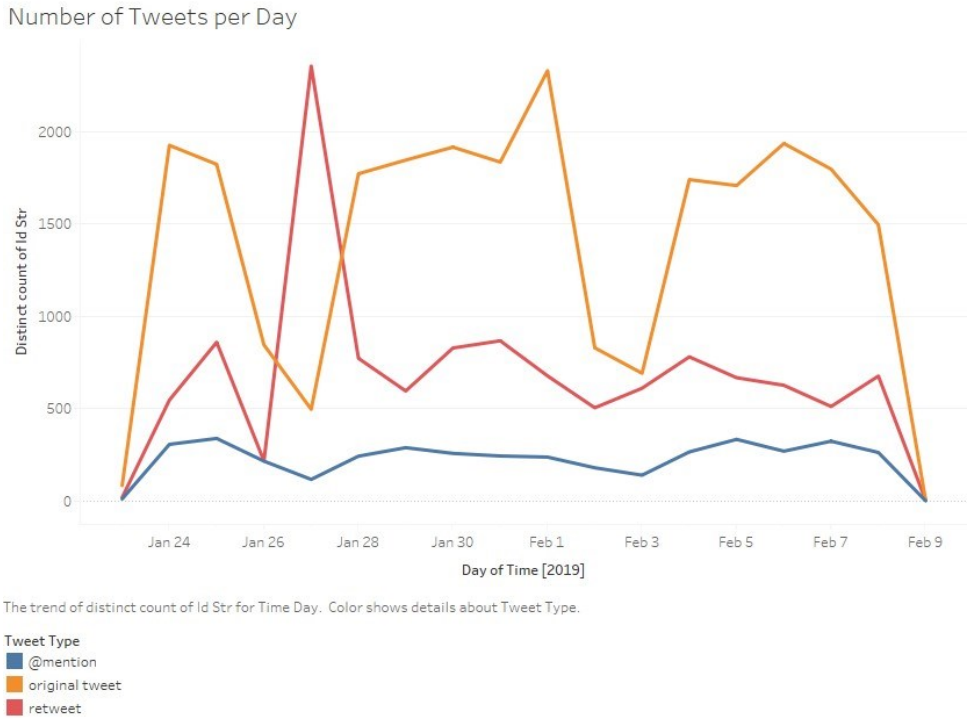

Fig. 2. Number of tweets per day from Jan 23 through Feb 9.

Fig. 3 shows the messages that Twitter users found most worthy of sharing. Re-tweets frequency chart indicates the frequency of retweeting activity on a tweet. This graph contains frequency of retweets ranging from minimum of 35 to the maximum of 2382. As an example, the most retweeted post by @shannonrwatts with nearly 2300 retweets would have good source of information related to our keyword “traffic accident”. This can be considered a good indication of interest and support from to-wards a tweet.

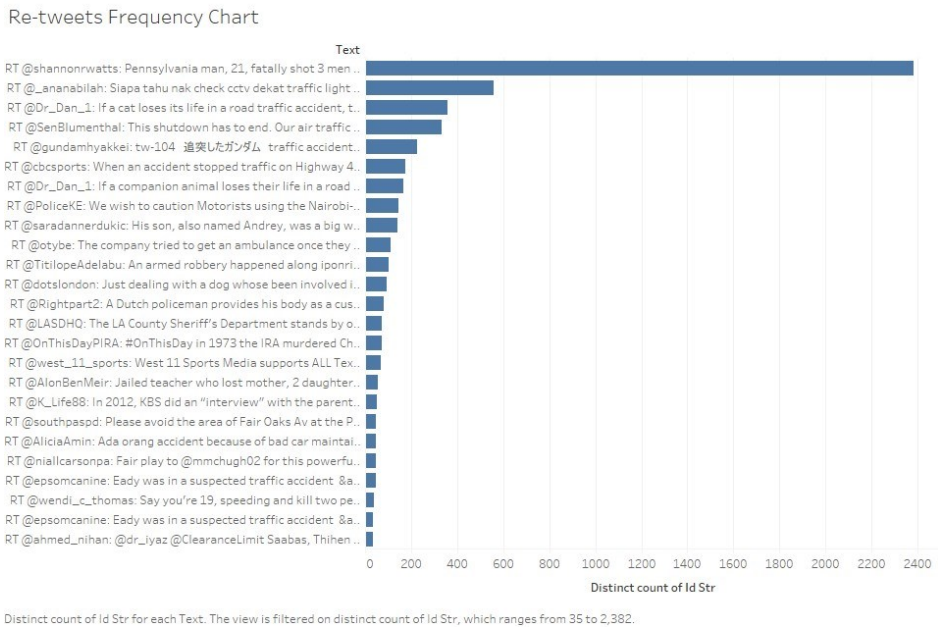

Fig. 3. Retweets frequency ranging from 35 to 2382 tweets.

Fig. 4 shows the most active users in our dataset. The user activity comparison chart shows all the information we need to examine the different activity patterns in our dataset. In our sample of "traffic accident" data, turns out to be dominated by several accounts that focus almost exclusively on original tweets, shown in orange. News accounts, spammers or individuals obsessively ranting at the world are the one usually posting these posts. Also, there are few accounts that post only @mention tweets, which is shown in blue. Rare accounts are retweets (shown in red color) that displays relatively balanced tweeting activities across all three tweet types.

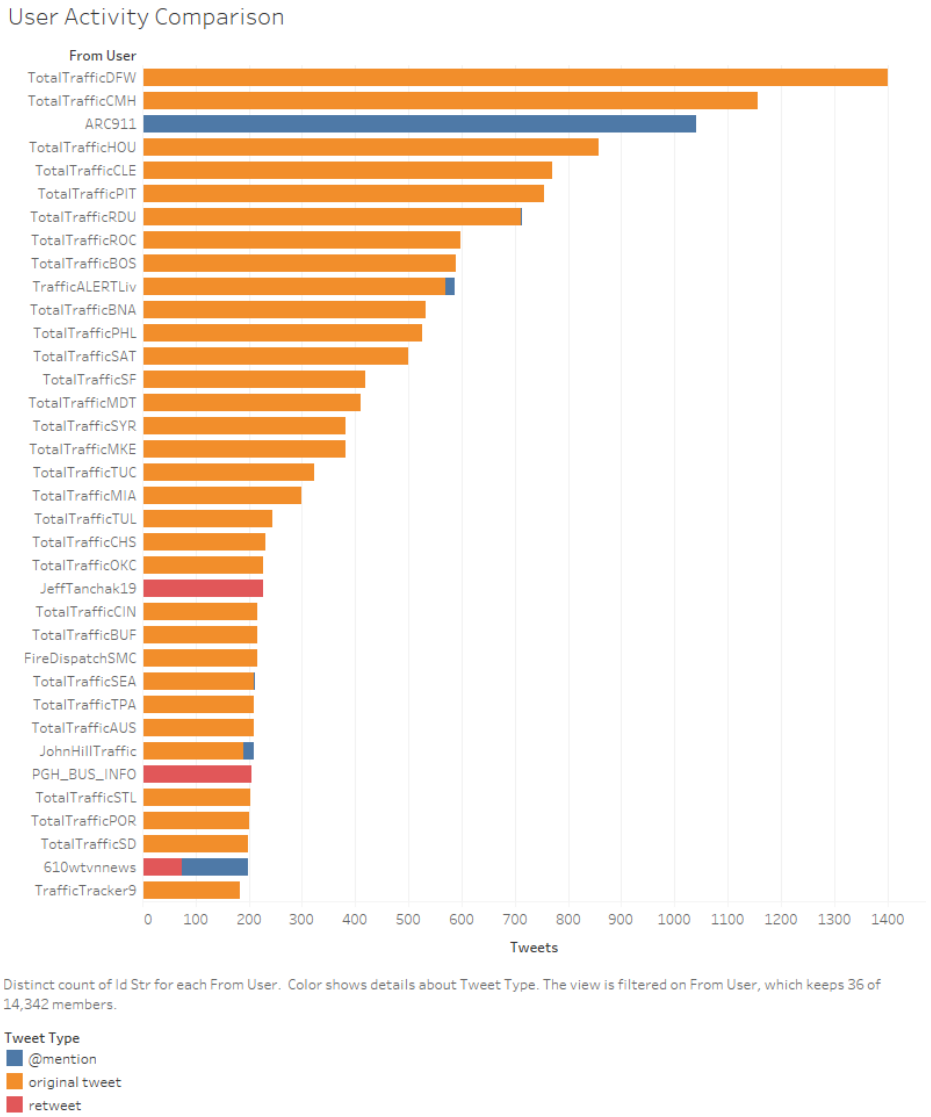

Fig. 4. Number of tweets per user.

Fig. 5 representing tweets (district count of Id str) indicating the most @mentioned and retweeted active users.

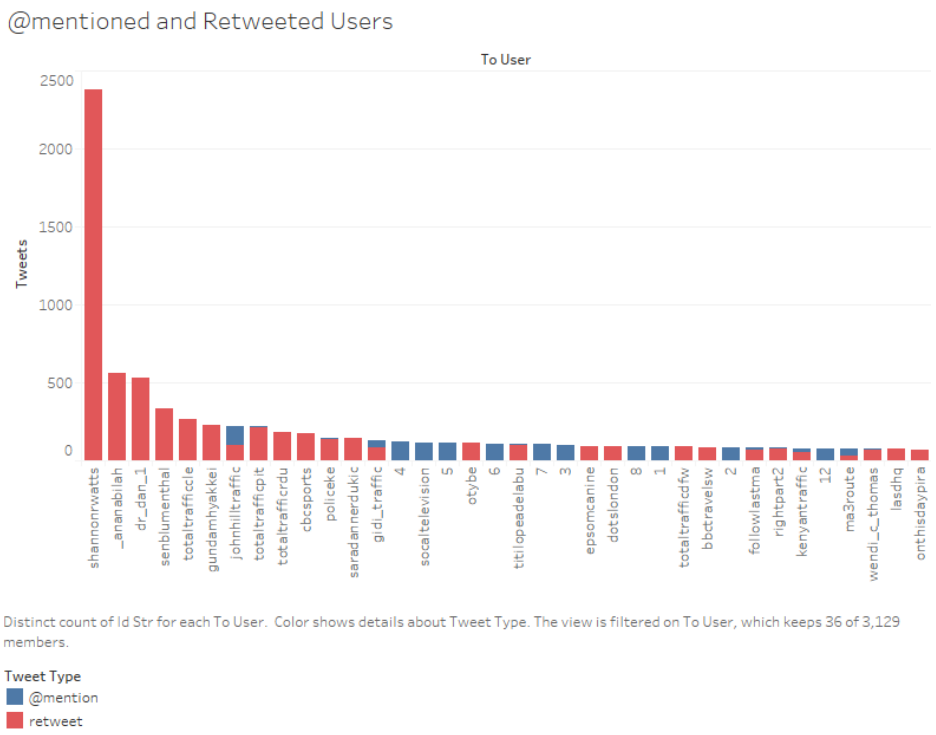

Fig. 5. Number of @mentioned and Retweets per user.

Though we used a keyword while acquiring tweets for this research, but there are users who would share their content before they decide which hashtag to use. And then links a hashtag later in retweets or after the text which won't appear in the hashtag hub are all known as secondary hashtags. The area graph on Fig. 6 lists the secondary hashtags that are contained in our dataset and indicates the hashtags with highest frequency are related to traffic accident representing the name of the cities where traffic related posts are linked to. There is a consistent usage of hashtag "traffic" observed at an average of 100 to 200 tweets per day (mostly in grey color), which proves the involvement of people with traffic.

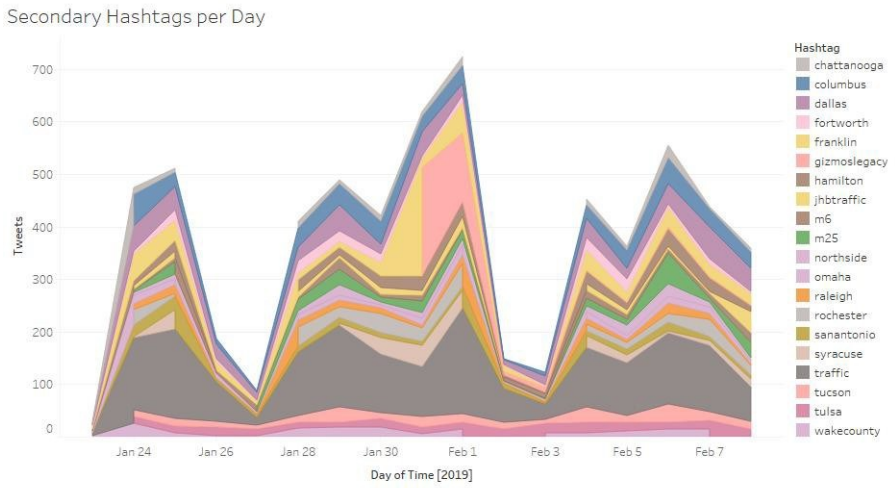

Fig. 6. Secondary hashtags frequency per day from Jan 23 through Feb 9.

Fig. 7 shows the word-cloud with 200 most frequently used words related to traffic accident. This word-cloud frequency chart was obtained using Rstudio to learn and understand the nature of accident related words in the tweets. The first step towards this process includes extracting the words possibly related to traffic accidents. Followed by another set of codes designed to remove all the excess words used to form a sentence (such as "I", "is", "are", "the", etc.) and other unrelated data from the sheet. Finally, the most frequently used 200 words related to traffic accidents are saved in a .csv file which is used to perform analysis on. The frequency of the word is the times that word is used in the data set.

Fig. 7. Word cloud shows frequency of words used with different font sizes.

5 Discussion

This research mainly explores the use of text mining tools to identify tweets related to traffic accident. Using data analysis, the traffic accident tweets are treated as a quantitative data in terms of timeline, frequency of tweets and types of user activities. Results indicates that twitter users shared information from two main sources- media reports or facts about traffic accidents, and common people's personal experience with traffic accidents. Results shows the tweets in the form of reports and news from media agencies are likely encountered with highest frequency. Many of these tweets are retweeted which apparently reflects its association with the users directly or indirectly, which ensures increased exposure to the same valuable information. Interestingly, three of the most frequently used terms (Fig. 7), "iphone", "href" and "https", relates itself with the use of technology which directly affects traffic accidents and its relationship with driving behavior.

In addition to that, words like "traffic police", "policeman", "sheriff's department", etc. are most frequently used in retweets (Fig. 3) which confirms that lawmakers are directly or indirectly associated with the term traffic accident. Terms like "animals", "shooting", "speed limit" and "traffic lights" are frequently used which as well confirms their involvement with traffic accidents. This proves the results from text mining shows some interesting patterns in the areas which require particular care. However, there are several hidden causes which indirectly influence the data which in this case is referred to traffic accident. As an example, the term "shutdown" which was at that time associated mainly with employees related to air traffic controllers and was being referred to accidents leading to an important emergency.

6 Conclusion

Twitter text analysis leaves a large imaginary gap between the actual data and the conclusions from the raw tweets. This is a very inexpensive method which can be used to reach as close to a conclusion as possible. That being said, twitter data can compliment other data sources in terms of timeliness, cost and its ability to perceive the social nature of factors affecting traffic accidents.

Using text mining and analyzing tweets proved to be a strong approach in learning the behaviors of traffic accident. The real time approach from people and their thoughts related to traffic accident can be rapidly studied by performing analysis on twitter text. This acts as a supporting tool to visualize the factors affecting as well as consequences with traffic accidents and distracted driving which moreover streamlines the future study.

References

1. ANNEX A: Summaries of Selected Health-Related SDG Indicators (World Health Organization), https://www.who.int/gho/publications/world_health_statistics/2018/EN_WHS2018_AnnexA.pdf?ua=1

2. Beshah, T., Hill, S.: Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. AAAI Spring Symposium: Artificial intelligence, pp. 14--19 (2010)
3. Abugessaisa, I.: Knowledge Discovery in Road Accidents Database Integration of Visual and Automatic Data Mining Methods. International Journal of Public Information Systems. 1, 59--85 (2008)
4. D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F.: Real-Time Detection of Traffic From Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems IEEE Trans. Intell. Transport. Syst. Intelligent Transportation Systems. 16(4), 2269--2283 (2015)
5. Fogue, M., Garrido, P., Martinez, F.J., Cano, J.-C., Calafate, C.T., Manzoni, P.: Using Data Mining and Vehicular Networks to Estimate the Severity of Traffic Accidents. Management Intelligent Systems, AISC. 171, 37--46 (2012)
6. Gostin, L.O., Jacobson, P.D.: Reducing Distracted Driving: Regulation and Education to Avert Traffic Injuries and Fatalities. Georgetown University Faculty Publications. 303(14), 1419--1420 (2010)
7. Gu, Y., Qian, Z., Chen, F.: From Twitter to Detector: Real-time Traffic Incident Detection Using Social Media Data. Transportation Research Part C. 67, 321--342 (2016)
8. Kumar, S., Toshniwal, D.: A Data Mining Approach to Characterize Road Accident Locations. J. Mod. Transport.. 24(1), 62--72 (2016)
9. Sarkar, S., Vinay, S., Maiti, J.: Text Mining based Safety Risk Assessment and Prediction of Occupational Accidents in a Steel Plant. In: International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). New Delhi, India: IEEE (2016)
10. Shanthi, S., Ramani, G.R.: Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques. In: Proceedings of the World Congress on Engineering and Computer Science, 1 (2012)
11. Traffic Safety Facts 2015 (U.S. Department of Transportation: National Highway Traffic Safety Administration), https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384
12. Twitter Business, https://business.twitter.com/
13. Zhang, Z., He, Q., Gao, J., Ni, M.: A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data. Transportation Research Part C: Emerging Technologies. 86, 580--596 (2018)