ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 2, February 2018

Efficient Analyses And Inference Of Geo-Social Media To Make Real Time Decision In Big Data

[1]Ms.S.Sindhuja, [2]Ms.M.A.Thansira banu, [3]Ms.M.Sowndharya, [4]Ms.C.Gokulapriya,
[1][3][5]Assistant Professor, Department of Information Technology, A.V.C College of Engineering, TamilNadu, India.
[2][4]UG Students, Department of Information Technology, A.V.C College of Engineering, Tamil Nadu, India.

Abstract: ***Geo-social Networks have major impact on human life and the Geo-social Network such as twitter, face book have fetched the interest of researchers for its enormous amount of user generated content including tweets, blog posts, and forum messages is created. This data can provide benefits to governments, normal citizens and business people. Geo-social Network data can be served as an asset for the authorities to make real-time decisions and future planning by analyzing Geo-social media posts. However, there are millions of Geo-social Network users who are producing overwhelming of data, called "Big Data" that is challenging to be analyzed and make real-time decisions. In this proposed architecture Twitter data are analyzed in order to identify current events or disasters. The proposed system consists of five layers i.e., Tweet collection, Data Preprocessing, Feature Extraction, Classification, and Decision Making. During the training phase, the tweets are used to classify the emotions depending on the words which are extracted and processed using Natural Language Processing (NLP) techniques. The posts, texts, tweets, comments, statuses, and smiley are analyzed using text analytics, statistical analytics, complex machine learning and data mining techniques in order to monitor and determine what is occurring, where and why. Such type of data can be used to predict future events based on the current user trends that correspond to various areas.***

Keywords:Geo- social Media; Big-Data;Hadoop; NLTK[Natural Language Toolkit]

I. INTRODUCTION

Geo-social networking is a type of social networking in which geographic services and capabilities such as geo- coding and geo-tagging are used to enable additional social dynamics. Geocoding is the process of transforming a description of a location such as a pair of coordinates, an address, or a name of a place to a location on the earth's surface. Geo-tagging is the process of adding geographical information such as digital photograph or video, a posting on a social media website can help people get a lot of specific information about where the picture was taken or the exact location of a friend who logged on to a service. Geo-location techniques can allow social networks to

connect and coordinate users with local people or events that match their interests. Geo-social network allows users to interact relative to their current locations by using Twitter, Facebook etc. The location of people posting, commenting, or uploading pictures on social media is recorded. Thus, by aggregating such type of location data from all network users, social networks produce warehouses of Geo-social data. For example, Marketing Campaigns in Twitter using a Pattern Based Diffusion policy introduce a novel methodology to achieve information diffusion within a social graph that activates a realistic number of users for effective marketing [1]. Social network data could be beneficial for many fields to analyze and make decisions[2]. Geo-social media data are also used to predict the number of spectators and TV ratings of football viewers, regards to football fans off-line and on-line behaviors[3]. Some other applications with limited functionalities, such as Spark which handle both batch, Real-time analytics and data processing workloads[4].

In addition Geo-social media, data analysis can help an airline company to understand the airline passengers and improve customer relationship management, using sina weibo posts means popular social media websites on china[5]. Manuel Rodriguez-Martrinez used twitter data for Twitter Health Surveillance (THS) System to allow end- users to monitor a stream of tweets and process the stream with a combination of built-in functionality and their own user-defined functions[6].Similarly, Andy Januar Wicaksono proposed Predicting US Presidential Election by Analyzing Sentiment in Social Media using data collection technique to make the prediction more accurate and also adjust to the actual situation[7].

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 2, February 2018

To support many real-world applications, such as public opinion monitoring for governments and news recommendation for websites are achieved by Sentiment Computing[8]. Storing and handling of large quantities of data with some features such as easy accessibility, fast performance, durability and security can be achieved using High Performance analytics system[9]. Large scale opinion

mining system used to collect and analyze data from social media [10]. Processing real-time geo-social media data is a very challenging task. Special computational environment and advanced computing techniques is needed to make real- time decisions.

II. PROPOSED COMPUTING MODEL

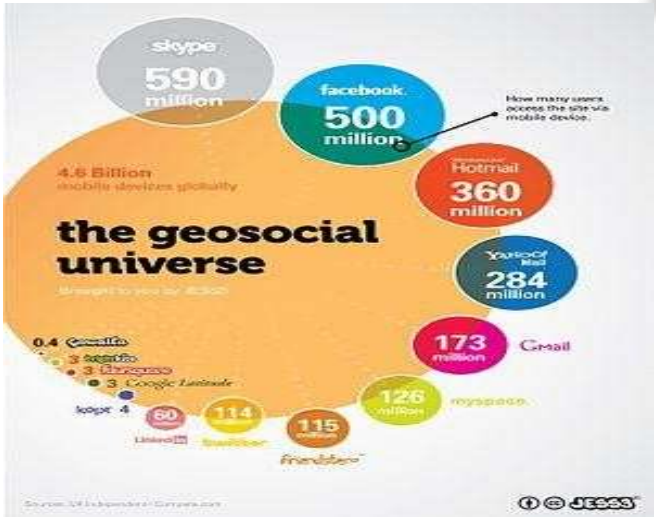
This section describes the proposed model including system overview, proposed architecture, python language and Natural language toolkit.

A. System Overview

Humans are the most reliable source for reporting events, activities and important issues. Geo-social Networks use humans as sensors for monitoring activities worldwide. When a user posts any activity that is related to some events, the user acts as a sensor that sends data to its station. All such information corresponds to user locations and can provide real time monitoring of disasters, fatal diseases or accidents. Moreover, user data and their location information can be used to recommend various systems based on the user's current location such as useful products, restaurants, hotels and transportation.

Types of Geo- social media data

1. Tweets
2. Facebook Posts
3. Blog posts
4. Discussion forums
5. Online reviews
6. Social publishing platforms such as Word Press and Blogger.



B. Proposed System

In our proposed system we use Python as our base programming language which is used for writing code snippets. NLTK could be a library of Python that plays really important role in changing the text to a sentiment either positive or negative. NLTK additionally provides different sets of knowledge that are used for classifiers. These datasets are structured and keep in library of NLTK which can be accessed simply with the assistance of Python.

The system deploys a Hadoop scheme for information processing. Apache Hadoop could be a ASCII text file implementation of frameworks for reliable, scalable, distributed computing and information storage.

It is a versatile, highly-available architecture for giant scale computation and processing on a network of artifact hardware. During this system, Natural Language process (NLP) is combined with Hadoop. This planned approach is to spot the options from unstructured matter reviews. Feature choice is associate activity that choose relevant options supported a selected measurement.

C. Python

Python is a high level, dynamic programming language. It is a mature, versatile and strong programming language. Python is a dynamic programming language, that is well framed for its practicality of process language knowledge, i.e. spoken English exploitation NLTK.

D . Natural language toolkit

NLTK is a library of python, that provides a base for building programs and classification of knowledge. NLTK provide graphical demonstration for representing various results or trends and it also provide sample data to train and test various classifiers respectively in Python.

NLTK contains functions to perform a large set of tasks within the field of text process. One function often used as the first natural language processing stage is tokenization, available in NLTK using the function word tokenize, which takes as input a string of text and returns a list of strings, where each element of the list is a word or a piece of punctuation. A tokenized text, i.e. a list of strings, can be passed to NLTK for part-of-speech tagging, either using a specialized tagger or the default offering.

NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them.

NLTK provides varied functions that are employed in pre- process information of the twitter. NLTK support various machine learning algorithms which are used for training classifier and to calculate the accuracy of different classifier.

II. PROCESSING STAGES

In this section, we are going to see the different stages in processing of twitter data.

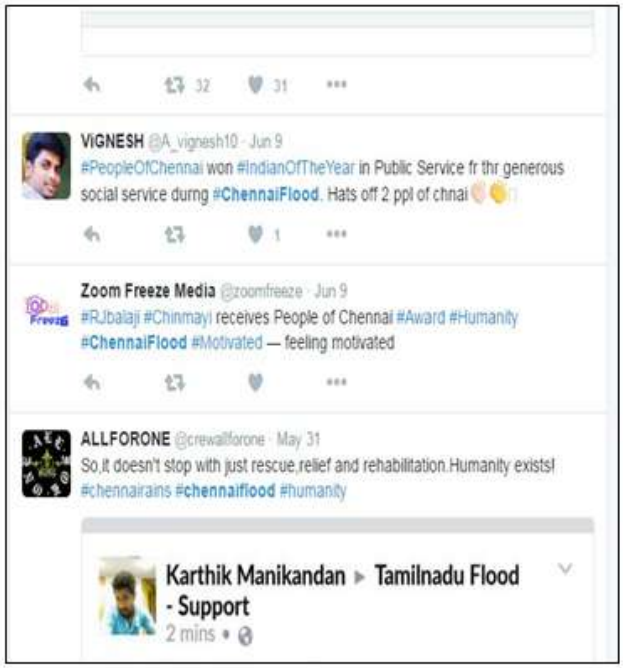
A. Collecting tweets

The first step in this proposed methodology was to collect a large set of texts from tweets. For this purpose, the Python package tweepy and json are used to retrieve tweets. The tweepy and json packages fetch tweets from Twitters streaming API in real-time. Tweepy is one amongst the open python library that permits python to speak with twitter and its API to gather knowledge. For retrieving the tweets, the hash tag data should be given as the input.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 2, February 2018

Hash tag

User can create hash tags by placing the hash character (or pound sign) # in front of a word or either in the main text of a message or at the end to retrieve exact information. Searching for that hash tag can then gift every message that has been labeled with it.



B. Preprocessing

Data obtained from twitter is not appropriate extracting options. Largely tweets consists of message beside user names, empty areas, special characters, stop words, emoticons, abbreviations, hash tags, timestamps, URL's ,etc. therefore to form this information appropriate mining we have a tendency to pre-process this information by NLTK. In pre-processing we have a tendency to first extract our main message from the tweet, then we have a tendency to take away all empty areas, stop words (like is, a, the, he, them, etc.), hash tags, continuance words, URL's, etc. we have a tendency to then replace all emoticons and abbreviations with their corresponding meanings like :-), =D, =), LOL, Rolf, etc. are replaced with happy or laugh. Once we have a tendency to are finished it, we have a tendency to are prepared with processed tweet that is provided to classifier for needed results.

We create a code in Python in which we define a function which will be used to obtain processed tweet. This code is used to achieve the following functions:

∑ remove quotes	Provides the user to remove quotes from the text
∑ remove @	Provides choice of removing the @ symbol, removing the @ along with the user name, or replace the @ and the user name with a word 'AT_USER' and add it to stop words
∑ remove URL	Provides choices of removing URLs or replacing them with 'URL' word and add it to stop words
∑ remove RT (Re-Tweet)	Removes the word RT from tweets
∑ remove Emoticons	Remove emoticons from tweets and replace them with their specific meaning
∑ remove duplicates	Remove all repeating words from text so that there will be no duplicates
∑ remove #	Removes the hash tag class
∑ remove stop words	Remove all stop words like a, he, the, and etc.., which provides no meaning for classification

C. Feature extraction

Training and testing data is collected from NLTK corpus. Both the training and testing data must be represented in same order for learning. One of the ways that data can be represented is feature-based. Attribute choice is that the method of extracting options by that the information are going to be delineate before any machine learning takes place. Attribute selection is the first task when one intends to represent instances for machine learning. Once the attributes are select, the information are going mistreated to the attributes. So attributes are the features. Although we used the entire data set in our selection of attributes, the representation of the data must be done as per instance (Twitter post) basis. Feature vector plays a very important role in classification and helps to determine the working of the build classifier. Feature vector also help in predicting the unknown data sample. There are many types of feature vectors, but in this process we used unigram approach. Each tweet words are added to generate the feature vectors. The presence/absence of sentimental word helps to indicate the polarity of the sentences. Once we have a tendency to extract the options from information, we have to pass these in our build classifiers. A script written in python which is used to pass training sets in classifier. Once, the classifier is trained we can also check the accuracy of each classifier by passing the testing set.

D. Classification

The goal of the classification stage is to assign two classes to each tweet, one describing the sentiment of the tweet and one Describing the subject matter discussed in the tweet. The assigned classes are later used as the basis of the response to The tweet that is generated by the program. The sentiment of a tweet can be classified as positive, negative, or objective. While the classes for the sentiment classification are rather general, the classes for the classification of subject matter must be chosen with consideration of the data sources. Since the number of classes is otherwise almost unlimited. With this in

mind, the subject matter of a tweet can be classified as scientific news, scientific opinion, popular science, political or other. To classify tweets in different class (positive and negative) we build a classifier which consists of several machine

learning classifiers. To build our classifier we used a library of Python called, Scikit-learn. Scikit-learn is a very powerful and most useful library in Python which provides many classification algorithms. Scikit-learn also include tools for classification, clustering, regression and visualization. In order to build our classifier, we use seven in-build classifiers which come in Scikit-learn library, which are:

∑ Naïve-Bayes Classifier
∑ MultinomialNB Classifier
∑ BernoulliNB Classifier
∑ Logistic Regression Classifier
∑ SGDC
∑ Linear SVC and
∑ Nu SVC

E . Decision making

Every person stores information on the blogs, various web applications and the web social media, social websites. For getting the relevant information we need to perform Big-data analytics to analyze and process data to return some of the useful results.

III. ANALYSIS AND DISCUSSION

In this section, Geo-social media data's are analyzed to detect various disasters and events.

A. Description of dataset

Twitter's stream grab the twitter data. Each Datasets are more than 40GB in size, has the tweets of more than a month and is classified by date. Data's are classified with respect to popular hash tags. Tweets are heterogeneous in nature.

B. Discussion

Tweets of events and disasters in varied Earth regions, like earthquakes, tsunami and other fatal diseases are analyzed. In the case of earthquake analysis, we have a tendency to found the majority of the tweets from various places.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 2, February 2018

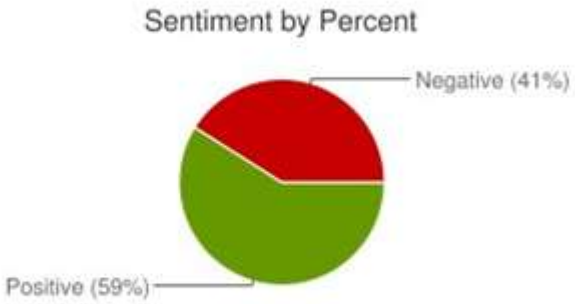


Fig. 1. Sentiment Classification of tweets in percent

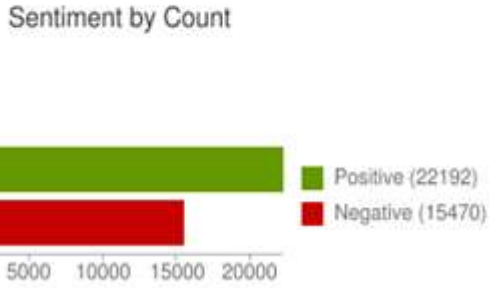


Fig. 2. Sentiment Classification of tweets by count.

Fig. 3. Graphical representation

IV. SYSTEM IMPLEMENTATION AND EVALUATION

For the proposed system implementation, we tend to directly collect the information from Twitter by making a Twitter application with the access token, key, and token secret. The real time twitter information is harvested ceaselessly in chunks of few second. Here we are using Python as our base programing language that is employed for writing code snippets. NLTK is a library of Python that plays a really necessary role in changing language text to a sentiment either positive or negative and also helps us to classify various tweets of twitter.

V. CONCLUSION

Geo-social Networks are beneficial for governments in terms of providing facilities and safety from disasters through correct management and reduction of the fear of spread of any infections. It is also beneficial to common citizens by the way of providing recommender system, safety, healthcare etc., and to entrepreneurs for launching new merchandise in numerous spaces by observing the Geo- social information of a specific area. Such benefits can be achieved only by better analytics of Geo-social network data. In Geo-social media humans acts as sensors for observing activities world wide. Once a user posts any activity that's associated with some events, the user acts as a sensor that sends knowledge to its station to search out data concerning what's occurring in the world. In this proposed system, we analyze posts, texts, tweets, comments, statuses and smiley using text analytics, statistical analytics, complex machine learning and data mining techniques in order to monitor and determine what is occurring, where and why. Therefore, in this paper, we proposed a system that uses Geo-social data for making better plans, safety from disasters and awareness, etc.,. The system not only harvest large amount of Geo-social network data, it can also process, analyze and make decisions in real time.

ACKNOWLEDGMENT

We wish to express our sincere gratitude to Ms.S.Sindhuja Asst. Professor, whose supervision & guidance in this investigation has been carried out. Without her guidance and constant supervision it is not possible for us to complete this research paper successfully.

REFERENCES

[1] Eleanna Kafeza, Christos, Makris, Pantelis Vikatos, “Marketing Campaigns in Twitter using a Pattern Based Diffusion policy”, Big Data (Big Data Congress), 2016 IEEE International Congress on 27 June 2016.

[2] Nadiya Straton, Raghava Rao Mukkamala, Ravi Vatrapu , “ Big Social Data Analytics for Public Health: Predicting Facebook Post Performance using Artificial neural network and Deep Learning”, Big Data (Big Data Congress), 2017 IEEE International Congress on 25June 2017.

[3] Nicolai H. Egebjerg, Niklas Hedegaard, Gerda Kuum, Raghava Rao Mukkamala and Ravi Vatrapu, “ Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data”, Big Data (Big Data Congress), 2017 IEEE International Congress on 25 June 2017.

[4] Adel Assiri, Ahmed Emam, Hmood Al-dossari, “Real-Time Sentiment Analysis of Saudi Dialect Tweets Using SPARK”, Big Data (Big Data), 2016 IEEE International Conference on 5 December 2016.

[5] Sien Chen, Yinghua Huang, Wengqiang Huang, “ Big Data Analytics on Aviation Social Media: The Case of China Southern Airlines on Sina Weibo” Big Data Computing Service and Applications (Big Data Service), 2016 IEEE Second International Conference on 29 March 2016.

[6] Manuel Rodriguez-Martinez, “Experiences with the Twitter Health Surveillance (THS) System”, Big Data (Big Data Congress), 2017 IEEE International Congress on 25 June 2017.

[7] Andy Januar Wicaksono, Suyoto, Pranowo, “A Proposed Method for Predicting US Presidential Election by Analyzing Sentiment in Social Media”, Science in Information Technology (ICSITech), 2016 2nd International Conference on 26 October 2016.

[8] Dandan Jiang, Xiangfeng Luo, Junyu Xuan, Zheng Xu, “Sentiment Computing for the News Event Based on the Social Media Big Data”, IEEE Access (Volume: 5), 15 September 2016

[9] Yanish Pradhananga , Shridevi Karande, Chandraprakash Karande, “High Performance Analytics of Big data with Dynamic and Optimized Hadoop Cluster”, Advanced Communication Control and Computing Technologies (ICACCCT), 2016 International Conference on 25 May 2016

[10] Nikos Tsirakis, Vasilis Poulopoulos, Panagiotis Tsantilas, Iraklis Varlamis, “Large scale opinion mining for social, news and blog data”, Journal of Systems and Software Volume 127, May 2017.cation With Randomized Tag”, IEEE Transactions On Information Forensics And Security, Vol. 12, NO. 3, March 2017.