Mohamad Hasan[1]

DOI: 10.35595/2414-9179-2021-2-27-233-240

USING SOCIAL MEDIA DATA TO MAP MORTAR SHELLS FALLING IN DAMASCUS, SYRIA

ABSTRACT

The paper analyzes the use of social media data in geographical information systems to map the areas most affected by mortar shells in the capital of Syria, Damascus, by using geocoded and parsed social media data in geographical information systems. This paper describes a created algorithm to collecting and store data from social media sites. For the data store both a NoSQL database to save JSON format document and an RDBMS is used to save other spatial data types. A python script was written to collect the data in social media based on certain keywords related to the search. A geocoding algorithm to locate social media posts that normalize, standardize and tokenize the text was developed. The result of the developed diagram provided a year by year from 2013 to 2018 maps for mortar shell falling locations in Damascus. These layers give an overview for the changing of the numbers of mortar shells falls or in hot spot analysis for the city. Finally, social media data can prove to be useful when creating maps for dynamic social phenomena, for example, mortar shells' location falling in Damascus, Syria. Moreover, social media data provide easy, massive, and timestamped data which makes these phenomena easier to study.

KEYWORDS: Data mining, Geocoding, NoSQL databases, Social Media.

INTRODUCTION

From the beginning of the Syrian war, the civilian people were struck by mortar shelling, bombing and artillery bombardment. Shelling of the city of Damascus started since the battle of Damascus in 2011 with the attempt of the armed opposition to take control of the city. There are multiple new articles since 2011 about the shelling neighborhoods of Damascus and the causalities of the shelling. These shelling became a characterization of the Syrian civil war, to the point where improvised artillery canons have emerged, which were modified to create maximum distraction[2]. In addition to the mortar and artillery shelling, the capital was struck by multiple rockets. With the rise of jihadists and radical groups in Syria after the civil war, a series of suicide attackers and car bombs exploded in the capital. Social media users whether news anchors or usual users often wrote about any of these effects either to question about it or to inform others.

Social media users can share messages, videos, pictures, and even news links on their own social media pages. Social media networks can be classified by many aspects, as a series of web applications that allow users to create blogs (e.g., Twitter), social networks (e.g., Facebook), or photo, audio, and video sharing services (e.g., YouTube and Instagram) [*Kaplan,* 2010]. Since web applications have emerged, Social media networks have allowed their users to create and exchange generated huge amounts of content. By the end of 2018, Twitter claimed to have about 326 million monthly active users worldwide[2], while Facebook claimed to have 2.32 billion monthly active users[2]. With such huge quantities of users, the data are being generated

[1] Saint Petersburg University, Sredniy Prospekt V.o., 41, 199034, Saint-Petersburg, Russia, *e-mail:* **mo-hasan89@hotmail.com**

[2] https://en.wikipedia.org/wiki/Improvised_artillery_in_the_Syrian_Civil_War

[3] Statista: url: https://www.statista.com/

continuously and dynamically to study different social phenomena. Also, the shared data can be used to study geographical phenomena since data on these social networks usually have geotags (i.e. coordinate of post location), and if not a geocoding process of the text content is necessary to get the coordinate for the data [*Hecht,* 2011]. Using sentiment analysis on social media data, many studies have been conducted on predicting disaster responses in cities [*Alexander,* 2014], tracking disease spread [*Hung,* 2015], monitoring special events [8], and measuring public opinions on political issues without a survey [*Sakaki,* 2010].

In times of disasters both natural (e.g. hurricanes) and man-made (e.g. terrorist attacks) the usage of social media increases which is followed by a larger than the usual number of posts, and with such increase researchers can study human behavior in times of disasters [*Soulis,* 2013]. Therefore, geolocated data from social media became a main source of data in times of disasters such as disaster response [*Simon,* 2015] and disaster relief [*Goodchild.* 2010], as well as in spatio-temporal crime analysis and crime predictions [*Barbier,* 2011]. Also, social media data have been used in counter-terrorism applications [*Alharith,* 2018]. With such wide applications, social media data can be used to analyze mortar shelling falls in Syria since more social media data in some areas means a larger effect for the mortar shelling on the population for that area.

Even though social media can provide dynamical, huge, and easy to obtain data with many advantages, but using these data still have it limitation and drawbacks. For example, it is hard to collect data that are representative of the whole population, due to the digital divide which is the inability or lack of desire to use or access the whole internet or write about certain topics. The digital divide affects the social media data since content creators are young people (especially, university graduates) who tend to share data on the internet [*Brake,* 2014]. Moreover, Social media data are subjected also to inequality in representation because certain topics such as politics are more shared on social media websites based on the creators [*Schradie,* 2011]. Researchers estimate that the number of who use and share data on social media cannot represent the full number of the population. The percentage of Twitter users represents about 66% of the USA [*Smith,* 2019].

Using social media data to analyze geographical activities is getting mere popular with the spread of using mobile devices that can get users' coordinates. Although geo-referenced data from mobile phones have quality issues, such as noises and spam spots, the data can prove to be crucial to some researchers. It is required; therefore, additional parsing for the data or implementation of NoSQL data stores. Finally, Social media sites do not provide full and direct access to all their datastores. Therefore, researchers often create their mechanisms to collect data via designated access interfaces provided by social media services.

Even though there are some challenges to using social media data in analyzing geospatial activities, these challenges should not prevent using social media data in GIS applications, since the benefits of using dynamic and timestamped data overcome these drawbacks. Moreover, the amount needed to collect a huge amount of data is much less than traditional methods which have their drawbacks as well. Therefore, Social media should not be excluded as a data source for GIS analysis.

MATERIALS AND METHODS

Although most social media services have commercial solutions to fully access their content in their databases, these commercial uses can prove to be costly. Most social media sites provide an application programming interface (API) that allows limited access to their data for free. Twitter's API is a free interface that allows extracting data from Twitter's datastore up to 7 days before the query, while Facebook's graph API allows only access on public pages using their Facebook ID. A solution to these problems is to use third-party applications which allow collecting data through requesting it using an internet browser.

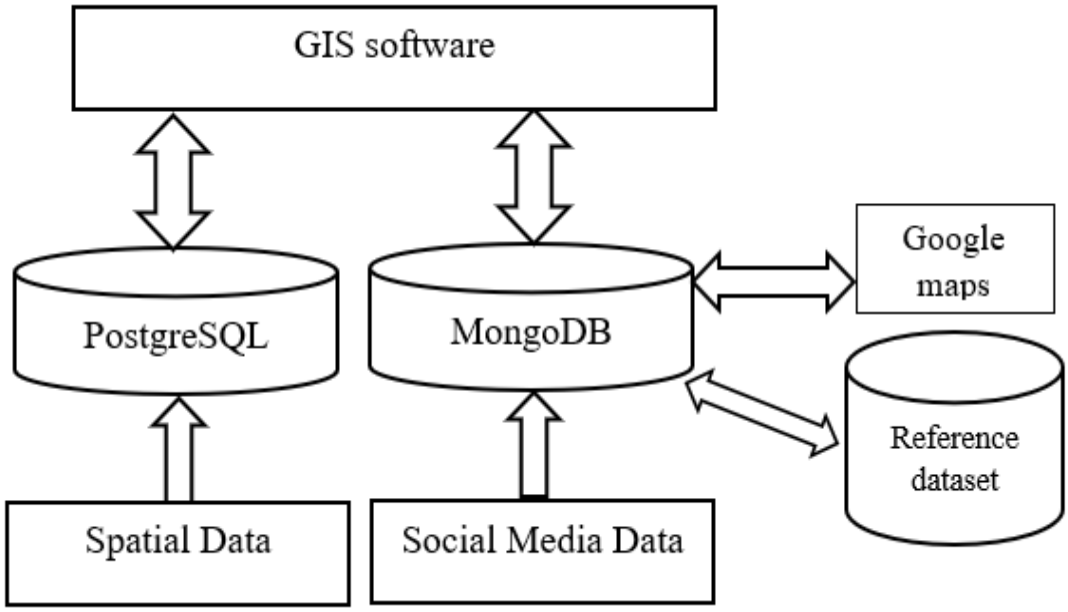

Fig. 1. Architecture for collecting and using Social media data

The designed architecture uses a python script to collect and store social media data in BJSON format in MongoDB which is an open-source NoSQL document database. Other geospatial data were stored in PostgreSQL, another open-source relational database. Google geocode API was used to get the location to the social media data. Google geocode API uses tokenized words as input to get the coordinates. The tokenized words are selected from a pre-selected set of words that have a location meaning.

The extracted social media data of Syria were approximately 14 million tweets and 1 million posts from Facebook gathered from 80 local news network pages. The social media data were generated for 6 years since 2013 and amassed over 15 GB in size. When using data from social media there is always a question about the validity of the data, since every user can spread information. The data mined from Facebook were collected from only trusted sources. These trusted sources included local news networks and pages with a high number of followings. These pages are assumed to have more accurate data. Data from Twitter collected from regular users without any restriction. Although in Syria laws incriminate the spread of fake news on social networks since 2012, but the overall accuracy of the data could not be guaranteed.

The data mined from social media have a JSON (JavaScript Object Notation) which is a key, value format to save data. Figure 2 shows an example of the data mined from Twitter. Data retrieved from Facebook have a similar data structure but with different key names. The "geo" key in the figure represents the geotag in Twitter's data. The key can have either a null value or an object with the latitude and longitude values. Although the geotag can prove to be useful in some applications, in this study the algorithm focuses on geocoding the text of the tweet of the Facebook post as the user could be writing about the event far from it.

Both the keys "Message" in Facebook's data and "Text" contain the text data of the post which will be used to feed the geocoding algorithm to obtain the coordinates. Another important key in the data is the "created_at" key which contains the date on which the data were created. This key can be used to provide a time series for mortar falling.

```
{
    "created_at":"Thur Apr 09 12:59:37 +0000 2015",
    "id":"1069533912605630464",
    "id_str":"1069533912605630464",
    "text":"دمشق : سقوط قذيفة هاون في منطقة العدوي بالقرب من جسر الحياة.
        #jaramana #جرمانا #سوريا #سورية #syria #اخبار #الغوطة #دمشق #ريف_دمشق #أخبار",
    "truncated":"False",
    "entities":"{hashtags: [list of hash tags], symbols: [],
            user_mentions: [list of user's menstions],
            urls: [{url: tweet's url, expanded_url: long link,
            display_url: long data, indices: []}]}",
    "metadata":"{iso_language_code: ar, result_type: recent}",
    "source":"null",
    "in_reply_to_status_id":"None",
    "in_reply_to_status_id_str":"None",
    "in_reply_to_user_id":"None",
    "in_reply_to_user_id_str":"None",
    "in_reply_to_screen_name":"None",
    "user":"user's information e.g. user's id, image link, status, etc.",
    "geo":"None",
    "coordinates":"None",
    "place":"None",
    "contributors":"None",
    "retweeted_status":[
       "list of retweets "
    ],
    "is_quote_status":"False",
    "retweet_count":"2282",
    "favorite_count":"0",
    "favorited":"False",
    "retweeted":"False",
    "possibly_sensitive":"False",
    "lang":"ar"
}
```


Fig. 2. Example of a mined tweet.

Geocoding forms a fundamental part of spatial analysis in a variety of research disciplines especially in extracting spatial coordinates from social media [*Golubovic,* 2017]. The algorithm used to solve geocoding the data is similar to the gazetteer algorithm [*Churches,* 2002]. The gazetteer algorithm cannot generate the coordinates itself, where google maps geocode API query is used on the output to have the coordinate needed. If Google mas API did not return any results by the algorithm the data is thrown away, since not every social media feed should have a geospatial location.

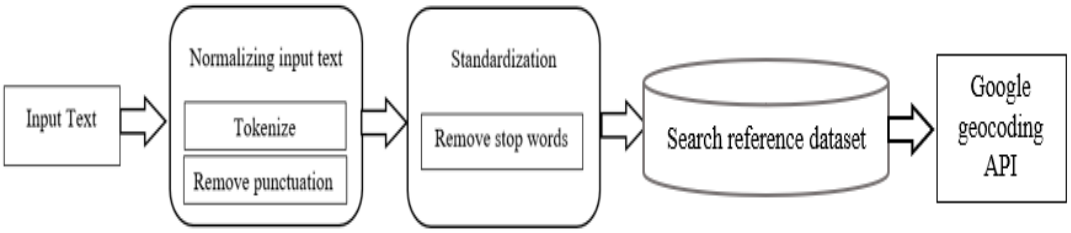

Fig. 3. Schematic of geocodeing social media data

Geocoding forms a fundamental part of spatial analysis in a variety of research disciplines especially in extracting spatial coordinates from social media [*Golubovic,* 2017]. The algorithm

used to solve geocoding the data is similar to the gazetteer algorithm [*Hill,* 2000]. The gazetteer algorithm cannot generate the coordinates itself, where google maps geocode API query is used on the output to have the coordinate needed. If Google mas API did not return any results by the algorithm the data is thrown away, since not every social media feed should have a geospatial location.

Figure 3 demonstrates the diagram of the algorithm in which the algorithm standardizes and normalizes the input to get the tokenized sentiment the algorithm removes stop words, corrects typos, and removes punctuation marks. The algorithm then makes a 2-gram word set since most cities, towns, street names and squares in Syria consist of two words. Then, the 2-grams and the words are compared with two datasets. The first dataset is the most important points in Syria, such as names of cities, towns, villages, city squares, etc. While the second dataset came from analyzing the social media datasets and picking the words which have a geospatial meaning, i. e. have in Arabic words such as school, road, hospital, etc. After the normalization and standardization to be compatible to be searchable on google maps geocoding API, the algorithm queries the text through the API to derive the final output coordinates. The results of the geocoding algorithm are shown in figure 4.

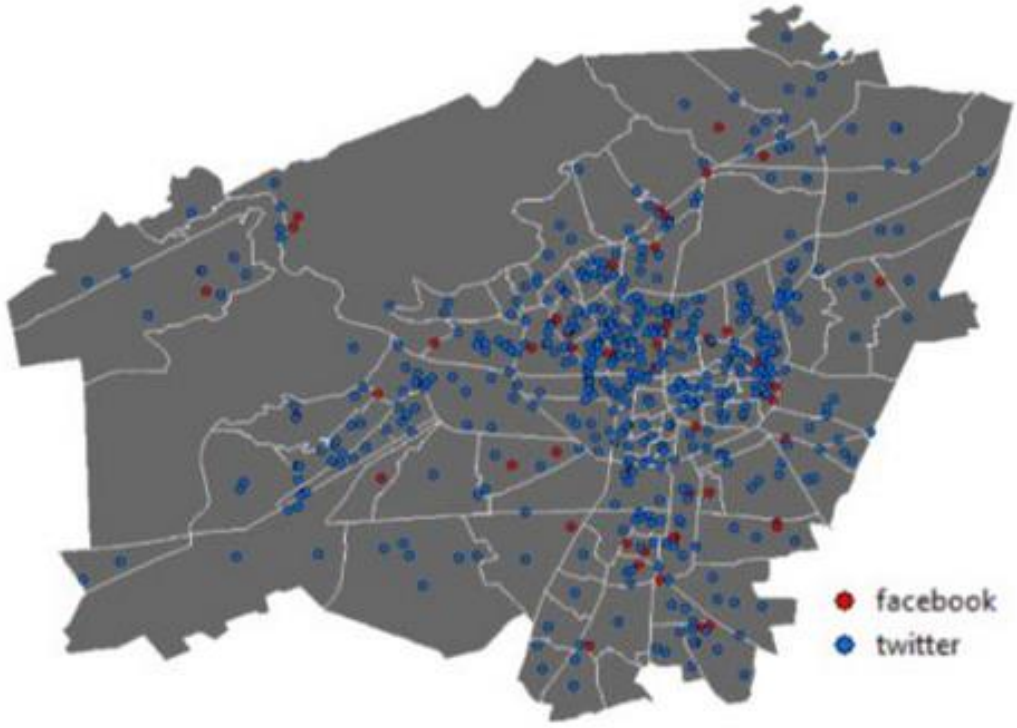

Fig. 4. Map of geocoded social media data in Damascus about Mortar falling

RESULTS AND DISCUSSION

The data which were processed by the geocoding algorithm mentioned in the algorithm used in geocoding was able to find the coordinates for 15% for Twitter data and 48% of the Facebook data. Some of the posts were found out the borders of Damascus, this can be attributed to the fact that both level 1 geocoding data which is neighborhood names, and other place names such as schools, streets, and mosques, are repeated in other cities, therefore, an additional

restriction that one of the words should be (Damascus "دمشق", Sham "شام", capital "عاصمة") in the posts and tweets to ensure that the geocoding results are within the city.

One hundred random posts and tweets were randomly selected to check the accuracy of the geocoding algorithm. Feeding these data to the geocoding algorithm returns the words used in the google maps API and by checking manually the location of these words. It was found that about 83 of the 100 inputs were placed correctly. It should be noted that if google maps API returned only the coordinates of the city the data were removed from the database, as such data does not add any additional information.

Many processes of the geocoding algorithm gave only the coordinates of the neighborhood center because many of the social media data are written only with the name of the neighborhood in the post. since most data posts mentions only the name of populations places without mentioning the whole address of these conflicts. The geocoded mined data from social media were divided between 36% of the data were geocoded using the first level, i.e. names of neighborhoods only, and 13% that were geocoded using the second level of geocoding, i.e. names of point of interests, such as schools and city squares.

Table 1. Percentage of geocoded social media data of the total mined data

Social Media Type	Only nebierhood centers	Nebierhoods+ points of intrests
Facebook	45 %	32%
Twitter	35 %	11%
Facebook+Twitter	36 %	13%

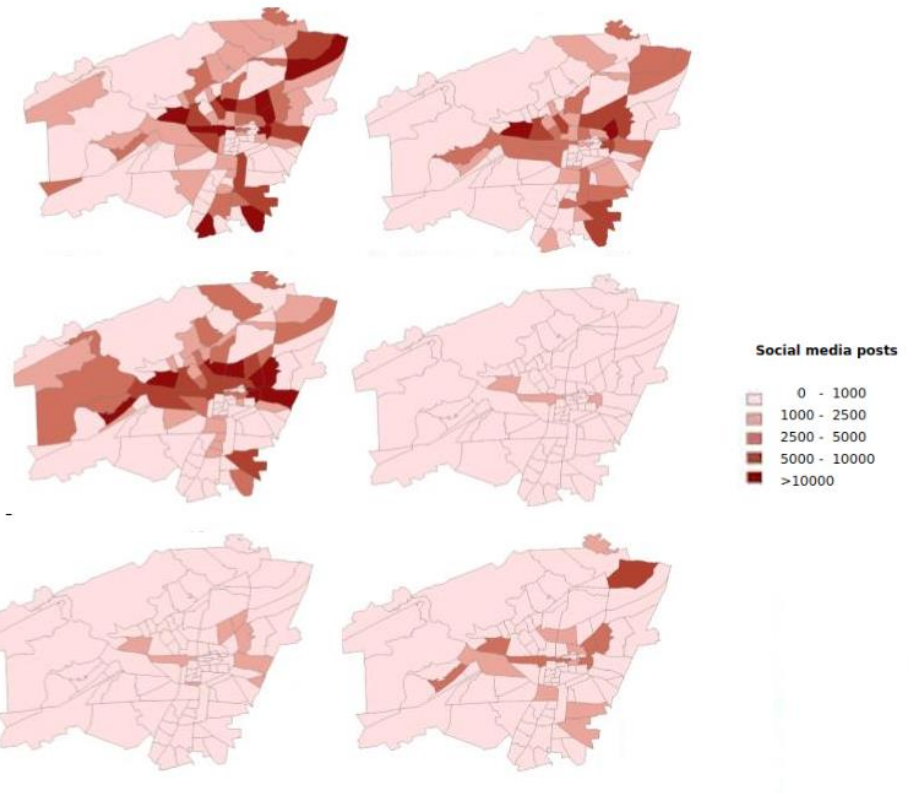

Fig. 5. Mortar shells posts by neighborhood between 2013- 2018 in Damascus, Syria

CONCLUSION

Social media sites provide a data source for GIS to butter analyze complex social and environmental systems. Social media structure which is massive and semi-structured requires nontraditional methods to process the data for GIS software, such as NoSQL and parsing JSON documents. The proposed architecture showed to be reliable when processing Twitter and Facebook data for GIS analysis. The study was conducted to analyze mortar shell effects on the population-based on social media data for revealing spatiotemporal patterns of risks. Such insights are difficult to acquire, especially when studying the effects of mortar falling on population, a topic that could not be hard to analyze using traditional methods of GIS data collecting and acquiring. Secondly, the results show that using social media data can be used in counter-terrorism and humanitarian proposes to study the area most hit by mortar shelling.

REFERENCES

1. *Alexander, D. E.,* Social media in disaster risk reduction and crisis management. Science and engineering. 2014. V. 20, issue. 3. P. 717–733.
2. *Alharith, A., Samak, Y.* Fighting terrorism more effectively with the aid of GIS: Kingdom of Saudi Arabia case study. American journal of geographic information system, 2018. V. 7, issue. 1. P. 15–31.
3. *Barbier, G., Goolsby, R., Gao, H.*, Harnessing the Crowdsourcing power of social media for disaster relief. IEEE intelligent systems, 2011. V. 26, issue. 3. P. 10–14.
4. *Brake, DR.*, Are we all online content creators now? web 2.0 and digital divides. Journal of computer-mediated communication, 2014. V. 19, issue. 3. P. 591–609.
5. *Churches, T., Christen, P., Lim, K., Zhu, J. X.*, Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making, 2002. V. 2. P. 1–16.
6. *Golubovic N., Krintz, C., Wolski, R., Lafia, S., Hervey, T., Kuhn, W.,* Extracting spatial information from social media in support of agricultural management decisions. Proceedings of the 10th Workshop on Geographic Information Retrieval, 2017, Burlingame, California: GIR '16.
7. *Goodchild, M., Glennon, A.*, Crowdsourcing geographic information for disaster response: A research frontier. International journal of digital earth, 2010. V. 3, issue. 3. P. 231–241.
8. *Hay, S.I., George, D.B., Moyes, C.L., Brownstein, J. S.,* Big data opportunities for global infectious disease surveillance. PLoS medicine, 2013. V. 10, issue. 4, doi: 10.1371/ journal.pmed.1001413.
9. *Hecht, B., Hong, L., Suh, B.,* Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. Proceedings of the 2011 annual conference on human factors in computing systems, 2011, P. 237–246.
10. *Hill L.,* Core elements of digital gazetteers: placenames, categories, and footprints. Research and advanced technology for digital libraries, 2000. V. 1923. P. 280–290.
11. *Huang, Q., Yu Xiao, Y.*, Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS international journal of geo-information, 2015. V. 4, issue: 3. P. 1549–1568.
12. *Kaplan, A., Haenlein, M.,* Users of the world, unite! the challenges and opportunities of social media. Business Horizons, 2010. V. 53. P. 59–68.
13. *Sakaki, T., Okazaki*, M., Earthquake shakes twitter users: Real-time event detection by social sensors. Proceedings of the 19th international conference on world wide web, 2010. P. 851–860.

14. *Schradie, J.,* The digital production gap: The digital divide and web 2.0 collide. Poetics, 2011. V. 39, issue. 2. P. 145–168.
15. *Simon, T., Goldberg, A., Adini, B.*, Socializing in emergencies, A review of the use of social media in emergency situations. International journal of information management, 2015. V. 35, issue. 1. P. 609–619.
16. *Soulis, K., Varlamis, I., Giannakoulopoulos, A., Charatsev, F.,* A tool for the visualization of public opinion. International journal of electronic governance, 2013. V. 6, issue. 3. P. 218–231.
17. *Smith, A.,* Why Americans use social media. Technical report, Pew Research Centre. URL: http://www.pewinternet.org/Reports/2011/WhyAmericans-Use-Social-Media.aspx, visit date: 12/4/2019.