Information Systems Frontiers (2018) 20:961–979
https://doi.org/10.1007/s10796-018-9837-8

Analysis and Early Detection of Rumors in a Post Disaster Scenario

Tamal Mondal[1] · Prithviraj Pramanik[2] · Indrajit Bhattacharya[1] · Naiwrita Boral[1] · Saptarshi Ghosh[3]

Published online: 9 March 2018

Abstract
The use of online social media for post-disaster situation analysis has recently become popular. However, utilizing information posted on social media has some potential hazards, one of which is rumor. For instance, on Twitter, thousands of verified and non-verified users post tweets to convey information, and not all information posted on Twitter is genuine. Some of them contain fraudulent and unverified information about different facts/incidents - such information are termed as rumors. Identification of such rumor tweets at early stage in the aftermath of a disaster is the main focus of the current work. To this end, a probabilistic model is adopted by combining prominent features of rumor propagation. Each feature has been coded individually in order to extract tweets that have at least one rumor propagation feature. In addition, content-based analysis has been performed to ensure the contribution of the extracted tweets in terms of probability of being a rumor. The proposed model has been tested over a large set of tweets posted during the 2015 Chennai Floods. The proposed model and other four popular baseline rumor detection techniques have been compared with human annotated real rumor data, to check the efficiency of the models in terms of (i) detection of belief rumors and (ii) accuracy at early stage. It has been observed that around 70% of the total endorsed belief rumors have been detected by proposed model, which is superior to other techniques. Finally, in terms of accuracy, the proposed technique also achieved 0.9904 for the considered disaster scenario, which is better than the other methods.

Keywords Microblogs · Twitter · Rumor · Disaster · Chennai floods

1 Introduction

Natural disasters in recent years (e.g., *Nepal Earthquake 2015, Chennai Flood 2015* etc.) have posed a great threat to mankind by claiming innocent human lives and causing huge loss of property. There have been continuous efforts in minimizing such losses. There lie many problems and constrains while performing post-disaster relief work. The intermittent unavailability of network infrastructure (*like GSM, Internet*) leads to the disruption of communication in a post-disaster situation and causes major hardship in gathering relevant information regarding the disaster in real-time. However, it is not often that the network infrastructure is completely disrupted in disaster situations. Even in case of large-scale disasters like the 2015 Nepal Earthquake[1], or the 2015 Chennai floods[2], it has been observed that the network connectivity still persists in small pockets. This selective availability of network infrastructure, along with the ubiquity of mobile phones and smart phones, has given rise to various alternative sources of information like Online Social Networks (OSNs) for understanding and analyzing post-disaster scenarios. Efficient relief work in post-disaster phase requires accurate up-to-date static and dynamic information. To obtain a complete picture of

✉ Tamal Mondal
tamalkalyanigov@gmail.com

Prithviraj Pramanik
prithvirajpramanik@yahoo.co.in

Indrajit Bhattacharya
indra51276@gmail.com

Naiwrita Boral
boralnaiwrita@gmail.com

Saptarshi Ghosh
saptarshi@cse.iitkgp.ernet.in

[1] Department of Computer Application, Kalyani Government Engineering College, Kalyani, India

[2] Department of Computer Science and Engineering, National Institute of Technology Durgapur, Durgapur, India

[3] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India

[1] https://www.pcworld.com/article/2914972/internet-steady-in-nepal-after-earthquake-but-lastmile-connectivity-an-issue.html.

[2] http://www.channelworld.in/news/after-nepal-earthquake%2C-internet-connectivity-is-key-moved.

the scenario, information from many different sources needs to be integrated and put together.

Nowadays, information from OSNs like Twitter, Facebook, etc. is extensively used for getting an overview in disaster situations, for relief and need assessment (Dhanjal et al. 2011). To that effect, the information content shared on the OSNs by general users is also being increasingly used. In this way, direct human observation is included, often from ground zero (actual site of the disaster), to obtain an enhanced environmental knowledge, and to further optimize the planning process using various multi-objective optimization techniques (Zhao et al. 2017). OSNs, especially Twitter, have made large-scale data collection from the general public and official sources very convenient and popular. Freely available, huge quantities of Twitter messages (tweets) have been mined and researched in the recent past for various disaster-related applications.

However, utilizing information posted on OSNs has some potential hazards, one of which is *rumor* or unverified information. In post-disaster situations, millions of tweets are posted from different verified/non-verified Twitter accounts. The people present in the disaster-affected sites post about what they are witnessing (Liu et al. 2014). But, the people, who are not present at the disaster site might propagate information posted by others and cite information links from other sources, without knowing the actual facts (Liu et al. 2014; Laniado et al. 2017). In a post-disaster scenario, due to lack of proper knowledge or news media intervention, many users post unverified information which makes up rumors (unverified claims). It has been observed that some people support those unverified claims by either re-tweeting them, or by manipulating the information (Zhao et al. 2015). Such behavior can provide a channel for the propagation of rumors in the network. In a disaster situation, it is often time-consuming to confirm various facts; also, people are naturally anxious and panicky during such times. Hence, people often re-tweet any disaster-related tweet obtained from the users they follow, or simply contrive the news in order to fill the blanks of the partial stories available to them (Nekovee et al. 2007; Doerr et al. 2012). Such behavior of users often leads to rapid propagation of rumors in OSNs like Twitter.

Note that, in (Sen et al. 2015), authors claimed that most of the situational tweets during any crisis situation are centered on a finite set of disaster-related content words. Rumors can be characterized as situational tweets, containing disaster-related content words. Thus, at any particular point in time, a limited set of situational tweets are propagating along an OSN due to the limited amount of knowledge among the users in the network. Some prominent features of rumor propagation have been discussed in (Kwon et al. 2013a, b) - several rumor features like *temporal, structural, linguistics, social ties* have been analyzed to characterize rumors. These distinct features of rumor propagation in OSNs motivated us to design a novel rumor detection model by implementing the features more systematically, for efficient detection of rumors at early stages in any post-disaster situation. Through individual coding of each feature, the proposed model seeks to check whether efficient rumor detection is possible at an early stage in the aftermath of a disaster.

The tasks performed by the proposed model are fivefold:

1. At any point of time, from the collected tweets related to a disaster situation, a set of informative and mixture tweets is extracted.
2. From the set of informative and mixture tweets,

(a) a cluster has been prepared by extracting tweets which satisfy rumor propagation features like *structural, linguistic, social ties (*Kwon et al. 2013a, b*)*, and
(b) two trained experts have been involved to detect actual rumors from among the clustered tweets.

3. In the next phase, the popular Oh et al. (2013) rumor detection model has been adopted to check the presence of five rumor causing factors in each tweet in the cluster. Other recognized lexicon factors (e.g., Word Relevance (WR), Relevance Factor (RF) and Global Relevance (GR)) have also been tested for each tweet in the cluster.
4. The rumor probability of each tweet in the cluster is computed, considering rumor causing factors as well as lexical factors in the content of each tweet.
5. The tweets with *high rumor probability* are extracted and compared with actual rumors (as identified by human experts, in Step 2b) to check the efficiency of the proposed technique.

In the present work, tweets from the devastating flood in November–December 2015 in the Indian city of Chennai, have been collected and analyzed for rumor detection. Different instances of the proposed model have been built upon modern advancement of NLP for microblogs (Corvey et al. 2010). It can be noted that the proposed model concentrates on both *prominent features* as well as *content-based analysis*. Thus, at first, for each informative or mixture tweet, the possibility of the posted tweet being a rumor is determined (Step 2a). If the possibility is high, then the rumor probability has been evaluated by utilizing the content-based features of that tweet (Step 3). By integrating various existing microblog data processing aspects, a real-time rumor detection scheme has been developed for post-disaster situation.

The rest of the paper is organized as follows. In Section 2, prior works on rumor identification and spreading have been discussed. Section 3 discusses the tweet collection procedure for the present work. In Section 4, the proposed rumor detection model has been described in detail. After analyzing the performance of the proposed technique, the results have been presented in Section 5. We finally conclude the paper in Section 6.

2 Related Works

Several studies have been performed on various dimensions of rumor analysis in microblogs. Till now, the studies on rumors in micro-blogs are mainly concentrated on three different dimensions as follows:

a) Various rumor theories have been studied to analyze different factors that cause rumors, factors that enhance a rumor to further propagate, and the trustworthiness of a tweet. Many variations of rumor have also been studied in the literature of communication studies.
b) Different rumor detection models have been proposed, using different features like social behavior, content-based network, etc.
c) Rumor control strategies have also been adopted, based on anti-rumor spreading, analyzing rumor life cycle, etc.

2.1 Rumor Definitions

While defining rumor, Rosnow and Kimmel defined a rumor as "unverified proposition or belief that bears topical relevance for persons actively involved in its dissemination" (Rosnow 1991; Kimmel 2013). According to them "any report, statement or story that one has heard for which there is no immediate evidence to verify its truth" can be considered to be a rumor. According to Buckner (Buckner 1965), rumor is defined as "Unconfirmed message passed from one person to another … that refers to an object, person or situation rather than an idea or theory". The authors in (Zhao et al. 2015) defined rumor as a controversial and fact-checkable statement.

These different rumor definitions show some characteristics of rumor are: (i) rumor generally arises in the context of ambiguity, (ii) any statement which is unverifiable, which does not have any valid proof can be considered as a potential rumor. However, uncertain truth value does not mean that rumors imply false information. All existing definitions designated rumor as unconfirmed statement or statement without any valid evidence. There is always a possibility that unconfirmed statements are true. But due to the absence of authentic knowledge, they might be interpreted as a rumour. In recent days, with the increasing use of OSNs & news media, a large number of micro-blog users and reporters actively engage in generating or propagating news/views about any particular trending topic (Asur and Huberman 2010) which might also include crisis situations. Therefore, any unconfirmed true statement might not take much time to be known to everyone. On the other hand, there might be certain stories about which news media or government sources fail to convey the actual information within a finite amount of time. These stories often lead people to manipulate or twist the information, question the credibility of the information and circulate various statements related to those stories. As a result, those unverified factual statements propagate rapidly and sustain for a longer period of time that unnecessarily creates nuisance, panic, etc. among the different group of people. From these insights and various definitions of rumors, it is evident that rumor is an "*unverified factual information propagating through the network, which is either accepted or rejected or questioned by the community*".

2.2 Theories on Rumor Propagation

In (Oh et al. 2010), the authors pointed out that information with credible sources contributes to suppressing the level of anxiety in the Twitter user community, which leads to rumor control and high information quality. Another work (Mendoza et al. 2010) pointed out the property of propagation of rumors in tweets. It was found that the propagation of tweets that correspond to rumor differs from that of tweets that spread news because rumors tend to be questioned more than regular information by the Twitter community. This observation implies that it is possible to detect rumors using aggregate analysis on tweets. The work (Oh et al. 2013) proposed a model that uses five rumor-causing factors - Anxiety, Source Ambiguity, Content Ambiguity, Personal Involvement and Directed Messages - to predict the probability of rumor generation and spreading. An extension to the above work has been performed in (Liu et al. 2014), where the authors used a slightly modified model on the retransmitted tweets. The properties of Online Social Networks that enable rumors to propagate faster have been studied in (Doerr et al. 2012). During simulation, it was observed that in Twitter, rumor related to a particular topic propagates to an extensive set of users within a very short duration.

In (Kwon et al. 2013a, b), several rumor propagation features were illustrated. From those feature illustrations, the following four properties of rumors can be concluded:

(i) Any rumor related to a certain topic is influential for a short time window when no news media is capable of capturing the actual facts about the topic. (*Temporal*)
(ii) During this short time frame, the rumor topic becomes *skeptic* when it enters any denser social network. (*Structural*)
(iii) Statements related to a rumor topic usually carry higher sentiments than statements related to non-rumor topic. (*Linguistic*)
(iv) People's trust in information obtained from their friends, and discussion about any unverified information within a group where people know each other, leads rumors to propagate faster. (*Social Tie*)

The rumor theories discussed above clearly points out several factors that cause/ enhance rumors to propagate in OSNs like Twitter. Lack of veracity and trustworthiness about any information primarily leads users to further propagate any unverified information. The lack of proper knowledge regarding any topic is actually why rumors are often able to influence a large number of members in an OSN.

It is clear that rumor detection models, which might be able to detect rumors propagating in OSNs at early stage in any crisis situation, are practically necessary. The rumor propagation theories can help in developing such models– it is possible to enact rumor propagation theories in rumor detection models in such a way that any tweet related to rumor can be flagged at an early stage.

2.3 Rumor Detection Models

A number of rumor detection models have been proposed in literature. We briefly discuss these models, along with some of their limitations in the application in a post-disaster scenario, which motivated us to develop the model in the present work.

The authors of (Zhao et al. 2015) proposed a rumor detection technique based on finding 'enquiry phrases', clustering similar posts together, and collecting related posts that do not contain these simple phrases. They also ranked the clusters by their likelihood of containing rumors. The method in (Zhao et al. 2015) mainly consists of identifying signal tweets (tweets containing verification or correction phrases) as they are the good indicators for rumors; however, indicators like verification or correction usually appear at later stages of rumor propagation, and hence have limited applicability in early detection of rumors.

The work (Dayani et al. 2015) was a retrospective analysis of a rumor dataset, which applied Machine Learning techniques for detection of tweets spreading rumors. The work mainly focuses on tracking the users who endorsed rumor tweets on some topics in the year 2009. The machine learning techniques used in this work for rumor detection are based on factors that are not applicable for rumor detection at an early stage in a post-disaster situation.

In (Yang et al. 2015), the authors mainly focus on automatic 'hot topic' detection by combining bursty term detection and sentence modeling. The proposed model first detects hot topics and then applies three rumor classifiers for detection of rumors. A method for exploring three categories of features for identification of rumors has been designed in (Qazvinian et al. 2011). This work shows that the content-based features are very effective for rumor identification. However, the model has been implemented on four rumor topics, three of which are not related to any disaster.

The authors in (Dayani et al. n.d.) focused on content-level analysis of rumors. A data repository was built, containing tweets related to various rumor topics. After that, a step-by-step procedure was used to find whether tweets posted on Twitter are related to any of the rumor topics stored in the repository; if a match is detected, and then the tweet's sentiment value is extracted and compared with the tweets on the same topic in the repository. Finally, tweets that are found similar are added to the repository, and so on. Due to the establishment of a repository, the technique in (Dayani et al. n.d.) can only detect tweets related to those rumors that are already present in the repository. However, in a post-disaster situation, new rumors might propagate with respect to time; hence this method would not be suitable for such scenario.

The work (Yang et al. 2012) studied the problem of information credibility on SinaWeibo, China's leading microblogging service provider. The authors collected an extensive set of micro-blogs that were confirmed to be false rumors, based on information from the official rumor-bursting service provided by SinaWeibo. Unlike previous studies on Twitter where annotators did the labeling of rumors manually, the official nature of this service ensures the high quality of the dataset. A large set of features was extracted from the microblogs, and a classifier was trained to automatically detect the rumors from a mixed set of true information and false information. Their experiments showed that some of the features considered in previous studies have different implications with SinaWeibo than with Twitter.

The authors in (Liang et al. 2015) presented a rumor detection scheme based upon various behavioral features of microblog users. Here, the authors observed behavioral divergence of rumor spreaders with normal users in the social network and tried to differentiate the rumor-mongers from normal users in context of some common behavioral features. Also, rumor posts get different kind of responses than normal posts. In a post-disaster situation, when almost all users are anxious and panicky, any verified or non-verified users can originate rumors; hence we preferred not to rely on user properties in the proposed rumor detection model.

A novel approach to capture the temporal characteristics of the life cycle of a rumor was presented in (Ma et al. 2015). The authors studied variations of various rumor features over time. It was observed that over time the propagation rate of any rumor decreases, as the actual facts become clear among the community; from that point of time, the rumor is contradicted or questioned by the community. But the verification/correction appears at later stages of rumor diffusion. In any crisis scenario, it may take a long time to confirm about any factual claim. In (Ma et al. 2016), the authors utilized the same phenomena as in (Ma et al. 2015) and developed a recurrent neural network based model to capture the feature variations of any rumor.

2.4 Rumor Control Strategies

Alongside rumor detection, some prior works have also developed ways to control the spread of rumors. In (Tripathy et al. 2010), the authors proposed an 'anti rumor' propagation strategy for combating the spread of rumors in social networks. The authors mainly studied the belief time (the time at which the people believe a rumor) and the point of decline (the time when the anti-rumor dominates the rumor). An impendent model was used to simulate the spreading of rumors. The “delay-start” and “beacon” model was used to combat the rumor spread through anti rumor spread. The authors in (Bao et al. 2013) proposed a technique for characterizing rumor propagation and spreading. The proposed technique was tested both through simulation and on a real dataset from the Sina Weibo micro-blogging site. These rumor control strategies clearly show the efficiency of combating rumors with anti-rumor propagation in real time.

The authors in (Tripathy et al. 2010) also pointed out the importance of the time-span that is the time difference between the start of rumor propagation and the start of anti-rumor propagation. The efficiency of the two proposed models (i.e., delayed start and beacon models) depends only on the value of this time-span, the lifetime of a rumor increases if the delay in detecting it increases (Tripathy et al. 2010). Therefore, for effective implementation of any rumor control strategy, delay in detecting any rumor should be minimized as much as possible. In this context, the question still remains about how a rumor can be identified at an early stage. The present work seeks to answer this question by designing a real-time rumor detection model for early detection of rumors.

2.5 Novelty of the Present Work

As discussed above, several rumor detection schemes have been proposed for micro-blogging sites like Twitter and SinaWeibo. However, most of these models are not particularly suitable for early detection of rumors in post-disaster situations. For instance, the authors in (Zhao et al. 2015; Ma et al. 2015, 2016) discussed the *skepticism* component in rumor propagation. But skepticism can occur only when the correct facts about the rumor topic become densely known in the social network. Therefore, skepticism is not likely to be a property through which early detection of rumors can be performed efficiently. The influence of a rumor can last longer in a post-disaster situation due to lack of information at early stage. Furthermore, at early stage in a post-disaster scenario, due to lack of news media interventions and anxiety among the people, most of the facts posted can be unverified. As a result, before the actual fact becomes densely known, there might be long time differences between the rumor related posts and skeptic posts. These issues motivated us to develop an early rumor detection model specifically for post-disaster situations.

In the present work, we utilize the four properties of rumor propagation identified in (Kwon et al. 2013a, b) – Temporal, Structural, Linguistic and Social Tie (as discussed earlier in this section). It can be noted that temporal and structural characteristics might not be independent of each other. During shorter time frames, it is possible that people doubt or question the authenticity of any unverified proposition. However, posting of verification statements about a fact does not mean that the fact is a rumor. Additionally, in a post-disaster situation, it might not always be the case that influence of a rumor lasts for only a short time window (as we observe in Section 5.3). In the present work, we have primarily considered the *Structural, Linguistic and Social tie* properties of rumor propagation, to extract rumor features. In addition, influence of rumor topics at different time frames has also been analyzed (in Section 5.3) in order to characterize the *Temporal* feature of rumor propagation.

3 Data Collection

We collected tweets related to the severe floods in the city of Chennai, India, during November–December 2015. A total of 452,544 tweets posted during December 01–10, 2015were collected using the Twitter Search API. The tweets were collected using several search keywords, as follows. The keywords“#chennaiRains”, “#chennaiFlood”, and “chennai” (case insensitive) were selected based on their popularity on Twitter during the said period. However, many tweets might contain information related to the flood, without containing any of the above terms. To capture such tweets, various keywords, phrases and their synonyms related to the topic ‘flood’ were considered, as obtained from *https://www.bangkokpost.com/learning/vocabulary/200842/flood-related-vocabulary*. Some of the terms in this lexicon are shown in Table 1.

After the initial data collection with the search keywords mentioned above, duplicate tweets (which were matched by more than one keywords) were removed; this resulted in a set of 197,497 distinct tweets related to the Chennai flood event.

Tweets posted during a disaster event are of three types – (i) informative, which contain situational information, (ii) non-informative, which do not convey any information about the situation, rather contain emotions and sentiment, and (iii) mixture, which contain both situational and non-situational information (Rudra et al. 2015). We extracted out the informative and mixture tweets using the *lexical and syntactic features*

Table 1 Some lexicons related to crisis type ‘flood’

Lexicons	Meaning
Downpour + Synonyms	A lot of rain in a short duration
Evacuate + Synonyms	Cause to leave an unsafe place
Stranded + Synonyms	Struck some with no way of going
Victims + Synonyms	People who are dead, injured etc.
Collapse + Synonyms	Suddenly fall down.

identified in (Rudra et al. 2015). The procedure for obtaining informative and mixture tweets has been discussed in Section 4.1 in greater detail.

The reason behind including both informative and mixture tweets are -(i) the mixture tweets may contain some valuable information along with some sentiments or opinions, and (ii) to make a fair analysis of the model in (Oh et al. 2013), the model has considered some sentiment or opinion related factors (*anxiety, personal involvement* etc.) as rumor-causing factors. Therefore, along with informative tweets, mixture tweets can also be strong candidates for generation of rumors. After extraction of informative and mixture tweets, we were left with 79,125 such tweets. All subsequent analyses were performed on this set of 79,125 tweets. Some examples of informative, mixture and non-informative tweets for considered scenario are shown in Table 2.

4 Proposed Model

In this section, we discuss the proposed model in detail. As discussed in Sections 1 and 2, it is necessary to extract genuine situational information at an early stage in any disaster response situation. In the proposed method, such a real-time mechanism has been adopted for the Twitter micro-blogging site, with an objective to identify genuine informative tweets and discard rumor tweets at any point in time. The workflow of the proposed model is shown in Fig. 1. Each step of the proposed technique has been described in this section.

4.1 Informative and Mixture Tweet Extraction

After collecting tweets relevant to a disaster event, using some specific constraints as mentioned in Section 3, the next step is to extract informative and mixture tweets and discard all non-informative tweets. Informative tweets generally contain situational information that can help in relief operations. Whereas, non-informative tweets contain sentiment, opinion, event analysis and charity related statements (Rudra et al. 2015). Mixture tweets contain both informative and non-informative statements.

To separate informative and mixture tweets from non-informative tweets, we have used some low level syntactic and lexical features that were found useful for this purpose in (Rudra et al. 2015). These features are as follows,

- **F-measure:** The formality measure of a piece of text (e.g., a tweet) is defined as,

$$F = (noun \quad freq + adjective \quad freq + preposition \quad freq + aricle \quad freq - pronoun - verb \quad freq - adverb \quad freq - interjection + 100/2) \tag{1}$$

Here, *freq* represents the number of words of a certain category, among the total number of words in the tweet. It has been observed that informative and mixture tweets generally have higher F-measure values than that of non-informative tweets.

- **Count of numerals:** informative and mixture tweets usually contain more numerals than non-informative tweets.
- **Count of personal pronouns:** Non-informative tweets are often written from a personal standpoint, and thus contain more personal pronouns compared to informative and mixture tweets.
- **Exclamations:** It has been observed that non-informative tweets often contain exclamatory words compared to informative and mixture tweets.
- **Intensifiers:** Intensifiers are generally used in non-informative tweets to convey stronger sentiment.
- **Modal verbs:** Modal verbs are generally used to give opinion regarding some issues. Thus, non-informative tweets contain more modal verbs than informative or mixture tweets.

We computed the values of all these features for the 197,497 distinct tweets we collected. It was observed that informative tweets are written in a more formal way, contain more numerals, and have larger F-measure values (Rudra et al. 2015). On the other hand, strictly non-informative tweets consist of more exclamations, modal verbs and intensifiers (Rudra et al. 2015). Hence, it can be inferred that, for any tweet to be considered as a mixture, all syntactic feature values must be in between those of informative and non-informative tweets, as mixture tweets contain both formal as well as informal statements in the tweet.

Next, the following tasks were performed to filter out non-informative tweets. Each tweet was considered along with the values of above specified syntactic features. The minimum and maximum values of each syntactic feature were extracted (across all tweets). The tweets were tested by processing various sets of queries on various ranges of values (which lie between the minimum and maximum values) of each syntactical feature. Based on this analysis, we decided to discard those tweets for which the feature values lie between the specified ranges stated in Table 3. The remaining 79,125 tweets are considered to be informative and mixture tweets.

After separating out informative and mixture tweets, we performed two parallel tasks using those informative and mixture tweets: (i) use experts' intervention to find out real rumor

Table 2 Informative, Mixture and Non-Informative Tweets

Informative tweets	Mixture tweets	Non-informative tweets
RT @ndtv: In flooded Chennai, airport shuts down as runway goes under water #ChennaiRains	Hi, all the theatres at chennai are open to provide accommodation. Please try to help the people of chennai #chennairains	Oh there is a torrential downpour in our chennai god please save the city and that people #chennairains
RT @Anushka_ASF: Heavy rains lash Chennai water logging in many parts of the city.	Electricity out in Pallavaram since 1 pm on the 01st of Dec! What is the EB doing! #chennairains #Chennai @Ndtv	RT @parvati786: Do you want to restore faith in humanity? See India and the chain of help in Chennai, then say.
@TwitterIndia Phoenix Marketcity Chennai is open for anyone seeking shelter #ChennaiRainsHelp@the_hindu	#Chennai Corp announced emergency contact numbers! Retweet Please! #ChennaiRains #ChennaiRainsHelp [url]	Standing by each other when it matters. No wonder we aren't newsworthy. #chennairains [url]

related tweets, and (ii) perform rumor hypothesis test for each tweet and form rumor cluster.

4.2 Experts' Intervention

We employed two human experts, (who are regular users of Twitter with good knowledge of English, and prior experience in studying tweets posted during disaster situations) to manually identify real rumor related tweets. We collected 10,417 tweets related to some of the rumors published in (Qazvinian et al. 2011). The experts were told to inspect each tweet (which are written in English), verify whether the information is true, using web search or from different popular news media like NDTV, AAJ TAK, BBC, etc. The experts were also shown the various rumor definitions as described in Section 2. After 120 h the experts

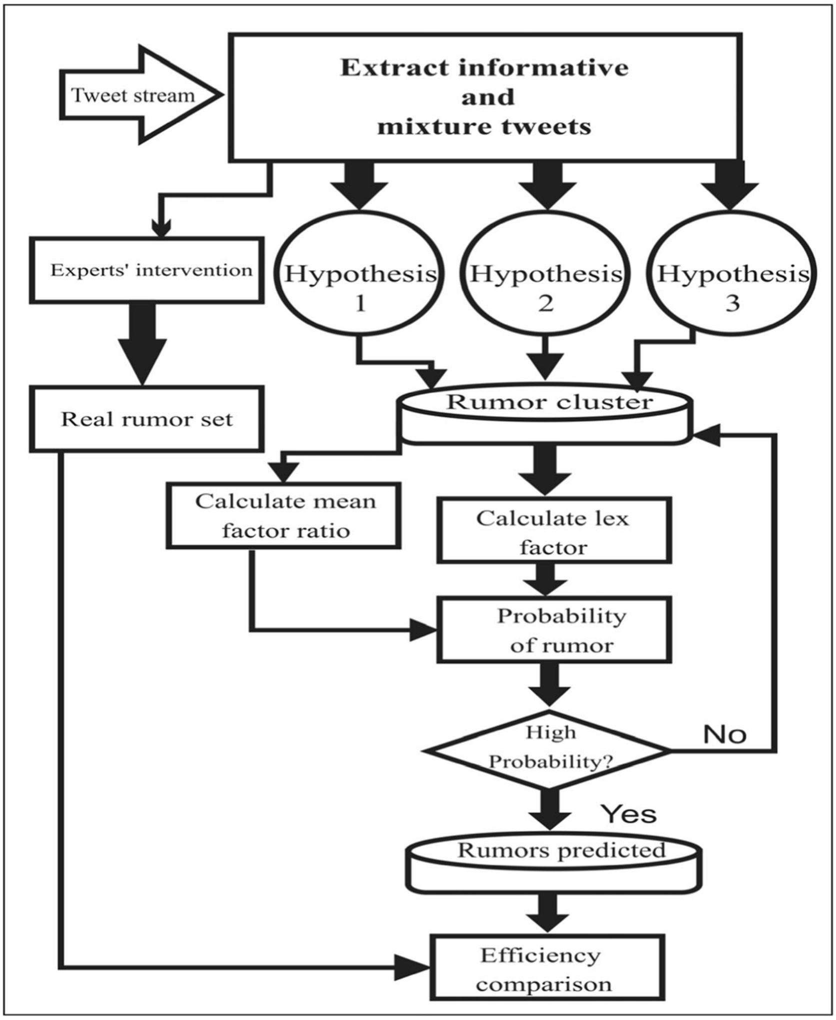


Fig. 1 Workflow diagram of the proposed rumor detection model

Table 3 Range of values of each feature for non-informative tweets

Features	Minimum value	Maximum value
F-measure	49.5	54.5
Count numerals	0	1
Personal Pronouns	0	3
Exclamation	0%	9%
Intensifiers	6%	17%
Modal verbs	0%	17%

finished their tasks and approximately 373 rumor tweets were identified related to two topics:

(1) *Crocodile escape from a zoo*: There was circulating news about crocodiles escaping from the Madras Crocodile Bank in Mamallapuram. But later, the crocodile bank confirmed that it was a rumor.
(2) *Nasa predicts El Nino*: A Whatsapp message asking Chennai people to leave the city as NASA had predicted a Tsunami in next 72 h. But later it had been found that NASA had not issued any such report.

The inter-expert agreement had Spearman Correlation (Myers and Sirois 2006) value of 0.85, which is satisfactory. Table 4 shows some examples of belief tweets (which say that the information is true) and disbelief tweets (which say that the information is false) related to the two rumors. Table 5 shows the number of belief and disbelief tweets identified by the experts. A total of 373 rumor related tweets were found, out of which 156 were related to "Nasa predicts El Nino" and 217 were related to "Crocodile escape."

Table 5 Number of beliefs and disbeliefs related to two rumor topics detected by experts

	Disbelief statements	Belief statements	Total
Nasa predicts El Nino	128	28	156
Crocodile escape	152	65	217

4.3 Analysis of Rumor Hypothesis and Cluster Formation

In this section, we test three distinct dimensions of rumor diffusion, on the informative and mixture tweets. Tweets that satisfy any of these dimensions are extracted to form a rumor cluster. Prior studies found that there are some key differences between dissemination of rumors and non-rumors (Kwon et al. 2013a). Besides, the key structural and linguistic differences in rumor and non-rumor were also identified in (Kwon et al. 2013b). In the present work, we evaluate the notions of these rumor propagation features.

During absence of information from verified sources (e.g., news channels), rumor may propagate due to lack of factual information (Zhao et al. 2015). In such times, some people might seek out stronger evidence about any factual information by interrogating or questioning. Furthermore, it has been observed that a rumor diffusion network usually collapses whenever the rumor topics are discussed in denser networks where people express doubts about the truthfulness of the information (Kwon et al. 2013b). In (Kwon et al. 2013a; Vosoughi 2015), it was also found that rumors are dominant over non-rumors in terms of certain types of sentiments *like anger, aggression, happiness,* etc. Therefore, it is imperative to evaluate overall sentiment value for each tweet. In (Kwon

Table 4 Examples of belief and disbelief statements related to two rumor topics

Rumor topic	Belief	Disbelief
Crocodile escape	RT@latasrinivasan: Must read When crocodiles had escaped from Madras Crocodile Bank in the rains	RT @Sibi_Sathyaraj: Guys pls stop spreading rumours about crocodiles invading Chennai! Ppl are already facing enough problems!#ChennaiFlood
	20 crocodiles missing from zoo in chennai	RT@iArafathh:Crocodiles news in Chennai is strong rumor ..created by some idiots ..so don't worry makkale #ChennaiRainsHelp
	@ibnlive:crocodiles are escaped from the park in Chennai [url]	RT @GabbbarSingh: The crocodile escaping in Chennai news is a Hoax.
Nasa predicts El Nino	Another disaster is coming for Chennai again said by nasa Pray for chennai	NASA Warned Chennai Will Suffer Very High Rainfall with a Hurricane: Hoax [url]
	RT @ImBharathan: NASA hints there might be a chances of El Nino in chennai If that's true then chennai has not undergone even its initial	So no heavy rainfall on following days in chennai news which circulated around about #Nasa forecast is rumour....... #feelfree
	@Sibi_Sathyaraj @Troll_Cinema a strom named as El Nino Storm wch attack in chennai It is expected too come up to 250 cm Verified by NASA	Fake WhatsApp Message Heavy Rains Chennai NASA Warning Don't Believe False News [url]

Table 6 Regular expression patterns for verification and correction

Regular expression	Type
is (that | this | it) true	Verification
wh[a]*t[?!][?]*	Verification
(real? | really? | unconfirmed)	Verification
(rumor | debunk | hoax)	Correction
(that | this | it) is not true	Correction

et al. 2013b), it was found that people's trust on information received from their friends, and discussion about any unverified information within a group where people know each other, leads rumors to propagate faster. Considering these *temporal/structural and linguistic* features of rumor propagation, three distinct dimensions about rumor propagation have been specified in the proposed model. These three distinct dimensions have been termed as rumor hypotheses, which are tested further to construct the rumor cluster. The considered rumor hypotheses (H) are as follows:

i. *Rumors generally contain words related to verification or correction* (H1)
ii. *Rumors are dominated by high sentiments, as compared to other types of information* (H2)
iii. *Rumors generally contain words related to social ties and actions like hearsay* (H3)

In the present work, we used the Python Natural Language Toolkit (NLTK) to implement the test of the three hypotheses. Packages like *tokenizer, wordnet, sentiwordnet,* etc. have been used to (a) tokenize the tweets, (b) extract synsets of content words in the tweets, and (c) evaluate sentiment scores of the tweets. Note, here 'content words' refer to words with parts of speech (POS) forms (noun, verb, adjectives, etc.) and numerals used to construct the tweet. We now describe how the hypotheses are tested.

H1 test

a) The regular expression patterns used in (Zhao et al. 2015) were considered to extract tweets related to verification and correction. The regular expressions are shown in Table 6. As verification or correction words signal that the topic discussed in the tweet might be a rumor, tweets that contain such signals are extracted and put into the rumor cluster. Here, synsets of each content word used in the regular expression patterns are also considered. For each tweet, if any sub-string of that tweet matches any of the predefined regular expressions, then the tweet is included in the rumor cluster.

b) Tweets that are *controversial* are detected and included in the rumor cluster. Any controversial event provokes a public discussion where people express their disbelief or opinion about the event. Controversial tweets often *instigate people to express their views, doubt about the fact mentioned in the tweet* and this may lead to rumors in future. Hence, controversial tweets are included in the rumor cluster. Given any informative or mixture tweet, the controversy score is generated using the following formula (Popescu and Pennacchiotti 2010),

$$TW\_CONT\_MIX = \frac{Min(|Pos|, |Neg|)}{Max(|Pos|, |Neg|)} \times \frac{|Pos| + |Neg|}{|Pos| + |Neg| + |Neu|} \qquad (2)$$

Where,

Pos	Positive sentiment associated with the tweet.
Neg	Negative sentiment associated with the tweet.
Neu	Neutral polarity associated with the tweet which is (1- (Pos + neg)).

To evaluate positive and negative polarity, the distinct *content words* present in a tweet are extracted. Now, for each distinct *content word*, the number of *synsets and synonyms* is evaluated. Some examples of informative and mixture tweets are shown in Table 7 along with their distinct *content words*. A numeric estimate of positive and negative polarity associated with every synset is then obtained using *sentiwordnet* (Denecke 2008). After that, *controversy* associated with a tweet is evaluated using (2). If *controversy* of *a tweet is greater or equal to the average controversy value of all tweets*, then the tweet is added to the rumor cluster.

Table 7 Some Informative & mixture tweets along with their distinct content words

Tweets	Distinct content words
RT @ndtv: In flooded Chennai, airport shuts down as runway goes under water #ChennaiRains	flooded, Chennai, airport, shuts. Runway, goes, water
Electricity out in Pallavaram since 1 pm on the 01st of Dec! What is the EB doing! #chennairains #Chennai @Ndtv	Electricity, Pallavaram, since, 1 pm, 01st, Dec, What, EB
RT @Anushka_ASF: Heavy rains lash Chennai waterlogging in many parts of the city.	Heavy, rains, lash, Chennai, waterlogging, many, parts, city

H2 test As mentioned earlier, rumors are dominated by high sentiment values. Though various sentiment classification algorithms have been proposed (Choi and Lee 2017), in proposed work, to test hypothesis 2 an algorithm specified in (Esuli and Sebastiani 2006) has been customized. The proposed modified algorithm is as follows,

Algorithm: Sentiment Extraction from Tweets

```
1.  Start
2.  retrieve tweets one by one
3.  For retrieved tweet
    3.1 tokenize the tweet and store it into token_list
    3.2 remove stop words, punctuation marks from token_list
    3.3 count total number of content words($n$) in token_list
    3.4 for each content word in token_list
        3.4.1    find out synsets ($sns$) of the content word
        3.4.2    for each synset($i$)
            3.4.2.1 use sentiwordnet [33] to find out positive and negative
                    sentiment for that synset.
            3.4.2.2 sum positive and negative sentiment value
        3.4.3    find out sentiment value ($sval$) for each content word:
                 $sval = \frac{1}{sns \sum_i (positivei + negetivei)}$
    3.5 find sentiment of the tweets using the formula:
        $sentiment = \sum_{j=1}^{n} (\frac{sval_j}{n})$
4.  if sentiment value of the tweet >= average sentiment value per tweet
    4.1 extract the tweet
5.  else
    5.1 go to step 2
6.  Exit
```

In the above algorithm, before calculating the sentiment score for any tweet, the stop words and punctuation marks are removed. After that, for each remaining term in the tweet, the set of synonyms (*synsets)* are obtained. Then for each *synset*, the sentiment value is calculated (step 3.4.3). Finally, the sentiment of the tweet is evaluated using step 3.5. Note that, only tweets with high sentiment values are included in rumor cluster. The sentiment value of a tweet is considered high if it is greater than or equal to the average sentiment value of all the tweets in the data set (see Step4).

Note that, though both H1(b) and H2 are evaluated based on sentiments associated with the tweet, they capture distinct dimensions of a rumor - H1(b) evaluates skepticism associated with the tweet, while H2 evaluates the actual sentiment associated with the tweet.

H3 test Social ties play a crucial role in disseminating information, irrespective of whether the information is false (or questionable) or true. Information received via friends or trusted people are often transmitted without knowing its authenticity (Friggeri et al. 2014). This often leads rumors to propagate faster within a network community where people know each other. The rumor spreaders often target these personal relationships as their source of rumor spreading (Kwon et al. 2013a). To detect such spreading, for each informative and mixture tweet, the *content words* related to *social relationship and hearsay* and their *synsets* are detected. Those tweets whose *content words* matches with at least one of social relationship or hearsay lexicons are extracted. Some content words related to social relationship and hearsay are shown in Table 8.

Note that one tweet can be matched by multiple rumor hypotheses; such tweets are considered once. As a result, a rumor cluster is obtained, containing a set of tweets that satisfy at least one of the rumor hypotheses. The percentage of

Table 8 Some content words related to social ties

Social relationship	Hearsay
Mate (plus synsets), Talk (plus synsets), child (plus synsets).	Listen (plus synsets), Hearing (plus synsets)

informative and mixture tweets satisfying each rumor hypothesis is shown in Fig. 2. After forming the rumor cluster two parallel tasks are performed to evaluate the probability of each tweet being a rumor. The tasks are:

- For each tweet in the cluster, the *mean factor ratio (Fr)* is computed using five rumor-causing factors of the model by Oh et al.
- For each tweet, *lex factor (LXF)* is computed.

4.4 Mean Factor Ratio

Mean factor ratio is calculated for a tweet by detecting the presence of five rumor causing factors of the model by Oh et al. (Anxiety, Source ambiguity, Content ambiguity, Personal involvement and Social tie) in the tweet. In this model, it has been claimed that if a tweet is a rumor, there is a high chance that at least three of the five rumor causing factors would be present. Hence, if a tweet is a rumor then it can be assumed that it must have high mean factor ratio of at least 0.60. We have used a modified version of RIAS (Bordia and DiFonzo 2004), a content analytic system, to predict the presence of each rumor-causing factor (of Oh et al. model) in the tweet. RIAS identifies rumors by introducing 14 categories of statements; for our proposed work, out of those 14 categories, 7 categories have been mapped with the five rumor causing factors of *Oh* et al. model (see Table 9). For a given tweet, if any factor is present then the value for that factor has been marked as 1 and 0 otherwise. Now, for each tweet, the Mean factor (Fr) ratio has been calculated as,

$$Fr = presenceof\,(Anxiety + Source\_ambiguity + Content\_ambiguity + Personal\_involvement + Social\_tie)/5 \qquad (3)$$

The value of *Fr* lies in [0, 1]. However, it is possible that not all rumors have the value of *Fr>=0.60*. Hence, another factor called *lexfactor* has been introduced.

4.5 Lex Factor (LXF)

Before calculating LXF the tweets in the cluster are preprocessed as follows. All URLs, stop words and punctuation marks are removed. All contracted forms like ‘ppl’, ‘cud’, ‘abt’, etc. are replaced with their expanded versions (‘people’, ‘could’, ‘about’).As a result of this preprocessing, each tweet in the cluster becomes a set of content words.

Now, for each tweet, the lex factor (LXF) is calculated as:

$$\left\{ \mathrm{LXF} = \frac{\mathrm{WR} + \mathrm{GR}}{\mathrm{RF}}, \text{if RF}\neq 0 \right\} \\ \mathrm{OR} \\ LXF = WR + GR, otherwise \qquad (4)$$

where,

WR (Word Relevance): Total number of distinct content words in the tweet/average number of content words per tweet in the cluster.

GR (Global Relevance): WR/Total number of content words in the tweet.

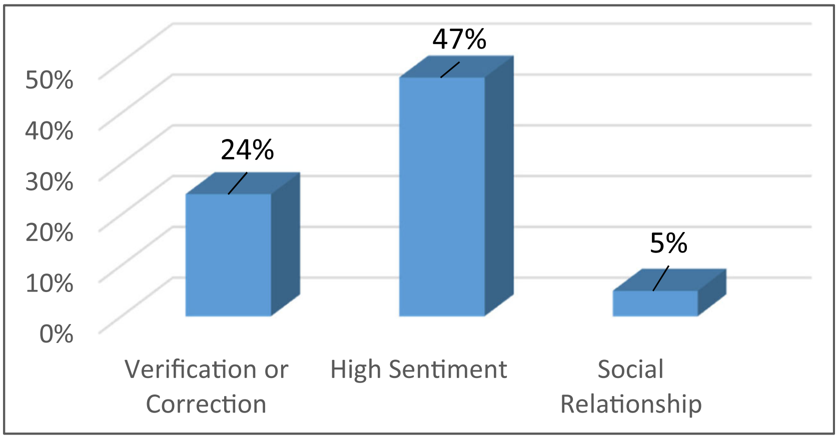


Fig. 2 Percentages of informative and mixture tweets that satisfy three rumor hypotheses

Table 9 Categories of statements in RIAS mapped with the five rumor causing factors of Oh et al. model

Statement type	Rumor causing factor of Oh et al.
Prudent Statements (usually refer statements like hearsay.)	Source ambiguity, Social Tie
Apprehensive Statements (express fear, dread, anxiety)	Anxiety
Interrogatory Statements (questions seeking information)	Source ambiguity, Content ambiguity
Directive Statements (suggests course of action)	Content ambiguity
Sarcastic Statements (ridiculed about the source of information)	Source ambiguity
Personal Involvement (person's involvement regarding the event)	Personal Involvement
Providing Information (Material relevant proofs like pictures, URLs etc.)	Source Ambiguity

RF (Relevance Factor): Total number of distinct disaster-relevant content words of particular disaster type (flood in this case). The value of LXF also lies in [0, 1] for a tweet.

After calculating Fr and LXF for each tweet, the probability *p* for each tweet (T_i) being a rumor has been calculated as follows:

$$Probability(rumor) = \sum_{T_i} Max(Fr, LXF) \tag{5}$$

After calculating the rumor probability of each tweet in the cluster, the tweets with high probability are separated. Tweets with high rumor probability value have been considered as predicted rumors by the proposed model.

It can be noted that the proposed model is an unsupervised learning model, which achieves its objective, i.e. rumor detection, through clustering/grouping of unlabeled data. The clustering algorithms are designed based on three rumor hypotheses, inferred from past literatures (Kwon et al. 2013a, b). In order to perform grouping, some patterns, similarities etc. among tweets have been used, which have also been inferred from past literatures. After grouping, real rumors have been extracted through a probabilistic analysis of *Mean Factor Ratio* and *Lex Factor* for each tweet. We would like to clarify that no training, development and test sets have been used in the present work, as our model is an unsupervised one.

5 Results & Analysis

The objective of this work is to identify tweets having high rumor probability. For this, we need to set up a critical point for the probability (*p*) in the experiment. We set the value in such a way that the proposed model can detect most rumors related to the two topics ("Crocodile escape" and "Nasa predicts El Nino"). We have compared our proposed technique at different probability threshold values as shown in Fig. 3. It can be observed from Fig. 3, that at $p> = 0.5$, a total of 983 rumors were predicted, out of which 185 real rumors related to two topics were detected. Out of the 185 real rumors, 84 tweets (24 beliefs and 60 disbeliefs) were related to "Nasa predicts El Nino" and 101 tweets (34 beliefs and 67 disbeliefs) were

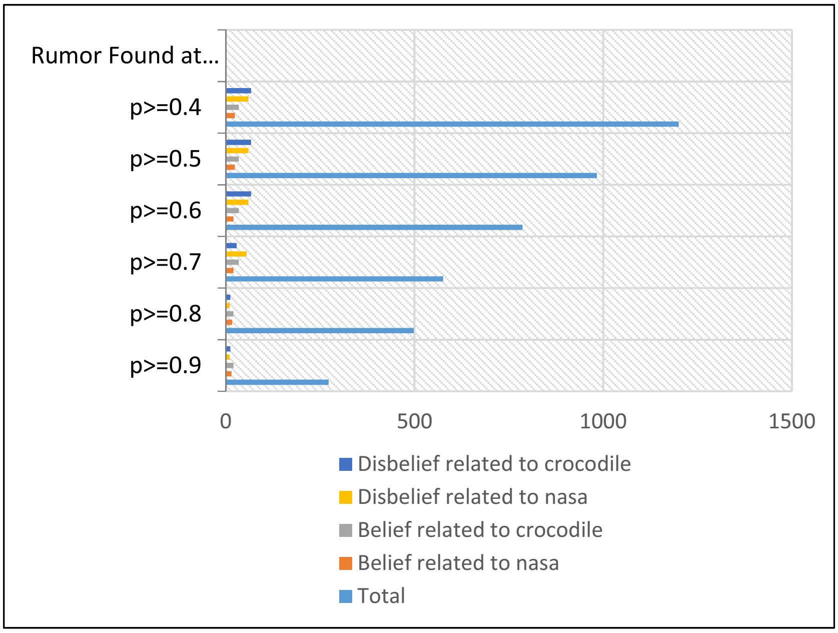


Fig. 3 Number of rumors related to the two topics and their probability (p) ranges

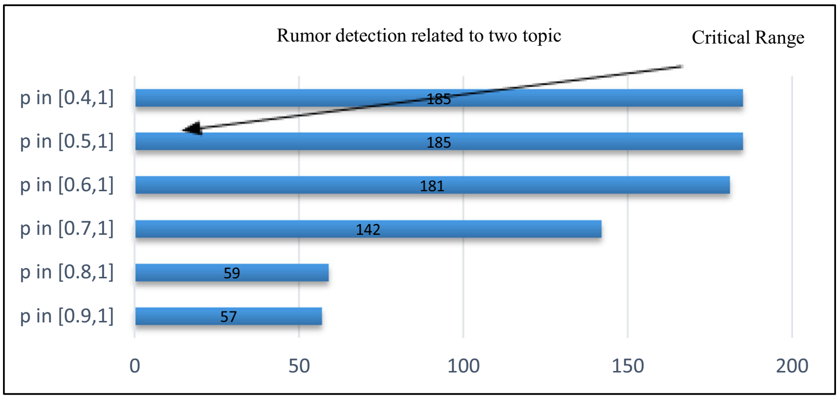


Fig. 4 Number of rumors related to the two topics and their probability (p) ranges

related to "Crocodile escape". It is also observed from Fig. 3, that the total rumor prediction value increases at $p > =0.4$ but the number of rumor tweets related to the two topics that are detected remains constant (see Fig. 3). This observation implies that; the model has predicted only a few spurious tweets (which are not related to either of the two rumor topics). So, it can be concluded that for the 'Chennai flood' dataset, the model provides a critical range of [0.5, 1] at which most of the rumors are detected (see Fig. 4). Hence, for 'Chennai flood' the proposed model achieves a critical point for rumors, which is 0.5 in this case.

However, the critical point for rumors may vary for different disaster events. As described in the previous section, the experts have detected 373 rumor tweets related to the two rumor topics. We have computed the relevance of experts' judgment and outcome of the proposed model, for the two rumor topics. It can be observed from Fig. 5, at the critical range, around 40% of experts' judgment, i.e., 126 rumors related to two topics matched with the prediction of the proposed model. As a result, the model is achieving a precision of 0.33. After implementing the proposed rumor model with different threshold values, an interesting outcome was noticed. Though the model has detected 185 rumors related to two rumor topics at critical point 0.5, only 126 of them matched with experts' judgment. 59 rumors remained unmatched, though they are clearly related to the two rumor topics. Hence, it can be concluded that our proposed model is able to detect more rumors related to the two topics than what the human annotators could identify.

5.1 Functional Completeness of Proposed Rumor Detection Model

As discussed earlier, the proposed model was able to detect 185 rumors related to two rumor topics. The model achieved a matching precision of 0.33 with experts' judgment, where 126 detected rumors matched with experts' judgments. However, the model has also missed 247 rumor-related tweets that were detected by the experts. Therefore, the proposed model achieves *partial functional completeness* with respect to experts' judgment while detecting rumors.

Now, the 185 detected rumors are considered as a *partially completed* (f_1) set. In addition, the 247 undetected rumors can also be as considered as another *partially completed* (f_2) set. Now, to make the set of rumors detected by the model *functionally complete,* we derive another partially *completed set* of 247 undetected rumors from the detected set using a supplementary operation Δ. Note, the additional Δ operation has to be performed by the proposed model to verify whether it is functionally complete or not. Therefore, Δ should be chosen in such a way that from $f_1$the model can achieve f_c by performing Δ. From the set f_2, four distinct sets named as '*undetected set (UDS)*' are prepared by extracting belief-tweets and disbelief-tweets related to the two rumor topics.

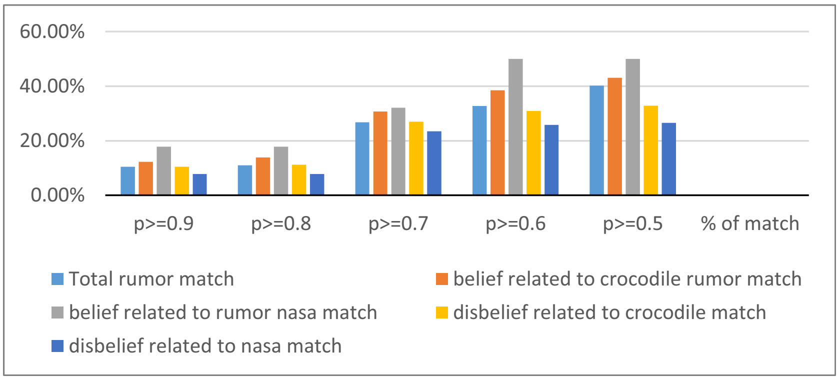


Fig. 5 Percentage of experts' judgment that matched with the model

Table 10 Closure rumor detection of DS(j)

Rumor type	DS(j) – detected by the model	Derived UDS(i) (Performing Δ operation on DS(j))
Belief-tweets related to 'crocodile escape'	People living near chennai crocodile park be alert 40 crocodile have been escaped from park Be safe [url]	RT @Darshan2206: 40 Crocodile out of the reserves Chennai people please be safe ChennaiFloodsChennaiRainsHelpchennaiairport @ibnlive: crocodiles are escaped from the park in Chennai [url] Crocodiles from chennai crocodile park 40 of them now on streets People be careful [url]
	20 crocodiles missing from zoo in Chennai	RT @BharathLodha2: Crocodile escaped from zoo in Chennai #ChennaiFloods 40 Crocodile escaped from zoo in chennai Please share and make awareness hope this should be fake [url]
Belief-tweets related to'NASA El Nino'	#ChennaiRainsHelp According to NASA Reports the chennai rainfall is mainly bcos of EL NINO which will get intense in coming days	RT @AMDIDX: It's EL Nino in Chennai Japan & NASA warns us, so there's possible for chennai will sumerge into ... [url] @RaiViveka BBC news is probably true The El nino one But it won't destroy chennai And NASA has nothing to do with this
	NASA announced that the big #tusunami was coming towards Chennai Mudunja escape peopleChennaiFloods	RT @IndianReporter: NASA image shows 2015 tsunami over Chennai [url]
Disbelief-tweets related to'crocodile escape'	No #crocodile I repeat NONE escaped the crocodile park in Chennai. Stop spreading rumors. #ChennaiFloods #ChennaiRainsHelp	Just to clear some existing rumor in chennai rain relief scenario. No crocodiles have escaped from crocodile ... [url] There Are No Crocodiles On The Loose In Chennai url #DailyScoopurl Don't go through rumours. There are NO crocodile open I'm chennai. [url]
	In Flooded Chennai, 'Crocodiles Have Escaped' Rumours Are Denied [url] via @ndtv	In Flooded Chennai, 'Crocodiles Have Escaped' Rumors Are Denied In Flooded Chennai, 'Crocodiles Have Escaped' Rumours Are Denied [url] #IndiaNewsndtv Ache Din Aane Wale
Disbelief-tweets related to 'NASA El Nino'	Dear people of Chennai,NASA hasn't predicted a hurricane or excess rainfall - Daily News & Analysis [url]	Dear people of Chennai, NASA hasn \ u2019t predicted a hurricane or excess rainfall [url] via @dna @dna No #NASA didn't predict more rains in Chennai [url]
	Fake WhatsApp Message Heavy Rains Chennai NASA Warning Don't Believe False News [url[[url]	WhatsApp message false, no warning by NASA regarding rain hurricane, high rainfall in Chennai [url]

That is, any *UDS(i)* contains undetected belief/disbelief set of rumors related to any of the two rumor topics. On the other hand, from the set f_1, four other distinct sets named as '*detected set (DS)*'are prepared using the same procedure. Each *DS(j)* contains the detected belief/disbelief set of rumors related to any of the two rumor topics.

After these steps, Δ operation is performed in terms of Jaccard similarity test (Niwattanakul et al. 2013) (over the set of distinct words contained in a tweet) between each belief set and disbelief set of *UDS* and *DS* in context of the same rumor topic. It is observed that for any rumor of *UDS(i)* there exists at least one rumor in the corresponding set of *DS(j)* whose Jaccard similarity value is greater than 0.60. Therefore, *UDS(i)* can be derived by performing a similarity test with *DS(j)*. Hence, it can be concluded that though the model was unable to detect some rumors (that were identified by the experts), an additional similarity operation can lead the model towards achieving *functional completeness*. Therefore, by introducing the Δ operation, the proposed model can achieve *functional completeness* in rumor detection (see Table 10).

Fig. 6 Comparison of proposed model with baseline rumor detection techniques in terms of rumor endorsement detection

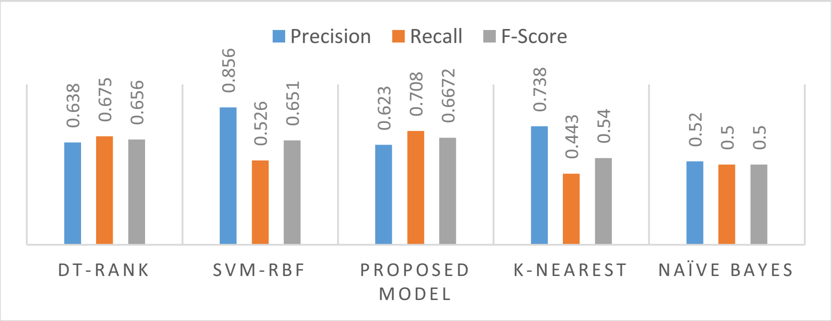


Table 11 Examples of rumors that were detected by the model, but undetected by human experts

Whatsapp message about crocodiles lose in #Chennai is FALSE. People need to stop spreading rumours, things are scary enough as it is.
RT @KishoreRT: People in perumbakkam area stay safe. Crocodiles in the water. Stay indoors. #ChennaiRescue #ChennaiRains #Chennai [url]
Now more than 20 Crocodiles all over chennai! It has escaped from the Crocodile Park! Breaking News all over! [url]
#ChennaiFloods ..declared a disaster zone already ..park broke and some 40 crocodiles on the loose...god save chennai...
Oh god 40 crocodiles escaped from the park! #Chennai please be alert, at this time u have to be strong @shrutihaasan
40 crocodile in chennai.... So be safe [url]
Dear nasa team please update about Chennai weather, we are getting more false alert... @NASA #chennairain

5.2 Early & Efficient Detection of Rumors

As mentioned earlier, the main objective of the proposed work is to detect rumor at early stage, i.e., at a time when no authorized source is involved in collecting news after any disaster. Manual verification is clearly not feasible for detecting rumors - as explained earlier, the two manual expert's needed120 hours to find out 373 rumor-related tweets. Some prior works relied on verification or correction posts, to identify potential rumors. However, *verification or correction statements appear in a later stage of a rumor's circulation.* If we discard all disbelief-tweets related to the two rumor topics, then out of 373, only 93 belief-tweets were detected by the human experts. Our proposed model has detected 58 belief-tweets out of the 93. In order to evaluate the quality of performance, some conventional evaluation metrics like *precision, recall, and F-score* are used. In terms of rumor endorsement detection, the proposed technique gained a precision of 0.623 by detecting 58 out 93 belief rumors (see Fig. 6). It can be observed from Fig. 6 that the performance of the proposed rumor detection technique is comparable with that of some popular baseline rumor detection techniques.

As mentioned earlier in this section, the proposed model was able to detect additional 59 tweets related to the two rumor topics, which the experts had failed to detect. In Table 11, some of those 59 tweets are shown. 27 out of these 59 tweets were belief-tweets (tweets that endorse the rumor). Now, for a fair evaluation of recall score, it is assumed that the experts have detected these 27 belief-tweets. Therefore, the number 27 has been added with 93 as well as with 58; this gives a recall score of 0.708. This observation implies that around 70% of total belief-tweets are detected by the proposed model, which is superior to the performance of prior baseline models for rumor detection, such as *DT-Rank* (Zhao et al. 2015) *and SVM-RBF* (Dayani et al. n.d.). This improvement of performance is due to the fact that, both these prior techniques featured only skeptics and temporal characteristics of rumor propagation. Hence, in large margin of time, these rumor detection models might be able to characterize the life cycle of a rumor (when the community has started questioning the rumor). But at early stage in post-disaster situation, due to lack of proper knowledge, users often believe information shared by their friends or by the people local to the affected region. As a consequence, they often retweet such information without questioning it. Hence, the performance of models like DT-Rank, SVM-RBF degrades at early stages. In terms of F-Score, which is the harmonic mean of precision and recall scores, the proposed model has achieved better result compared to other baseline approaches.

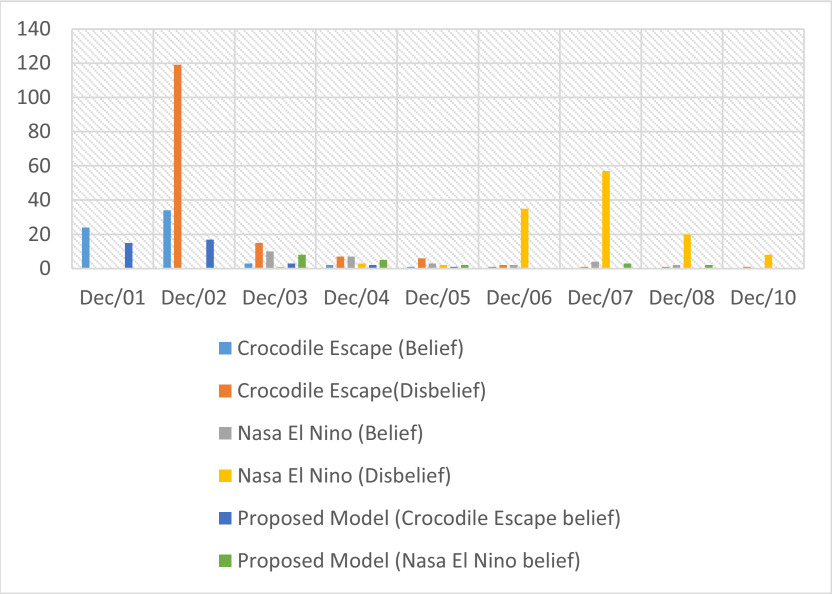
5.3 Behavioral Aspects of two Rumor Events with Respect to Time & Proposed Model

It has been mentioned earlier that from the collected 79,125 information and mixture tweets, a total of 373 tweets related to two rumor topics ("*Crocodile escape from zoo*" and "*NASA predicts El Nino*") were detected by the experts. We now analyze the temporal behavior of these two rumor events. To

Table 12 The chronologically earliest belief and disbelief tweets encountered in our dataset

First belief-tweet related to 'crocodile escape'	20 crocodiles missing from zoo in Chennai[Dec 01 17:29:58 2015]
First disbelief-tweet related to 'crocodile escape'	Crocodiles in d city and all other nonsense are just rumours. Plz Stop spreading thse in watsapp and here. Will add more cho… [Dec 02 06:54:40 2015]
First belief-tweet related to 'NASA El Nino'	Another disaster is coming for Chennai again…said by #nasa …. [Dec 03 08:18:44 2015]
First disbelief-tweet related to 'NASA El Nino'	Nasa Message regarding El Nino, which is going viral is a fake one… [url] [Dec 03 17:08:51 2015]
Tweets predicted by the model to be related to the rumors (belief-tweets)	20 crocodiles missing from zoo in Chennai [Dec 01 17:29:58 2015] 140 crocodiles escaped from #Crocodile_park ECR,moving around in #OMR #Velachery beware of it My Chennai [Dec 01 19:20:12 2015] #ChennaiRainsHelp According to NASA Reports the chennai rainfall is mainly bcos of EL NINO which will get intense in coming days [Dec 03 19:30:02 2015]

Fig. 7 Temporal aspects of the two rumor events (based on tweets detected by experts) and rumor detection of proposed model with respect to date



this end, we consider all belief-tweets and disbelief-tweets related to the two rumor events in the chronological order in which they were posted on Twitter. We then analyze the time difference between the *first belief-tweet* and the *first disbelief-tweet (related to the same rumor event) that were posted in the Twitter system.* Note that this analysis was only based on the tweets that were identified by the experts.

It can be observed from Table 12 that, for the 'crocodile escape' rumor, the time difference between the first belief-tweet and the first disbelief-tweet was approximately 14.25 h. For the 'El Nino' event, the time difference between first belief-tweet and the first disbelief-tweet is 8.40 h approx. These time differences create fertile ground for rumors to propagate rapidly. For instance, it can be observed from Fig. 7 that, on December 1st, 2015 along with first posted belief-tweet, 23 more belief-tweets related to 'crocodile escape' were posted by various users. On December 2nd, 34 more belief-tweets were posted. Though on December 2nd, more people started posting verification or correction tweets (see Fig. 7) as they found the topic as skeptic, and belief-tweets disappeared gradually in subsequent periods. It can also be observed from Fig. 7 that, the propagation of belief-tweets related to 'NASA El Nino' started from December 3rd, 2015 and disappeared gradually from December 5th, 2015 when more disbelief-tweets countered the rumor topic.

These observations clearly signify the argument about the delayed occurrences of verification or correction statements about any rumor event in Twitter network. Clearly, to restrict rumors from propagating in absence of any verification or correction statement, an effective rumor detection model is needed which can detect rumors at an early stage of its propagation. The proposed model could identify 32 belief-tweets related to 'crocodile escape' (see Fig. 7) that were posted during December 1–2, 2015. Whereas, 13 belief-tweets related to 'NASA El Nino' were detected during December 3–4, 2015. Some of the tweets matched at early stages of the events have been shown in Table 12.

Now, to evaluate the *accuracy* of the proposed model with respect to experts' decision, the following observations can be made:

i. The experts' detected a total of 75 belief-tweets related to the two rumor events at early stage.
ii. Out of the 75 belief-tweets detected by the experts, the proposed model detected 32 + 13 = 45 belief-tweets related to the two rumor events (see Fig. 7). The model could not detect the other 30 belief-tweets.

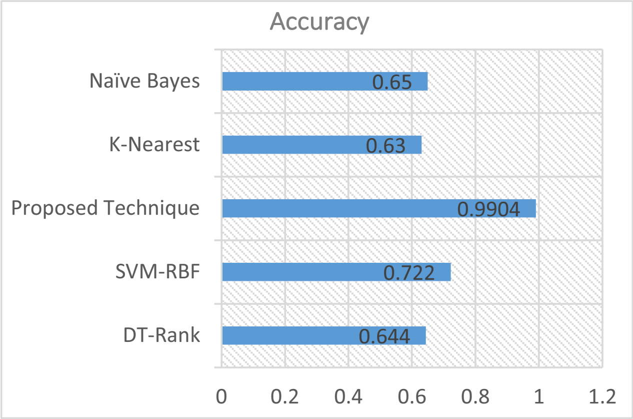


Fig. 8 Accuracy of proposed model with respect to baseline techniques in the time duration of 2 h in terms of rumor endorsement detection

iii. The proposed model predicted 784 tweets as related to the rumor events at early stage, which were not identified by the experts.
iv. Both proposed model and experts' have analyzed 78,773 situational and mixture tweets at early stage (the tweets appearing in later stages, when many disbelief-tweets are being posted, are not considered in this analysis)

Now, we calculate the *accuracy* of the proposed model, using the *confusion matrix* shown below (combined, for the two rumor events). It can be observed that the proposed model wrongly predicts 784 tweets as belief-tweets, whereas for 77,989 cases both experts and proposed model have found the tweets to be unrelated to rumors. Hence, the accuracy of proposed rumor detection model can be evaluated to be 0.9904 for the task of early detection of rumors, which is higher than those of several baseline models (see Fig. 8). The observations altogether show the necessity of proposed rumor detection model for early stages in the aftermath of any disaster.

Prediction of the Model			
	TRUE	FALSE	Experts' Opinion
TRUE	45	30	
FALSE	784	77,989	

6 Conclusion

Rumor detection from Online Social Network at early stages after a disaster is of much importance. In the proposed work, we have designed a technique for efficient detection of rumors at early stages in the aftermath of a disaster. The performance of the model has been analyzed based on the tweets posted during Chennai flood 2015.We found that our proposed rumor detection technique performs well and is able to find out rumors at early stages, even before contradicting or interrogating posts are posted. In future, we plan to evaluate the performance of our proposed model for other types of disaster scenarios. We also plan to devise effective rumor control strategies that can be adopted after early detection of a rumor.

Acknowledgments This research was partially supported by a grant from the Information Technology Research Academy (ITRA), DeITY, Government of India (Ref. No.: ITRA/15 (58)/ Mobile/DISARM/ 05).

References

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent technology (WI-IAT), 2010, vol. 1, pp. 492-499. IEEE, 2010.

Bao, Y., Yi, C., Xue, Y., Dong, Y. (2013). A new rumor propagation model and control strategy on social networks. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1472-1473. ACM.

Bordia, P., & DiFonzo, N. (2004). Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly, 67*(1), 33–49.

Buckner, H. T. (1965). A theory of rumor transmission. *Public Opinion Quarterly, 29*(1), 54–70.

Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers, 19*(5), 993–1012.

Corvey, W.J., Vieweg, S., Rood, T., Palmer, M. (2010). Twitter in mass emergency: what NLP techniques can contribute. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, 23–24. Association for Computational Linguistics.

Dayani, R., Chhabra, N., Kadian, T., Kaushal, R. (2015). Rumor detection in twitter: An analysis in retrospect. In Proceedings of IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 1-3. IEEE.

Dayani, R., Chhabra, N., Kadian, T., Kaushal, R. (n.d.). Rumor: Detecting Misinformation in Twitter, 3rd Security and Privacy Symposium, 2015. IIIT Delhi, Poster Session

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. Data engineering workshop, 2008 with IEEE ICDEW.

Dhanjal, C., Blanchemanche, S., Clemençon, S., Rona-Tas, A., Rossi, F. (2011). Information diffusion within social networks.

Doerr, B., Fouz, M., & Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM, 55*(6), 70–75.

Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." In Proceedings of LREC, vol. 6, pp. 417–422. 2006.

Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J. (2014). Rumor Cascades. In Proceedings of AAAI ICWSM.

Kimmel, A.J. (2013). Rumors and rumor control: A manager's guide to understanding and combatting rumors. Routledge.

Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y. (2013a). Prominent features of rumor propagation in online social media. In proceedings of IEEE international conference on data mining (ICDM), 1103-1108. IEEE.

Kwon, S., et al. (2013b). Aspects of rumor spreading on a microblog network. In Proceedings of International Conference on Social Informatics. Springer International Publishing.

Laniado, D., Volkovich, Y., Scellato, S., Mascolo, C., & Kaltenbrunner, A. (2017). The impact of geographic distance on online social interactions. *Information Systems Frontiers*, 1–16. https://doi.org/10.1007/s10796-017-9784-9.

Liang, G., He, W., Xu, C., Chen, L., & Zeng, J. (2015). Rumor identification in microblogging systems based on users' behavior. *IEEE Transactions on Computational Social Systems, 2*(3), 99–108.

Liu, F., Burton-Jones, A., Xu, D. (2014). Rumors on social media in disasters: Extending transmission to retransmission. In PACIS, 49.

Ma, J., Gao, W., Wei, Z., Lu Y., Wong, K-F. (2015). Detect rumors using time series of social context information on microblogging websites. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1751-1754. ACM.

Ma, J., Gao, W., Mitra, P., Kwon S., Jansen B.J., Wong, K.-F., Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In Proceedings of IJCAI.

Mendoza, M., Poblete, B., Castillo, C. (2010). Twitter under crisis: Can we trust what we RT?. In Proceedings of the first workshop on social media analytics (SOMA '10). ACM, 71–79.

Myers, L., and Sirois, M.J. (2006). *Spearman correlation coefficients, differences between*. Wiley StatsRef: Statistics Reference Online.

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications, 374*(1), 457–470.

Niwattanakul, S., et al. (2013). Using of Jaccard coefficient for keywords similarity. Proceedings of the international multi conference of engineers and computer scientists. Vol. 1.

Oh, O., Hazel, K.K., Rao, H.R. (2010). An exploration of social Media in Extreme Events: Rumor theory and twitter during the Haiti earthquake. In Proceedings of *ICIS* 2010*.* 231.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *Research Article MIS Quarterly, 37*(2), 407–426.

Popescu, A-M., and Pennacchiotti, M. (2010). Detecting controversial events from twitter. Proceedings of the 19th ACM international conference on information and knowledge management. ACM.

Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, 1589–1599.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychology, 46*(5), 484–496.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., Ghosh, S. (2015). Extracting situational information from microblogs during disaster events: A classification-summarization approach. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 583-592. ACM.

Sen, A., Rudra, K., Ghosh, S. (2015). Extracting situational awareness from microblogs during disaster events. In Proceedings of Social Networking Workshop, with IEEE International Conference on Communication Systems and Networks (COMSNETS).

Tripathy, R. M., Bagchi, A., Mehta, S. (2010). A study of rumor control strategies on social networks. In Proceedings of the 19th ACM international conference on information and knowledge management, pp. 1817-1820. ACM.

Vosoughi, S. (2015). Automatic detection and verification of rumors on twitter. PhD dissertation, Massachusetts Institute of Technology.

Yang, F., Liu, Y., Yu, X., and Yang, M. (2012). Automatic detection of rumor on Sina Weibo. In proceedings of the ACM SIGKDD workshop on mining data semantics (MDS '12), article 13, 7 pages.

Yang, Z., et al. (2015) Emerging rumor identification for social media with hot topic detection. In Proceedings of Web Information System and Application Conference (WISA). IEEE.

Zhao, Z., Resnick, P. Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In Proceedings of the 24th International Conference on World Wide Web, 1395–1405.

Zhao, W., Zeng, Q., Zheng, G., & Yang, L. (2017). The resource allocation model for multi-process instances based on particle swarm optimization. *Information Systems Frontiers, 19*(5), 1057–1066.

Tamal Mondal is a Ph.D Scholar at Department of Computer Science & Engineering, National Institute of Technology, Durgapur funded by ITRA Media Lab Asia at Department of Computer Application, Kalyani Government Engineering College, India. His research areas include NLP, Social Networks, UAV path planning. His publications include ACM SIGSPATIAL, IEEE CISS, Peer to Peer Networking, Springer etc.

Prithviraj Pramanik is a Visvesvaraya Ph.D Fellow at National Institute of Technology Durgapur in India. His research interests include Post Disaster Management Systems, Air Pollution Monitoring & Data Science. He has been trying to implement ICT based disaster management solutions in India. He has publications in IEEE PerCom, IEEE CISS. He loves travelling to new places, understand new cultures & have new cuisines.

Indrajit Bhattacharya is an Assistant Professor at Kalyani Government Engineering College (KGEC) in West Bengal, India. He has completed his Ph.D from Jadavpur University in West Bengal, India in 2014. He obtained his Master in Computer Science & Technology from the University of Calcutta, West Bengal. He has a teaching and research experience of more than 14 years in different institutes of repute. He is the Principal Investigator of the Project titled DiSARM (Post Disaster Situation Analysis and Resource Management using Delay Tolerant Peer-to-Peer Opportunistic Networks) at KGEC, funded by Information Technology Research Academy, Media Lab. Asia, Govt. of India. His research interests include delay tolerant networks, wireless networks, sensor networks, and radio frequency identification.

Naiwrita Boral is a Final Year MCA Student at Kalyani Government Engineering College in India. Her research areas include Soft Computing Techniques, Disaster Resource Management, Data Mining etc. She is also a member of KGEC ITRA Research Lab. She has a publication in IEEE CISS.

Saptarshi Ghosh is an Assistant Professor in the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India. He has also been an Assistant Professor at the Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, India. He has received Ph.D from the Department of CSE, Indian Institute of Technology, Kharagpur, India in 2013. He was a Humboldt Post-doctoral research fellow at the Max Planck Institute for Software Systems, Germany. His research areas include Online Social Media, Information Retrieval, Data Mining, Natural Language Processing, and Complex Network Analysis.