A Review of Models for Hydrating Large-scale Twitter Data of COVID-19-related Tweets for Transportation Research

Mahmoud Arafat[1]

[1]Affiliation not available

April 27, 2020

Abstract

In response to the Coronavirus disease (COVID-19) outbreak and the Transportation Research Board's (TRB) urgent need for work related to transportation and pandemics, this paper contributes with a sense of urgency and provides a starting point for research on the topic. The main goal of this paper is to support transportation researchers and the TRB community during this COVID-19 pandemic by reviewing the performance of software models used for extracting large-scale data from Twitter streams related to COVID-19. The study extends the previous research efforts in social media data mining by providing a review of contemporary tools, including their computing maturity and their potential usefulness. The paper also includes an open repository for the processed data frames to facilitate the quick development of new transportation research studies. The output of this work is recommended to be used by the TRB community when deciding to further investigate topics related to COVID-19 and social media data mining tools.

A Review of Models for Hydrating Large-scale Twitter Data of COVID-19-related Tweets for Transportation Research

Mahmoud Arafat, (Corresponding Author)
Research Assistant, Ph.D. Candidate
Department of Civil and Environmental Engineering
Florida International University
10555 West Flagler Street, EC 3730, Miami, FL 33174
Tel: (305) 450-0740; Email: mhela001@fiu.edu

ABSTRACT

In response to the Coronavirus disease (COVID-19) outbreak and the Transportation Research Board's (TRB) urgent need for work related to transportation and pandemics, this paper contributes with a sense of urgency and provides a starting point for research on the topic. The main goal of this paper is to support transportation researchers and the TRB community during this COVID-19 pandemic by reviewing the performance of software models used for extracting large-scale data from Twitter streams related to COVID-19. The study extends the previous research efforts in social media data mining by providing a review of contemporary tools, including their computing maturity and their potential usefulness. The paper also includes an open repository for the processed data frames to facilitate the quick development of new transportation research studies. The output of this work is recommended to be used by the TRB community when deciding to further investigate topics related to COVID-19 and social media data mining tools.

Keywords: COVID-19 Pandemic, Twitter Data, Machine Learning, NLP, Social Media Mining.

Word count: 3,223 words.

Submission Date*: April 24th, 2020.*

Disclosure statement*:* This is to acknowledge that there is no financial interest or benefit arisen from the direct applications of this research.

INTRODUCTION

Human behavior is considered a key element of understanding the significant impacts of Coronavirus (COVID-19) on transportation systems and traffic operations. Transportation researchers, in fact, have always been concerned with all aspects pertinent to the influence of human behavior on transportation networks. In the past few decades, researchers have developed conventional tools derived from regular transportation surveys and simulation results to analyze the impacts of human behavior on transportation safety and mobility. Recently, social media platforms have received widespread use around the world as a new communication paradigm for social interaction and information exchange in large global virtual communities. Approximately, two-thirds of American adults (65%) use social networking sites (*1*), with nearly 3.8 billion people around the world. With the increasing usage of such platforms combined with the powerful computations of machine learning techniques, scientists have adopted them as a pivotal source of generating diverse data that offers strengths and opportunities for studying human behavior. As a result, it becomes increasingly important for agencies, practitioners, and governments to utilize the benefits of such big datasets especially during extreme events such as Coronavirus (COVID-19) pandemic.

In 2019, Twitter has reported monthly active users of approximately 330 million people (*2*). Lately, with the evolution of the smartphone and cellular technologies, it was, therefore, no wonder that scientists utilized Twitter's large datasets to reveal trends and patterns related to human's social and psychological needs. However, the process of collecting and processing Twitter datasets is time-consuming and requires basic programming knowledge based on the insights that a researcher wants to draw from the data. The motivation of this paper is to extend the work that was done on social media data mining and to provide a review of contemporary open-source tools available for extracting Twitter datasets, including their computing maturity and their potential usefulness. The methodology is applied on a ready-to-use COVID-19 Twitter dataset processed for open research for further analysis promptly.

Researchers Tekumalla and Banda, (2020) published a Twitter data mining tool called Social Media Mining ToolKit (SMMT) to employ social media data for research purposes. The tool consists of various codes and scripts that are bundled for fetching and fusing the data to consequently facilitate the research development on social media data mining. The SMMT is designed as a bundle of full Python Programming codes and instructions enabling social media data acquisition while also allowing a user to delve deeper through data preprocessing, annotation, and standardization. There are three main components for utilizing the SMMT tool: Data acquisition tools, Data preprocessing tools, and Data annotation and standardization tools. More details about these components can be found in (*3*). Twitter data acquisition is the process of collecting data from Twitter streams and hydrating these data into a meaningful format. This process opens several opportunities to derive a variety of benefits, such as generating a Comma-separated (.csv) and JavaScript Object Notation (.json) file formats that can be processed with the SMMT Preprocessing tools to extract several attributes of interest.

The Twitter Application Programming Interface (API), provided by the developer, is an immensely useful tool for data mining applications. Data scientists and researchers use API to extract Twitter content under restricted use. Twitter API allows the user to perform complex queries such as extracting Tweets related to a certain topic within a specific period of time. In order to use Twitter API, simply a researcher has to have an active Twitter user account. However, such use is contingent upon a set of rules and regulations. Based on the policies and conditions of the Twitter developer agreement (*4*), Twitter contents are available for non-commercial research purposes as long as the data are shared in the form of Tweet Identifiers (Tweet IDs). For protecting the privacy of all Twitter users and their personal information, sharing Twitter's data comes only in the form of a spreadsheet that includes Tweet IDs. A Tweet ID is a unique integer representation generated solely for every Tweet. Therefore, a hydration process is required for converting this Tweet ID into meaningful Tweets with metadata. The Twitter API is used to hydrate a dataset of Tweet IDs to extract information such as the Tweet text, language used, date created, URLs, and other user-related information. The hydration process of Twitter datasets can be done by utilizing programming languages such as the Python scripts developed as part of the SMMT tool (*3*). Also, an alternative approach to hydrate Twitter datasets is by using prebuilt online released Hydrator such as DocNow Hydrator, free open-source hydration software that can be installed on any desktop with versions for different operating systems (*6*).

Since DocNow and SMMT tools are free, open-source software models, there is no commercial intent or conflict of interest out of this review to promote any model. Thanks to the developers who provide the codes and scripts to support the research society. In sum, the main goal of this paper is to prioritize the aforementioned tools based on criteria such as the tool's processing time, ease of modeling, and the likelihood of software deployment to facilitate the process of extracting more social media data for future research purposes.

LITERATURE REVIEW

The previous research efforts argue on the benefits of integrating social networking sites as a tool for generating enormous data related to human behavior. For instance, the transportation community has started utilizing web-based applications and social media platforms for collecting unprecedented datasets that can be used in research topics related to traffic management, travel demand modeling, and traffic safety. Transportation scholars explicitly consider social media platforms as a new source for extracting data related to human behavior. Apart from the big data benefits that social media platforms offer, these applications provide features that allow users to link their daily social posts to other mapping applications such as Google Maps through location-based services. The increasing usage of social media applications on the smartphone along with the geo-tagged features of these applications provides useful information that can be used as a new source for traffic information systems to support urban transportation networks and reduce traffic congestions.

Rashidi et. al., (2017) explored the evolution of social networking sites and its impact on transportation engineering. The researchers discussed the possible extraction of transportation-related data from these sites such as modes of transport, trip attributes, and other socio-demographic information. In addition, the study conducted a qualitative survey on the applicability of social media data on modeling travel demand. The study reported that the data generated from such platforms are useful and considered a rapidly growing area in transportation research studies (*7*). During disasters and extreme events, users of social media applications have a useful feature to report their safety condition, for example, the "mark-as-safe" feature on Facebook. Moreover, app users have additional features such as location-based services that are used for posting their daily "check-in" activities. The use of social media generates a huge amount of data that can be used to understand human behavior in real-time. The Big data generated from these platforms are of interest to travel demand modelers and transportation agencies (*7*). Chaniotakis et. al., (2015) developed a methodology that can be used for extracting and analyzing Twitter data. The study utilized the geo-tagged feature of Twitter and applied spatial analysis to the collected data. The study reported the usefulness of Twitter data and results showed the increased usage of Twitter especially during non–working hours (*8*).

To make the best use from the generated Twitter datasets, Maghrebi et al., (2015) reported that advanced Natural Language Processing (NLP) and data mining techniques are useful tools that can be used to drown out the common noise found in these platforms for extracting and monitoring human behavior (*9*). An effort was done by Das et. al., (2019) to draw and understand different patterns related to bike-commuting based on Twitter data mining results. The researchers collected eight years of bike commuting hashtags using a Twitter database. The study showed a positive sentiment towards bike-commute trips and noted that the majority of Twitter posts were related to bike events (*10*). Lee et al. (2016) studied the benefits of using geo-tagged Twitter data in developing trends and patterns for travel demand modeling. The researchers compared the OD matrix output from Twitter with the OD matrix output from a four-step model to measure the degree of correlation between the two matrices. The study reported the promising results of using location-based features in social media platforms in travel demand modeling (*11*). Roy et. al., (2019) utilized a Twitter dataset to model the human evacuation behaviors during hurricanes such as evacuation decision-making, evacuation time and destination. The researchers validated the model based on ground truth data of 1.81 million tweets related to hurricane Irma. The results showed the model's capability and usefulness in modeling different evacuation behavior in real-time (*12*).

METHODOLOGY

At the time of writing this paper, researchers have started extracting a tremendous amount of Twitter datasets related to the COVID-19 topic. Lamsal (2020) collected more than 34 million tweets on the topic and published the data for NLP researchers (*14*). Banda et. al., (2020) initially released a Twitter dataset of more than 179 million tweets related to the COVID-19 topic. This dataset was preprocessed, annotated, and standardized using the SMMT tool. Furthermore, the dataset was processed by the researchers and converted into a clean version of approximately 37.63 million Tweets mostly in English, French, and Spanish languages (*5*).

Before utilizing the aforementioned datasets in transportation research, the datasets need to be hydrated first into their original form of Tweets with metadata. To extend this effort, hopefully making the analysis accessible to a broader audience, the present paper utilized contemporary online open-source tools for fetching Twitter datasets to examine their computing maturity. The hydration process of these datasets can be done either through the Python scripts of SMMT or by using prebuilt Hydrator, with ensuring the ethical collection, use, and preservation of the media content. The DocNow Hydrator, a free open-source tool (*6*), along with the data adopted from Banda et. al., (2020), were used in this paper for the hydration process.

Tweet IDs Hydration

Two updated versions of DocNow Hydrator (V.0.0.4 and V.0.0.5) are installed for extracting the original Tweets from Twitter datasets. The Hydrator itself is coded in Python scripts. The user does not need to use the Application Programming Interface (API) or any programming code to utilize the tool since it is compiled in an executable form. The Hydrator is an executable file that could be installed and run on different operating platforms. Thus, the user only needs to interface with the tool through the graphical user interface as shown in Figure 1. This methodology was proven effective by previous researchers (*13*) who used the same technique in developing a tool for estimating the system performance and the new technology impacts of intelligent transportation systems.

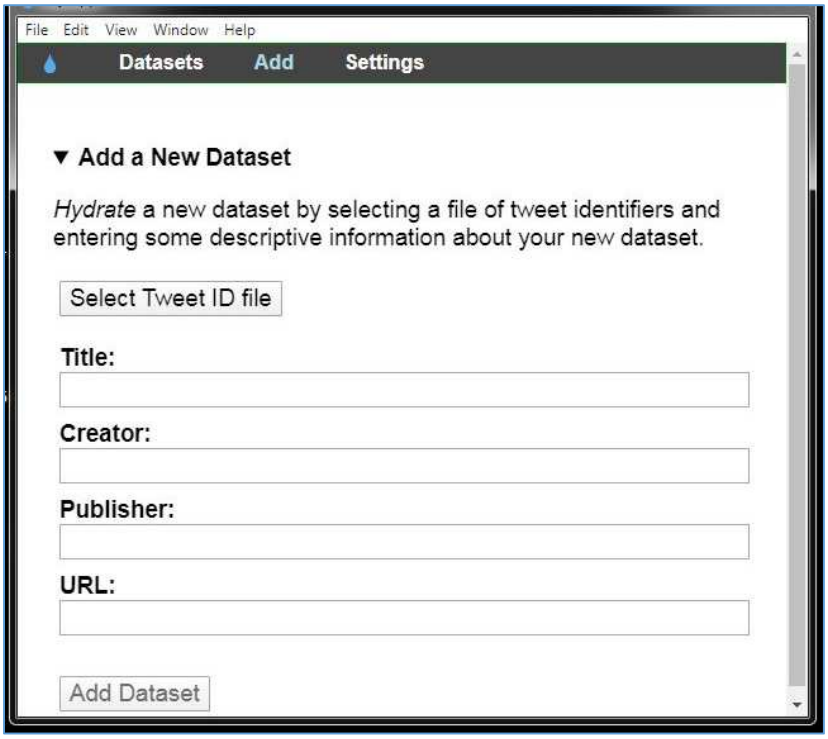

Figure 1: DocNow Hydrator Graphical Interface

To start using the Hydrator, first, the user needs to sign-up for a free Twitter account. By doing so, Twitter will grant access to authorize the use of its content under the developer's statement of policies and agreements. The Hydrator generates a "user key" from the setting screen to authorize the application and to allow the connection between the user's Twitter account and the tool. The Tweet IDs should be placed in a text spreadsheet with no headers in the first line, or otherwise, the Hydrator script will retrieve a message that shows error in reading the first line as shown in Figure 2. Also, it should be noted that only one column of Tweet IDs is allowable in the spreadsheet and this column should have the Tweet IDs in full integer number format and not text format or exponentially written.

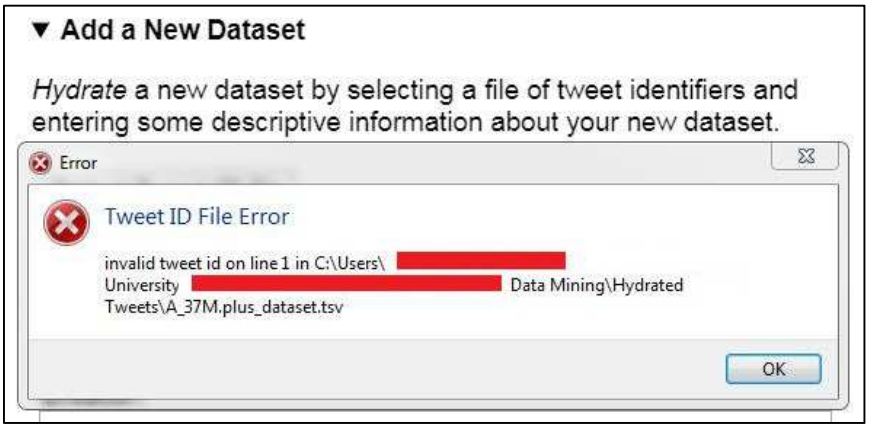

Figure 2: Tweet ID File Error Message

After preparing the spreadsheets that contain Tweet IDs, the "Add" tab is used to upload the files in the software to start hydrating the datasets. The user can also provide additional information such as the project title, creator name, publisher, and URL as shown in Figure 1. To ensure the datasets were added successfully, the Hydrator will provide information about the loaded file path and the number of Tweet IDs. Furthermore, the user will be directed to a new screen to start the Hydration process and save a JavaScript Object Notation (.json) file format with metadata that can be processed using NLP.

In the case of raw datasets, the Hydrator might ignore some IDs when the Tweet itself is no longer available in the database. During this process, the user can start and stop the hydration as needed. Also, the interface can be reloaded to refresh the tool during working with big text files. Finally, the user can export the hydrated Tweets in a (.csv) file format. After exporting the original metadata from the Twitter stream, it becomes necessary to reduce the noise in these big data files and process the (.json) file to facilitate analyzing the fields of interest. At this stage, Tekumalla and Banda, (2020) published a set of Python codes and script as part of the SMMT tool for Twitter data preprocessing, annotating, and standardizing the previously collected datasets, more details about these components can be found in (*3*).

Preliminary Results and Discussion

In this paper, thirty-seven new spreadsheets were generated from the originally released dataset (*5*) using the R programming language to compare the output files of SMMT with the output files of DocNow Hydrator. Each dataset includes the captured Tweet IDs out of one million Tweets related to COVID-19. Regarding DocNow Hydrator, there was not much difference observed between the two examined versions (V.0.0.4 and V.0.0.5). The Hydrator is a user-friendly software for beginners and does not require much experience. The key observation pertains to the tool performance in large-scale datasets, is that the process of extracting the original Tweets was time-consuming. As a case in point, the hydration of a dataset of one million Tweets takes approximately two to four hours, depending on the user's internet speed and the desktop's computing power. Besides, the software can only capture around 1 % of the imported Tweets more or less, which means for every one million Tweet IDs, around one thousand of them were captured by the tool. As a consequence, further analysis will be performed to test the coding computations behind SMMT and compare it to the output from DocNow Hydrator. The thirty-seven datasets will be uploaded through Harvard open data repository (*15*) for validating the results compared to the original data repository. Moreover, the datasets were cleaned and processed with common transportation research keywords that can be beneficial for disciplines of transportation engineering such as human behavior analysis, pedestrian safety, sustainable infrastructure, logistics, traffic network analysis, traffic safety, infrastructure resilience, public transportation, traffic system operations, public transportation, and transit. Currently, research is underway to explore the possibilities of deriving trends and patterns from the datasets of Tweets related to

COVID-19 using machine learning and advanced Natural Language Processing. The new processed datasets will be updated in (*15*). For future releases, a user will need to search the Harvard repository with the present paper title to access the new processed datasets.

NOTES

The topic presented in this preprint is part of research in progress about transportation engineering and pandemics. The current research will focus on methods to estimate the pivotal benefits of large-scale data obtained from web-based applications in supporting research opportunities during outbreak events such as pandemics and national crisis.

The review presented in this paper is not endorsing or against any software. The products, services, or software cited herein and any trade name that may appear in the work has been included only for research purposes. The views, thoughts, and opinions expressed in the presented text belong solely to the author and not necessarily any other entity or organization.

Data Availability

All data, and models generated or used during the study appear in the text of the article. Datasets related to Tweet IDs are deposited in Harvard data repository.

REFERENCES

1. Perrin, A. (2015). Social media usage. *Pew research center*, 52-68.
2. Twitter-First quarter, 2019 Earnings Report https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf
3. Tekumalla, R., & Banda, J. M. (2020). Social Media Mining Toolkit (SMMT). *arXiv preprint arXiv:2003.13894*.
4. https://developer.twitter.com/en/developer-terms/agreement-and-policy
5. Banda, Juan M., Tekumalla, Ramya, Wang, Guanyu, Yu, Jingyuan, Liu, Tuo, Ding, Yuning, & Chowell, Gerardo. (2020). A Twitter Dataset of 179+ million tweets related to COVID-19 for open research (Version 5.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3749360
6. Documenting the Now Project DocNow - Website (https://www.docnow.io/) Source: (https://github.com/DocNow/hydrator/releases)
7. Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, *75*, 197-211.
8. Chaniotakis, E., & Antoniou, C. (2015, September). Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (pp. 214-219). IEEE.
9. Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, T., 2015. Complementing travel diary surveys with Twitter data: application of text mining techniques on activity location, type and time. In: IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), 2015. IEEE
10. Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., & Elgart, Z. (2019). Extracting patterns from Twitter to promote biking. *IATSS research*, *43*(1), 51-59.
11. Lee, J.H., Davis, A., Yoon, S.Y., Goulias, K.G., 2016. Activity Space Estimation with Longitudinal Observations of Social Media Data. In: Transportation Research Board 95th Annual Meeting, 16-0070
12. Roy, K. C., & Hasan, S. (2019). *Modeling the dynamics of hurricane evacuation decisions from real-time Twitter data* (No. 19-04479).
13. Hadi, M., Xiao, Y., Iqbal, M. S., Wang, T., Arafat, M., & Hoque, F. (2019). Estimation of System Performance and Technology Impacts to Support Future Year Planning.
14. Lamsal, Rabindra. (2020). Corona Virus (COVID-19) Tweets Dataset. IEEE Dataport. http://dx.doi.org/10.21227/781w-ef42
15. Harvard Repository: https://doi.org/10.7910/DVN/LJWIGZ