iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35571

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

A System to Predict an Outbreak of a Disease using Twitter Data

Sneha M. Yadav[1], Dr. Bijith Marakarkandy[2]
Thakur College of Engineering and Technology Kandivali[E] , Mumbai,, Maharashtra 400101

Abstract: ***Social media (Twitter) data can be analyzed by the system to provide data regarding the diseases by checking the location from the tweets retrieved from Twitter. This will be done by searching for the disease in the system. Using tweets most used hash tag is retrieved to find out which disease is discussed to spread the awareness and it is helpful to prevent spreading of communicable Disease. Detecting the diseases that are widely spread in the society and predicting future stages of the diseases has become very important in modern life. The System is to make people aware of the Epidemic diseases in a particular location. The people can also search for other non-communicable diseases to check the tweets helpful to them in providing information regarding the disease. As many people are using social media to share different information we are utilizing this fact to make our system more efficient and reliable by taking the data from Twitter and analyzing various factors like sentiments, location. This can be utilized by anyone who needs information regarding the disease and can be utilized to support clinical providers, public health experts, social scientists, healthcare providers and Government.***
Keywords: Twitter API, TextBlob, WordCloud, Tweets, Sentiment Analysis, Statistical Analysis, ANOVA, NetworkX.

I. INTRODUCTION

Public posts on a social website such as Twitter include personal status, opinion sharing, discussions, marketing, campaigning. Twitter channel presents gold mine of opportunities with massive potential to increase coverage and awareness. Among the material users share on Twitter are tweets related to health and healthcare. Some users share information and updates on their health or the health of loved ones. This is a common behavior by many people who seek support during difficult times. These social posts can be of tremendous value if detected and monitored by healthcare providers. They provide indicators about the general public health, they can help in early detecting of an epidemic, or can alert about ongoing concerns with healthcare. In addition, people often share their experience with healthcare facilities like hospitals, clinics, and health centers[1].

Twitter is one of the social media that is gaining popularity. Twitter is a gold mine of data. There are 330 million monthly active users and 140 million daily active users on twitter. Top three countries by user count outside U.S. are Japan(49.1 million users),India(17 million) and Brazil(15.7 million).There are 63% of users who are in between 35-65 years old Unlike other social platforms, almost every user's tweets are completely public and pullable. This is a huge plus if you're trying to get a large amount of data to run analytics on. Twitter data is also pretty specific. Twitter's API allows you to do complex queries like pulling every tweet about a certain topic within the last twenty minutes, or pull a certain user's non- retweeted tweets [2].

Social media platforms are also useful epidemiological data sources. Twitter is quickly becoming a favorite social media data source for epidemiologists because it can provide more contextual information than search queries. Frequently , people who are experiencing symptoms or have been confirmed to have a specific disease post about it on social media [3].

II. LITERATURE SURVEY

Twitter tweets are analyzed by many researchers as they are good source of information about a particular disease. They have used different techniques and Method to extract insights from it.

Method and Technique	Author	Disease Name	Results
Prediction of death rates using R-Studio	P Amrutha Vali 2017	Cancer	Sates of US
Influenza Detection and Surveilience	Kenny Byrd 2016	Influenza	Canada
FluTrack platform for flu related tweets	Karolos Talvis 2014	Flu	USA
A tool for monitoring Healthcare tweets forgovernment or Healthcare providers	Ahmed Ali	–	Platform is tested and demonstrated in London, Dublin and Boston

Table I :Methods of Tweet use

III. PROBLEM DEFINITION

Social Networks connect people without the barrier of geographical locations. They exchange their views and opinions without hesitation. Sometimes seeks mental support by searching people who have same problem and gets advices. Health care systems, medical practitioners and government systems can identify this trends to get useful insights and those used to spread relevant information about particular disease and related medicines. Here, as data or tweets considered from twitter we take tweets as exposed population to carry out our research. It is the scientific, data driven and systematic study of the frequency, patterns, causes and risk factors of health related states and events in specific locations country, state, global. This encourages to make application that locate the outbreak of disease by analyzing tweets using the hashtags, sentiments of tweets and the location.

IV. RELATED THEORY

A. Twitter API

Twitter's API allows to do complex queries like pulling every tweet about a certain topic within the last twenty minutes. Collecting tweets for the entered key words from twitter requires the API using which will access the tweets and fetch them to perform further operation. To do this one has to do request to twitter for developer access using proper and valid use case along with legal access as client application. Request has been made need to be approved by Twitter and once approved they provide with different set of keys. Those keys used with weepy API to retrieve tweets from the Twitter database. Authentication is handled by the tweepy.AuthHandler class. . Twitter's API allows to do complex queries like pulling every tweet about a certain topic within the last twenty minutes[4][5].

B. TextBlob

It is python (2 and 3) library which process textual data. It provides a simple API for diving into different NLP tasks such as sentiment analysis. It supports complex analysis and operations on textual data. TextBlob gives Subjectivity and Polarity of a text. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Polarity reversed by negation words [6].

C. WordCloud

A WordCloud is collection of or cluster of words depicted in different sizes. The bigger and bolder the word appears the more often it's mentioned within given text and the more important it is. It is also known as tag clouds or text clouds, ideal ways to pull out most relevant parts of textual data from blog posts to database[7].

D. ANOVA

ANOVA test is a way to find out if survey or experiment results are significant. It helps to figure out if you need to reject null hypothesis or accept the alternate hypothesis. One way Anova referred to one factor Anova. The reason behind it even though there are 3 or more groups being tested , those groups are under one categorical variable and the name is referring to the number of variables in the analysis and not the number of groups. Anova is parametric test used to test for statistically significantly different than the others[8].

V. METHODOLOGY

The proposed system designed in four modules. It initiates with collection of tweets then preprocessing it and getting a keywords which gives required insights. These tweets are further analyzed using TextBlob to generate sentiment polarity of tweets and display it using Pie Chart. WordCloud is used to display the most often used hashtag ,disease name and locations. Then data is stored into the database. Stored data is further statistically analyzed using ANOVA to prove the assumption we made is totally satisfied. Network Analysis is also done using NetworkX to check the network density of sentiments and locations. Hence the tweets will be used in future as crucial data to make prediction about epidemic diseases. The Figure 1 explains the step by step methods described in proposed system along with the necessary techniques used to perform the operations on the data going through each steps. The diagram explains the proposed flow of data from the initial level as a raw data till the results received as output and how output is depicted using different visualization form. It help to give basic structure of the system .

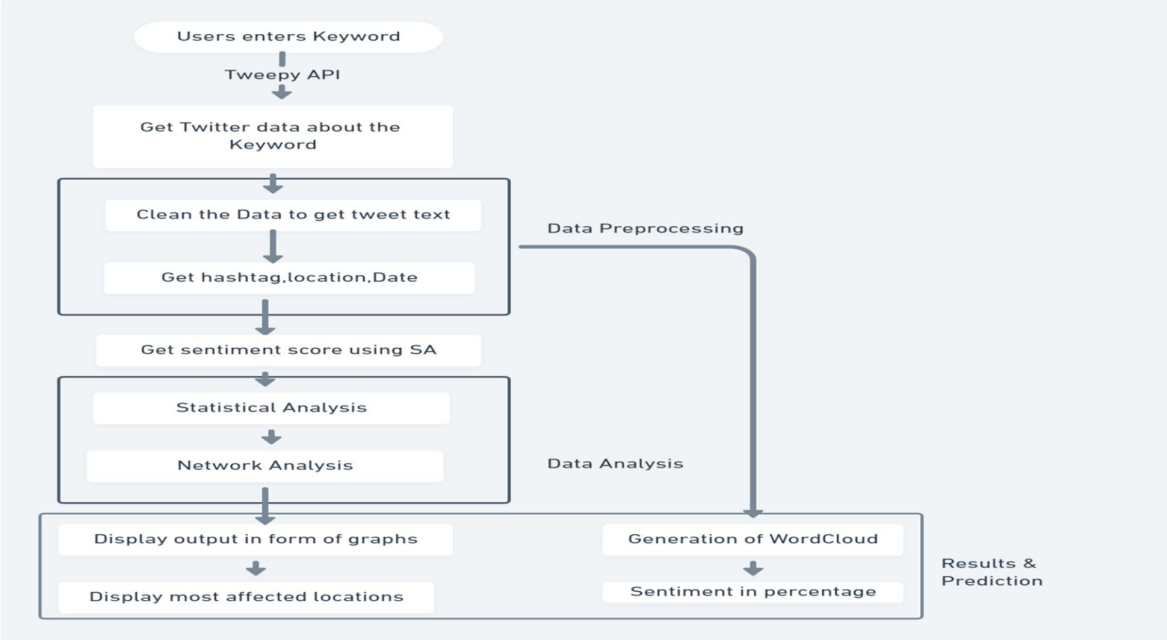


Fig 1: System Architechure

A. Data Collection

Tweepy is the python client for the official Twitter API .Open source python package that gives convenient access to the Twitter API with python like an interface. It includes set of classes and methods that represents Twitter's model and API endpoints and handles implementation like Data encoding and decoding transparently. It gives access to most Twitter's functionality.

1) *Application Registration:* Log in to the Twitter developer portal(https://developer.twitter.com/), all we need to do is point our browser to the Application Management page from Apps to menu and create the new app.
2) *Twitter API Client Selection:* Once the app is registered, under the Keys and Access Tokens tab, we can find the information we need to authenticate our application.

a) *Consumer API key:* Key associated with the application
b) *Consumer API secret key:* Password used to authenticate with the authentication server
c) *Access token key:* given to the client after successful authentication of above keys.
d) *Access token secret:* Password for the access key

B. Sentiment Analysis in python with TextBlob

TextBlob is package of python library applies to sentiment analysis. It is rule based and requires pre- defined set of categorized words. Once the initial step is done and python model is fed by the input data. User can obtain obtain sentiment scores in the form polarity and subjectivity[23].

```
from textblob import TextBlob

    #get tweets text
    tw_text = tweet.full_text
    tweets.append(strip_links(tw_text.lower()))

    #get Sentiment of text
    tweet_blob = TextBlob(strip_links(tweet.full_text))

    if tweet_blob.sentiment.polarity < 0:
        sentiment = 1
    elif tweet_blob.sentiment.polarity == 0:
        sentiment = 2
    else:
        sentiment = 3

    tweet_Sentiments.append(sentiment)
```

Fig 2: SA using TextBlob

C. Data Analysis (Statistical and Network Analysis)

The preprocessed data need to be stored for Statistic and Network Analysis purpose. Database software helps in storing data in structured format and also allows retrieving data as per the requirement of users.

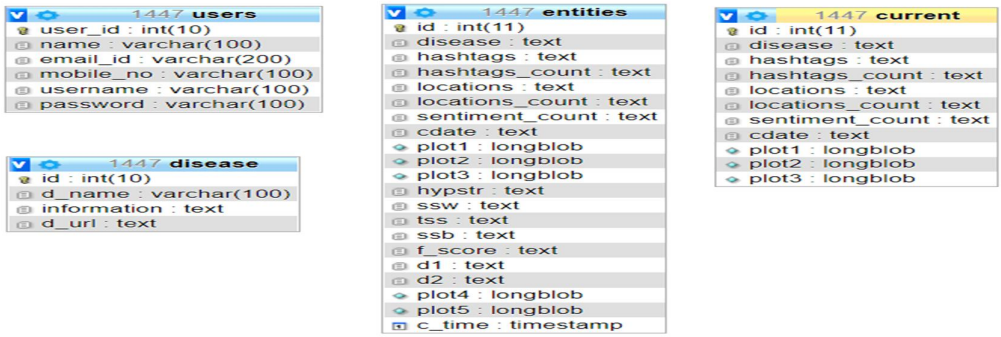


Fig 3: Database

D. Statistical Analysis

ANOVA test is a way to find out if survey or experiment results are significant. It helps to figure out if you need to reject null hypothesis or accept the alternate hypothesis.

The hypotheses of interest in an ANOVA are as follows:

- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$
- H_1: Means are not all equal.

where k = the number of independent comparison groups.

Sum of Squares between the groups , Sum of squares within the groups ,Total sum of squares within the groups and F_{score} calculated considering tweets collected from last three days and sentiment of tweet and time of tweet.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
Between samples	SS_B	$k-1$	$MS_B = \frac{SS_B}{(k-1)}$	$\frac{MS_B}{MS_W}$
Within samples	SS_W	$n_T - k$	$MS_W = \frac{SS_W}{(n_T-k)}$	
Total	$TSS = SS_B + SS_W$	$n_T - 1$		

Table II: Anova Table

Mean of each group is calculated to get SS_B and SS_W. Degree of freedom is calculated for between samples and within samples to get the value of $\mathbf{MS_B = 2}$ and $\mathbf{MS_W = 6}$. This helps in calculating F-Statistics. **F-Statistics = $\mathbf{MS_B}$ / $\mathbf{MS_W}$ Which is 5.14 .**
When API sends a request for tweets regarding disease, collected tweets are analyzed to give the above values for the particular disease. These values helps to find out the assumption we made is significant. The calculated values for particular disease related tweets falls below calculated F-Statistics value then the null hypothesis is rejected.

E. Network Analysis

NetworkX is a Python language software package for the creation , manipulation and study of the structure , dynamics and functions of complex networks. The two parameters are processed to plot a graph namely sentiment and location of a particular disease. A graph is collection of nodes and edges. In NetworkX , nodes are hashable object like text string, an image, another graph or a customized node object.
The Python code uses matplotlib.plot to plot the graph. Pandas DataFrame is used where edges are specified easily.

```
import pandas as pd

import networkx as nx
import matplotlib.pyplot as plt


def plot_density(Sentiment_new, Location_new, filename):
    """
    save the density map
    """
    df = pd.DataFrame({ 'from': Sentiment_new, 'to': Location_new})#pandas Dataframe
    G=nx.from_pandas_edgelist(df, 'to', 'from')
    density = nx.density(G)
    nx.draw(G, with_labels=True, arrows = True)
    plt.savefig(filename)
    file = imageToByte(filename)
    return density,file
```

Fig 4: NetworkX Function

Graph of sentiment and location for specific disease using the all tweets retrieved in one request. This gives information about the density of network and how strongly the discussion is made using tweets. It can be utilized to locate hotspot of affected disease.

Figure 5 shows working of the system from the entering a disease name by user till getting the relevant results. It checks for the mentioned conditions and after satisfying them data is handled to next phase. Each module clearly states the technique used with the operations. From collection of raw data to predict the disease on the basis of relevant attributes all steps are performed for the analysis purpose to get values those helps to find out the outbreak and calculate the factors that will allow to perform the detailed analysis.

A website designed to implement a proposed methodology and execute the expected results. Different techniques are collaborated together and linked with each other for smooth functioning and data handling.

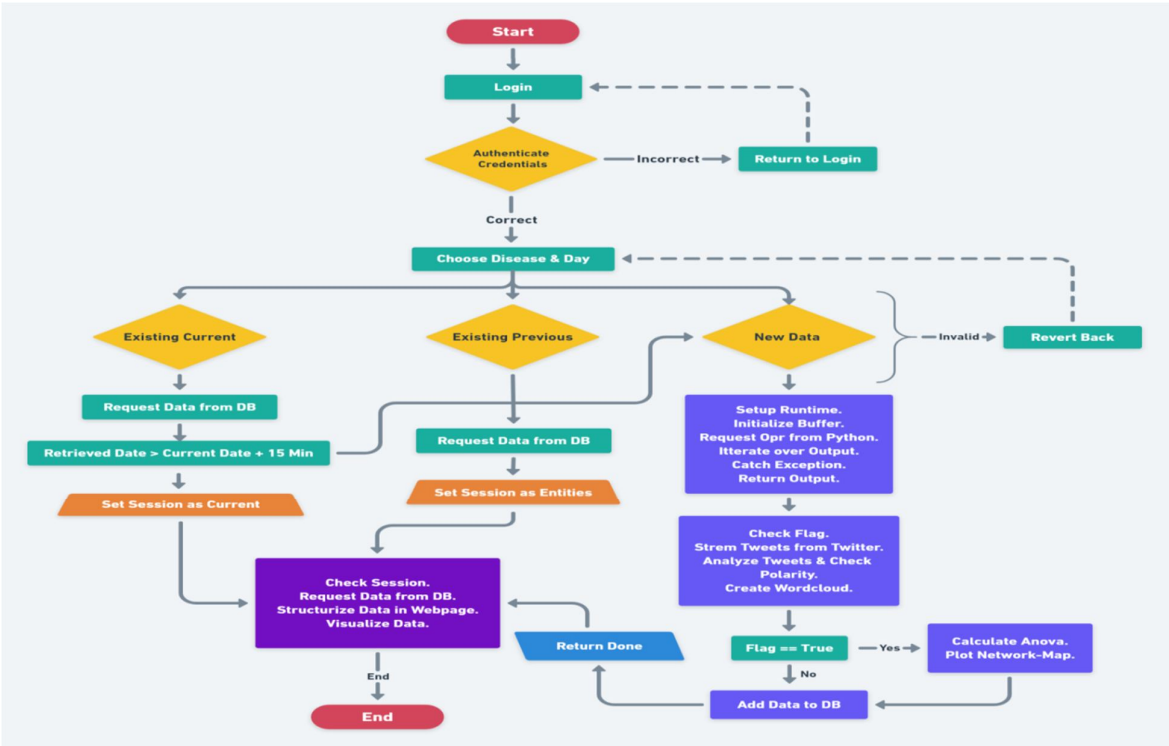


Fig 5: Work flow of the System

VI. RESULT AND ANALYSIS

A user need to enter the username and password to use a system and after validating the credentials authorized access is given. Registered user gets the access to move to next page where he/she can enter the disease name and day. This page is for entering the disease name (keyword) and the particular day. Request for tweets made upon the category selected by the user. And the related tweets are retrieved using the API. Category is either current day or previous day. Tweets gathered are processed using mentioned techniques in last chapter. And the results are displayed using Bar Graphs, Pie Charts and WordCloud.

For example below are the results upon entering disease name as Swine Flu and choosing day as current. The tweet retrieved is for the current date and time. Different Hash Tags collected for the entered disease are depicted using Bar Graph. Along with hashtags we retrieved the location of the tweets to see the most affected location.

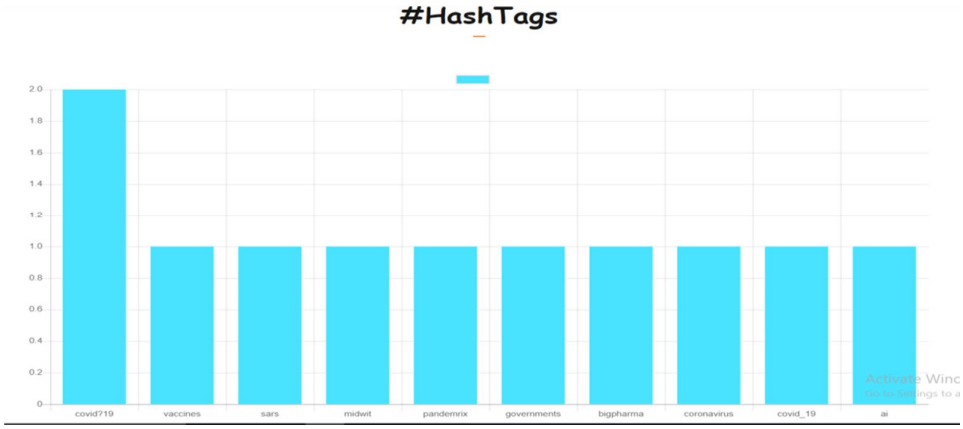


Figure 5: Hashtags of Swine Flu

Figure 6: Locations of Swine Flu

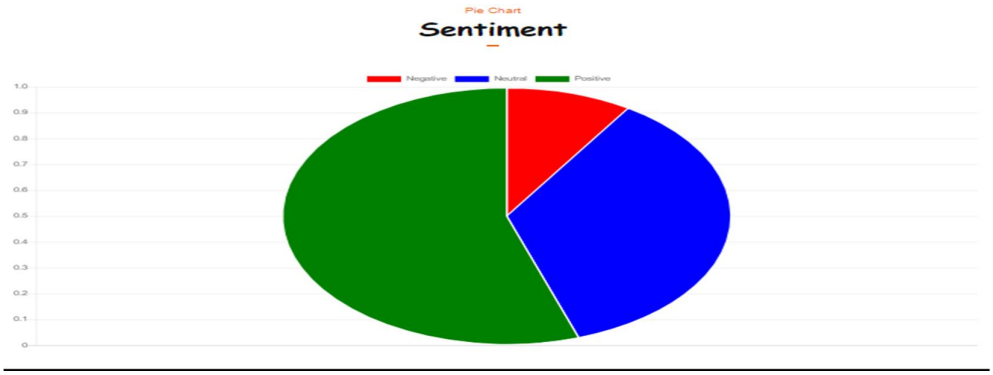


Figure 7: Pie Chart of Sentiments

For the above example , Sentiments are calculated depending upon the number of tweets received for entered disease name using TextBlob. Values generated for Positive , Neutral and Negative are depicted using Pie Chart.
WordCloud is used to hastags, locations and tweet text to give most frequently used words in tweets

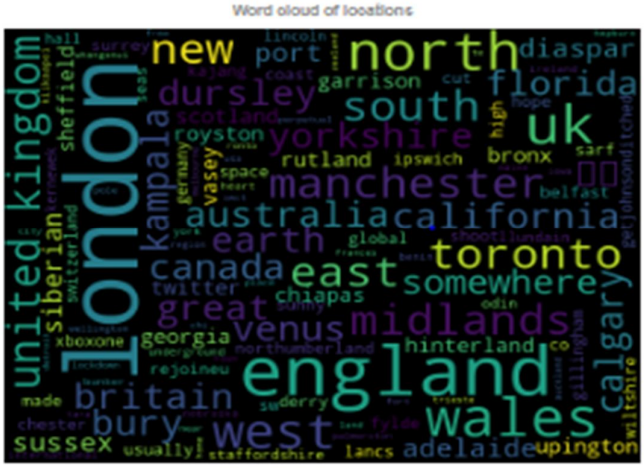


Figure 8: WordCloud of locations

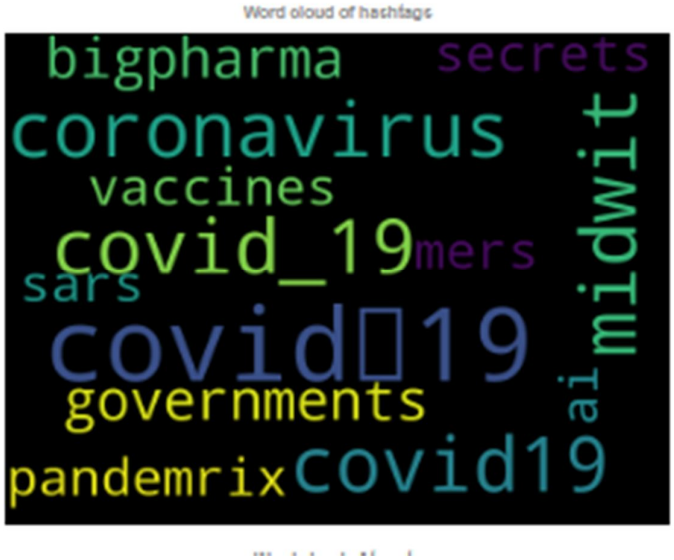


Figure 9: WordCloud of hashtags

A. Statistical Analysis

ANOVA tests for a difference overall between all groups. One way ANOVA is one factor ANOVA used to test for a statistically significant difference between 3or more groups. It is parametric test which make assumptions about the parameters of the populations distribution from which the sample is drawn. Here the groups are nothing but the number of days and one categorical variable is disease as one is number of variable in the analysis.

A hypothesis test is rule that specifies whether to accept or reject the claim about a population depending on the evidence provided by sample of data. Prediction of disease can be made upon the tweets as now days people share their immediate reaction on social network to the things specially by tweeting or following the health campaign.

Null Hypothesis for the above claim we made is rejected which includes that there are difference between the groups which are analyzed. Values are varying hence helps to get an idea about spread of particular disease.

Test Report

#	Title	Value
1	Hypothesis test	Null Hypothesis Rejected
2	Sum of squares within groups	930335.3333
3	Total sum of squares within groups	934896.0000
4	Sum of squares between groups	4560.6667
5	F score	0.0147
6	Density between location and sentiments	0.0863
7	Density between Location and date of creation of tweets	0.0863

Table 10: Test report for ANOVA

B. Network Density

The following figure indicates locations and date of creation of tweets. This is the result where we fetch the data on 17th June. As the selction is for previous day hence we can see the nodes with numbers 16,14 and 15 and they are connected with the different places from tweets are generated regarding the selected disease in this case it is swine flu. All the locations are taken to generate the map and mapped with the Date of Creation.

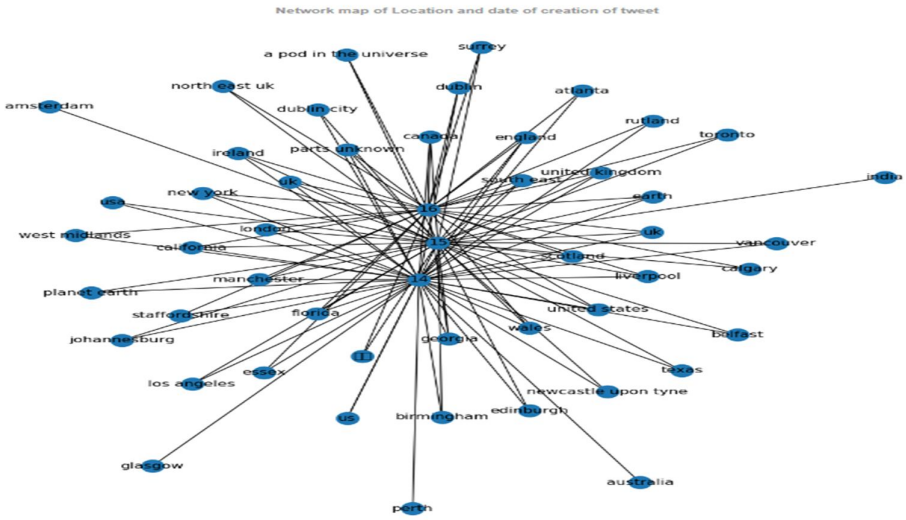


Figure 11: Network Map of Locations and Date of creation of tweets

VII. CONCLUSION AND FUTURE WORK

The result depicted are clearly states the relationship between sentiment of tweets and location which represents the different affected places by the disease. Test report represents the accuracy of the data used in the system. This states that the data used and processed is satisfying all the conditions and generating the values according to rules stated.

The future scope is to create alerts for the locations which helps the social bodies to take necessary actions immediate basis. That results in saving time and prevent the losses. The future development may include separate results regarding the current coronavirus disease and also researcher get basic framework to advances the result and present the more precise finding which helps in Health care system and Medical Expert System..

REFERENCES

[1] Xiang Ji, Soon Ae Chun, James Geller "Monitoring Public Health Concerns Using Twitter Sentiment Classifications" *IEEE International Conference on Healthcare Informatics 2013*

[2] Purva Grover, Arpan Kumar Kar,Gareth Davies " Technology enabled Health" – Insights from twitter analytics with a socio-technical perspective *ELSEVIER/International Journal of Information Management* Volume 43, December 2018, Pages 85-97

[3] Amir Karami, Alicia A. Dahl, Gabrielle Turner-McGrievy, Hadi H K Kharrazi, George Shaw "Characterizing diabetes, diet, exercise, and obesity comments on Twitter" ELSEVIER/*International Journal of Information Management 38(1):1-6 · February 2018*

[4] Hideo Hirose, Liangliang Wang "Prediction of Infectious Disease Spread using Twitter: A Case of Influenza" *Fifth International Symposium on Parallel Architectures, Algorithms and Programming* 2012

[5] Kruti Nargund ; S. Natarajan " Public health allergy surveillance using micro-blogs" *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* 2016

[6] Kaur, Chhinder & Sharma, Anand. (2020)."Twitter Sentiment Analysis on Coronavirus using Textblob" https://www.researchgate.net/publication/339998775_Twitter_Sentiment_Analysis_on_Coronavirus_using_Textblob

[7] F. Heimerl, S. Lohmann, S. Lange and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," IEEE 2014 47th Hawaii International Conference on System Sciences, 2014, pp. 1833-1842, doi: 10.1109/HICSS.2014.231.

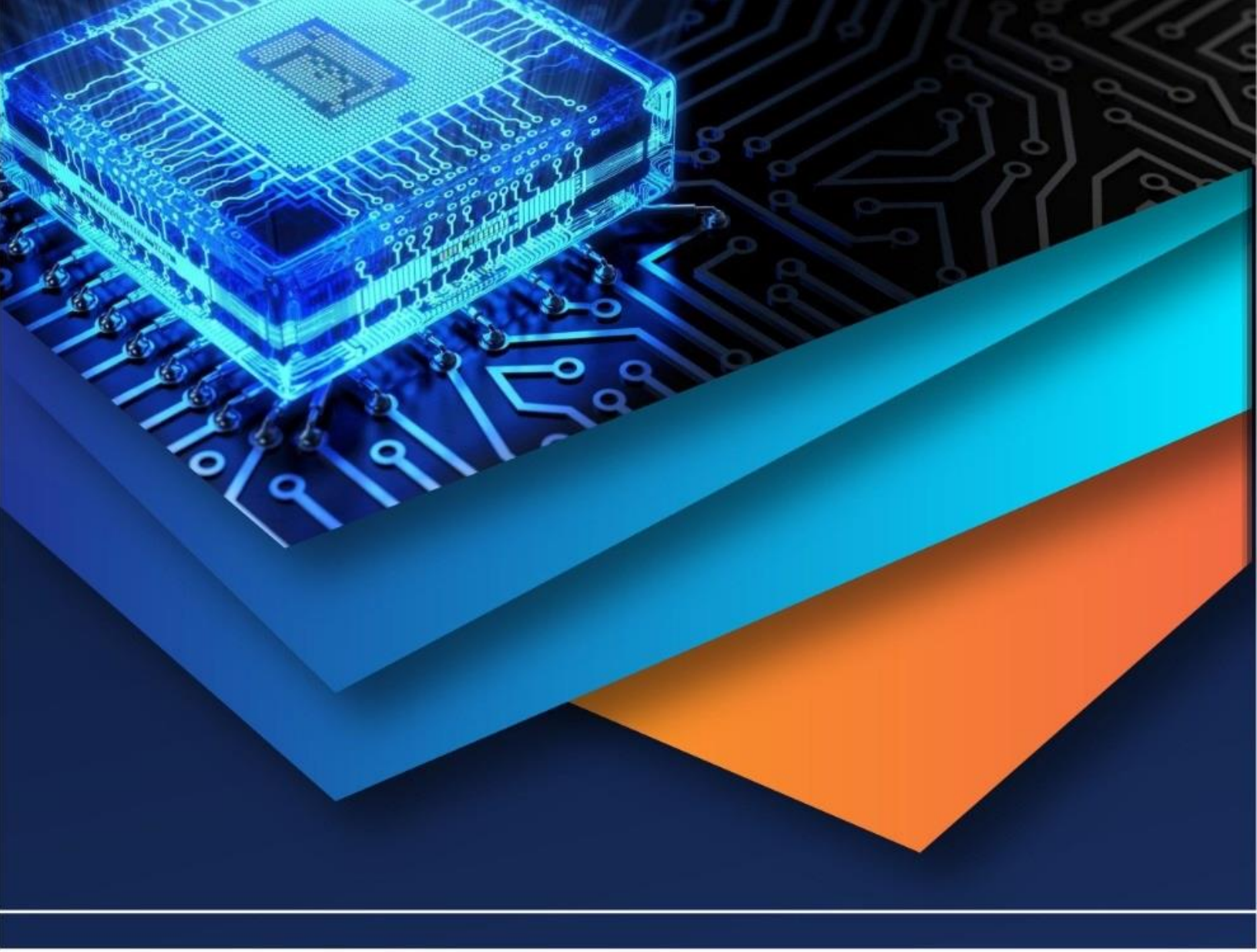
[8] Manar Alassaf, Ali Mustafa Qamar,"Improving Sentiment Analysis of Arabic Tweets by One-way ANOVA",Journal of King Saud University - Computer and Information Sciences,2020.

[9] P. Gupta, S. Kumar, R. R. Suman and V. Kumar, "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2020.3042446.

[10] I. N. Dewi, R. Nurcahyo and Farizal, "Word Cloud Result of Mobile Payment User Review in Indonesia," 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA), 2020, pp. 989-992, doi: 10.1109/ICIEA49774.2020.9102048.

[11] Giachanou, Anastasia & Crestani, Fabio. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys. 49. 1-41. 10.1145/2938640.

[12] L. Wang, J. Niu and S. Yu, "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 10, pp. 2026-2039, 1 Oct. 2020, doi: 10.1109/TKDE.2019.2913641.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY