Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

Detection of Social and Newsworthy events using Tweet Analysis

G. THILAGAVATHI , G. PRIYADHARSHINI, AKSHATHA M, BOOPIKA A M, SWETHA S V,
Coimbatore Institute of Technology,
Coimbatore, Tamil Nadu, India.
thilagavathi.g@cit.edu.in , priyadharshini.g@cit.edu.in

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544726

Abstract* — Social media play a vital role in this information era. Twitter is one of the important microblogging platform where people can share information known to them. Often these tweets are about local events. News agencies report on local events, but the time taken for an agency to analyse, investigate and report on the event can be substantial. Twitter users share their views and information about a particular event by posting tweets. These tweets can be used to identify whether the event occurred or not. Event detection from twitter data has gained importance nowadays. Our proposed system analyses tweets from a given geographical region to determine if an event occurred. The system then report the most descriptive tweet associated with an event occurred in that particular region. By the proposed system, it would be a quick way to alert people about an event occurring in their locality. In this, we split data into clusters based on location, identifies the tweet which exceeds the threshold, and then group the tweets based on similarity. The clustering models DBSCAN and HDBSCAN are employed to eliminate noise from the data and cluster similar tweets. Our system converts each tweet into a vector and normalise using TF-IDF technique. Finally, tweets which are similar on the same event will be analysed and collected. People can be notified of local events occurring before news outlets can report them when it is implemented in real time. The application varies on the type of event detected using our system. The News stations can also be intimated about the event so that they can explore further.

***Keywords* — Event Detection; Social Media; Twitter; Microblog; Clustering.**

I. INTRODUCTION

Microblogging, as a form of social media, is fast emerging in this decade. One of the best examples for this is Twitter which allows 280-character limit for a tweet. It is a powerful way of communicating with the surrounding people to provide updates of life. Twitter serves as an online platform that exists as a specific form of blogging.

In the traditional blog, the contents are large and are less frequent. But in micro blog , the contents are smaller where people can share small messages, videos, pictures. All these kinds of small messages are called as micro posts. This is the main reason for its fame. Moreover, Data mining tools are developing at a faster rate because of the availability of huge volume of data [14]. These unexploited and unstructured data is converted in a way that is useful for analysis using data mining techniques.

Twitter provides real-time information about worldwide events and generate huge amount of different data which can be used to find out the happenings around the world. In this information era, interest towards analysing these data are growing at a faster rate. Researchers are curious to explore the social media data [15] to bring out some better solution to find the happenings and events in and around the world. One of the important characteristic of microblogging is its real time nature. Blog users usually post or update their blog once in a week or once in several days accordingly whereas Twitter users post many times in a particular day itself. This leads to huge volume of data in Twitter.

By posting tweets, users can know what other users think about a certain issue or an event and also agree with their opinion by retweeting. Some important instances are earthquake in Haiti in which many pictures were shared in Twitter and an airplane crash in New York were reported first in Twitter and Tumblr. These type of continuous update in twitter results in numerous reports and data related to the events. They include social events such as gatherings, birthdays, sports, weather reports. Disastrous events such as hurricanes, storms, bursting of fire, traffic jams, heavy rainfall, landslides and earthquakes are also included. Basically, Twitter is used for various real-time notification such as that is necessary for help during a fire emergency or live traffic updates.

To handle these challenges, this paper proposes a novel method to use the twitter data effectively to detect social and newsworthy events. The tweets are collected from same area which contain longitude and latitude information also. The

next step is to divide the data into clusters by location. Clean the text body of each tweet by removing URLs, retweets, and hashtags. Then we stem and tokenize each tweet. Next, we convert each tweet into a vector and normalize by document word frequency (TF-IDF). Finally, DBSCAN clustering model is used to cluster similar tweets.
Then determine which tweet provides useful information about an event. To achieve this, we tokenize (split into a list of words) each tweet and find content words from each tweet. At last, we find the tweet that gives the details of an event using TF-IDF. TF-IDF is an effective normalization technique [16].

In the rest of this paper, we first present a brief survey of related work. Next, we give a concise description of clustering technique, anomaly detection and finding similar tweets. Then we also discuss the techniques to find the summary tweet that describes the event. Finally, we conclude with directions for future work.

II. RELATED WORK

Detecting local events in real-time from the Twitter data has gained more attention in recent years from researchers in and around the world. Event detection has gained more interest in the recent years and several researches emerged [10]. Event detection system [1] that involves specialized inverted indices and an incremental clustering approach to detect both major and minor events in real-time from the Twitter data has been suggested. The main objective was to detect important events through low computational cost. Researchers have also attempted topic detection using Twitter [9]. An extensive parameter sensitivity analysis has been done to fine-tune the parameters used in Twitter News+ to achieve the best performance. In a recent work on twitter data [2] , a web application called BlockShame has been designed and deployed for on-the-fly muting/blocking of shamers attacking a victim on the Twitter. The authors [3] also present an expert and intelligent system that helps the entrepreneurs who are in need of selling their products in marketing. In order to make popularity in their products they will tweet their product to reach out people. It also allows to get feedback about the product to reach the real world using twitter media. One of the paper presented [4] the effect on public infrastructure and in human beings. It also depicts the size of earthquake, Mercalli scale and GPS coordinates are used when the magnitude of earthquake becomes different in different . It also produces precise information in case of emergency and reports are done by observing the graphical areas. In one of the research paper [5], they started analysing disruptive events like earthquakes, tsunami happens often. Here, they detected political views, urbanization and Flooding events are done using graph concept and relevant tweets are clustered using the technique K-medoids. For easy search hashtags and keywords are used N-gram technique in BN-gram are used upto six-grams. In a paper [6], natural disaster can destroy human life and the disasters may be heavy rainfalls, deforestation and it harms them in various ways like emotional and physical level. This paper protect them from more dangerous by post disaster policies. Dynamic Query Expansion technique is used to track the level of disaster happening. The authors in one paper depict[7] to get more tweets in Arabic language because Arabic content in web are not upto the level as other languages do. Latent method can be used for distributing words in a document and semantic analysis method also help into it. It also helps to reach more Arabic content and information to reach people around the world.
One of the literature survey paper [8] extends the structural literature, where tweets can be identified and addressed and also it can be updated by making relevant solution. Clustering technique has gained importance even in medicine field [13]. Also many supervised learning techniques like Support Vector Machine is involved in tweet analysis to bring out better outcomes and insights from unstructured data [12]. TwIt is revealed in the literature survey that has huge volume of data and important challenges are faced by researchers. It brings a clear view of how to use the social media data wisely to detect events, natural calamities like earthquakes [11] and other potential problems. Statistical and sentiment analysis are used to make better solutions.

III. PROPOSED WORK

Our proposed systemaims in detecting both social and newsworthy events that occur in a particular locality so that the news stations can be benefited in such a way that they are able to identify an event at a faster rate. There are mainly 3 stages in our systemand finally stemming and tokenisation is involved to select an appropriate tweet that describes an event better than others.

A. ARCHITECTURE OF THE PROPOSED SYSTEM

There are mainly 3 stages as mentioned above and they are Stage 1 (Clustering module), Stage 2 (Event occurrence probability), Stage 3 (Integrating identical tweets), Determining the appropriate tweet which describes the event (Summary tweet).

1. STAGE 1 (CLUSTERING MODULE)

Dataset is simple and unstructured containing the tweets posted by various users in UK region. For simplicity, the system is tested with the tweets in London area alone. The twitter data is given as an input of the clustering module.

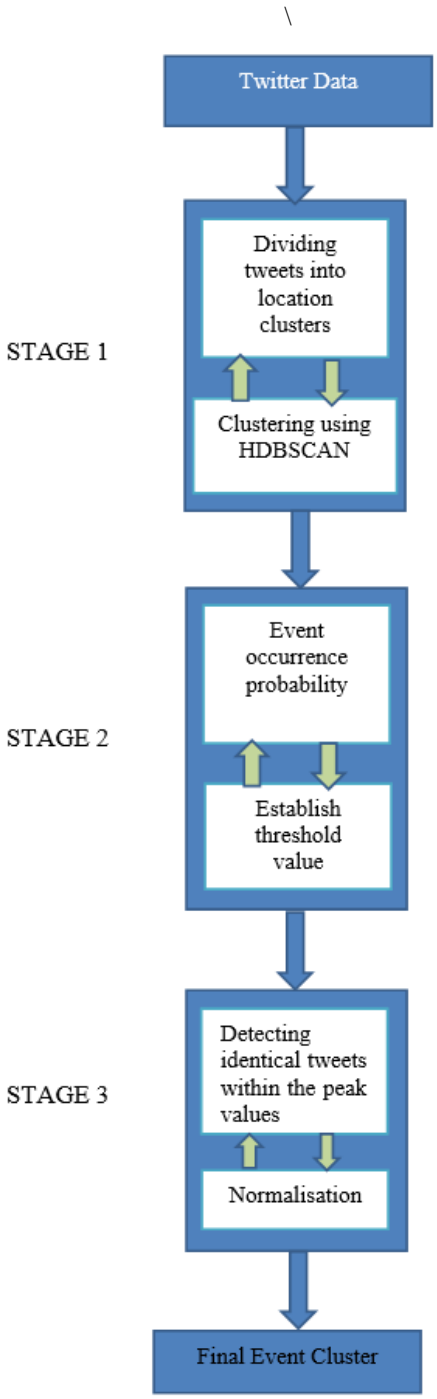

Fig.1 Architecture of the system

Twitter data often contain many irrelevant information which consumes lot of time while processing. It is necessary to process the tweets to avoid inaccuracy of event detection. In order to reduce the complexity in detecting events, pre-processing is done in three ways. Since twitter is a worldwide microblogging platform, there were tweets in many languages (i.e) many non-English tweets were present. Our system eliminated the other language tweets except English for better understanding. The algorithms used in clustering cannot analyse the fields in the twitter stream which are missing. To tackle this, we filled all the missing attributes of the twitter stream. The twitter data stream is reduced in and around London area based on latitude and longitude.

Following the pre-processing of twitter stream, clustering of data is done based on location. There are many clustering algorithm in the literature. Classification of clustering algorithm is a tedious process since the models may overlap each other. The fundamental clustering methods can be classified into partitioning methods, hierarchical methods, density-based methods and finally grid based methods. In our system, density-based methods are employed in order to cluster tweets based on latitude and longitude. Hierarchical clustering is a technique which is used to create subject hierarchy. Hierarchy clustering algorithm can be classified into divisive (top-down) and agglomerative(base-up).

Divisive methodology uses similar root cluster and work with all documents by iteratively splitting each groups or clusters into various small ones until the end rule is met for each cluster(end rule is a point). Partitional clustering algorithm are appropriate for large amount of data and information. These require low computational cost and linear quantity of documents. They are more preferable than agglomerative algorithms. Hierarchical clustering can be utilized from partitional clustering by rehashed use of partitional algorithm like bisecting K -means

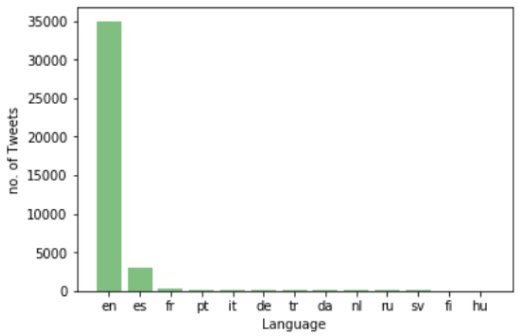

Fig.2 Tweets before pre-processing language

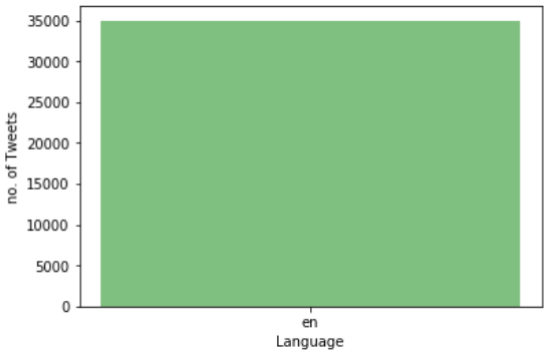

Fig.3 Tweets after pre-processing

To find clusters of arbitrary shape, clusters are modelled as dense regions in the data space, separated by sparse regions. In this density-based clustering methods, clusters of non-spherical shapes can be identified. It includes DBSCAN and HDBSCAN.

A) Hierarchical Density Based Spatial Clustering of Application with Noise (HDBSCAN)

HDBSCAN is a density-based clustering model that helps in extracting most significant clusters. Our system clusters the location and groups the data points accordingly. HDBSCAN is merely an extension of DBSCAN clustering model. In all cases, Real data is usually containing outliers, corrupt data and also messy. The Prime aim of clustering algorithm is to combine or cluster the similar data points and eliminate noise. Our clustering model is robust against noise, and it performs DBSCAN over various epsilon values and finally the results are integrated to provide best stability over epsilon values. There are various distance metric which has to be set before clustering. Haversine distance metric is used to cluster points in sphere which involves latitude and longitude. The haversine formula is an accurate way of computing distance between two points in a sphere and is the reformulation of spherical law of cosines. The haversine function is

$$\text{Haversine}(\theta) = \sin^2(\theta/2)$$

In K-means clustering the expected number of cluster will be known prior. K-means is a parametric algorithm(K cluster centroids) which performs best when the clusters are spherical, equal dense, no noise, equally sized and the density is more in center. The primary use of these clustering algorithm is data exploration.

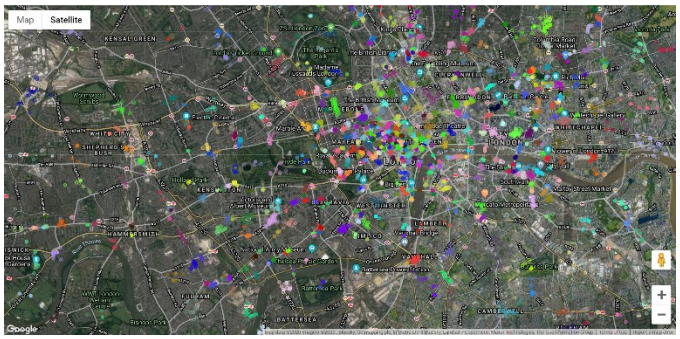

Fig.4 Satellite view of clusters

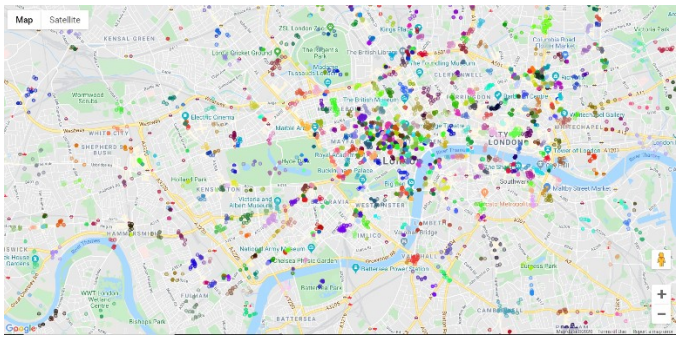

Fig.5 Map view of clusters

2. STAGE-2 (FIXING THERSHOLD TO FIND EVENT)

After clustering the data, our system aims to find the tweets which gives the probability of an event happening in the locality. In this stage, we divide the tweets by a timestamp of one hour to determine the tweets which gives the highest probability of an event. Mean and standard deviations are used to compute normal distribution of tweets per hour. Normal distribution also known as Gauss distribution or bell curve is a continuous distribution having the following algebraic expression for the probability density,

By setting threshold the most probable events can be detected using flagging, by using the normal distribution we can differentiate the normal tweet hours by event occurrence peaks. Certain clusters are identified with peak which exceeds the threshold value, and it corresponds to an event.

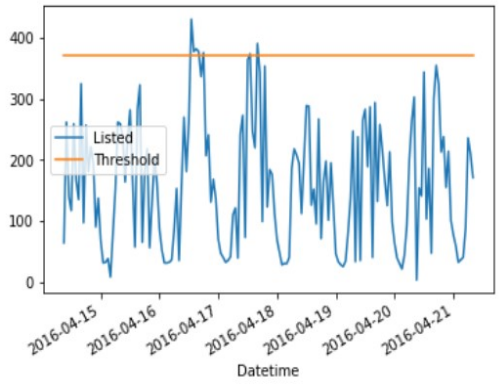

Fig.6 Threshold vs Hourly tweet occurence

3.STAGE-3 (DETECTING IDENTICAL TWEETS)

The peak values above the threshold level is identified by stage-2 statistical analysis. Then the system detects whether the peak is caused by an event, or any other irrelevant tweets posted by the user in that particular time period. These irrelevant tweets are considered as noise because it is not considered as an event since our system focuses only on the tweets which signifies an event occurrence. To eliminate noise, segregation of unnecessary parts like emotions, URL, hashtags and retweets are done using data cleaning. Then the proposed system performs stemming and tokenization technique. The popular technique TF-IDF is employed to do the above. TF-IDF stands for *Term Frequency - inverse document frequency* and is a weight-based methodology used mainly in text mining. It is used to extract information from data by measuring the weight of the document which gives the importance of a word from the document. The two main components of this technique are: Term frequency and Inverse document frequency.

A) TF: Term Frequency

It is a measure of the frequency of a term present in the tweet. A term occurs more frequently in a document which is larger in size than that of smaller one. Thus the TF measure is calculated using length of the document.

$$TF = \frac{\text{Number of times(frequency) a term appears in a document}}{\text{Total number of terms in the document}}$$

After obtaining the TF weight , the system checks if each of the term is found in all the document and also count of the total number of document.

B) IDF: Inverse Document Frequency

Importance of a term is measured using IDF. Equal importance is given to all the terms. The term which appears more frequently are given less importance and the one which appears rare are given more importance.

$$IDF = \frac{log_e(\text{Total number of documents})}{\text{Number of documents with the term}}$$

The higher occurrence of a word in documents yields higher term frequency(TF) and less occurrence of a word in document will give higher importance(IDF). TF-IDF is the product of TF and IDF.

$$TF\text{-}IDF = TF * IDF$$

HDBSCAN is used to cluster similar tweets and only few tweets correspond to the cluster. If there are any clusters inside clusters, the largest among them is chosen and it indicates that sufficient tweets about a particular event is obtained.

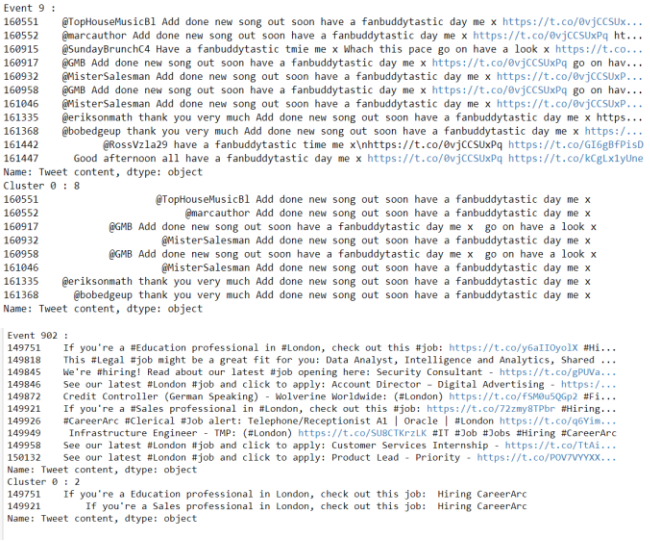

Fig. 7 A list of Similar tweets

.

The functioning of all the modules can be well understood by portraying the Pseudo code as follows.

- *Input(T)* // T-Twitter data.
- *Output()* => A group of similar tweets representing an event.
- *Step 1:* Pre-process the data twitter data by filling in the missing value and removing non English tweets.
- *Step 2:* Clusters the tweet based on location.
- *Step 3:* Convert to Date Time format.
- *Step 4:* Iterate through clusters and find hourly count of the tweets.
- *Step 5:* Calculate threshold to flag the hours with more tweets.

- *Step 6:* Find similar tweets from the flagged tweets.
- *Step 7:* Clean the tweets for improving efficiency.
- *Step 8:* Cluster of similar tweets that represent the event is detected.

IV. CONCLUSION

The approach of this system aims in detecting social and newsworthy events using twitter. The proposed system clusters the tweets based on location to identify events occurring in a locality. Finally, the most summarized tweet is obtained which is the representative. In future, the system can be implemented for more data on real time so that the user can be intimated about the happenings around the area at a faster rate.

V. FUTURE WORK

In future, the system can be implemented for more data on real time so that the user can be intimated about the happenings around the area at a faster rate. The system can also be improved by selecting a summary tweet from the set of similar tweets obtained. This summary tweet can be given to the news station earlier which reduces their time on investigation. Finally, defining events in real time can have meaningful business impact.

References

[1] Mahmud Hasan,Mehmet A. Orgun,Rolf Schwitter, “Real-time detection from the twitter data stream using the Twitternews+ framwork” , Information Processing and Management ,vol.3 ,2018.

[2] Rajesh basak,Shamik Sural,IEEE,Niloy Ganagaly and Soumya K.Ghosh, “Online public shamming on Twitter :Detection, Analysis and Mitigation” , IEEE Transactions on computational social systems ,pp 2329-924x , 2019.

[3] Lim, Sunghoon, and Conrad S. Tucker. "Mining Twitter data for causal links between tweets and real-world outcomes." Expert Systems with Applications*: X* 3 (2019): 100007.

[4] Mendoza, M., Poblete, B., & Valderrama, I. (2019). Nowcasting earthquake damages with Twitter. EPJ Data Science, *8*(1), 3.

[5] Winarko, Edi, and Reza Pulungan. "Trending topics detection of Indonesian tweets using BN-grams and Doc-p." Journal of King Saud University-Computer and Information Sciences 31.2 (2019): 266-274.

[6] Mehdi Jamali,Ali Nejat , Souparno Ghosh,Fang Jin , Guofeng Cao, “Social media data and post diataster recovery”, International Journal of Information Management, vol.44,2018.

[7] Saptarasi Gowsami,Sanjay Chakraborty, Sanhita Ghosh ,Amlan Chakrabarti , “A review on application of data mining tecniques to combat natural disasters” , Science direct, Science Direct - Ain Shams Engineering Journal, vol.9,pp 365-378,2018.

[8] Rafea, Ahmed, and Nada A. GabAllah. "Topic Detection Approaches in Identifying Topics and Events from Arabic Corpora." Procedia computer science 142 (2018): 270-277.

[9] Stafan Stieglitz, Milad Miebabaie, Bjorn Ross, Chirstoph Neuberger , “Social media analystics – challenges in topic discovery,data collection and data preparation ” Elsevier international journal of IM ,vol.66, pp 156-168,2017.

[10] Doulamis, Nikolaos D., Anastasios D. Doulamis, Panagiotis Kokkinos, and Emmanouel Manos Varvarigos. "Event detection in twitter microblogging." IEEE transactions on cybernetics 46, no. 12 (2015): 2810-2824.

[11] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Tweet analysis for real-time event detection and earthquake reporting system development." IEEE Transactions on Knowledge and Data Engineering 25.4 (2012): 919-931.

[12] Zgheib, Wassim A., and Aziz M. Barbar. "A Study using Support Vector Machines to Classify the Sentiments of Tweets." *International Journal of Computer Applications* 975 (2017): 8887.

[13] Sandhiya, R., and M. Sundarambal. "Clustering of biomedical documents using ontology-based TF-IGM enriched semantic smoothing model for telemedicine applications." *Cluster Computing* 22.2 (2019): 3213-3230.

[14] Padhy, Neelamadhab, Dr Mishra, and RasmitaPanigrahi. “The survey of data mining applications and feature scope.”arXiv preprint arXiv:1211.5723 (2012).

[15] M. Sarah, C. Abdur, H. Gregor, L. Ben, and M. Roger, “Twitterand the Micro-Messaging Revolution,” technical report, O’ReillyRadar, 2008.

[16] Kim, Sang-Woon, and Joon-Min Gil. “Research paper classification systems based on TF-IDF and LDA schemes.” Human-centric Computing and Information Sciences 9, no. 1 (2019): 30.