Identifying Touristic Interest Using Big Data Techniques

Maritzol Tenemaza[1,2(✉)], Loza-Aguirre Edison[1], Myriam Peñafiel[1,2], Zaldumbide Juan[1,2], Angelica de Antonio[2], and Jaíme Ramirez[2]

[1] Departamento de Informática y Computación, Escuela Politécnica Nacional, Quito, Ecuador
{maritzol.tenemaza, edison.loza, myriam.penafiel, juan.zaldumbide}@epn.edu.ec
[2] Laboratorio ETSI, Universidad Politécnica de Madrid, Madrid, Spain
{angelica, jramirez}@fi.upm.es

Abstract. The objective of this paper is to identify the most visited places through a sentiment analysis of the tweets posted by people who visited a specific region of a city. The analyzed data were related to preferences and opinions about tourist places. This paper outlines an architectural framework and a methodology to collect and analysis big data from twitter platform.

Keywords: Big data · User's interest · Sentiment analysis · Harvesting

1 Introduction

Big Data techniques are widely used in data harvesting studies. The amount of data traveling on the Internet today is large, complex and interesting. Big data is the way that information is handled. The processing of large quantities of data is complex, nevertheless, there are many predictive analytics tools that control data volume, velocity and variety. The value of data or quality and veracity or consistence of data are additional issues for a big data approach [1].

Twitter contains massive human – information. Nowadays, Twitter has 350 million users geographically distributed in all world. A twitter user has little geospatial information, because the users disable the user's location in their smartphone. Twitter user tracking by associating the longitude and latitude. Microblogging today has become a very popular communication tool among Internet users. Millions of messages are appearing daily in Twitter. The users share opinions on a variety of topics and discuss current issues. Microblogging Twitter become valuable sources of people's opinions of sentiment analysis [2].

In this paper, we propose the identification and validation of the most popular tourist places in a city by using Big data Techniques and Twitter as the data source. In our case, our interest in detected the best places to visit in a specific city. We use Microblogging of twitter for the following reasons: In Microblogging platforms, the people express their opinion and sentiments. This site is constantly updated in real time and grows moment by moment. The tourist's audience tweet in regular form,

T. Ahram (Ed.): AHFE 2019, AISC 965, pp. 169–178, 2020.
https://doi.org/10.1007/978-3-030-20454-9_17

this audience is representative; it is possible to collect information of individual tourists, familial tourists, and group tourists. The tweets contain positive, negative and neutral sentiments. The geo-location Twitter users are possible to detect.

Our proposal could be applied to any city, furthermore, in this article we collected data from New York, Paris, and London. We harvested and analyzed more than 16 million tweets to find better places to visit in these cities. These results are important because any recommendation system required information of the best places previously identified by other users.

The remainder of this paper is organized as follows: In Sect. 2, an overview of different twitter analysis is mentioned. Next, in Sect. 3 the data recollection, framework proposal, analysis, and results are described. Third, in Sect. 4 the results are analyzed. Finally, conclusions and future work are discussed in Sect. 5.

2 Literature Review

The growth of online environments has made the issue of information search and selection increasingly cumbersome [3]. The recent explosion of digital data is so important because using big data, managers can measure, and hence know, radically more about their business, and directly translate that knowledge into improved decision-making and performance [4]. As of 2012, about 2.5 Exabyte of data are created each day. More data cross the internet every second than was stored in the entire internet 20 years ago. This gives companies an opportunity to work with many petabytes of data in a single data set.

However, for some companies, Velocity is more important than volume. Real Time o nearly real time information makes it possible for a company to be much more agile. Additionally, big data takes the form of messages, updates and images posted to social networks [4]. Thus, variety is another characteristic to consider when we discuss Big Data.

Twitter is a popular microblogging service where users create status messages called tweets. Twitter is mainly characterized by social functions [5] These tweets sometimes express opinions about different topics [6]. Millions of people are using social network sites to express their emotions, opinions and disclose about daily lives. However, people write anything such as social activities or any comment on products. Through the online communities provide, an interactive forum where consumers inform and influence others. Moreover, social media provides an opportunity for business that giving a platform to connect with their customers such as social media for connecting with the customer's perspective of products and services [7]. Microblogging websites have evolved to become source of varied of information on which people post real time messages about their opinions on a variety of topics discuss current issues, complain and express positive sentiment for products they use a daily life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product [8].

The amount of information about travel destinations and their associated resources such as accommodations, restaurants, museums or events among others is commonly searched for tourists in order to plan a trip [9]. Additionally, tourists visiting urban destinations require identifiers the most interesting attractions [10]. For this reason,

we observe the opportunity of analyses a data set based in microblogging Twitter, where the user's express opinions of their visit specific cities. We will analyze the sentiment expressed in a tweet the objective will be determining the most interesting places evaluated by the tourist. This information we called the opinion of other people. These results will be important for the tourism enterprises.

3 Method

3.1 Architecture

To use the Twitter API, a virtual machine was implemented. Elasticsearch, Kibana, Cerebro as servers have been used, additionally scripts Phyton were necessary to apply the harvesting architecture. The virtual machine was defined in a web server (Fig. 1).

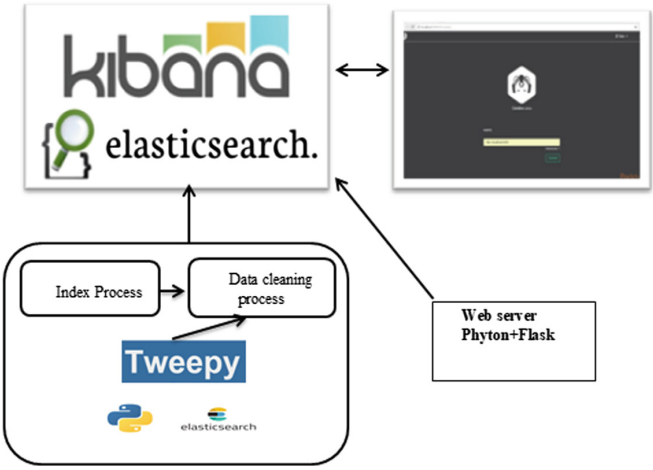

Fig. 1. Big data architecture

3.2 Methodology

For a recommender based in content, it is necessary the analysis of other tourist interest. Thus a Harvesting Methodology is applied. It defined: Data structuration, Data collection, Sentiment detection, Sentiment classification and the Presentation of results.

3.3 Data Structuration

The objective is to identify and map the attributes of Twitter that Elastic Search needs to collect. The tweets are collected in JSON format. They are send later to a NO-SQL database.

3.4 Data Collection

16 million of tweets were collected, by one month. The not structured data was structured. It was necessary to ensure that data is correct and representative. Tweepy library collected the information.

Data were collected in Paris, London and New York. Each city was segmented, Fig. 2 shown the segmentation on Paris, Fig. 3 on New York and Fig. 4 on London.

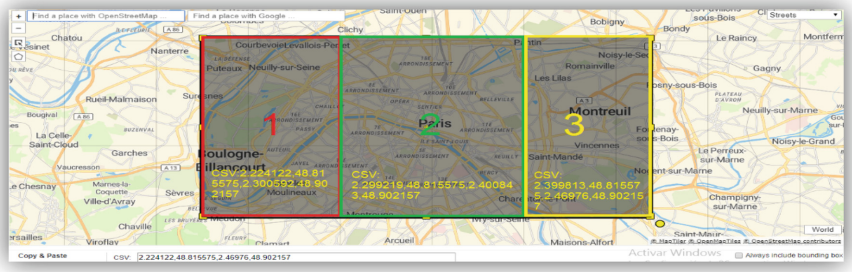

Fig. 2. Segmentation on Paris

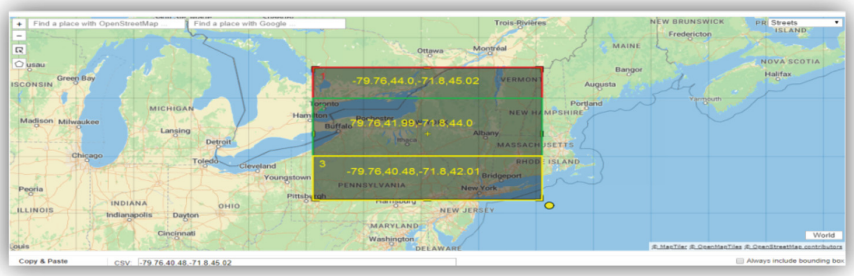

Fig. 3. Segmentation on New York

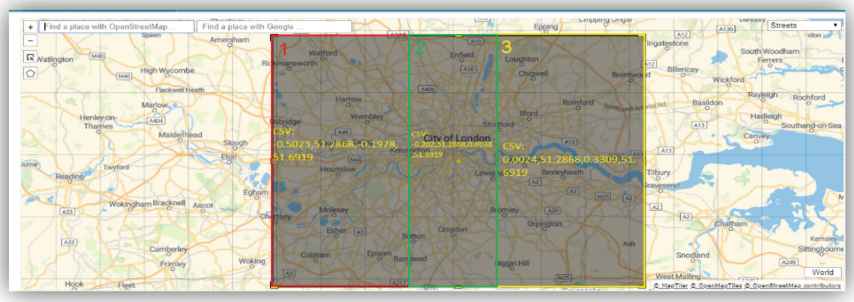

Fig. 4. Segmentation on London

The data collected from twitter is necessary transform to format JSON. Then, it was stored in Elastic Search.

3.5 Sentiment Detection

The sentiment detection of a tweet is based in the analysis of the emotional charge makes it possible to distinguish the polarity (positive, negative, neutral), intensity (positive, negative) and emotion (happy, sad and others).

The sentiment analysis is detected by Text Blob library of Phyton, for that purpose is necessary import the code to the Text Blob, that is observed in Fig. 5.

```
import tweepy
import json
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
from textblob import TextBlob
from elasticsearch import Elasticsearch
# impor er
import re
# import twitter keys and tokens
from configLnl import *
```

Fig. 5. Importation of tweets to the Text Blob

In Fig. 6 we observe the code to obtain the polarity of every tweet and writing the polarity positive, negative o neutral.

```
print ("tweet capture -ID"+ str(dict_data["id"]))

# output sentiment polarity
print (tweet.sentiment.polarity)
```

Fig. 6. Code for tweet sentiment detection

3.6 Sentiments Classification

A learning supervised algorithm was applied, the machine was trained. The polarity (positive, negative, neutral) and the subjectivity (objective/subjective) is getting from the learning algorithm.

The informal nature of twitter requires the tweet pre-processing for demand the need to correct the colloquial expressions of the texts. The actions are: (a) eliminate the

Uniform Resource Locator (URL) of message, it is observed in Fig. 7 (b) tokenizer the words of twitter, (c) delete the stop words, whitespace and lines breaks, (d) replace smileys with their corresponding categories: happy, sad, tongue, wink and others, (e) exclude terms belonging to certain morph syntactic categories that are not significant for the analysis of feelings.

For natural language pre-processing, we use Text Blob for detection and classification of feelings. The classification is observed in Fig. 8.

```
cleantext = re.sub(r'^https?:\/\/.*[\r\n]*','',dict_data["text"],flags=re.MULTILINE)

cleantext = re.sub(r'^http?:\/\/.*[\r\n]*','',cleantext,flags=re.MULTILINE)
```

Fig. 7. Clean text of twit

```
# determine if sentiment is positive, negative, or neutral
if tweet.sentiment.polarity < 0:
    sentiment = "negative"
elif tweet.sentiment.polarity == 0:
    sentiment = "neutral"
else:
    sentiment = "positive"
```

Fig. 8. Sentiments classification

Because not all tweets have coordinates. It was necessary to separate the storage of tweets that have coordinates (Fig. 9) and those that not have coordinates (Fig. 10).

```
# add text and sentiment info to elasticsearch
es.index(index="londres_coord1",
         doc_type="test-type",
         body={"author": dict_data["user"]["screen_name"],
   "date": dict_data["created_at"],
   "tweet": dict_data["text"],
   "text": cleantext          ,
   "polarity": tweet.sentiment.polarity,
   "subjectivity":tweet.sentiment.subjectivity,
   "sentiment":sentiment,
   "location":dict_data["user"]["location"],
   "coordinates":dict_data["coordinates"],
   "geo_enabled":dict_data["user"]["geo_enabled"]})

print(f"Ingreso En elasticSearch paris_coord1")
```

Fig. 9. Code for storage the tweets in elastic search

```
        else:
          print(f"Ingreso En elasticSearch newyork_all1")
          # add text and sentiment info to elasticsearch
          es.index(index="newyork_all1",
                   doc_type="test-type",
                   body={"author": dict_data["user"]["screen_name"],
              "date": dict_data["created_at"],
              "tweet": dict_data["text"],
              "text": cleantext          ,
              "polarity": tweet.sentiment.polarity,
              "subjectivity":tweet.sentiment.subjectivity,
              "sentiment":sentiment,
              "location":dict_data["user"]["location"]})
          return True
    # on failure
```

Fig. 10. Code for storage the tweets without coordinates

4 Results

The general purpose of the analysis is to transform the data obtained from twitter into meaningful information. The first step to ending the process is together the segments in each city. It is observed in Fig. 11, this process is known as re-indexation (Fig. 12).

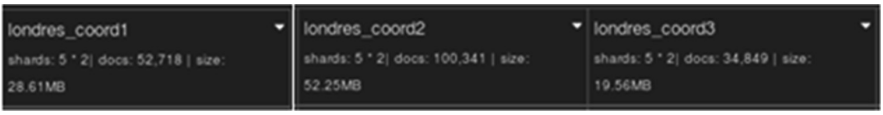

Fig. 11. London segments


```
{ -
  "cordlondres": { -
    "settings": { -
      "index": { -
        "creation_date": "1532207983503",
        "number_of_shards": "5",
        "number_of_replicas": "1",
        "uuid": "LtDMxc8xTPKuYVW0q1qD_w",
        "version": { -
          "created": "5060999"
        },
        "provided_name": "cordlondres"
      }
    }
  }
}
```


Fig. 12. Query to configure the London index

4.1 Scenarios

The three scenarios are Paris, New York and London. The data was collected from 20 May 2018 to 12 August 2018. In total, 16'064,840 tweets was collected. For this study only was analyzed the tweets with coordinates (Table 1).

Table 1. Scenarios by city

City	Tweets with coordinates	Tweets without coordinates
Paris	95,696	982,116
New York	439,996	4,121,340
London	187,908	2,109,975
World	7,139,185	-

Figure 13 shown the number of tweets in New York, The 15 July 2018 there are the greatest number of tweets. Figure 14 shown the best places in New York. The red area represents 50% of positive tweets. And the orange area represents 26% of positive tweets. The other areas son referenced in the rest of tweets. 90% of tweets refer to Times Square, Fifth Avenue, 40 theaters that make up the Broadway circle. Others references are the World Trade Center among others.

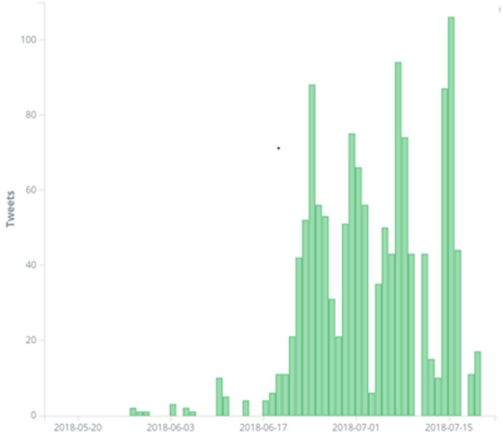

Fig. 13. Tweets in New York by date

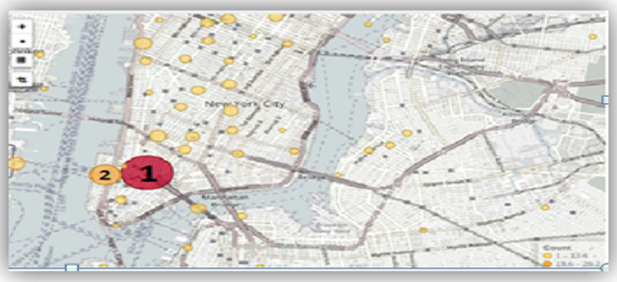

Fig. 14. The best places in New York

The same process was applied to Paris and London. With a map of coordinates all the tourist places mentioned in tweets of Paris, London or New York were appreciated. The results are observed in Fig. 15 All these results are offered in a REST web service.

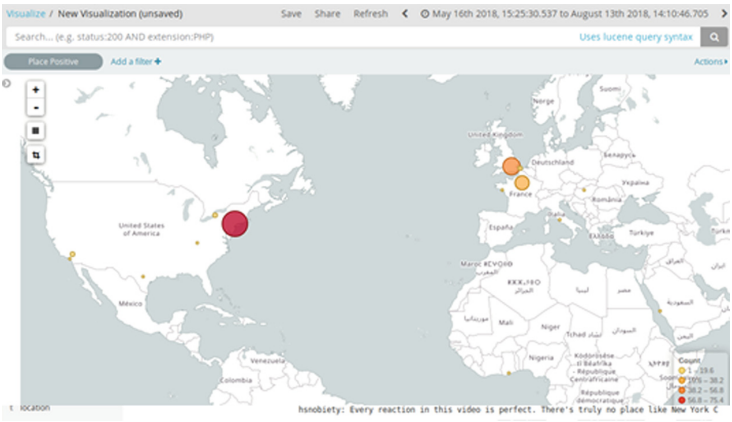

Fig. 15. Touristic places mentioned in tweets

5 Conclusions

Our results are coincident with the world barometer where the most visited city is New York and the most visited places in that city.

Our work had shown the potential of big data tools in the sentiment analysis of tweets. It is possible analyses emoticons, hashtags and others, it shows the potential of twitter information.

The big data – natural language processing tools used are useful and easily the processing of text, the polarity and translate of emoticons.

Microblogging data like twitter, on which users post real-time reactions to and opinions about specific places in different cities is the best material to define a tourism recommender.

Our future work will be developing a touristic recommender based in the other user's information based in analysis of twitter using Big Data tools.

References

1. Tole, A.A.: Big data challenges. Database Syst. J. **4**(3), 31–40 (2013)
2. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc (2010)
3. Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: Mobile recommender systems in tourism. J. Netw. Comput. Appl. **39**, 319–333 (2014)
4. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J., Barton, D.: Big data: the management revolution. Harvard Bus. Rev. **90**(10), 60–68 (2012)
5. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
6. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. Entropy **17**, 252 (2009)
7. Sarlan, A., Nadam, C., Basri, S.: Twitter sentiment analysis. In: 2014 International Conference on Information Technology and Multimedia (ICIMU). IEEE (2014)
8. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics (2011)
9. Borràs, J., Moreno, A., Valls, A.: Intelligent tourism recommender systems: a survey. Expert Syst. Appl. **41**(16), 7370–7389 (2014)
10. Gavalas, D., Kasapakis, V., Konstantopoulos, C., Pantziou, G., Vathis, N., Zaroliagis, C.: The eCOMPASS multimodal tourist tour planner. Expert Syst. Appl. **42**(21), 7303–7316 (2015)