International Journal of Road Safety

Journal homepage: www.ijrs.my

OPEN ACCESS

Data Gathering and Preparation for Social Media Data Utilisation in Road and Traffic Safety

Hazqeel Afyq Athaillah Kamarul Aryffin[1], Sharifah Sakinah Syed Ahmad[1,*], Mahathir Muhammad Rafie[2] & Zulhaidi Mohd Jawi[3]

*Corresponding author: sakinah@utem.edu.my

[1]*Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia*
[2]*Aibots Sdn. Bhd., Menara Zurich, Taman Abad, 80300 Johor Bahru, Johor, Malaysia*
[3]*Malaysian Institute of Road Safety Research (MIROS), 43000, Kajang, Malaysia*

ABSTRACT

Social media has altered the communications environment. Now people are more and more acquiring news and road safety as well as traffic related info from the internet, particularly the social media. Social media portals function as new sources of prolific observational data for studies based on opinions. Although the number of research works that deploy social data is increasing speedily, only a limited number of these works clearly describe their approaches for gathering and refining the data. Search filters as well as keywords used for social data constitute the basis on which researchers might see what and how individuals communicate regarding a specific subject. Such technologies are also beneficial for enabling them to externalise their individual experiences as well as views, and share them with other people. In this work, we come up with the generic process of how to gather social media data related to road safety and traffic concerns, utilisation and setup through direct scraping and Application Programming Interfaces (APIs). Furthermore, we deliberate the sampling technique and ethical concerns encompassed in data gathering through social media. The contribution of this work lies in dealing with the technical issues when it comes to mining social media information from the perspective of road safety and traffic, setting up the bases for more exploration in this domain.

ARTICLE INFO

Article History:
Received 24 July 2021
Received in revised form
11 Oct 2021
Accepted
25 Oct 2021
Available online
01 Nov 2021

Keywords:
Social media
Road safety
Filtering data

1. Introduction

Road traffic injuries (RTIs) are a public health issue wherein the overall civic society and decision makers have to unavoidably face major fatalities and disabilities. RTIs are a primary cause of fatality within the age group of 5 to 29 years. As per the 2018 World Health Organization (WHO) report, 1.35 million individuals die every year due to RTIs while 20 to 50 million are hurt. These are the most significant reasons of fatalities in low- and middle-income nations. Around 93% of road traffic fatalities take place in these nations. The monetary effects of RTIs are also substantial. As per WHO, RTIs constitute 2% to 7% of gross domestic product (GDP) in low- and middle-income nations (Collaboration, 2011). At a young age, men are more likely to be part of road traffic accidents compared to women. Around 73% of all road traffic fatalities take place among young men under the age of 25 and they are nearly 3 times as likely to die in a road traffic accident as compared to young women (WHO, 2018). This indicates that a significant percentage of young drivers are a part of road traffic mishaps or occurrences. Most of the times, such mishaps result in grave harm or misfortune. Young drivers are certainly associated with the social media and it is this channel several regional and state safety authorities are looking forward to. The majority of these road safety campaigns aim at young males. This is because the rate of mishap within this demographic is very high, and disparate as against other age groups of drivers. Irrespective of what triggers a high rate of mishaps in young men, education is surely crucial. Secure driving habits among young individuals should be encouraged, and social media is a potent tool influencing this essential change.

Social media “reliance” is mounting each day, particularly to the youths. Road safety administrations in several nations are moving to and espousing this means, to increase coverage, aim at particular demographics and convey robust messages to a restricted audience. Generating of “real life” situations which endorse secure driving practices, like “do not drink and drive” and “slow down”, usually has been done by means of mainstream media channels. However, there is a change in this trend.

Furthermore, the human conducts and exchanges on social media have upheld themselves as extremely vigorous real-time social systems signifying individual social mindfulness at fine spatial, digital, and temporal resolutions. Social media has progressed during the past decade to emerge as a primary driver for obtaining and disseminating information in diverse fields like education (Klar et al., 2020), business (Wibowo et al., 2021); (Sivarajah et al., 2020), crisis management (Saroj & Pal, 2020) (Bridgman et al., 2020) and politics.

Utilising social media is a valued opportunity to bond individuals around a mutual topic (Amirmokhtar Radi & Shokouhyar, 2021). Social media platforms, like Twitter, Facebook, LinkedIn, YouTube, Foursquare, or Flickr, facilitate sharing of knowledge by means of the internet, along with idea conception, interaction with individuals with similar thinking, and distribution of information. It allows users to express their view on any subject which impacts their life. A research by (Jin et al., 2014) indicated that individuals, particularly young grown-ups, are attached to social media constantly or quite often in their day-to-day lives. The developments in social media utilisation present fresh prospects for examining numerous features of communication patterns. For instance, social media information could be scrutinised to obtain understandings into concerns, fashions, important actors and other types of data.

Extraordinary public trust with social media has offered feasible and culturally pertinent bases for use in road safety involvements. Nevertheless, there is an argument that means of assessing engagement in these involvements have not stayed in line with their execution. Of late, the upsurge in social media analytics (SMA) and web-based marketing has driven the growth of analytic tools, which look promising for these kinds of tasks. The main objective in this study is to implement the best method for data gathering and preparation with regards to social media data utilised in traffic and road safety. An effectual technique for quality data is quite vital so as to obtain a better reaction in terms of public mindfulness for road safety. The practice of crawling, processing and refining public tweets in real time would be discussed comprehensively. The tweets are then scrutinised to mine incident information by means of an easy process based on Natural Language Processing methods.

This paper is structured as follows. Section 2 reviews social media studies associated with traffic and road safety. Section 3 outlines the theoretical background of data collection and quality assessment regarding road safety topics. Section 4 elucidates the approach behind tweets-related data. Section 5 presents the conclusions and discussed them.

2. Social Media and Their Significance for Road and Traffic Safety

Of the different reasons of mishaps, the human aspect is the foremost one. As per the multiple causation theory, within the MVE (man/vehicle/environment) structure, a human behavioural aspect is seen in 90% of accidents, an ecological aspect in 30% and a vehicle aspect in 10%. Thus, we need to certainly act on the behavioural aspect, but the question is "how?" (Assailly, 2017).

Road Safety Education (RSE) raises the public's mindfulness about the significance of road safety. Education is a foremost approach of traffic safety, and also one of the "four Es", with the others being engineering, enforcement, and emergency systems. Education is definitely not the approach delivering the faster gains: when you convert "X crossroads" into a roundabout, you see instant positive impacts on accidents with the mechanical decline of speeds and the containment of lateral collisions, while an awareness act might render impacts over the long term. Nonetheless, we cannot install roundabouts all over. We cannot convert a nation into a huge secure playground; we cannot have a single cop watching each driver due to inadequate police resources and social tolerability. Thus, even though not the most effectual, we require active cognizance approaches to formulate a stable and all-inclusive traffic safety programme.

On the social media, individuals share their opinions and sentiments regarding happenings in the tangible world. These kinds of happenings could be openly stated or indirectly mentioned within their posts. For example, few social media posts may openly comprise a link to a news article associated with road safety they might want to discuss, while other posts may convey the users' approach. Actions as well as interactions in the digital format along with regular status updates could express themselves as extremely vigorous real-time social setups that allow the government to frame proper policies for the pertinent groups and targeted publics (Shi et al., 2014). The electronic trails and perceptions left by social media consumers and products of intricate social networks could be used to improve the outline of location-based services (Ye & Lee, 2016). Therefore, there have been growing demands for charting and scrutinising social media data, which entail more inventive technological and conceptual developments in computational and visual approaches. These exploratory challenges and prospects could enable a paradigm change in the extensive social science fields in this new kind of data setting. Social media messages could portray the interrelated patterns and associations between physical space and cyberspace, and could also be disseminated promptly to a huge number of users worldwide, who might belong to diverse virtual groups (Shelton et al., 2015).

The influence of media on human lives is enormous. It is not merely a source of entertainment but it aids in making us shrewder from our opinions on key concerns of social significance. Electronic and print media aid in mindfulness about current affairs in this age of information; the effect of media, be it electronic or print, on our life, cannot be overlooked. The outburst of social media has highlighted the fruitfulness of new media for effectual communications to involve young folks. It has emerged as a huge global platform for sharing information in the real time. Social networking portals have abundant potential from the point of view of education and mindfulness, and are already being deployed for traffic and road safety.

There are worries that approaches for assessing social media-based involvements, mainly with regards to how partakers connect with them, have not stayed in line with their execution (Perski et al., 2017). On the whole, as people utilise social media, they generate original content as pictures and posts; intermingle with the content by other consumers through reactions, comments, or shares; and form network connections. Collectively, this data generates time-stamped digital records of user activity, which confer researchers the capability to pursue real-time reactions to the involvement. For instance, for group-based social media intermediations, the number of comments, posts, and reactions by every partaker could be utilised as instruments of their individual study engagement (Haines-Saah et al., 2015).

Nonetheless, social media data cannot be obtained without a price. The uncovering of road safety mindfulness based on Twitter is taxing. The advanced text mining methods cannot be implemented directly to extract tweets as the tweet language differs significantly from day-to-day language. Messages on Twitter are short (140 maximum characters) and could frequently have grammatical errors, typos, and cryptic abbreviations. Figure 1 depicts the outline for social media analytics (SMA). This is the most acknowledged one in information systems, on the basis of citations of the paper in IS related studies (Stieglitz et al., 2014). This triggers the following four-step outline:

- Discovery: The "revelation of dormant patterns and structures" (Chinnov et al., 2015)
- Tracking: This stage encompasses decisions about the data source (e.g. Facebook, Twitter), technique, approach, and output. A thorough sub-division of this stage can be observed in Stieglitz et al. (2014). In many research works, the comprehensiveness of diverse Twitter sources was matched
- Preparation: Further than this, the original outline does not expound on the preparation steps required.
- Analysis: Based on the objective, there are many approaches available, such as opinion mining and social network analysis.

The majority of the vastly cited articles handle particular methods. For instance, Latent Dirichlet Allocation (Blei et al., 2003) is quite extensively utilised in the domain of topic modelling. They all offer event detection prototypes and deliberate challenges and complications that surface through data velocity, volume, and variety. Notably, three publications will be termed pertinent as per the norms utilised before (Weng et al., 2011). All three stimulate or deliberate their models from the perspective of every challenge, from the ever growing quantity of real-time data to the flexible and vigorous nature of the data along with the noise in tweets disguised as “futile babbles” (Weng et al., 2011). Nevertheless, quite few of the publications explicitly handle data gathering or preparation (Stieglitz et al., 2018). Apparently, researchers encountering challenges in these domains have no extensively acknowledged sources to refer.

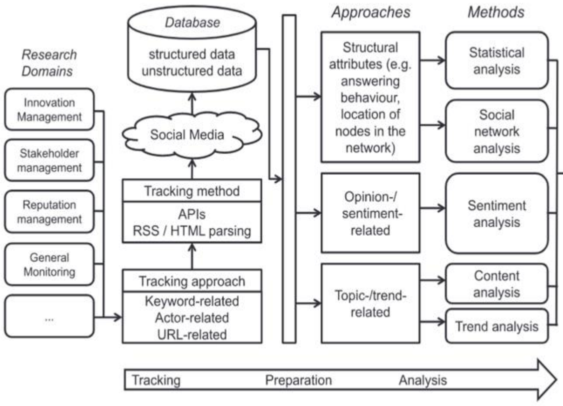


Figure 1: Social media analytics framework (Stieglitz et al., 2014)

Usually, there are three means of transportation: land, water, and air. Presently, the most utilised means is land transportation. There are different alternatives for land transportation: road or railways or off road. Examples of land transportations are motorcycle, car, lorry, bus, bicycle and train. This study would focus on land transportation which takes place on the road situated in Malaysia and their safety linked concerns. As the attribute of these events is typically fickle and could happen haphazardly anywhere, the data gathering procedure would be tough and consume lot of time. However, the advent of social media like Twitter, Facebook and Instagram has presented a novel platform for gathering views or ideas of individuals and establishments. Such platforms have seen vast rise in usage, hence driving accrual of data. In view of this, we require an apt mode of gathering this data by utilising API presented by matching platforms.

There are more than 400 million tweets posted each day on Twitter globally, with more than 100 million regular active users. It is a mega resource for individuals to share info they notice or see. Tweets are basically short messages, restricted to 140 characters. Twitter consumers comprise authoritative users (like public organisations) and individual users. Individual consumers can tweet to share direct (actual) info about events. Few instances are:

1. *Rindulah x dapat pandang kereta sendiri, macam nie ke rindu kalau tak pandang buat hati ðŸ˜…*

2. *Dia punya penat sampai pening. Dia punya pening sampai terlelap bawak kereta tadi hmmm*

3. *Nama je #TheFastSaga, takde kereta Saga pun dalam trailer. Scammer! https://t.co/n7GwADrptc*

4. *@amiirulhaqiim @magmalaya Sebab tu wujudnya ABS pada sistem keselamatan kereta sekarang. Jadi, kereta yang nak langgar dari belakang tu akan dapat elak takpun kurangkan impact perlanggaran.*

5. *Malasnya aku nakgi cuci kereta ðŸ˜’*

6. *baru ambik kereta lepas buat chemical wash, dan2 dia hujan pulak.. burit la*

7. *@adibrazak14 Orang yang pakai kereta ni mesti dia dah ready untuk hadap segala jenis kos haha*

Instagram and Facebook APIs entail workaround and possess variable privacy settings. Twitter offers an API which allows developers to obtain data streams effortlessly with a specific limit within an hour and only tweets from past seven days. Twitter API can be utilised with or without keywords, since keywords would aid in narrowing down the data which would be recovered. We are utilising the API with keywords to make sure the retrieved data pertains to road safety concerns. The procedure of choosing keywords has to be carried out sensibly as the quality of recovered data relies heavily on the keywords. If the chosen keywords are very expansive, the majority of the retrieved data would be immaterial to the subject, whereas if the chosen keywords are very narrow, the data would not echo present state of affairs appropriately and might trigger bias.

Although the keywords are selected judiciously, few of the recovered data might be immaterial to the subject. For instance, when we utilise the keyword ‘drive’ for retrieving data, some of the text might be deliberating about ‘google drive’ or ‘you drive me crazy’, and this is not related to the topic. Since the tweets are focused on Malaysia, the majority of the tweets are multi-language, hence needing the keywords to be in English and Malay. Moreover, the majority of Twitter users do not mention their location. This has to be filtered out since the location is vital for ascertaining the tweet’s location. Besides, few of the keywords might correspond with the Twitter user’s username rather than the text, and hence this has to be taken into account during the procedure of filtering tweets.

3. Conceptual Framework for Social Data Collection and Quality Assessment

The framework recommend by us comprises three main steps which warrant that the search filters can seize the most pertinent text of the chosen subjects. The recommended framework entails access to the data streams as the users have to obtain certain data understanding.

Table 2: In-text citation and reference examples.

Step	Details
Develop Search Filter	1. Build a list of keywords that are related to the topics 2. Search twitter with the keywords and then: Observe the tweets that may have suitable keywords for the topic Observe tweets that have words that are similar to keywords Observe tweets for alternate spelling or short-form of the keywords Remove keywords that displayed many out of context tweets Add list of words that usually irrelevant from the keywords 3. Repeat step 1 and 2 until no more new keywords that can be obtained 4. Filter tweets that does not specify any location (we need to ensure the tweets are located in Malaysia) 5. Filter tweets that do not have keywords (sometimes, the keywords match with the username, not with the tweets) 6. Filter tweets that have unwanted words
Apply Search Filter	1. Pre-process the retrieved data 2. Go through pre-processed retrieved data and divide text into ‘not noise’ and ‘noise’ text
Create Noise Classification Model	1. Train a fastText model using the splitted data 2. Get accuracy and recall of the model

4. Methodology

4.1. Developing Search Filters

The initial stage in formulating search filters is choosing the list of keywords. The keywords created should be associated with the subjects, like selecting common terms which are typically utilised when discussing the subject. After constructing about five to ten keywords, put those keywords on Twitter Search. Next, scan through the resulting tweets from the search to obtain fresh apt keywords regarding the subject. Furthermore, through the resultant tweets, similar keywords could be acquired. For instance, *'kemalangan'* in Malay denotes 'accident'; however, occasionally, *'langgar'* also has a similar connotation, and hence the word *'langgar'* could be added as fresh keyword.

Moreover, Twitter users typically tweet in short form because the tweets have a word limitation. Hence, any viable short form can be included as a fresh keyword. For instance, *'jalan'* in Malay denotes 'road' whereas the most utilised short form for *'jalan'* is *'jln'*. Furthermore, few of the short forms are not viable as it could denote something else; for instance, 'org' is the short form for *'orang'* in Malay, whereas 'org' could also denote 'organization' in English. Therefore, 'org' cannot be added as a fresh keyword. In case the retrieved outcomes from the keywords are generally irrelevant, the keywords can be eliminated from the list.

Lastly, look for words which typically pertain to another subject within good keywords. For instance, when employing the keyword 'drive', the majority of the retrieved tweets are regarding driving a car; however, occasionally, it could retrieve 'Google Drive' as the outcome. One more instance is *'jalan'*, which denotes 'road'; however, it could also be utilised to talk about *'jalan cerita'* which denotes 'story plot'. Therefore, 'Google Drive' and *'jalan cerita'* have to be included in the list of unsolicited words so that if the tweets contain these words, it would be cleared out.

Then, carry on with those steps till no more fresh keywords could be attained. After concluding the list of keywords, sort out the tweets which do not have particular location, since we have to make sure the tweets belong to Malaysia. Furthermore, a geocode has been utilised for making sure the tweets belong to Malaysia and adjacent nations to simply the filtering procedure. The tweets which do not include the keywords are filtered; this could occur when the keywords are matched with the username rather than the tweets. Lastly, filter tweets which contain words from the unsolicited words list.

4.2. Apply Search Filter

The API entails a list of keywords to carry out queries. We can present an early group of keywords, viz. a preliminary dictionary. Notably, the recall and accuracy are typically low as those keywords fail to cover all language signifying incidents. To warrant the finest quality and highest number of TS tweets, which can be feasibly obtained, we have to augment the dictionary for retrieving additional TS tweets. One approach is to pick up from all words of the probed tweets which do not form part of the dictionary and choose those words pertinent to incidents. We can add these fresh words to the dictionary and conduct fresh queries to attain additional TS tweets in a fresh iteration. This is called as adaptive data attainment since this approach of data attainment could be executed over time to acclimate with the most modern Twitter language. In every iteration, we would manually tag TS tweets, and calculate the frequencies of fresh words. It is assumed that the higher the frequency of a word (or a mixture of words) utilised in all TS tweets, the higher the chance of incident and geo-location info. The procedure of adaptive data attainment is outlined in Figure 2.

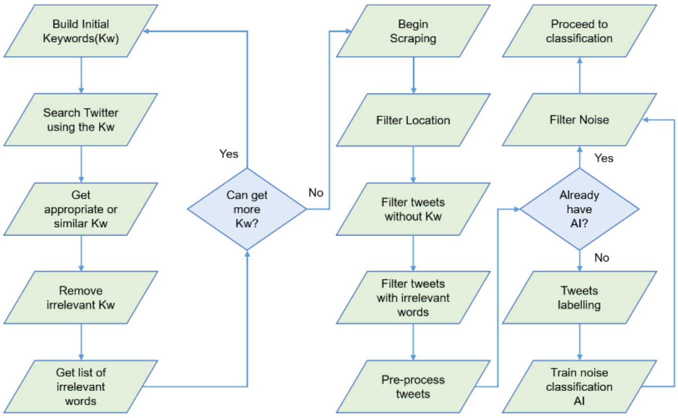

Build Initial Keywords(Kw)
Search Twitter using the Kw
Get appropriate or similar Kw
Remove irrelevant Kw
Get list of irrelevant words
Can get more Kw?
Yes
No
Begin Scraping
Filter Location
Filter tweets without Kw
Filter tweets with irrelevant words
Pre-process tweets
Proceed to classification
Filter Noise
Yes
Already have AI?
No
Tweets labelling
Train noise classification AI

Figure 2: Flow chart of adaptive data acquisition

The overall procedures are displayed in Figure 2 where the cleaned data can be used for classification to get meaningful insights. These steps ensure to get the most topic related data thus avoiding misinterpretation of current situation.

Twitter API allows developers to scrape statistics only from prior seven days, and therefore the retrieval procedure has to be carried out weekly. The retrieved info can either be kept in a database or a comma separated value (csv) file. The retrieved data then has to be pre-processed so that it could be marked and then input into the FastText model. Pre-processing of data also comprises sifting out tweets which does not belong to Malaysia.

Pre-processing aids in decreasing the corpus mass of data, thereby decreasing training time and enhancing the precision of the model. Post retrieving about ten thousand data, it can be divided into two classes: 'noise' and 'not noise'. In case the gathered data is quite huge, random sampling could be carried out as per every keyword to make sure data from every keyword is encompassed.

4.3. Create Noise Classification Model

Post ensuring the 'noise' and 'not noise' data are balanced, they have to be divided into 'valid' and 'train' files. Then, the ConceptNet Numberbatch word vector is obtained and 'Malay, English and Chinese' word vectors are mined since the majority of the Malaysians utilise these three languages when they tweet. Lastly, the data is input into the fastText model which uses the pre-trained word vectors from ConceptNet Numberbatch. Following the training procedure, the model is tested with 'valid' file and its precision and recall are obtained. In case these are low, more data labelling has to be carried out.

For acquiring the pertinent info from tweets to the maximum extent, it is necessary to formulate a mixture of keywords which renders the finest recall and realistic accuracy. Recall and accuracy as elementary scores of the data acquisition structure are described as follows:

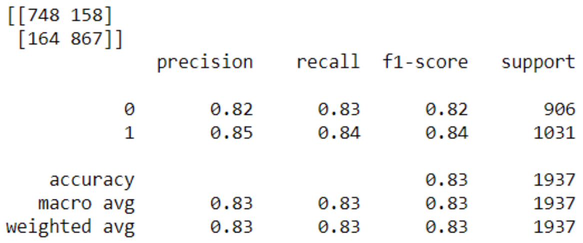
$$Recall = (A \cap B)/A \tag{1}$$
$$Precision = (A \cap B)/B \tag{2}$$

Where A is the group of all traffic safety (TS) tweets in a time span, and B is the group of all obtained tweets during the same time span. The best objective is to attain as much accuracy and recall as possible concurrently, i.e., all obtained tweets are just TS tweets and all TS tweets within the pool are obtained. Nevertheless, it is typically not possible to attain a 100% recall, or to accurately project the recall, considering the drawbacks of the overall number of tweets available through API. Moreover, there is barely any ground truth of a whole group of TS tweets. However, it could be projected by utilising the ratio of E over Q wherein Q is signified by the number of TS tweets among all the tweets of arbitrarily chosen test users from Twitter, and Q is signified by the number of tweets crawled on the basis of API (with a particular group of query factors) crossing the tweets of those test users. A 100% accuracy is tough to attain as there are always few tweets similar to the query keywords but do not pertain to traffic events. Thus, the objective of the data acquisition procedure is to attain judiciously high recall and accuracy, so that we attain as many TS tweets as possible by means of free-of charge Twitter APIs (Gu et al., 2016).

5. Results and Discussion

The tweets were combed for the period 29 November 2020 to 25 February 2020, returning over a million retrieved tweets. Furthermore, the majority of the tweets are eliminated in the pre-processing phase due to duplication, location not being from Malaysia, and categorisation as 'noise' by the fastText model, returning just 81,969 tweets for data scrutiny. There are 60 chosen keywords and 18 unsolicited words for retrieving the tweet.

The noise model utilises 9,681 tweets, wherein 5,081 are 'not noise' and 4,600 are 'noise'. When divided, 7,744 data were utilised for training and 1937 for testing. The accuracy for 'not noise' was 85%; for recall, it was 84%. The recall and accuracy plunged a bit for 'noise', valued at 82% and 83%, respectively. The precision of the model was 83%.



```
[[748 158]
 [164 867]]
              precision    recall  f1-score   support

           0       0.82      0.83      0.82       906
           1       0.85      0.84      0.84      1031

    accuracy                           0.83      1937
   macro avg       0.83      0.83      0.83      1937
weighted avg       0.83      0.83      0.83      1937
```

The challenges in the procedure of gathering data differs according on the kind of data required. Although the remarkable growth in users of social media drives a massive volume of data on the internet, several ethical concerns are still there. Therefore, the finest alternative is to utilise the API offered by respective social media. Furthermore, the scraping procedure has to be reorganised each time the API changes. This could trigger monotonous work and lead to breaking up of the present project as there could be alteration in which data could be retrieved or volume of data which could be retrieved or the layout of data retrieved. The only path for developers to evade this is by scraping the data straightforwardly; however, this could lead to ethical concerns.

Furthermore, the swift progress of data in social media is characterised by noises which could impact the quality of data gathered. Different filtering techniques could be espoused to make sure the majority of the noises could be sifted out; either entirely through human process or through machine learning. Data quality is vital to make sure the analysis outcomes echo the most precise depiction of the problem. Hence, the most effectual technique to deal with the problem could be produced easily.

6. Conclusion and Recommendations

This study recommends mining tweet texts for extracting info about traffic and road safety as an effectual and cost-effective substitute to prevailing data sources. We provide a method to crawl, process and refine tweets which are available to the public for free. Tweets are obtained from Twitter servers by utilising the REST API. The data attainment happens as per an iterative procedure. It begins with queries to APIs with a dictionary of 'preliminary keywords' and iteratively augments the dictionary as per a simple Natural Language Processing (NLP) process till the obtained tweet data set congregates. The procedure of adaptive data attainment chooses the most significant keywords and their mixtures to create a feature space which is instructive and non-redundant. Tweet texts are then charted into a high dimensional binary vector within this feature space described by the dictionary. The recommended technique entails humans to carry out physical labelling for noise data in machine learning. In case the data is quite huge, clustering could be chosen for decreasing the time taken. Besides, the recommended framework runs with text-based data, and hence it could be utilised for other blogging or social media website even though few configuration changes have to be carried out based on the data gathered like length of sentences and language.

Acknowledgements

The authors express their deepest gratitude to ASEAN NCAP Collaborative Holistic research (ANCHOR) III - ANTARABANGSA-ANCHOR/2020/FTMK-CACT/A00032 for backing this research.

References

Amirmokhtar Radi, S., & Shokouhyar, S. (2021). Toward consumer perception of cellphones sustainability: A social media analytics. *Sustainable Production and Consumption*, *25*, 217-233. https://doi.org/10.1016/j.spc.2020.08.012

Assailly, J. P. (2017). Road safety education: What works? *Patient Education and Counseling*, *100*, S24–S29. https://doi.org/10.1016/j.pec.2015.10.017

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4–5), 993-1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, *1*(June), 1-18. https://doi.org/10.37016/mr-2020-028

Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., & Trautmann, H. (2015). An overview of topic discovery in Twitter communication through social media analytics. *2015 Americas Conference on Information Systems, AMCIS 2015*, 1-10.

Collaboration, U. N. R. S. (2011). Global plan for the Decade of Action for Road Safety 2011–2020. *Geneva: WHO*, 25. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Global+Plan+for+the+Decade+of+Action+for+Road+Safety+2011-2020#0

Gu, Y., Qian, Z., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321-342. https://doi.org/10.1016/j.trc.2016.02.011

Haines-Saah, R. J., Kelly, M. T., Oliffe, J. L., & Bottorff, J. L. (2015). Picture me smokefree: A qualitative study using social media and digital photography to engage young adults in tobacco reduction and cessation. *Journal of Medical Internet Research*, *17*(1):e27. doi: 10.2196/jmir.4061

Jin, Y., Liu, B. F., & Austin, L. L. (2014). Examining the Role of Social Media in Effective Crisis Management: The Effects of Crisis Origin Information Form, and Source on Publics' Crisis Responses. *Communication Research*, *41*(1), 74-94.

Klar, S., Krupnikov, Y., Ryan, J. B., Searles, K., & Shmargad, Y. (2020). Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. *PLoS ONE*, *15*(4), 1–15. https://doi.org/10.1371/journal.pone.0229446

Perski, O., Blandford, A., West, R., & Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational Behavioral Medicine*, *7*(2), 254–267. https://doi.org/10.1007/s13142-016-0453-1

Saroj, A., & Pal, S. (2020). Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, *48*(March), 101584. https://doi.org/10.1016/j.ijdrr.2020.101584

Shelton, M., Lo, K., & Nardi, B. (2015). Online Media Forums as Separate Social Lives: A Qualitative Study of Disclosure Within and Beyond Reddit. *IConference 2015 Proceedings*, 1-12. https://www.ideals.illinois.edu/handle/2142/73676

Shi, Y., Luo, Y. L. L., Yang, Z., Liu, Y., & Cai, H. (2014). The Development and Validation of the Social Network Sites (SNSs) Usage Questionnaire. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8531 LNCS*(June), 113–124. https://doi.org/10.1007/978-3-319-07632-4_11

Sivarajah, U., Irani, Z., Gupta, S., & Mahroof, K. (2020). Role of big data and social media analytics for business to business sustainability: A participatory web context. *Industrial Marketing Management*, *86*(July 2018), 163-179. https://doi.org/10.1016/j.indmarman.2019.04.005

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics. *WIRTSCHAFTSINFORMATIK*, *56*(2), 101-109. https://doi.org/10.1007/s11576-014-0407-5

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*(December 2017), 156-168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Weng, J., Yao, Y., Leonardi, E., & Lee, B. S. (2011). Event detection in twitter. *HP Laboratories Technical Report*, *98*, 1-21.

WHO (World Health Orgainisation). (2018). Global Status Report on Road. *World Health Organization*, 20.

Wibowo, A., Chen, S. C., Wiangin, U., Ma, Y., & Ruangkanjanases, A. (2021). Customer behavior as an outcome of social media marketing: The role of social media marketing activity and customer experience. *Sustainability (Switzerland)*, *13*(1), 1-18. https://doi.org/10.3390/su13010189

Ye, X., & Lee, J. (2016). Integrating geographic activity space and social network space to promote healthy lifestyles. *SIGSPATIAL Special*, *8*(1), 20-33.