Generation of Twitter Information Databases: A Case Study for the Mobility of People Infected with COVID-19

Alicia Martínez-Rebollar, Pedro Wences-Olguin, Gilberto Palacios-Gonzalez, Hugo Estrada-Esquivel, Yasmin Hernandez-Perez

Tecnológico Nacional de México,
National Center for Research and Technological Development,
Mexico

{alicia.mr, d15ce096, m20ce067, hugo.ee, yasmin.hp}@cenidet.tecnm.mx

Abstract. Social networks are fundamental tools today for common day life but also for research purposes. The main objective of social network platforms is allowing the users to interact among them using internet. This user interaction generates a significant amount of data daily. However, accessing this information can be quite difficult for inexperienced developers. The objective of this paper is to detail the process for extracting unstructured information from the social network Twitter, and structuring it in a database. which can be used for analysis in data mining processes. This process is presented through a case study that analyzes possible places of contagion of the pandemic disease COVID 19 in Mexico.

Keywords: social networks, twitter extraction, user mobility.

1 Introduction

Social networks generate a large amount of data as result of users interaction. Two goods examples of this big data generation through the social networks are Twitter and Facebook [1]. Several research groups in the world has found the advantages that can be obtained by analyzing the different data generated in social networks within the category of big data [2].

The large amount of data generated by the users' interaction in social networks has given rise to new disciplinary fields, such as data science, computational social sciences and even other disciplinary initiatives [3].

As an example of the large amount of data produces by social networks, the social network Twitter has more than 300 million users that produce an average of 7,000 tweets per second [4]. In this way, Twitter has become one of the most conducive virtual environments for collecting large volumes of data. However, one of the main issues in the use of big data coming from social networks is the access to the information, which could be complicated for inexperienced developers.

This paper presents the process of extracting unstructured information from the social network Twitter and structuring it to be stored in a database. In order to demonstrate the proposed process, a case study is presented that use Twitter data for the analysis of possible places of contagion of the pandemic disease COVID 19 in Mexico. The paper is organized as follows: Section 2 shows the background and related works. Section 3 presents the description of our proposal. Section 4 details the generation of databases for data mining. Section 5 shows the tests carried out, and finally section 6 shows the conclusions and future work.

2 Background and Related Work

2.1 Social Networks: Twitter

Social networks are defined as web-based services that allow individuals to construct a public or semi-public profile within a bounded system, to articulate a list of other users with whom they share a connection, and also enables the use to view and traverse their list of connections and those made by others within the system [5].

One of the advantages of social networks is the possibility of use the location of user for monitoring the changes that exist in human mobility. The extraction of information from social networks allows exploring a wide range of fields of study, including public health, surveillance, migration, among others [6].

Twitter is a social network to share ideas and information in short messages of up to 280 characters. This social network has 322 million active users, and it is based on the publication and display of user content for followers as well as open publications that can be seen for any Twitter user [4]. Currently, there are two endpoints that can be used to access user tweets: Filtered stream and Search Tweets

2.2 Related Work

This section presents research works focus on obtaining information from the Twitter social network, as well as works related to the mobility of people considering the information obtained from Twitter. Some authors have focused on how to structure the data considering the mentions, retweets and replies that are generated, hashtags used and what resources have been shared, be they images, web pages or other resources [7].

Some other authors apply techniques to obtain the information from the Twitter social network to make an automatic linguistic analysis of the Twitter texts [8]. However, the social network Twitter makes updates that require modifications to the way for accessing the twitter data. Other related works are focused on the use of Twitter data for analyzing mobility of people through the information they publish in the app.

For example, in [9], the authors proposed an approach to analyze and predict the regularities in human mobility. In research work [10], the authors propose the use of Twitter to obtain valuable information on human mobility. Therefore, their objective was to discover the patterns and mobility of Location-Based Social Networks, such as Twitter. In the research work [11], the authors addressed the exploitation of data

from social networks, such as Twitter, to understand human mobility in an urban site. This research work predicts the next locations of users, using geolocated tweets. In research work [12], new data acquisition and evaluation methods were studied, with an approach aimed at reducing the transmission rate of SARS-COV-2 in the population and evaluating the geographical spread.

3 Description of the Proposal

This section presents an overview of the proposed approach to generate a database from the information collected from the Twitter social network. Our proposal is composed of four main processes (Figure 1): definition of the problem in which the user defines the information to be analyzed in tweets, request authorization to access to Twitter data, development of the app to extract data, and, finally information validation. In order to illustrate our proposal, a case study of the COVID 19 Pandemic was performed, where data is extracted from users of Twitter in Mexico referring to current pandemic.

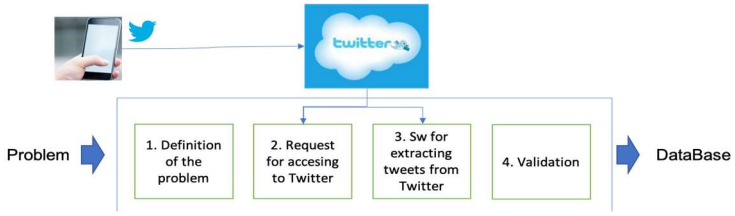

Fig. 1. Overview of the proposed approach.

4 Generation of Databases for Data Mining Purposes

Following, the main four phases of the proposed approach are presented in detail.

4.1 Definition of the Problem

The first step to generate a database on a specific subject is to establish a research objective and to define the data that are needed to achieve that objective. In the problem addressed in this paper, a database needs to be created that allows to identify the mobility of people suffering from COVID-19 over a period of time.

The data obtained from Twitter should allow us to trace the places visited by persons who use Twitter and posted that they were infected with COVID-19 virus. This information permits the identification of the possible places of infection. To do this, it is necessary to obtain the states, municipalities or places of interest where the tweets were published. Figure 2 shows the database that was used to identify the mobility of people commenting they are suffering from the COVID disease using twitter.

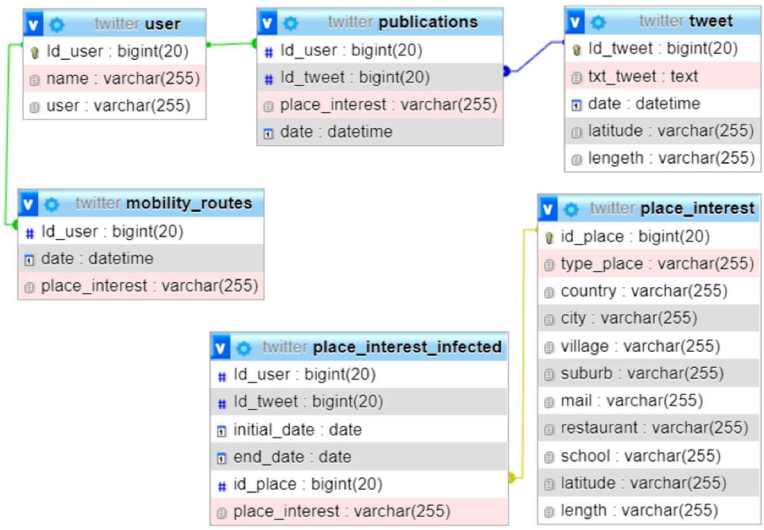

Fig. 2. Database for the COVID monitoring case study.

This database contains six tables representing users, publications, tweets, routes, interest points and the point of interest visited by people infected.

4.2 Request for Accessing to Twitter API

The second step of our proposal is the request process to access the API Twitter. A registration form must be completed, which is made up of two sections. In the first section, the user must answer the question: Which best describes you?.

The second step consists of filling out a form that is made up of four steps or blocks, which are: Basic info, Intended use, Review, and finally, Terms. The information requested in basic information field is some easy data to be provided.

In the Intended use section, the Twitter API tries to discover the use that will be given to the information. In this section, the user must answer questions related to the use of the information and functionalities of twitter. In addition, the user must explain in depth the intended use of the data. This explanation is relevant because it will the factor that use Twitter staff to approve or decline the account request. In the Review section, the user must check that the information provided is correct. Finally, in Terms section, the terms for the use of the Twitter API need to be accepted.

4.3 Software for Extracting Tweets From Twitter

The third step of our proposal is the development of the application for extracting tweets from the social network twitter. The twitter API allows access to its functions or endpoints. There are two ways to extract tweets from Twitter: *Search Tweets endpoi*nt

and the *Filter real-time tweets endpoint*. The *Search Tweets endpoint* option allows the user to programmatically access filtered public tweets posted over the last week. The option *Filter real-time tweets endpoint* allows to receive posts at the moment they are generated, following a real-time approach. The application will generate, for the first time, the keys and tokens needed to access the Twitter information. These keys and tokens must be kept in a safe place.

All the information on the twitter APIs can be consulted at [13]. The endpoints offered by twitter are grouped into three categories: Tweets, User, Spaces and these can be consulted in [14]. In our case study, the python programming language and the Tweepy [15] library was used for coding. Tweepy looks to be the best-known open source library to access the API from Python.

A class was developed that implements *tweepy.StreamListener* to be instantiated from the main method. This class defines the attributes to be obtained from the twitter API in real time. Figure 3 shows the code used to extract data from twitter. A filter is made by the keyword "COVID", the tweets, the date and time of publication are extracted; and they are stored in the variables text and datetime respectively.

```
import tweepy
from autenticate import get_auth
class MyStreamListener(tweepy.StreamListener):
    def on_status(self, status):
        if status is not False and status.text is not None:
            try:
                texto = status.extended_tweet["full_text"]
                fechahora = status.created_at
            except AttributeError:
                texto = status.text
    def on_error(self, status_code):
        print(status_code)
        return False

if __name__ == '__main__':
    print("===== Captador de tweets =====")
    auth = get_auth()  # Retrieve an auth object using the function 'get_auth' above
    api = tweepy.API(auth)  # Build an API object.
    myStreamListener = MyStreamListener()     # Connect to the stream
    while True:
        try:
            myStream = tweepy.Stream(auth=api.auth, listener=myStreamListener)
            myStream.filter(track=['COVID'])
        except:
            print('Ocurrió un conflicto')
            continue
```

Fig. 3. Main code to extract data from twitter.

4.4 Validation

The fourth step is the validation of the code functionality. This validation consisted of executing the code and verifying that the tweets displayed in the console matched the filter keywords. The filter used was the following: myStream.filter (languages = ['es'], track = ['COVID, COVID-19, coronavirus']). Figure 4 shows some tweets that match the keyword filter. In this Figure, the keywords were highlighted in those tweets.

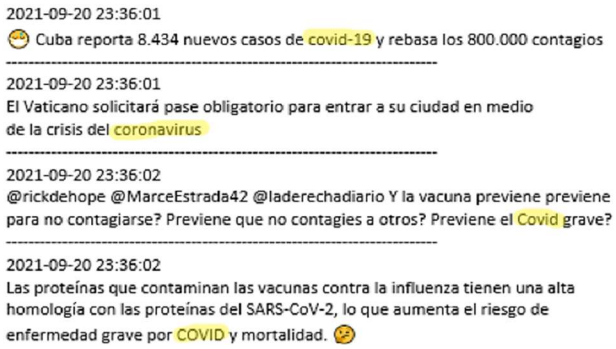

Fig. 4. Partial view of Tweets that matched keywords in validation.

Another validation carried out was made to verify that some tweets published by ourselves that contained the filter keywords were shown in the console. For example, we publish the following tweets at runtime: *"Creo que tengo coronavirus mañana me hago la prueba", "Lamento informarles que di positivo a covid", "El covid-19 sigue mutando espero que mi vacuna me proteja".* The console of our application received correctly these published tweets.

5 Test and Results

This section shows the tests and results. The tests carried out consisted of executing the process for tweets extraction for 30 minutes in 3 different times. Different filters were applied in each of these extractions. We choose to run the process for 30 minutes to preserve the resources provided by Twitter. However, the process can be executed until Twitter closes the connection when allowed resource quota is reached.

In the first run, 19 tweets were obtained and the following filter was applied: myStream.filter (track = ['September Mexico']) which means that the twitter API stream would search for those tweets having the keywords: *september* and *México*. In this case, a conditional "and" was applied between these two terms, because there is a blank space between the two terms. The filtering parameters to make requests can be consulted at the following link: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters

The twitter API shortens the URL's in the tweets resulting in some search words not appearing in the tweets because they are part of the URL's. Figure 5 shows some of the feedback that was obtained from the first run.

In second run, 5778 tweets were obtained applying the following filter: myStream.filter (track = ['September, Mexico']), which means that the twitter API stream would search for those tweets that had the keywords: September or México, that is, a conditional OR would be applied because "the comma" means OR. For this reason, a higher number of tweets was obtained compared to the first run. Figure 6 shows some of the comments that were obtained during the second run. The third test carried out

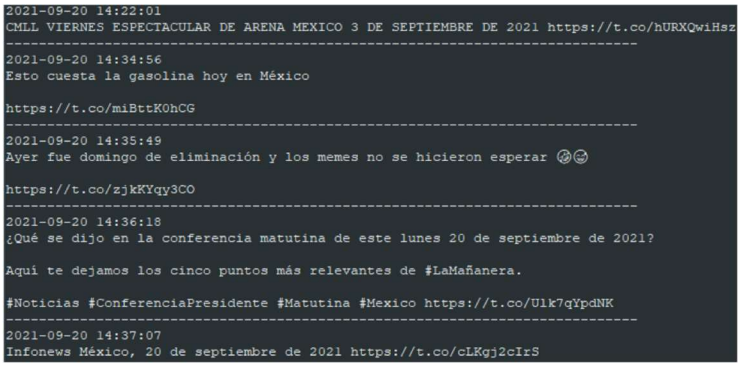

Fig. 5. Partial view of the results obtained with terms filtered with an and.

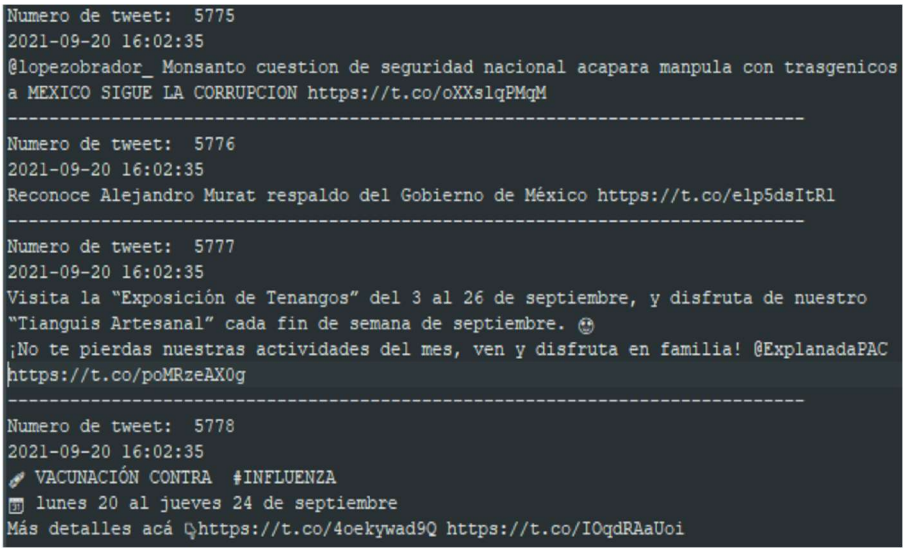

Fig. 6. Partial view of the results obtained with terms filtered with an or.

obtained 1352 tweets and this was applied to the case study presented in section 4.1. The following keyword filter was applied: myStream.filter (languages = ['es'], track = ['COVID, COVID-19, coronavirus']).

6 Conclusions and Future Work

Currently, the collection of data that can be obtained from social networks provides an excellent source of data that can be used and analyzed to discover information. This paper presents the process for generating a database of information extracted from the social network Twitter. Therefore, the steps that an inexperienced developer must take

to carry out the information extraction are discussed. This process will allow her/him to avoid common mistakes when carrying out this extraction process from Twitter.

Each of the steps presented is detailed using a case study. The main challenge in the work presented as future work is to develop a tool that allows us to analyze the mobility patterns that Twitter users have, facing the fact that currently, Twitter has restricted the exact location where a tweet is published for user safety reasons. The next objective of the system will be to produce statistics and draw reliable routes that allow the behavior and interaction of sick people to be analyzed.

References

1. Puyol, J.: Una aproximación a Big Data. Revista de Derecho UNED. (14), pp. 471–503, (2014)
2. Song, X., Shibasaki, R., Jing, N., Xing, X., Li, T., Adachi, R.: DeepMob: Learning Deep Knowledge of Human Emergency Behavior and Mobility from Big and Heterogeneous Data. ACM Transactions on Information Systems, 35, pp. 1–4 (2017)
3. Rocha, M., Elena, M.: Grandes datos, grandes desafíos para las ciencias sociales. Revista Mexicana de Sociología, 80, pp. 2–6 (2018)
4. Statista, https://es.statista.com/estadisticas/636174/numero-de-usuarios- mensuales-activos-de-twitter-en-el-mundo, last accessed 2021/09/20
5. Boyd, D., Ellison, N. B.: Social Network Sites: Definition, History and Scholarship. Journal of Computer-Mediated Communication. Journal of Computer-Mediated Communication, 13, pp. 210–230 (2007)
6. Osorio, J.: Análisis de los patrones espacio-temporales de eventos a partir de datos de Twitter: el caso de la World Pride 2017 en Madrid. Estudios Geográficos, 81, pp. 1–12, (2020)
7. Domingo, M., Minguillón, J.: Modelado, extracción y análisis de información del flujo de datos de Twitter. Universitat Oberta de Catalunya, (2012)
8. Blasco, E.: Aplicación de técnicas de minería de datos en redes sociales/web, Master. Universidad Politécnica de Valencia, (2015)
9. Comito, C., Human Mobility Prediction through Twitter. Procedia Computer Science, 134, pp. 129–136 (2018)
10. Al-Jeri, M.: Towards Human Mobility Detection Scheme for Location-Based Social Network. pp. 1–7, (2019)
11. Comito, C.: Minería de la movilidad humana de los medios sociales para apoyar la informática urbana. pp. 514–521 (2019)
12. Gao, S., Rao, J., Kang, Y., Liang, Y., Kruse, J.: Mapping county-level mobility pattern changes in the United States in response to COVID-19. 12, pp. 16–26 (2020)
13. Twitter API, https://developer.twitter.com/en/docs/twitter-api, last accessed 2021/09/20
14. Twitter API v2, https://documenter.getpostman.com/view/9956214 / T1LMiT5U#4a6cc2e6-5c99-421a-b1f0-ba9f170dce97, last accessed 2021/09/20
15. Tweepy, http://www.tweepy.org, last accessed 2021/09/20