Quantifying Information Dissemination Rate during Crisis and Location Detection Using Online Social Streams

Bhuvaneswari A (bhuvana.cse14@gmail.com)
VIT University - Chennai Campus https://orcid.org/0000-0001-6651-2031

Research Article

Keywords: Online Social Streams, Twitter, Event Detection, Shannon Entropy, Information Dissemination, Geo-tags, Geographic Information Systems

Posted Date: June 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-528819/v1

Quantifying Information Dissemination Rate during Crisis and Location Detection Using Online Social Streams

Bhuvaneswari Anbalagan*

*Assistant Professor, School of Computing Science Engineering,

Vellore Institute of Technology, VIT , Chennai, Tamil Nadu, India.

***Email ID:** bhuvana.cse14@gmail.com , **Mobile :** +91-9894007258

Abstract: The widespread practice of Online Social Networking leads to the diffusion of trending information and exchanging various opinions with socially connected people online. Social media steams data extracted from Social Networks has become a vital communication tool and also turn up as an eventual informative platform to catch real human voices at the time of emergency events like disaster. An effective underlying quantification model is proposed in this paper which uses change point detection algorithm to detect events based on the relative streaming tweet density - ratio respectively. A morphological time-series analysis is carried out determine the dissemination of information about crisis events using Information Entropy. Further, the Event - Link ratio (ELR) is estimated to obtain meaningful patterns in events been identified. This paper focus to empirically quantify the information dissemination of the events based on user's tweeting activities. The proposed quantification method is compared with state-of-art techniques in terms of event detection rate, the entropy of information spread. It is found that the accuracy of the proposed method is up to 94% with event detection after 75 seconds. K-Center Clustering (KCC) is used which results in the location detection accuracy of 85%.

Keywords: Online Social Streams; Twitter; Event Detection; Shannon Entropy; Information Dissemination; Geo-tags; Geographic Information Systems;

1. Introduction

In latest era of online technology, online social networking data is encapsulated with a variety of huge informative blocks about various real-world events and publics' opinion, which is insightful during disasters to provide public safety. Recently, social media has been successfully utilized to be a major replacement to measure the impacts of disaster events in online stream real time. It provides many user-friendly services with user-generated content towards the overwhelming amount of information on hand. Moreover, the events categorize across diverse temporal and dynamic spatial scales with respect to geographical information science. In specific, social networking sites namely Twitter, Weibo and Facebook, used as a vital social sensor of disasters like an earthquake, flood, landslides, etc to provide immediate response and recovery [1]. In addition, geographically located social streaming data is accepted to be a trustworthy objective for sensing disasters via online and examining reactive action after

the mass emergency events [2]. On the other hand, Online Social media mining is widely used in typical disaster scenarios; one of the most vital aspects to identify with social responses is to measure people's opinion for improved disaster support model. In particular, the users associate their impulsive reaction during disasters in terms of Social Media Timeline (SMT) , news feeds where the user- post (PT), tweet (TW), reply (RE), retweet (RT), share (S), and mentions (MT) along with images and videos to yield the attention of others to broadcast the sources of disaster information. However, the absolute volume of social information streams produces huge commotion that ought to be sifted. By recognizing patterns in the surge of messages and data stream a change point can be identified in the typical progression of streaming tweets [3]. Disaster events can be perceived as spikes in action, simultaneously as importance can be interpreted over the span of changes in content.

It become very difficult to acquire inside and out hints of data spread during events such as people's real social connections during disasters, their implicit behavioral profiles, and situation roles in social related activities. The extent of information spreading process depends on the core heterogeneous Social Networks and basic behavioral profiles of the individual user [4]. The user behavior activity is measured as “temporal series” with respect to data. The “popular” Twitter users are focused in this paper, who retweet activities are considered [5]. A new approach is proposed that incorporates the following information to understand user behavior during critical times. Tweets and retweets of the targeted users who eventually follow temporal patterns are monitored. The web-based online social media information contains more data about continuous occasions, yet additionally delicate issues that stay undetected. It goes to be exceptionally muddled for this situation to recognize the spreading [6] data of most importance. In this paper, we basically center around the Twitter information streams and think about the accompanying target. The primary objective of this paper is to plan a novel quantifying model to identify occasions identified with disaster that (i) To identify event detection rate which is observed at multidimensional scales, specifically, events that take place in diverse location and temporal timelines by computing Event Link Ratio(ELR), (ii) are influential beside the uncertain and unfiltered insights extant in the data between dynamic time-interval using change-point-detection algorithm, (iii)an information theoretic approach to classifying Twitter users entropy based on homogeneity in tweeting activities including user sentiment polarity, and (iv)providing a novelty evaluation method to identify event involved in intervals using the Z-score and Local-to-global Ratio.

This paper is organized as follows. First, the related state-of-art work of various quantification models related to disaster events is discussed in Section 2. The proposed work is discussed in Section 3. The experimental results and research findings are discussed in Section 4. The event detection rate results are given in Section 5. The location detection is discussed in Section 6. The conclusion and future work is deliberated in Section 7.

2. Related Work

Several approaches to quantify social media data to form support model for predictive analysis have been proposed over time. The tweeting dynamic activities on Twitter are made to understand the structural properties of information flow during disasters. The insight study is done on Twitter data during the Tohoku earthquake [7] in 2011. An automatic technique to find relevant corpus for tracking

disasters was investigated thoroughly as an early warning system. The paper identified how quickly Japanese people's concern returned to a stable level after the disasters. Twitter user's tweeting, answering, and retweeting exercises, distinguished a technique to separate Twitter clients [8] based on their exercises. The paper broke down all likelihood to programmed client grouping and sifting dependent on requirements. The test results with information from Twitter when the Japan Earthquake, their proposed strategy could characterize clients relying upon their characteristics with a precision of the examinations and with high exactness contrasting and the old techniques.

Tweets having URLs are analyzed by their combined `retweeting' dynamics [9]. The paper achieved a separation of different activities using two features to categorize content based on the user response it generates. Among them, the spreading processes of specific pieces of information, including studies on the corpus sequence and viral diffusion behaviors, are most related to our work. It is distinguished, numerous classes of retweeting action on Twitter: bots action, newsworthy data spread, publicizing and advancement, political, crusades, and promotions advertisement.

Sentiment characterization of user posts gathered from Twitter during the Hurricane Sandy[10]. The paper pictured the estimations on a geological guide which is fixated on the typhoon calamity occasion. It center around removed data and the handiness of associated emergency guides to keep up the crisis reaction. A strategy implied for influenced populace's reaction to a calamity can be estimated over the span of a feeling examination [11-13] and afterward planned corresponding to the fiasco in existence. Spatiotemporal factual examination is done on boisterous data web-based media information. Besides, different conventional ways to deal with distinguish occasions for fixed transient and spatial terminations, while truly occasions of various scales regularly happen simultaneously was proposed. A multi-scale occasion recognition [14] to process an information similitude chart at suitable scales and distinguish occasions of various scales by a novel diagram based grouping technique.

Quantitative research perceptions and strategies for evaluation of mass streams of data lead to ideal scaling for occasions dissected during catastrophes in Japan. The first and second Hayashi techniques [15]are applied to the situation where an outer standard is available and are utilized to anticipate the impacts of variables considered. The Hayashi work used to develop a spatial design to acquire the common relationship of the information for comparable clients [16] and occasions. The word order characteristics quantified used Hayashi's quantification method type III (HQM). It was examined using the first and second component of HQM and MDS results. The natural language processing using WordNet [17] lexicons for Twitter datasets are studied.

The political leaning inference [18] was framed to maximize tweet-retweet covenant with average mean error and user match technique with regularization duration. The convex optimization problem is solved by for Romney and Obama-bashing tweets circulated by the networking sites online during the process of election events. The three-class classification [19,20] problems is modified as two binary classification problems using the Senti-Strength algorithm [21]. In their experiments, the polarity of tweets are classified as positive, negative and neutral using machine-learning classifiers trained on bi-

grams , tri-grams and lexicons based features, and their combination [22]. An entropy-based metric is reported to represent sentiment limited to social media data. Various events are detected and visualized using Twitter micro blogs during certain natural hazards events [23], Crisis Mapping [24], Emergency situation awareness from twitter [25], Predict Disasters on Twitter Data [26], entropy based event detection [27], Real-time event detection for online behavioral analysis [28],[29], Twitter-based traffic event detection [30], Emerging topic detection in twitter stream [31] was proposed to provide situational awareness through social media. Various papers [32]-[34] discussed about quantifying event information spread using online social networks. Eagle & Pentland introduced a system for sensing complex social systems with data collected from 100 mobile phones. Bluetooth-enabled mobile telephones were used to measure information disseminated from different context through Shannon entropy. Moreover, the social patterns were recognized in daily user activity, infer relationships and identify socially significant locations [35]. Shetty & Adibi proposed entropy models to study information flow in an organization on keyword graphs are relevant or not. The results review with two different experiments which are based on entropy models [36]. The Entropy model identifies the most interesting and important nodes in a keyword graph which is partially adopted in the proposed work.

3. Proposed Work

3.1. Identifying Change Point Detection

Event burst distribution can be identified by observing the Twitter continuously using RuLSIF - change point detection. Moreover it proceeds with the method of setting up an instantaneous alert for immediate attention as soon as a real-time unexpected event is detected. It can be measured based on density of similar tweets that bursts out in a specified time interval t and **Δ**t. In particular, the overall frequency of words w_i tweeted in time-interval is directly proportional to the burst of an event E_j for the interval. The density of tweet exceeds certain peak period of threshold and attain a saturation point can be identified as change-point. The mathematical formulation for relative tweet density-ratio estimator is mentioned as $\hat{f}(Y)$, the α – relative Pearson(PE-divergence) can be approximated using Eq.(1).

$$\widehat{PE}_{\alpha} = -\frac{\alpha}{2n}\sum_{i=1}^{n}\hat{f}(Y_i)^2 - \frac{1-\alpha}{2n}\sum_{j=1}^{n}\hat{f}(Y'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\hat{f}(Y_i) - \frac{1}{2} \quad (1)$$

For our experiments, RuLSIF- based change point detection algorithm is used to directly estimate the relative tweet density-ratio where $0 \leq \propto < 1$ is a parameter.

3.1.1.Tweet Analyzing Parameters

The Document Incidence(DI), is the number of Tweets in which the word appears. The Global Frequency Rate(GFR), is the total number of times the word appears within the tweet dataset between First Interval(FI) and Last Interval(LI). Burst Ratio(BR) It can be obtained by calculating Z-Score, in which how frequently the word appears in the chosen interval relative to its average frequency across all intervals. A high z-score means that the word is unusually more frequent and therefore likely to be a good descriptive word for being novel or else rare topics discussed within the interval Novelty measures the percentage value that represents the degree to which tweets across surrounding intervals discuss a

novel topic. The novelty of 0% indicates that every term in the selected interval was the same as other intervals. Inversely, 100% indicate that every term in that interval was different than other intervals as defined by the burst term model selection. Homogenity is the percentage value that represents the degree to which tweets within that interval use the same keywords.

Burst Ratio(BR) It can be gotten by figuring Z-Score, in which how habitually the word shows up in the picked stretch comparative with its normal recurrence across all spans. A higher value of Z-Score implies that the word is uncommonly more successive and subsequently liable to be a decent engaging word for being novel or probably uncommon themes talked about inside the stretch Novelty estimates the rate esteem that addresses how much tweets across encompassing spans examine a novel subject. The Document Incidence(DI), is the quantity of Tweets where the word observed in the post. The Global Frequency Rate(GFR), is the quantity or number of times the word observed inside the tweet dataset between First Interval(FI) and Last Interval(LI). The uniqueness or the novel rate of 0% shows that each term in the chose span was equivalent to different interval period. Subsequentely, 100% demonstrate that each term in that stretch was unique in relation to different spans as characterized by the burst term model choice. Homogenity is the overall percent of data that addresses how much tweets inside that span utilize similar watchwords. $H(t, \Delta t)$ below 30% indicates every tweet has distinct content in the given interval. $H(t, \Delta t)$ between 31% to 70% indicates every tweet has similar content in the given interval. $H(t, \Delta t)$ between 71% to 100% indicates heavy retweeting activity of similar content.

$$\mathrm{H}(\mathrm{t}, \Delta \mathrm{t}) = \begin{cases} 0\% - 30\% \text{ , tweet has distinct content} \\ 31\% - 70\% \text{ , tweet has similar in content} \\ 71\% - 100\% \text{ , high retweeting activity of similar content} \end{cases} \quad (2)$$

Event Link Ratio (ELR) is the ratio amid total numbers of tweets containing URLs linked to disaster and the aggregate quantity of tweets in the given interval. The ELR ranges between 0 and 1. The less value of ELR, URLs linked with event-of-interest are minimum and maximum for highly linked events with a value reaching value one. The tweets that spread false news in large range regarding an event is identified as the false panic rate, which is classified to be bots in our experiments. A non-trivial parameter in detecting the interconnected event during a disaster is Temporal Burst Ratio (TBT). It is the ratio between Novelty and Burstiness for the time interval(t,**Δ** t).

$$\mathrm{TBR}(\mathrm{t}, \Delta \mathrm{t}) = \frac{\text{Homogeneity}}{\text{Burstiness}} = \frac{\text{Tweets in time discuss same topic words}}{\text{Topic word suddenly becomes popular}} \quad (3)$$

The sentiment or polarity of the tweet can widely provide the polarity of the event. The tweets are categorized as SENT+ and SENT- takes value ranges and {+1 to +5} {-1 to -5} appropriately.

3.1.2.Probability Distribution of Tweets

With the reference of the parameters mentioned above, the probability of the Twitter event burst is identified to follow a binomial distribution during disasters using Eq.4. In order to calculate the probability of the aggregated value of tweets that hold the lexicons word w_k at time $T(w_j)$, can be denoted as $P(n_{j,k})$, as mentioned below:

$$P(n_{j,k}) = \binom{N}{n_{j,k}} {p_k}^{n_{j,k}} (1 - p_k)^{N - n_{j,k}} \ni W_i \quad (4)$$

where N is the tweets count in given period of time series evaluation. Although N_i is the number of tweets that vary in each time-interval t_i, and it can be re-scaled in all time-interval by uniformly normalizing the frequency of words responsible for the event burst. From the above distribution,$p_k$is the anticipated probability of the tweets that contain a word w_k in a randomly chosen time-interval. Hence the mean of the detected probability of word w_k among given intervals comprehending word w_k, which are defined as

$$p_k = \frac{1}{C} \sum_{i=0}^{C} P_0(n_{j,k}) \ni W_k \quad (5)$$

where C is the count of intervals comprising word w_k and $P_0(n_{j,k})$We determine whether a word w_k is bursty or not by comparing the actual probability of the word w_k take place in the interval $T(w_j)$ against p_k of the lexicon word w_k occurring in a random interval (t,**Δ**t). If the calculated value of$P_0(n_{j,k})$ is greater than $w_k(p_k)$, then it visibly shows that word w_k exhibits an anomalous behavior in given time $T(w_j)$. Furthermore, it is evident the word w_k as, a bursty word(tweet) in time $T(w_j)$.

3.2. Entropy Based Quantification- Mathematical Model

Twitter user activities can be converted into an information speculative technique by scheming entropy for time-interval. Additionally, the user's entropy for a particular URL and Mentions on tweets embedded with *images* and *videos* are measured in our experiments. This can be used to characterize similar tweets and users based on the characteristics of tweets they post. The retweeting activity is redefined as the Twitter Retweet(RT) and Reply(RE) which focus on tweets containing only URLs. With this concept as a reference, we compute entropy by selecting unique features with respect to topic words that uses modified Ghosh's method. Our aim is to analyze trace of an event $T(E_i)$ due to the spread of topic words W_i and quantify the expected amount of information disseminated in the time-interval (t,**Δ**t) by calculating entropy. The time-interval can be referred as "Novelty Evaluation Epoch" where the resemblance in word-occurrence amid the chosen time-period and middling value of *K* foregoing interval can be determined.

For given N traces how do we dynamically characterize and categorize the tweets is more significant in real-time. The procedure includes computing time-interval entropy, user entropy, hash-tag entropy, similar user entropy and sentiment score for a specific event with the topic word in trend. Let the set of all posts on a specific topic word be represented as P_{word}, the set of users who tweet that topic word can be represented as U_{word}. The set of posts of a user $u \in U$ with a topic word is represented as $P_{u,word} \in P_{U,word}$ and $n_{u,p}$ be the total number of posts by users. Let b is used to representing the type of post where b=1 indicate normal tweets, b is 2 specify retweet (RT) and b is 3 specify reply (RE). The time-interval of particular user u's posts p with topic word W from (t-1) to t this expressed as $\Delta t_{u,p,t}$ where $\Delta t_{u,p,0}$ assigned as zero. The process of tweets trace is estimated utilizing time period entropy on subject words. The recurrence of word (W) in given time-stretch $\Delta t_{u,p,t}$ is determined as follows.

$$N_{u,p,t} = \frac{\Delta t_{u,p,t}}{r} \ni W, m = 1,2, \dots, n_{u,p} \ni W_i \quad (6)$$

where m is the constraint to decide the unit of time-interval. The time-interval entropy on topic words can be derived using following equations.

$$H_{T_{u,p}} = -\sum_{k=0}^{m_{u,p}} p_{\Delta T}\left(\Delta N_{u,p,t}\right)\log_2(p_{\Delta T}(\Delta N_{u,p,t}))\ni\ W \tag{7}$$

In order to measure entropy, the extent of user distribution is used to determine user entropy on topic words. Let random variable D represents a distinct user in trace T_i with all possible values$\{d_1,d_2,\ldots d_{nD}\}$. Let the number of retweets from user U_i in the trace T_i. Then p_F represents the probability density function of D, such that $P_F(d_i)$ provides the probability of every retweeting activity taken by the online user f_i, then the frequency of user a∈ U, tweet which is retweet by b to user c ∈ U is expressed by the following equation.

$$H_{User} = -\sum_{c=1}^{n} p_F\ (n_{abc})\log_2(p_{\Delta FT}(n_{abc}))\ \ni\ W \tag{8}$$

The frequencies of occurrence of the particular hashtag mention by various users over a continuous set of time-interval by j, Ch(j), we can calculate normalized hashtag entropy on topic words as follows.

$$H(X) = -\frac{\sum_{j=0}^{X} P(X_j)\log_2(P(X_j))}{\log_2(|X|)}\ni\ W \tag{9}$$

where $$P(x_j) = \frac{Ch(j)+ 0.01}{\sum_{j=0}^{X} Ch(j)+ 0.01|X|} \tag{10}$$

The above equation ensures the probability calculations are normalized so that entropy is finite for hashtags. The value 0.01 is used to normalize the entropy for fuzzy crisp dataset values. In our analysis, the dataset results high-entropy on hashtags are considered to be a significant long-running phenomenon which appears with uniform frequency over time.

In addition *Similar User - Entropy* is calculated using the burst words involved in tweets that typically retweeted more than 100 times within the same group of users. It is important to measure the similarity between different tweets corresponding to the same event. In our baseline event detection process, we measure the similarity between every pair of tweets T_a and $T_b$as:

$$Sim(T_a, T_b) = \begin{cases} sim_{tf-idf}(T_a, T_b)\ , & \text{if } time(T_a, T_b) \le T_t \text{ and } dist(T_a, T_b) \le T_d \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $time(Ta, T_b)$ and $dist(T_a, T_b)$ are the time difference and locality difference between tweet pairs (T_a, T_b) respectively. The thresholds T_t and T_d represent the event location and enforce the constraints of tweets restricted to a spatio-temporal boundary. The function $sim_{tf-idf}(T_a, T_b)$represents the similarity in text of T_a and T_b where the cosine angle between the vector representations of any similar tweets using the term frequency inverse document frequency (tf–idf) weighting method.

3.4. Event Detection and Quantifying Models

The proposed work identifies the disaster-related event depending upon tweet lexicons and evaluates the re-tweeting activity dynamics during the disaster. The proposed work model is shown in Figure 1. The overall frequency of the tweet corpus is analyzed in a particular time-interval. The system is

pre-configured with disaster corpus word library for reference to detect burst words. The system is automated to filter tweets and change point can be detected using RuLSIF's Change-Point Detection technique. The Senti-Strength algorithm which is a lexicon-based sentiment analyzing package is applied to the tweets. The tweets are filtered separately based on their sentiment score threshold. The system calculates three vital metrics namely Temporal Burst Ratio, Homogeneity Index and Event Link Ratio of the tweets corpus to precisely detect the event and its location. The threshold of time-interval, Sentiment score, locality and hash tag is set to a non-zero value. When the density of tweet exceeds the temporal tweet threshold, the event is identified along with the geographical location. The sentiment pulse of the event is determined using the sentiment threshold in which the positive and negative sentiments are noted. The user tweet frequency is correlated with time interval and location to classify the users as human and bots. The similarity between the tweets is calculated and Markov clustering is applied to cluster various events.

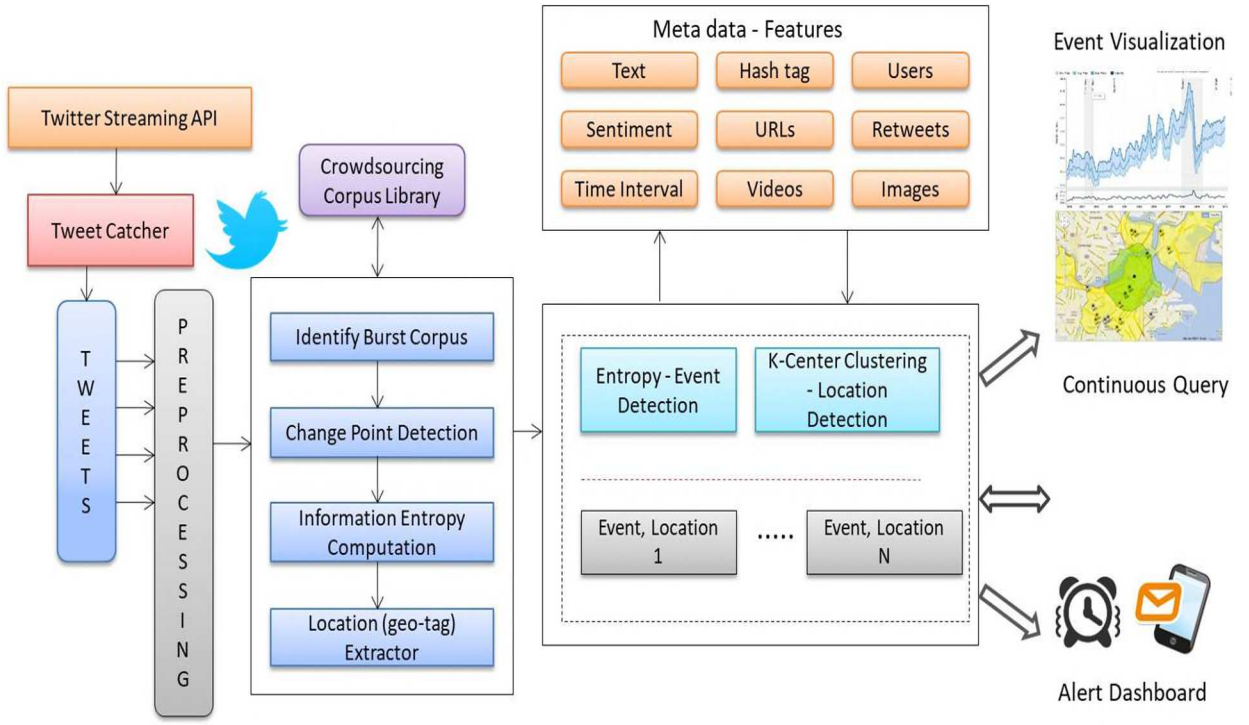

Figure 1.Proposed Event Detection Model

4. Implementation and Results

For implementation, the training dataset is collected which contains 20000 tweets captured during various disasters in the year 2014 and 2015 is utilized to pre-process initially. The overall statistical topic words diffused during most terrible disaster events in India. The tweets are gathered utilizing Twitter Streaming API to distinguish the corpus or word that blasts out to recognize a disaster event. In fact, their collection consists of tweets in native Indian languages. From the word index statistics shows the word is important in event clusters. The proposed system performs effective analysis for exploring and examining the spatial distribution of Twitter users. The time stamp joined to each message shows the posted season of the messages, demonstrating the time estimation of an event. The features namely

location, hashtag is first extracted from the tweets between a specified time interval. It is handled in MapReduce system for occasion identification utilizing the accompanying Mapper and Reducer method. The change point detection using RuLSIF is calculated with relative Pearson divergence α ranges value 0.7 to 0.9. It proves the change point is detected which inclusively shows the dynamics in social streams. The dataset collected with an unbalanced sample of 19,313 tweets. We use stratified sampling to get a balanced data-set of 10127 tweets. The total volume of tweets spiked out related to Assam flood and the proportion of tweets embedded with the URL are considered. The change point detection to identify flood in Assam, Twitter is monitored for disaster-related topic words. The tweets having sentiment score range lesser than tweet threshold($S < TW_s$) is filtered and considered for sentiment analysis processing. The data sets are analyzed and the experimental result shows the flood tweets(Assam) estimates 3% Novelty, 85% Homogeneity, Event Link Ratio of 0.81.

Table 1. Quantifying Metrics relevant to Assam Event Dataset

Event	**Quantifying Metrics**	
Assam Floods Dataset	Homogeneity	0.90
	Event Link Ratio (ELR)	0.85
	Temporal Burst Ratio (TBT)	0.66
	Burst Ratio(BR)	0.86
	Z-score	0.56
	Local-to-Global Ratio	0.79

The empirical values of sentiment show the tweets are highly negative sentiment during the disaster where people use most of the negating words. From the values, it shows 3% novelty, there exist an overall unique topic similarity across and between the intervals considered. The similarity of vocabulary in topic words is 85%, indicates twitter users tweet regarding flood using identical terms. The value 0.78 is Event Link Ratio of 81% shows more than 80% of the total volume of tweets is related to flood at Assam. The performance of experiments in event burst distribution detection was evaluated using two metrics: temporal burst ratio and false panic rate.

The experiments show, our burst detection mechanism achieves an overall burst detection rate (BDR) of 91% and a false panic rate (FPR) of 4.1%. The tweets are classified using Naïve Bayes classifier as two classes namely Class M (posts without multimedia content) and Class N (posts with multimedia content). The entropy values from Table 2 shows users retweet for more multimedia data rather than raw text tweets without multimedia data. The popular Twitter user page, shows high entropy whereas bot user page namely Bloombrg with very low entropy. The low-entropy hash tags are lightly specific and moreover related to short-term events.

Table 2. Evaluation of Entropy (Class **A** Vs Class **B**)

User Name	**T. Int. Ent.**		**RT- T. Int. Ent.**		**User Ent.**		**RT User Ent.**		**HashTag Ent.**		**HashTag RT Ent.**		**Similar User Ent.**		**Sentiment Score SENT +**		**SENT-**	
CLASS	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**	**M**	**N**

Sovinko	2.89	4.02	6.10	8.04	3.20	5.21	2.84	3.20	2.45	4.00	3.10	5.68	5.26	7.20	3	4	1	1
News7	2.45	4.25	1.04	3.56	2.74	3.84	3.74	5.36	0.89	1.25	0.98	2.01	0.00	1.36	2	3	1	2
CNN News	1.75	2.04	3.41	5.95	5.12	7.51	3.41	4.12	2.10	4.26	4.52	6.24	1.98	3.20	1	1	1	2
Bloombrg	0.04	0.12	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	2	3

For better examination on how the likelihood of a hashtag is important, we arranged hash labels twoly – entropy under 50% (Entropy M) and equivalent to or more prominent than 50% (Entropy N). As the time rots, the likelihood of significance hash labels with higher entropy stays to drift while the low entropy hash labels disappear with time. Evaluation results show that the entropy value of Class M (with multimedia content with a tweet) is relatively higher than the entropy value of Class N (without multimedia content). From our experiments, the system can predict important news sources of disaster event in Twitter and quantifies how much important they are used for disaster news diffusion. On the other hand, the users filtering based on entropy decided whether a user is human (unbiased activity) or robotic (biased to automate) activity during the time of disaster.

The Ghosh evaluation model outcomes in moderate exactness as 73% while the proposed technique gives 94%.This is apparent to show that the proposed measurement model outcomes high precision in distinguishing the catastrophe occasion identified with flood is 96%.The trial is directed on a personal computer with unobtrusive equipment (one i5 – processor, 8 GB memory), to gather the tweet datasets of size 3 GB with various tweets n=12000, and needs non-linear time to calculate word index statistics. Our implementation results can reasonably scale large problem sizes at a higher rate of theoretical O(n log n).

5. Event Detection Rate

The processing delay time needed to start monitoring and confirming event detection is evaluated from the data stream simulation of the real time dataset in Twitter namely Assam Flood (2016), Uttrakand Flood(2016), Chennai Flood(2015) is collected. The size of the incoming tweet is set as 750 tweets per minute. For the three sampling interval are considered as 2, 4 and 6 minutes respectively. The data stream is sent as input to the Hadoop MapReduce framework by fixing the location threshold as λ=300 and the hashtag threshold as τ=700. The proposed MapReduce framework detected the event in 75 seconds which is more efficient with an average time delay on single node. The proposed MapReduce shows better detection rate when compared to Nguyen et al. 2015 which takes 200 seconds for confirming an event.

Table 3. Evaluation of Event Detection Rate

Time Interval	Proposed MapReduce Framework		Nguyen et al. 2015
	Processing Start Time	Event Confirmation Time (Seconds)	Event Confirmation Time (Seconds)
2 minutes	121.24	192.4	324.7
4 minutes	240.8	283	456.1

6 minutes	361.4	401.4	566.1

Table 3 shows that the time taken to confirm an event between the start of event detection and the time at which the event detection is confirmed. The proposed MapReduce framework detected the event in 75 seconds which is more efficient when compared to Nguyen et al. 2015 which used 200 seconds for confirming an event. The results show that the proposed MapReduce framework take a time delay for extracting the events from the initial time interval to confirm the event detection. In the proposed framework, the average of processing time delay for processing the initial data stream containing tweets is after 192 seconds. The system averagely confirms the event detection after 75 seconds in comparing with its real timestamps.

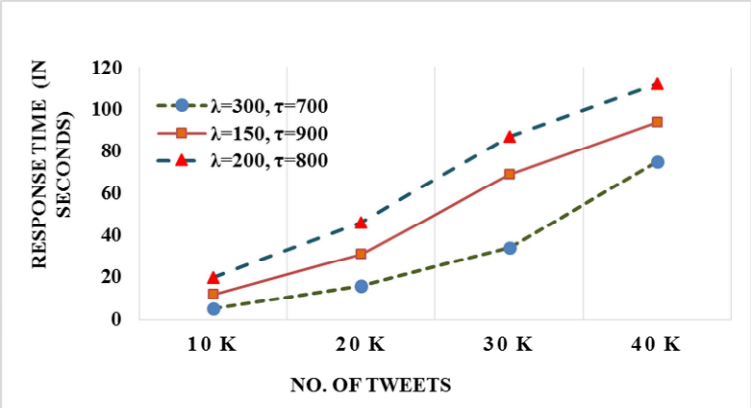

Figure 2 Performance evaluation – Event Detection Rate

6. Event Location Detection Method

We use the following measures to evaluate the efficiency of location detection models. In the experiment, we used Arbitrary Method (AM), Frequency based Method (FBM), K-Center Clustering (OK-CC). For each user, we select the center of the circle with maximum location references within the circle. The radius of the circle is the tolerance value and is same as N in Accuracy@N and 25 miles for tolerance value N as zero.

- **Average Error Distance (AED).** The normal distance between the real areas to the anticipated area of a user with Precision (ACC). The level of appropriately anticipated location among every one of the clients at a city level. The resilience esteem is zero (Tolerance value).
- **Accuracy within N miles (ACCURACY@N).** The level of anticipated areas that are inside N miles of the real area. For instance, ACC@75 estimates the level of anticipated areas that are inside 75 miles of area from the user location profile.

Table 4. Comparison of Accuracy of Various Location Ident cation Models to Identify Location of a User

Method	Unambiguous Location References				All Location References			
	ACC	ACCURACY @25	ACCURACY @50	ACCURACY @75	ACC	ACCURACY @25	ACCURACY @50	ACCURACY @75
SAMPLE - 1500 Tweets								
AM	0.453	0.513	0.524	0.532	0.551	0.561	0.570	0.581
FBM	0.515	0.539	0.561	0.582	0.592	0.621	0.629	0.678
OK-CC	**0.568**	**0.677**	**0.698**	**0.779**	**0.790**	**0.810**	**0.822**	**0.837**

Table 5. Effect of Time on Location Detection

Time after event detection (In Secs	15 Secs	30 Secs	45 Secs	60 Secs	75 Secs
ACCURACY - AM	0.447	0.565	0.616	0.756	**0.560**
ACCURACY - FBM	0.567	0.675	0.766	0.856	**0.671**
ACCURACY – OK-CC@25 Miles	0.554	0.745	0.781	0.831	**0.854**
% of tweets with location detail	33%	54%	66%	79%	**81%**

We use ACC@25, ACC@50, and ACC@75 to calculate accuracy within 25, 50 and 75 miles, respectively. N is the tolerance value is computed and noted in Table 4. It is observed with respect the unambiguous location references, the accuracy is obtained reasonably between 50 miles to 75 miles. Comparing with all location references, the accuracy of 83% is obtained maximum at 75 miles. The effect of time on location detection and percentage of location detail is shown in Table 5. It is observed that the location is detected with highest accuracy for KCC algorithm over time increase eventually. However, the accuracy of RDM and TVM methods shows less accuracy in detecting crisis event location. The normal distance blunder increments from 79 miles for N =25 to 120 miles for N =75. Online K-Center clustering algorithm proves to be an efficient no-regret online algorithm which detects the event location cluster with 85% accuracy.

7. Conclusion and Future Work

This paper focused on the even detection in real time social media and reported the levels of information spread during disasters using Twitter activities with time bounds. We characterize dynamics of tweeting activity associated on social media by calculating the entropy in time-interval distributions, similar user, hash tags and sentiment score. The results show classification of users as real-humans and bots. The proposed entropy-based quantification method identified popular users, hash tags which help us to analyze real human voices in form of tweets during disasters. Indeed, it exhibits the perception of transforming social media into a news media platform. Our system detected three major flood events during 2015-2016 showing disaster event detection rate 94% which is acceptably high. The proposed MapReduce framework detected the event in 75 seconds which is more efficient when compared to state-of-art method which used 200 seconds for confirming an event. From experimental results, the quantification method strategy considering different capabilities with similarly high accuracy contrasting and the customary procedure while keeping up the presentation to recognize the cautions of sudden disaster events. It is observed with respect the unambiguous location references, comparing with all

location references, the accuracy of 83% is obtained maximum at 75 miles. The effect of time on location detection is identified with highest accuracy of 75% for KCC algorithm.

In future work, the proposed model is planned to automate for identifying various natural disaster such as earthquake, Tsunami, and also man-made disasters such as the terrorist attack in distributed real-time environment.

DECLARATIONS:

This statement is to certify that all Authors have seen and approved the manuscript being submitted. I warrant that the article is the Authors' original work. I warrant that the article has not received prior publication and is not under consideration for publication elsewhere. We have no conflict of interest to declare. No Funding for this work is allotted. The data used in the research work is private data and used.

AUTHOR CONTRIBUTIONS

*Funding (information that explains whether and by whom the research was supported) – NOT APPLICABLE

*Conflicts of interest/Competing interests (include appropriate disclosures) - NONE

*Availability of data and material (data transparency) – Twitter Public data

*Code availability (software application or custom code) - Customized code to achieve efficiency in Python

*Authors' contributions (optional: please review the submission guidelines from the journal whether statements are mandatory) Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing-original draft preparation by Dr. Bhuvaneswari Anbalagan

References

[1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys(CSUR).***2015**,vol. 47, no. 4, pp. 67-105.

[2] Hughes.A.L., Palen.L. Twitter Adoption and Use in Mass Convergence and Emergency Events.*Int. J.of Emergency Management.***2009**,vol6,pp. 248–260.

[3] Liu, S., Yamada, M., Collier, N., & Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*,**2013**, vol.*43*, pp.72-83.

[4] Kim, M.; Newth, D.; Christen, P. Modeling Dynamics of Diffusion Across Heterogeneous Social Networks: News Diffusion in Social Media. *Entropy* **2013,** 15, 4215-4242.

[5] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski,D. S. Ebert, T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, **2012**, pp.143-152.

[6] Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, Albert-Laszlo Barabasi. Information Spreading in Context. *In Proceeding 20th Int. Conf. on World wide web*, **2011**, pp. 735-744, ACM New York, NY, USA.

[7] Son Doan, Bao-Khanh Ho Vo, Nigel Collier. An analysis of Twitter messages in the 2011 Tohoku Earthquake. *eHealth 2011 conference, Social Informatics and Telecommunications Engineering*, **2012**, vol 91, Part 4, 58-66 Malaga (Spain).

[8] Hiroki Kawaguchi , Shimpei Matsumotoy and Fujio Toriumiz. A Method to Quantify Twitter User's Posting Activities for Constructing Disaster Information Support System. *IEEE 7th International Workshop on Computational Intelligence and Applications*,**2014**, Hiroshima, Japan.

[9] Rumi Ghosh, Tawan Surachawala, Kristina Lerman. Entropy-based Classification of 'Retweeting' Activity on Twitter. *Social and Information Networks ,Computers and Society*,**2011**, arXiv:1106.0346v1.

[10] Cornelia Caragea, Anna Squicciarini, Sam Stehle, Kishore Neppalli, Andrea Tapia. Mapping moods: geo-mapped sentiment analysis during hurricane Sandy.*In Proceedings ofInt. SCRAM Conf*, **2014**,Pennsylvania, USA.

[11] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A. Sentiment strength detection in short informal text",*J. Assoc. Inf. Sci. Technol.*, **2010**, Vol.61(12), 2544–2558.

[12] Wei Gao, Fabrizio Sebastiani. Tweet Sentiment: From Classification to Quantification. *In Proceedings of ACM Int. Conf. on Advances in Social Networks Analysis and Mining*,**2015**, pp. 97-104.

[13] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. SentiFul: A Lexicon for Sentiment Analysis.*IEEE Trans. On Affective Computing*, **2011**, Vol. 2, No. 1,pp. 22 – 36.

[14] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, Pascal Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery, Springer*,**2015**, vol. 29, pp.1374–1405.

[15] Terumasa Ehara. Word order characteristics analyzed by Hayashi's quantification method type III. *Association for Natural Language Processing*, **2014**.

[16] F. Al Zamal,W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. *In Proc. Int. Conf.Weblogs SocialMedia*,**2012**, pp. 387–390.

[17] G.A. Miller. WordNet: An On-line Lexical Database.*Int. J. Lexicography*, **1990**, vol. 3, no. 4, pp. 235-312.

[18] Felix Ming Fai Wong,Chee Wei Tan, Soumya Sen, Mung Chiang. Quantifying Political Leaning from Tweets, Retweets, and Retweeters.*IEEE Trans. Knowl. Data Eng.* **2016**, vol. 28, No. 8, pp.2158 – 2172.

[19] Yafeng Lu, Xia Hu, Feng Wang, Shamanth Kumar, Huan Liu, Ross Maciejewski. Visualizing Social Media Sentiment in Disaster Scenarios (Short Paper). *ACM Int. WWW Conf. Committee(IW3C2)*, **2015**, Italy.

[20] Wei Gao, Fabrizio Sebastiani. Tweet Sentiment: From Classification to Quantification.*In Proceedings of ACM Int. Conf.on Advances in Social Networks Analysis and Mining*.**2015**, pp. 97-104.

[21] Senti-Strength Algorithm. Available Online :http://sentistrength.wlv.ac.uk/ (accessed on August 10,2016).

[22] A. Andreevskaia and S. Bergler. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses.*Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics*, **2006**.

[23] Vieweg, S., Hughes, A., Starbird, K. and Palen, L.Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *Proceedings of the CHI*. **2010**, Atlanta.

[24] Bhuvaneswari Anbalagan; C. Valliyammai. #ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters using Social Media, I*n proceedings IEEE International Workshop on Foundations in Big Data Computing*. **2016**, Hyderabad, India.

[25] Cameron, M., Power, R., Robinson, B. and Yin, J.Emergency situation awareness from twitter for crisis management. *In Proceedings of Int.Conf. companion on WWW*,**2012**, Lyon, France.

[26] Bhuvaneswari. A., J.Timothy Jones Thomas, P.Kesavan, "Embedded Bi-directional GRU and LSTM Learning Models to Predict Disasters on Twitter Data", Procedia Computer Science, Elsevier, Volume 165C. pp. 101-106, Jan 2020 ISSN 1877-0509.

[27] Bhuvaneswari, A., and C. Valliyammai. "Information entropy based event detection during disaster in cyber-social networks." Journal of Intelligent & Fuzzy Systems 36, no. 5 (2019): 3981-3992.

[28] Nguyen, DT & Jung, JE 2017, 'Real-time event detection for online behavioral analysis of big social data', Future Generation Computer Systems, vol. 66, pp. 137-145.

[29] Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. Information Processing & Management, 56(3), 1146-1165.

[30] Dabiri, S., & Heaslip, K. (2019). Developing a Twitter-based traffic event detection model using deep learning architectures. Expert systems with applications, 118, 425-439.

[31] Choi, H. J., & Park, C. H. (2019). Emerging topic detection in twitter stream based on high utility pattern mining. Expert systems with applications, 115, 27-36.

[32] Xu, S., Fu, X., Cao, J., Liu, B. and Wang, Z., 2020. Survey on user location prediction based on geo-social networking data. World Wide Web, 23(3), pp.1621-1664.

[33] Saroj, A. and Pal, S., 2020. Use of social media in crisis management: A survey. International Journal of Disaster Risk Reduction, p.101584.

[34] Bothorel, C., Lathia, N., Picot-Clemente, R. and Noulas, A., 2018. Location recommendation with social media data. In Social Information Access (pp. 624-653). Springer, Cham.

[35] Eagle, N & Pentland, A 2006, 'Reality mining: sensing complex social systems', Personal and ubiquitous computing, Vol.10, Issue. 4, pp.255-268.

[36] Shetty, J & Adibi, J 2005, 'Discovering important nodes through graph entropy the case of enron email database', In Proceedings of the 3rd international workshop on Link discovery, pp. 74-81, ACM.