A Real-Time Spatio-Temporal Bigdata System for Instant Analysis of Twitter Data to Monitor of Advertising Campaigns; Case Study New York City

Seyed Ali Hoseinpour*

* PHD student, GIS Dept, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
s.alihosein.p@ut.ac.ir

Abstract. Advertising monitoring is useful for detecting and determining user reaction, campaign performance and their challenges. Popular social networks especially Twitter are one of the best resources that could been analyzed for this purpose. Twitter's data analyzing needs different methods and tools because Twitter is one of the bigdata sources. In this paper, a system is developed dealing with Twitter data streams and can monitor advertising keyword spatio-temporal trends, real time. Also, it can predict the location of later tweets using MLP model. The prediction accuracy is 0.78.

Keywords. Location Based Services, Geotagged Big Data, Twitter, Advertising Campaigns, Spatio-Temporal

1. Introduction

Advertisement plays an important role in marketing, branding and business developing (Frolova 2014; Dib 2016), but it's not useful alone and to increase the performance and choose the best strategy, users' reactions and feedbacks must be controlled (Dib 2016). For example, after an advertise teaser, which group of customers, when and where were attracted (Frolova 2014; Dib 2016)?

Today, most of people share information on social media in different subjects such as products purchases (Varol et al. 2017). Twitter is one of the most important of them and is considered to be very important in the bigdata era (Shirdastian et al. 2019). There are 500 million tweets sent each day (Alotaibi et al. 2020). Twitter has an impact on marketing, so that, 40% of Twitter users purchased something after seeing it on Twitter (Alotaibi et al. 2020). Twitter contains spatial-temporal information so it can be used to spatio-temporal analytics (Martín et al. 2019).

The main purpose of this article is designing a real-time system for monitoring the spatial and temporal distribution of a particular keyword on

LBS 2021
16th Conference on Location Based Services
24-25 November 2021 Online

Published in "Proceedings of the 16th International Conference on Location Based Services (LBS 2021)", edited by Anahid Basiri, Georg Gartner and Haosheng Huang, LBS 2021, 24-25 November 2021, Glasgow, UK/online.

https://doi.org/10.34726/1783 | © Authors 2021. CC BY 4.0 License.

Twitter. Second purpose is adding a predicting model to it. So, this scenario is considered: A company has done an advertising with one keyword, and now, it needs this information to identify spatial and temporal hotspots to find the best location for a new store or billboard. What users have used this keyword in their tweets? How is the spatial and time sequence of these tweets on the map? What other keywords have been used with this keyword? Which hashtags are frequent in these tweets? What is the time distribution of tweets per hour? At next hour, how many tweets and where will be tweeted?

There are some challenges in this research that this paper tries to solve them:

- Twitter is one of the big data sources and has data stream. So, there are difficulties related to managing the 4V[1] characteristics(Alotaibi et al. 2020). Therefore, some special technologies and architectures (eg particular indexing) must be used.
- The base data in Twitter is text, so we need NoSQL databases, preprocesses and NLP[2].
- Many of Twitter's data don't have exact position. Also, although Twitter has an API to geocode request, but there is limitation for making requests number. So, we need location interpolation or estimation.

Martín et al (2019), with Python, MongoDB and Apache Pyspark, developed an architecture and work flow (*Figure 1.a)* to evaluate and locate the activity in the city of Valencia by Heatmap visiualization (Martín et al. 2019). Alotaibi et al (2020), developed a big data analytics tool (named Sehaa) for healthcare symptoms and diseases detection (*Figure 1.b)* using Twitter, Python and Apache Spark (Alotaibi et al. 2020). Sehaa uses machin learning method to detect various diseases in the Saudi Arabia.

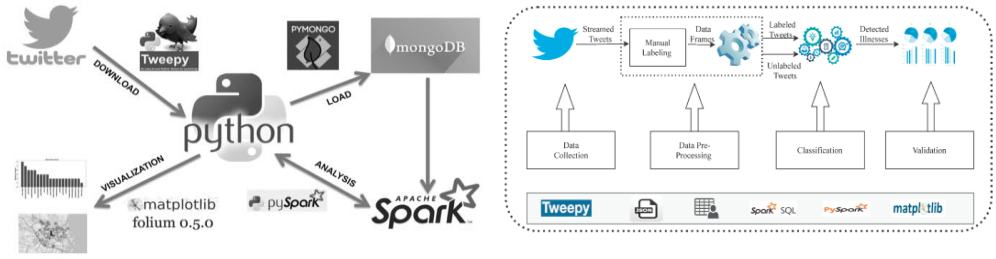

Figure 1. System architecture. **Left.** (Martín et al. 2019). **Rigth:** (Alotaibi et al. 2020)

2. Methodology

Figure 2 shows System's architecture; this system has 3 sections.

[1] i.e., volume, velocity, variety, and veracity

[2] Natural language processing

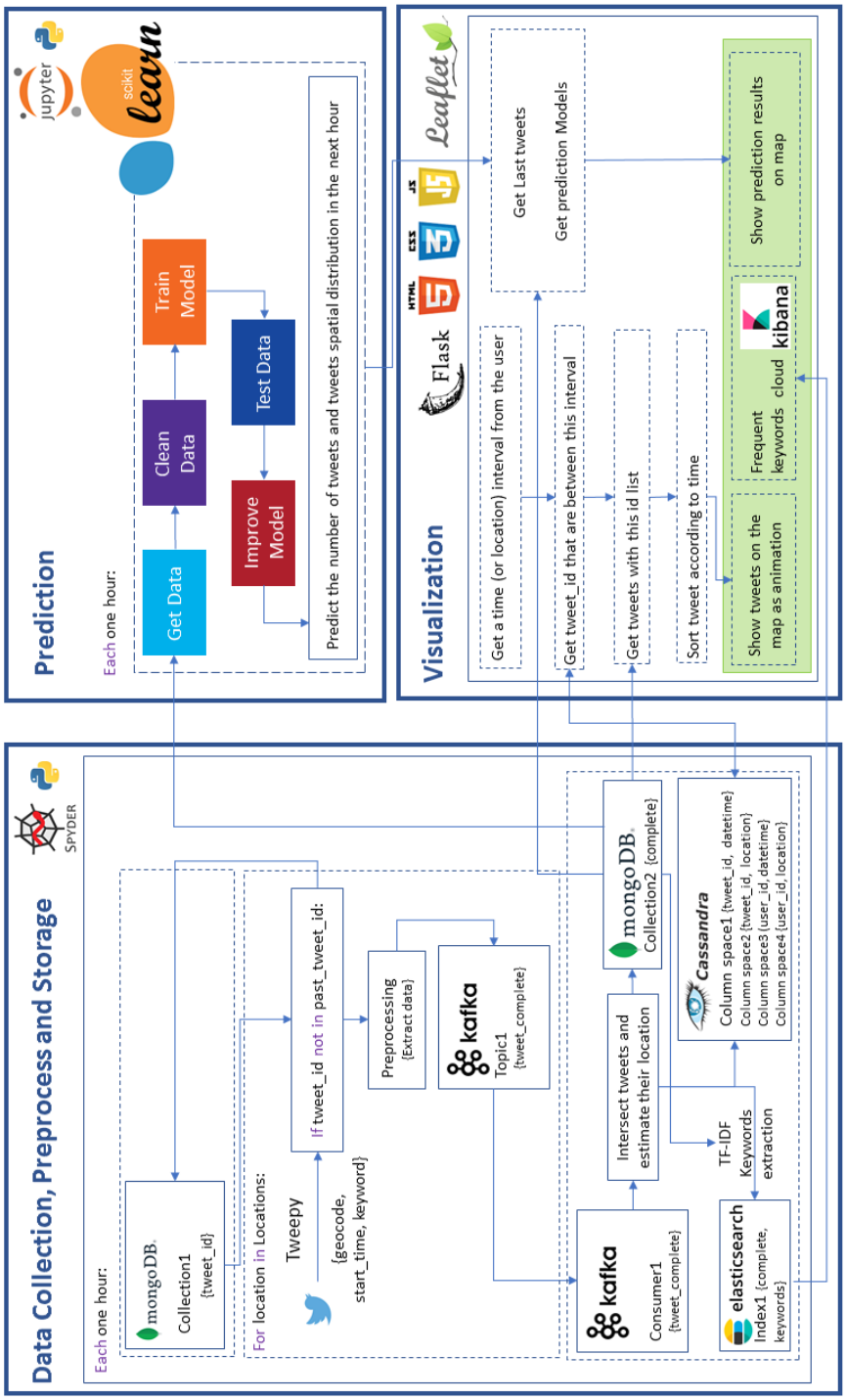

Figure 2. Architecture of proposed system

2.1. Data Collection, Preprocess and Storage

First a list of locations is defined as *Figure 3* according to them, geocode request with Twitter API by Python tweepy library is sent. Then a Python script code, sends a request with 4 parameters per hour for all locations: location geographic coordinates – keyword - start date - radius of search

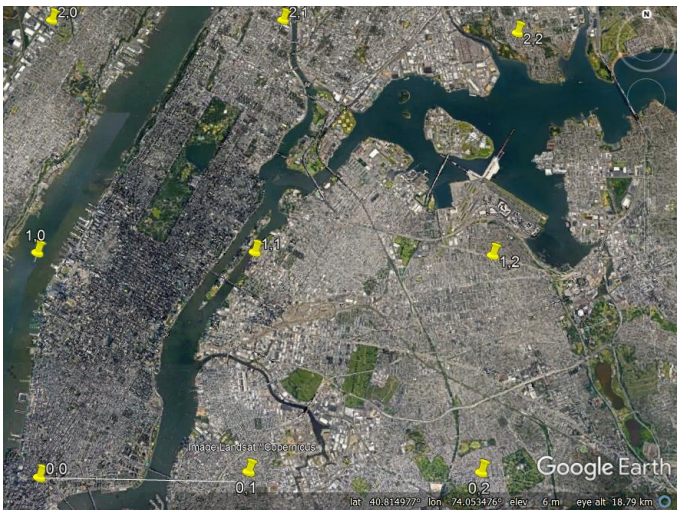

Figure 3. Locations list, New York. Distance between two neighbor points is 5km.

This paper used 'fire', '2021-09-01' and '5km'. Then all tweets with these features were returned. Then, for each tweet, if 'tweet-id' is not duplicated:

1. 'tweet-id' is saved in MongoDB first collection (MongoDB is optional).
2. **Preprocess**. All of tweet's essential data (eg hashtags) are extracted as a json.
3. The json is sent to Kafka topic using producer (Apache Kafka is an open-source distributed event streaming platform for high-performance data pipelines and streaming). Therefore, there is no need to worry about system disconnecting and reconnecting, and Kafka handles it. Kafka lets several consumers to use data separately.

When all tweets's json was sent to Kafka topic, a Kafka gets them. Then for each one:

1. **Spatial intersection.** It is possible that one tweet is returned by several requests. It means that this tweet is in the intersection of neighbors, as shown in *Figure 4*. The average of their coordinate is considered as tweet location and is added to json and the json is sent to MongoDB's other collection.
2. Tweet’s id and datetime is sent to Cassandra column space. Cassandra has been used for indexing and fast retrieval.
3. According to tweet’s text and other tweets's text, keywords are extracted by TF-IDF (Term Frequency — Inverse Document Frequency) technique and are added to json and the json is sent to Elasticsearch index. Elasticsearch has been used for text query and visualization dashboards.

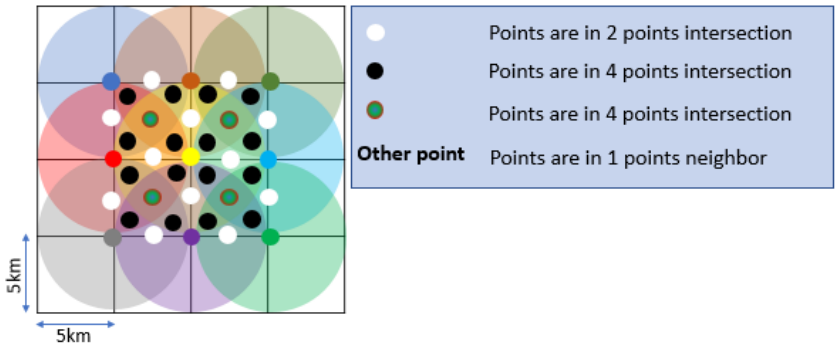

Figure 4. Spatial Intersection

2.2. Prediction

In another Python script code, each hour, all saved tweets are taken and after feature extraction and normalization, with sckit-learn library, a Multi-layer Perceptrons (MLP) model and a Random Forest model is trained for computing next hour tweets's count (function approximation) and distribution of 3 later tweets (classification). Then, the model's parameter is saved.

2.3. Visualization

A web application is been developed with FLASK framework. The user can select datetime interval, then tweets's spatio-temporal trend is shown on the map as animation. Also, users can see visualization dashboards (keywords' cloud, tweets count per hour, most frequent hashtags). Users can predict next hour tweets's count and 3 later tweets location on the map.

3. Results

This research has achieved its goals and has responded to scenario 'questions. Results are shown as several snapshot in *Figure 5-7*. *Figure 5* shows two snapshots of web app at different times while spatial distribution of tweets showing on the map. For each tweet, one red circle is added to the map in location that estimated for tweet. Also, some of tweet data such as id, text, user name, and tweet time are shown at same time. So, user can see tweets release direction and also can detect spatial and temporal hotspots that have more eager audience. *Figure 6* shows results of prediction on the map. *Figure 7* shows three dashboards on the map for most frequent Hashtags, keywords 'cloud and Tweets per hour. The full result can be seen at https://youtu.be/KOFikjf-vJ4. The MSE[3] of tweets' count prediction model is 0.00069. The average accuracy of 3 later tweets' location prediction model using MLP is 0.78 and using Random Forest is 0.7. So, MLP model is more suitable for this purpose. Research is ongoing to improve the accuracy of forecasting models using other ensemble and fusion methods,

[3] Mean Square Error

implement them using the Apache Spark and simultaneous use of Telegram data.

Figure 5. Tweets' spatio-temporal distribution for specified interval: **Top**.snapshot-1 **Bottom**.snapshot-2

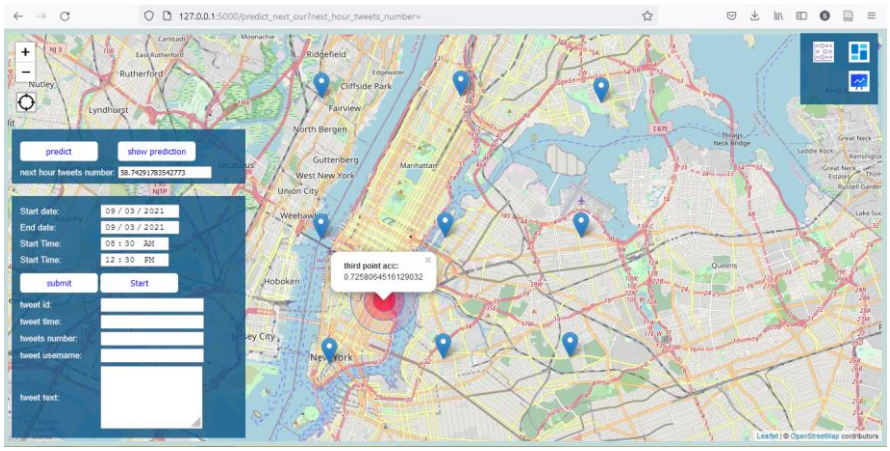

Figure 6. Prediction results. Predicted location for next three tweets.

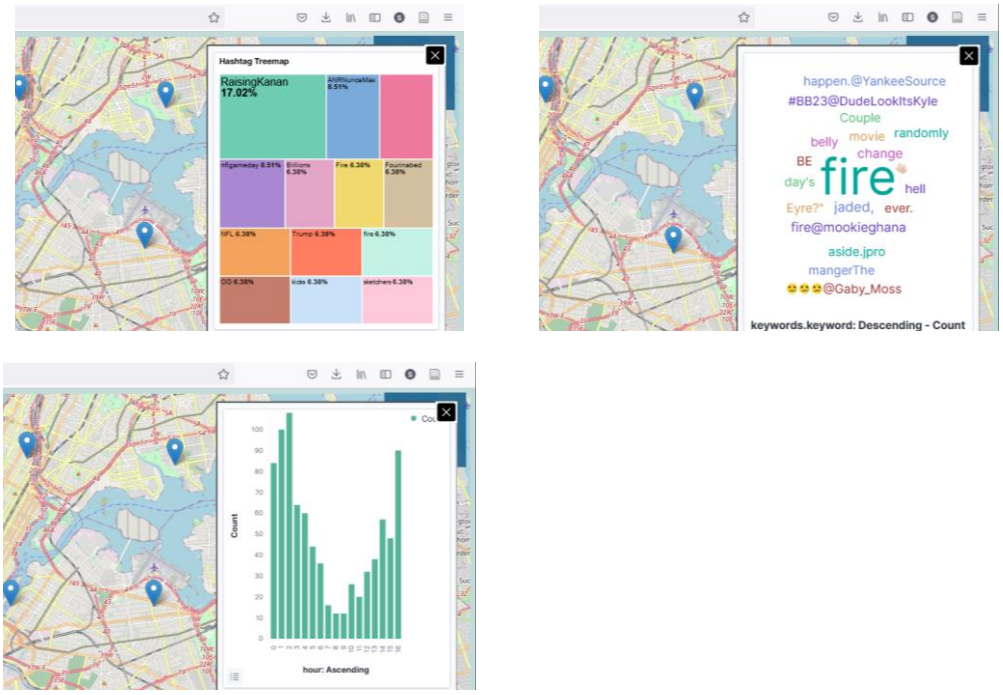

Figure 7. Elasticsearch dashboards that added to the web app: **Top-left**. Most frequent Hashtags **Top-right**. keywords 'cloud **Bottom**. Tweets-per-hour

References

Alotaibi S, Mehmood R, Katib I, et al (2020) Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning. Applied Sciences 10:1398. https://doi.org/10.3390/app10041398

Dib A (2016) The 1-Page Marketing Plan: Get New Customers, Make More Money, And Stand Out From The Crowd. Successwise

Frolova S (2014) The role of advertising in promoting a product

Martín A, Julián ABA, Cos-Gayón F (2019) Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain). Cities 86:37–50. https://doi.org/10.1016/j.cities.2018.12.014

Shirdastian H, Laroche M, Richard M-O (2019) Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. International Journal of Information Management 48:291–307. https://doi.org/10.1016/j.ijinfomgt.2017.09.007

Varol O, Ferrara E, Menczer F, Flammini A (2017) Early detection of promoted campaigns on social media. EPJ Data Science 6:1–19