@Twitter Credibility Ranking of Tweets on Events #breakingnews

Aditi Gupta, Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology, Delhi, India
{aditig, pk}@iiitd.ac.in
precog.iiitd.edu.in

ABSTRACT

Large amount of content is generated on online social networking and micro-blogging services daily; Twitter is one such micro-blogging service. Twitter has evolved from being used for conversing with friends and expressing opinions into a medium to share and disseminate information about current events. Events in the real world creates a corresponding spur of tweets in Twitter. In this paper, we analyzed tweets corresponding to fourteen major news events of 2011 around the globe. We empirically show that the properties of information diffusion (via retweets, and URLs) on Twitter differs during crisis and non-crisis events. Using supervised machine learning and relevance feedback approach, we show that ranking of tweets based on Twitter features can aid in assessing credibility of information in messages posted about an event. We found that both message and source based features help in predicting the rank of the tweets. The performance of ranking algorithm was significantly enhanced by using reranking strategy as it provided context specific (unigrams) features to the algorithm. To this best of our knowledge, this is the first work to study credibility of content on Twitter at the tweet level and exploring an automated ranking framework to predict rank of tweets according to their credibility.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information storage and retrieval—*Information Retrieval*; K.4.1 [**Computing Milieux**]: Computers and society—*Public policy issues*

General Terms

Experimentation, Measurement

Keywords

Crisis management, Credibility, Online social media, Trust

1. INTRODUCTION

With the evolution of online social networking and micro-blogging mediums, two major changes have occurred in the landscape of the Internet usage. Firstly, the Internet is replacing traditional media like television and print media as a source for obtaining news and information about current events [15]. Secondly, the Internet has provided a platform to common people to share information and express their opinions. The quick response time and high connectivity speed have fueled the propagation and dissemination of information by users on online social media services like Facebook, Twitter, and YouTube. Twitter is a micro-blogging service, which has gained popularity as a major news source and information dissemination agent over last few years. Users on Twitter, create their public / private profile and post messages (also referred as tweets or statuses) via the profile. The maximum length of the tweet can be 140 characters. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user (source) who posted the tweet. Studies have explored and highlighted the role of Twitter as a news media and a platform to gauge public sentiments [15, 18].

One major difference between dissemination of news or information through traditional media and Twitter is that, Twitter is a crowd sourced medium. Users on Twitter act like its sensors, and fill in the information gap about an event. In contrast to television, print or news websites where the source of information are few and known (i.e. credible). Due to the anonymous and unmonitored nature of the medium, a lot of content generated on Twitter maybe incredible. During an event, when a user types a query on the Twitter search (e.g. UK riots) or clicks on a related trending topic (e.g. #ukriots) , all tweets matching the query words are displayed to the user. The above search results display tweets ordered from top to bottom, in ascending order of time difference between the query and the time when the tweet was posted. When an event of sizable magnitude and impact occurs, thousands of tweets are posted per hour. Due to the large amount of content generated on Twitter, it is hard to identify the tweets with credible information manually. We propose an automated ranking scheme to present the user a ranked output of tweets according to the credibility of information in the tweet.

In this paper, we consider two kinds of news events – crisis and non-crisis. The credibility and quality of information often plays more critical role in extreme circumstances (crisis events) than daily life news events (non-crisis); therefore,

we specifically consider crisis events for analysis and use the non-crisis events to compare and contrast the trends in both. Previous research work have explored role of Twitter during certain singleton crisis events like forest fires, earthquakes, etc. [17, 23]. There has not been work which highlights and contrasts the nature and behavior of online social media during crisis v/s non-crisis events. Emphasis has been to utilize the information during events to increase awareness about the event. The effective utilization of any information from tweets, would remain subjective to filtering out spam and other forms of non-informative tweets like those expressing personal opinions. A snapshot of sample tweets captured by us for the event 'Hurricane Irene' are shown in Figure 1. All three tweets contain the words matching to the event and were posted while *Hurricane Irene* was the trending topic on Twitter. The top-left tweet provides correct and credible information about the event. The top-right tweet, is related, but contains no information about the event, it expresses the personal opinion of the user. The bottom tweet is spam even though it contains the related words, it includes an URL to an advertisement to sell a product.

Figure 1: Sample tweets in our dataset for the event 'Hurricane Irene'.

We envision, understanding the credible (incredible) information on Twitter to be useful for devising strategies to mitigate the wide-spread of incredible information (like fake news or rumors). To the best of our knowledge, this is the first work to study credibility of content on Twitter at the level of tweets and using an automated ranking technique to compute rank of tweets according to credibility. The main contributions of this paper are:

- We empirically show that information diffusion (via retweets, and URLs) differs between crisis and non-crisis situations.
- Using supervised machine learning and relevance feedback approach, we show that ranking of tweets based on Twitter features can aid in assessing credibility of information in messages posted about an event.
- We conclude that both message and source based features help in predicting the rank of the tweets.

The rest of the paper is organized as follows: Section 2, describes the closely related work, exploring the role of Twitter as a news media, quality of information on Twitter and relevance ranking in Web 2.0. Section 3 explains the methodology that we used in collecting data, events selection, and the coding scheme used for annotating the tweets. Section 4 describes the feature sets, ranking algorithm and evaluation metric. Section 5 discusses the statistical analysis, network analysis and ranking performed on the tweets for the events. Section 6 summarizes the results from our analysis and highlights the implications of our results. The last section presents the limitations, and future work of the paper.

2. RELATED WORK

Three prior research directions form the basis for our paper – role of Twitter during news events, factors that affect the quality of information on Twitter and automated mechanisms for relevance ranking of documents on Web 2.0.

Role of Twitter during news events

Prominence of Twitter as a news media, was established by Kwak et al., according to their work, 85% topics discussed on Twitter are related to news [15]. The patterns extracted by them showed relation between user behavior and tweet activity properties like number of followers and followees to the tweeting / re-tweeting numbers. Zhao et al. in their work, used unsupervised topic modeling to compare the news topic from Twitter versus New York Times, a traditional news dissemination medium [26]. They showed that Twitter users are relatively less interested in world news, still they are active in spreading news of important world events.

In recent years, many researchers have explored the role of online social media during crisis situation. Mendoza et al. used the data collected during 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity [17]. Their results showed that propagation of tweets related to rumors versus true events differed and could be used to develop automated classification solutions. Also, the tweets related to rumors contained more questions versus news tweets spreading correct news. Longueville et al. analyzed Twitter feeds during forest Marseille fire event in France and showed information from location based social networks can be used to acquire spatial temporal data that can be analyzed to provide useful localized information about the event [7]. Sakaki et al. investigated on how tweets can be used as social sensors, to predict the epicenter and impact area for earthquakes [22]. They used Kalman and particle filtering for location estimation in ubiquitous / pervasive computing.

Another closely related work, was done by Oh et al., they analyzed Twitter stream during the 2008 Mumbai terrorist attacks [19]. Their analysis showed how information available on online social media during the attacks aided the terrorists in their decision making by increasing their *social awareness*. A team at NICTA has been working on developing a focused search engine for Twitter and Facebook that can be used in humanitarian crisis situation. [1] Hughes et al. in their work on analyzing behavior of Twitter users during emergencies, compared the properties of tweets and users during an emergency to normal situations [12]. Their results showed that the use of URLs in tweets increase and @-mentions decrease during emergency situations. An automated framework to enhance situational awareness during emergency situations was developed by Vieweg et al. They extracted geo-location and location-referencing information from users tweets that

[1] http://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/

helped in increasing situation awareness during emergency events [24]. Another very similar work to the above was done by Verma et al., who used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [23].

Quality of information on Twitter

One of the major concerns about quality of information on Twitter, is due to the presence of spam and no restrictions in creating content online. Techniques to filter out spam and phishing from Twitter has been studied and various effective solutions have been proposed [1, 6, 10, 25]. A system called Truthy [2], was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a set of tweets to detect political smears, astroturfing, misinformation, and other forms of social pollution [21]. In their work, they presented certain cases of abusive behavior by Twitter users. Castillo et al. showed that automated classification techniques can be used to detect news topics from conversational topics and assess their credibility based on various Twitter features [4]. The precision and recall of 70-80% was achieved using J48 decision tree classification algorithms. Canini et al. analyzed usage of automated ranking strategies to assess credibility of sources of information on Twitter for any given topic [3]. They observed that content and network structure act as prominent features for effective ranking of users. Gupta et al. in their work on analyzing tweets posted during the terrorist bomb blasts in Mumbai (India, 2011), showed that majority of sources of information are unknown and with low Twitter reputation (less number of followers) [11]. This highlights the need to develop automated mechanisms to assess credibility of information on Twitter.

Relevance ranking in Web 2.0

Ranking techniques have been used widely to rank URLs, content and users on various Web 2.0 platforms. Page et al. developed a PageRank algorithm for webpages on the Internet, they used the number of out-links and in-links of a webpage to calculate its relative relevance to a query [20]. Duan et al. in their paper proposed a supervised learning approach for ranking tweets based on certain query inputs [9]. They used content and non-content features (like authority of users) to rank tweets according to their relevance to a topic. Their work used Rank-SVM technique and extracted the best features, that resulted in good ranking performance. The three prominent features were: whether a tweet contains URL, the length of tweet (number of characters), and authority of user account. In the paper, on URL recommendation for Twitter, Chen et al. built a tool called *zerozero88*, [3] which recommends URLs that a particular Twitter user might find interesting [5]. They showed, how topic relevance and social voting parameters help in effective recommendations. Dong et al. worked on using inputs from Twitter to improve recency and relevance ranking for search engines using Gradient Boosted Decision Tree (GBDT) algorithm [8]. They showed how in addition to existing features used to rank URLs on web, additional information from Twitter can be used to enhance the ranking of URLs on the Web.

[2]http://truthy.indiana.edu/

[3]http://zerozero88.com/

Data exploration and characterization of Twitter activity during individual events (crisis and non-crisis) have been studied. The work done to assess credibility of information on Twitter, have explored credibility of trending topics and users. In this paper, we use automated ranking techniques to assess credibility at the most atomic level of information on Twitter, i.e. at a tweet level. Using supervised machine learning and relevance feedback approach, we show that ranking of tweets based on Twitter features (topic and source) can aid in assessing credibility of information in messages posted about an event.

3. METHODOLOGY

We collected data posted and shared on Twitter. In this section, we discuss the data collection setup, the process in which we selected the events (crisis and non-crisis), and the coding scheme that we used to annotate the tweets. Figure 2 describes the architecture of the methodology and analysis performed in our work.

3.1 Data Collection

We collected data from Twitter using the *Streaming API*. [4] This API enables researchers to extract tweets in real-time, based on certain query parameters like words in the tweet, time of posting of tweet, etc. To obtain query terms, we used, *Trends API* from Twitter, which returns top 10 trending topics on Twitter. [5] We queried *Trends API* after every 3 hours for the current trending topics, and collected tweets corresponding to these topics as query search words for the *Streaming API*. We collected tweets corresponding to a topic until the time it remained as a trending topic. For assessing credibility of news topics on Twitter, Castillo et al. also used a similar framework to collect tweets using current trending topics [4]. We considered both worldwide and local trending topics from Twitter. We collected data of over 35 million tweets by more than 6 million users in the time period from 12^{th} July, 2011 to 30^{th} August, 2011. Table 1 gives the descriptive statistics of the data collected.

Table 1: Descriptive statistics of the Twitter dataset.

Total tweets	35,748,136
Total unique users	6,877,320
Tweets with URLs	4,973,457
Number of singleton tweets	22,481,898
Number of re-tweets / replies	13,266,238
Trending Topics (unique)	3,586
Start date	12^{th} July, 2011
End date	30^{th} August, 2011

3.2 Events Selection

Using the methodology described in Section 3.1, in total 3,586 unique trending topics were obtained. We shortlisted 14 major events that occurred all around the globe between July 12^{th} and August 30^{th}, 2011. We sorted all trending topics by descending order of number of tweets in each topic; we selected topics corresponding to fourteen events. Each

[4]https://dev.twitter.com/docs/streaming-api. API stands for Application Programming Interface.

[5]https://dev.twitter.com/docs/api/1/get/trends

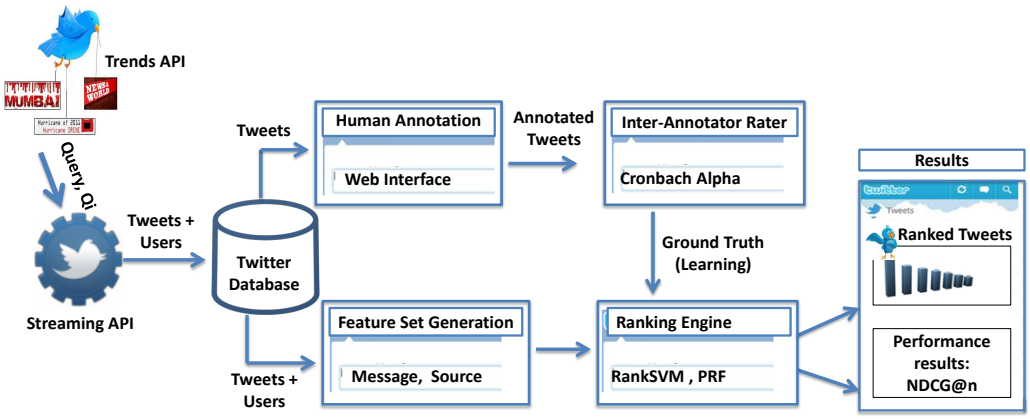


Figure 2: Describes the architecture of the methodology and analysis performed in this paper.

event had one or more trending topics associated with it, for example trending topics related to *Debt and downgrading crisis in US* were *AAA to AA, S&P, etc.* For each event, we consider tweets containing the words in trending topics to be the set of tweets for that event. Table 2 describes the fourteen events that we selected, the corresponding trending topics for the event and the number of tweets for each event. To ensure that we select events with high impact and relevance, we applied following minimum criterion for selecting an event for analysis:

- An event which had minimum of 25,000 tweets were considered. For example, one of the events was the *Riots in UK* in August, 2011, our system collected 542,685 tweets for the event.
- Topics corresponding to the event which were trending for minimum 24 hours as a country or worldwide trends on Twitter. For example, *Hurricane Irene* was a trending topic on Twitter for 150 hours.

We selected seven events that represent crisis [6] events like natural hazards and bomb blasts, and seven events which represented generic news events like political and financial news. We selected events covering various domains of news events like political, financial, and entertainment news. Similarly while choosing crisis events we chose events from natural disasters, terrorist strikes, political and financial crisis. Figure 3 shows the fourteen events analyzed in this paper; the date in the figure represents the first tweet that we have for the respective event in our dataset.

3.3 Coding Scheme

This section describes how we annotaed the set of tweets for each event. We took help from human annotators to obtain the ground truth regarding the presence of credible information in tweets related to a news event. Human annotation for understanding the ground truth is a well-established research methodology [4]. For the fourteen selected events, we picked a random sample of 500 tweets per topic. We restricted our analysis to tweets in English language; we selected tweets by those users who had set English as their language on Twitter. Though, there were some tweets by users which were in languages other than English, we provided the annotators with a *Skip tweet* option to avoid such tweets. For the purpose of annotation, we developed a web interface (See Figure 4) and we provided each annotator with an authenticating login and password.

To assess the presence of credible information, if any, we asked the human annotators to select one of the following options for each tweet:

- Tweet contains information about the event. Rate the credibility of information present:
  - Definitely Credible
  - Seems Credible
  - Definitely Incredible
  - I can't Decide
- Tweet is related to news event but contains no information
- Tweet is not related to news event
- Skip tweet

We provided the annotators with the definition of credibility [7] and then explained the above mentioned categories using an illustrative example. For each of the events, we provided a 5-10 line description of the event along with two URL links to news articles on the event featured in premier news websites likes CNN, Guardian and BBC (See Figure 4). We got tweets in each event (500 tweets) annotated by three different annotators.

To check the reliability of results obtained via annotation, we computed the Cronbach Alpha score. The overall Cronbach alpha value for inter-annotator agreement for all 7000 (14 events * 500 tweets) tweets was 0.748. Value for alpha greater than 0.7 is considered good [16], it implies there is a high agreement between the annotators of our experiment. For the final scores for each tweet, we selected the majority

[6]Crisis is defined by Webster dictionary as *a paroxysmal attack of pain, distress, or disordered function* (http://www.merriam-webster.com/dictionary/crisis).

[7]Oxford dictionary defines the term credibility as "the quality of being trusted and believed in". In the context of this research, we aim to assess the credibility of the information in the content of a tweet (message) by a user on Twitter. A tweet is said to contain credible information about a news event, if you trust or believe that information in the tweet to be correct / true.

Table 2: Fourteen crisis and non-crisis events selected for the time period from 12 July to 30 August, 2011.

Label	Crisis Events	Tweets	Trending Topics	Description
C1	UK Riots	542,685	#ukriots, #londonriots, #prayforlondon	Riots in United Kingdom caused 5 deaths, 16 civilian and 186 police injuries
C2	Libya Crisis	389,506	libya, tripoli	Rebels opposing Col. Qaddafi seized Zawiyah
C3	Earthquake in Virginia	277,604	#earthquake, Earthquake in SF	Earthquake of magnitude 5.8 hit the Piedmont region of the U.S. state of Virginia.
C4	US Downgrading	148,047	S&P, AAA to AA	Debt crisis in the US, led Standards & Poor to downgrade it from AAA to AA-plus
C5	Hurricane Irene	90,237	Hurricane Irene, Tropical Storm Irene	Hurricane Irene in US caused 55 deaths and a damage of US $10.1 billion
C6	Indiana State Fair Tragedy	49,924	Indiana State Fair	Five people died and 40 were injured in a stage accident at the Indiana State Fair.
C7	Mumbai Blast, 2011	32,156	#mumbaiblast, Dadar, #needhelp	Three bomb blasts in Mumbai (India) on 13th July, 26 people died and 130 injured
Label	**Non Crisis Events**	**Tweets**	**Trending Topics**	**Description**
NC1	JanLokPal Bill Agitation	182,692	Anna Hazare, #janlokpal, #anna	An anti-corruption movement against the Government of India.
NC2	Apple CEO Steve Jobs resigns	158,816	Steve Jobs, Tim Cook, Apple CEO	Apple's stock dropped 7% when Steve Jobs resigned as its CEO
NC3	Google acquires Motorola Mobility	68,527	Google, Motorola Mobility	Google buying Motorola Mobility in a $12.5bn cash deal, was a huge acquisition
NC4	News of the World Scandal	67,602	Rupert Murdoch, #murdoch	The News International phone hacking scandal exposed Rupert Murdoch
NC5	Abercrombie & Fitch stocks drop	54,763	Abercrombie & Fitch, A&F	Abercrombie & Fitch stocks drops 9% after a controversy
NC6	Muppets Bert and Ernie were gay	52,401	Bert and Ernie	Rumors circulated that muppet pair Ernie and Bert, are a gay couple
NC7	New Facebook Messenger	28,206	Facebook Messenger	Facebook launched a new messenger for mobile users

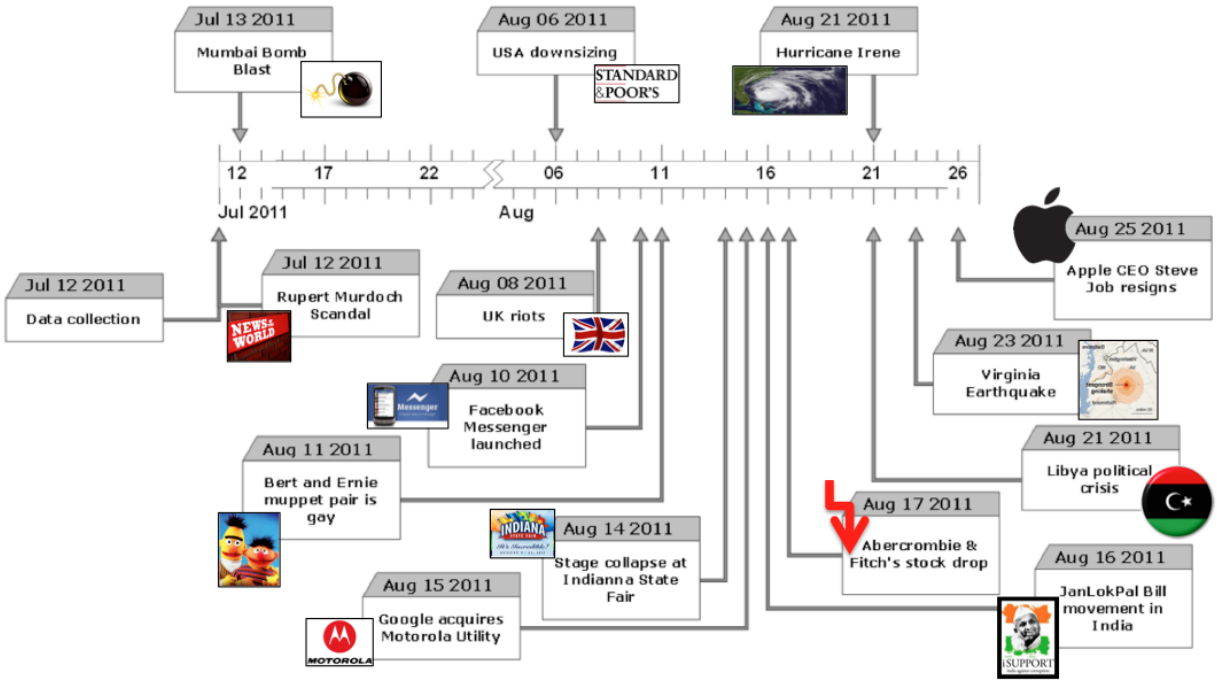


Figure 3: Timeline representing the fourteen events analyzed in this paper. We selected events from various categories like financial, political, natural hazards and technological news. The date in the figure represents the first tweet that we have for the event in our dataset.

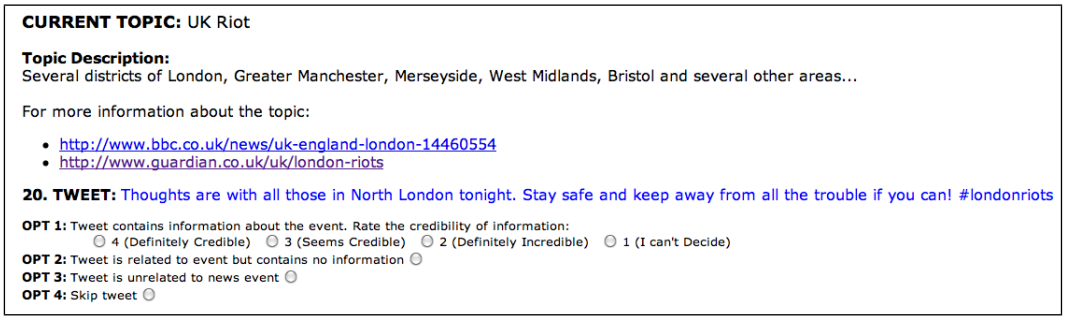


Figure 4: Options presented for each tweet by the web interface developed for the annotation task. [http://precog.iiitd.edu.in/annotation/login1.php]

score for a tweet (that is, value given by at least 2 annotators); we discarded all tweets for which all three annotators gave different scores. After removing tweets that had all three annotators giving different ranking score and tweets which annotators decided to skip, in total we obtained 5,578 tweets in our final annotated dataset.

4. RANKING

We propose an automated ranking framework for identifying credible information related to news events on Twitter. When a user types a query on the Twitter search or clicks on a trending topic, the tweets matching the query words are displayed for the user. The tweets are ordered from top to bottom according to time difference between when the query was fired and the time when the tweet was posted. As an event of sizable magnitude and impact occurs, it results in thousands of tweets being posted on the topic per hour. Out of the total volume of content being generated, it is tough to identify the tweets with credible information. We propose an automated ranking scheme to present to the user, a ranked output of tweets ordered according to the credibility of information provided in them. We use a combination of supervised machine learning and relevance feedback approach to rank tweets. We analyze the effectiveness of Twitter based features (message and source level) to rank tweets according to their information quality. As a next step, we propose an enhancement to the above ranking technique by using pseudo feedback relevance reranking scheme. We use SVM ranking algorithm to build a model for credibility of information in tweets. Ranking SVM algorithm is an extension of SVM classifier traditionally used for the classification task [14]. We used the SVM^{Rank}implementation code provided by researchers at Cornell University. [8] SVM^{Rank} trains a Ranking SVM on the training set, and outputs the learned rule to a model file. The ground truth for the task is obtained from the human annotated tweet scores. Based on the learned model, it predicts a ranking score that are written to the output file. We perform four-fold cross validation of our results.

4.1 Features

Any post or tweet on Twitter is characterized by two basic characteristics, the features of the message itself, and the properties of the user who posted the message. Table 3 presents each feature in the message and the user. We consider following two types of features as input to the ranking algorithm:

- **Tweet or message level features [FS1]**: The 140 character messages posted by users contain data and meta-data related to it. The text of tweet contains the words, URLs, hashtags and various other properties which are considered as a feature of the tweet. Meta data related to tweet includes whether the tweet is a reply or retweet or number of retweets of a tweet. We do not consider text semantic features here.
- **Source or user level features [FS2]**: The features representing the source of tweets or users of Twitter are properties such as number of friends, followers and status messages of the user.

4.2 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) also known as Blind Relevance Feedback, is a prominent reranking technique used in information retrieval tasks to improve the performance of ranking results [2]. The basic idea of PRF is to extract K ranked documents and then rerank those documents according to a defined score. In our algorithm, we extract most frequent unigrams from the top K tweets and use the text similarity between the most frequent unigrams and K tweets to rerank them. The improvement achieved by reranking using PRF is highly dependent on the quality of top K results given by the ranking algorithm; hence we apply PRF to the final set of results obtained by all Twitter features. We calculate the text similarity metric BM25 between a tweet T and the query set Q (formed with the most frequent unigrams extracted from top K tweets) for each event. Each word in query set Q is represented by q_i. The BM25 metric is given by:

$$BM25(T,Q) = \Sigma_{q_i \in Q} \frac{IDF(q_i).tf(q_i,T).(k_1+1)}{tf(q_i,T)+k_1(1-b+b\frac{Length(T)}{avg_{length}})}$$

where $tf(q_i)$ is the frequency of occurrence of word q_i in Tweet T, $Length(T)$ denotes the length of T and avg_{length} represents average length of tweet in corpus. The variables k_1 and b are constants. For this experiment, we take their standard values as k_1=1.2 and b=0.75. [9] The value of $IDF(q_i)$, Inverse Document Frequency for a query term q_i, is calculated as follows:

[8] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[9] http://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html

Table 3: Two types of Twitter based features (message and source based) considered for ranking. Enumerates the list in both message and source.

Message based features
Length of the tweet
Number of words
Number of unique characters
Number of hashtags
Number of retweets
Number of use of swear language words
Number of use of positive sentiment words
Number of use of negative sentiment words
Tweet is a retweet
Number of special symbols [$, !]
Number of emoticons [:-), :-(]
Tweet is a reply
Number of @-mentions
Number of retweets
Time lapse since the query
Has URL
Number of URLs
Use of URL shortener service
Source based features
Registration age of the user
Number of statuses
Number of followers
Number of friends
Is a verified account
Length of description
Length of screen name
Has URL
Ratio of followers to followees

$$\text{IDF } (\text{q}_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where, $n(q_i)$ represents the number of documents (tweets) containing q_i, and N is the total number of documents.

The algorithm (Algorithm 1) describes all the above steps of extracting *top k* ranked tweets. The function *Extract-Features*(T) computes all the above mentioned message and source based features for each tweet t_i from the set of tweets T. The $RankSVM(F, T)$ function, takes the feature set matrix F and the column vector A containing the ground truth annotation value for each of the n tweets. The $SortAsc$ and $SortDsc$ functions sort the tweets according to their given score value in ascending and descending value respectively. The function to extract the frequent L word unigrams from the top K tweets is called $FreqLUnigrams(\text{T}_K)$. The $BM25$ method computes the similarity score between the top L unigrams and each tweet t_i in T.

4.3 Evaluation metric

For evaluating the ranking results, we use the standard metric of NDCG (Normalized Discounted Cumulative Gain) [13]. As NDCG captures data with multiple grades well, we chose this measure over other information retrieval measures like

Algorithm 1 Ranking (T[1..n], A[1..n])

1: **for** i <- 0 to $n-1$ **do**
2: F$_i$ <- $ExtractFeatures$(T[i])
3: **end for**
4: FeatureRank <- $RankSVM(F, A)$
5: T$'$ <- $SortAsc$ (FeatureRank)
6: **for** i <- 0 to $k-1$ **do**
7: T$_K$[i] <- T$'$[i]
8: **end for**
9: W$_L$ = $FreqLUnigrams$(T$_K$)
10: PRFRank <- $BM25$ (T$_K$, W$_L$)
11: TweetRank <- $SortDsc$ (PRFRank)
12: return TweetRank[1..k]

MAP (Mean Average Precision). Given a rank-ordered vector V of results $< v_1, \ldots, v_m >$ to query q, let label(v_i) be the judgment of v_i (4=Credible, 3=Maybe credible, 2= Incredible, 1=Relevant but no information, 0=Spam). The discounted cumulative gain of V at document cut-off value n is:

$$DCG@n = \Sigma_{i=1}^{n} \frac{1}{log_2(1+i)}(2^{label(v_i)} - 1)$$

The normalized DCG of V is the DCG of V divided by the DCG of the "ideal" (DCG-maximizing) permutation of V (or 1 if the ideal DCG is 0). The NDCG of the test set is the mean of the NDCG's of the queries in the test set.

5. ANALYSIS

We present our observations and results in this section. First, we analyze how behavior of Twitter, its users and their activity differ or remain same in crisis v/s non-crisis scenarios. Next, we evaluate the ranking algorithm presented in previous section for the feature level (message and source), PRF, and type of event analysis.

5.1 Twitter during Crisis and Non-crisis

We consider all tweets posted for the fourteen events in the following analysis. Figure 5 shows the CDF (Cumulative Density Function) of number of tweets to the number of hours after the event. We consider the tweets for the first 72 hours (3 days) for an event. The time is measured starting from the first tweet about the event in our dataset. We plot the log of the CDF to remove the effect of differences in number of tweets per event. From the graph, we see that events C5 and NC4 show different characteristics than the rest of the events. *Hurricane Irene* and *Murdoch scandal*, were events which had new information about the events first in the initial hours and then after an interval of few hours. The CDF of these two events, have a plateau region from 10-35 hours. From 5, we also see that independent of it being crisis / non-crisis, all events plateau after 20 hours. This shows Twitter is a medium where the event is short lived.

To understand the nature of information diffusion during crisis and non-crisis events, we analyze presence of URLs in tweets, retweets and replies. Presence of URLs indicate, sharing of external resources (links) by users in their messages. We observe that the percentage of tweets containing

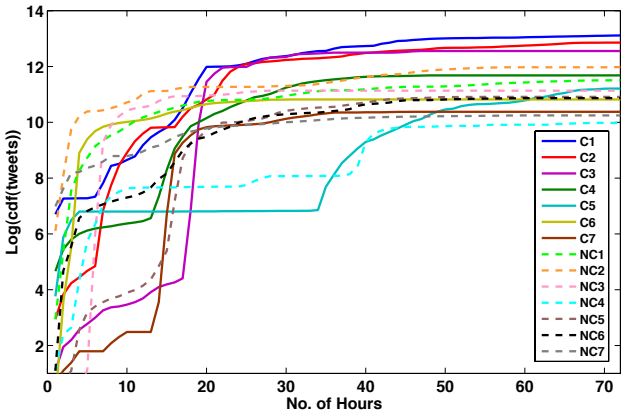


Figure 5: The above CDF graph shows the cumulative distributing of total number of tweets per hour for all events.

URLs in non-crisis situations is 20% greater than crisis situations. This can be attributed to the fact that, most URLs shared on Twitter are links to external news and other resources on the web. For crisis situation, due to the sudden nature of the events, there are not many external resources having related information about the event. For non-crisis situations, often the news breaks simultaneously on social media and news media, websites and blogs. Hence, more links are shared by users in such situations. Our results contradict results by Hughes et al [12]. The number of retweets during crisis situations are more than non-crisis situation by 6%, though the difference is not much, we may infer that people tend to disseminate on information more during crisis. We did not see a large difference between number of replies.

Table 4: Percentage distribution of URLs, retweets and replies during crisis and non-crisis events.

Type	Crisis	Non Crisis
Number of URLs	31.48	50.51
Number of Retweets	42.12	36.3
Number of Replies	6.01	4.66

To study the presence of credible information on Twitter, we analyze the annotated ground truth scores we received for the tweets. Figure 6 shows the percentage distribution of tweets during crisis and non-crisis as per the different categories of coding scheme. The chart depicts, on an average, 50% of tweets on an event are composed of tweets which are related to the event but provide no useful information about it. This observation is quite in-line with the nature of Twitter as a micro blogging website. These tweets generally express the personal opinion or reactions of Twitter users on the event. We also found 13.5% spam tweets in our dataset about the events, i.e. the tweets contained the words belonging to the trending topics but were not related to the event. We found that 30% of tweets contained information about the event, but only 17% of the tweets had information that was credible. We observe that the percentage of credible tweets (definitely + seems credible) is significantly higher for non-crisis events (30% = 12%+18%) as compared to crisis events (20% = 13%+7%). The above observations reinstate that even though number of messages posted on Twitter during important events maybe high, the amount of information content is very limited.

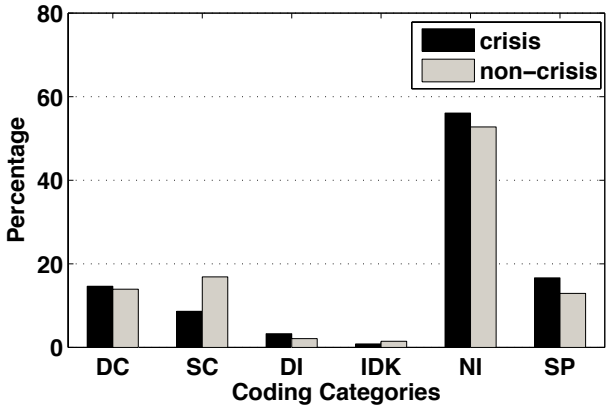


Figure 6: Distribution of tweets in different categories of coding scheme. DC: Definitely Credible, SC: Seems Credible, DI: Definitely Incredible, IDK: Can't Decide if information is credible or not, NI: No information, SP: Spam, N= 5,578.

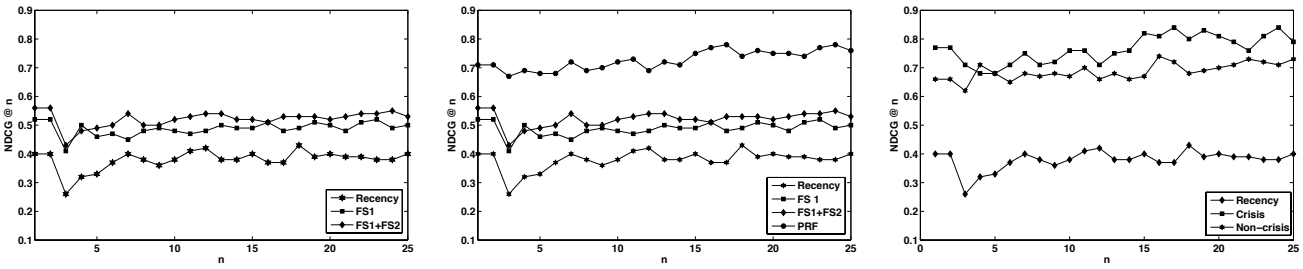
In order to study the difference in network propagation of credible / incredible information during different kinds of events (crisis / non-crisis), we construct a tweet-retweet graph for each scenario. In Figure 7, each node represents either a tweet or the corresponding retweet, each edge links the tweets to the retweets. Comparing subgraphs (a) with (b), and (c) with (d), we see that the credible messages are retweeted more than incredible messages. We also found that the incredible tweets shown in subgraphs (b) and (d) are also retweeted; these tweets are mainly fake news and rumors and spreading of these can be can lead to chaos and panic situations. Comparing subgraphs (a) with (c), and (b) with (d), we found that during non-crisis events, the retweet activity is comparatively less to crisis events, this implies that the diffusion of information is less. This result is in-line with the conclusion that we drew from Table 3 which was for the entire dataset.

5.2 Evaluation of Ranking

As described earlier, using the ground truth about credibility of tweets obtained, and feature sets extracted, we apply Rank-SVM algorithm to learn and rank tweets. In this section, we evaluate the performance of Rank-SVM using the NDCG evaluation metric. We evaluate the ranking framework at the feature level, PRF and type of events considered. We evaluate the ranking performance for top 25 tweets in each of the following analysis.

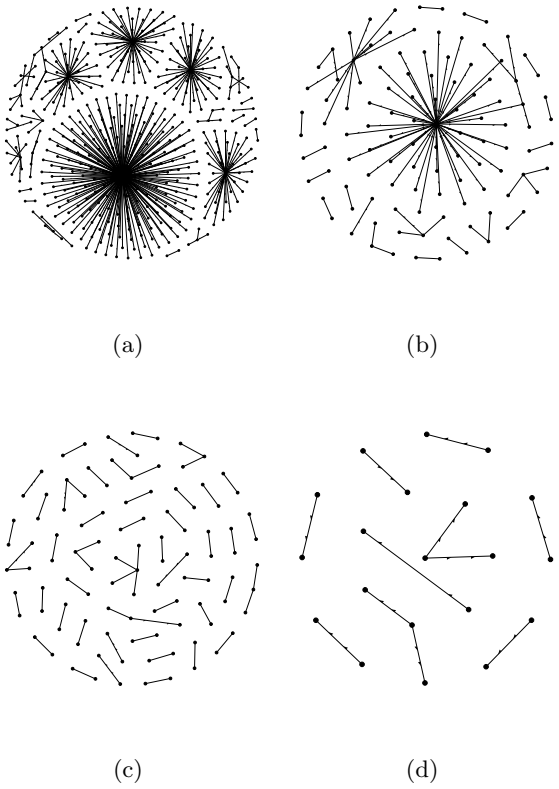
5.2.1 Baseline Analysis

For baseline evaluation of proposed ranking framework, we consider time recency as our baseline. Twitter orders posts according to ascending order of recency since time of query (on Twitter search). On an average for top 25 tweets, we were able to achieve 0.37 NDCG. Figure 8 (a) shows the cumulative gain over baseline using Ranking SVM.



(a) (b) (c)

Figure 8: Performance Evaluation of ranking algorithm. (a) Baseline results v/s FS1 and FS2; (b) Improved performance using PRF technique; (c) Performance of PRF reranking for crisis and non crisis events.



(a) (b)

(c) (d)

Figure 7: Tweet-Retweet graph of credible versus incredible information for crisis and non-crisis events. (a) Credible tweets during crisis; (b) Incredible tweets during crisis; (c) Credible tweets during non-crisis; (d) Incredible tweets during non-crisis

5.2.2 *Feature Level Analysis*

We use ranking SVM to train a model based on the ground truth obtained from annotations. We analyzed two types of Twitter based features to rank tweets as per the credibility of information contained in the tweet. We first perform ranking only with message based features (FS1) and then we add source level features (FS2). We found statistically significant difference in the performance of ranking between using only FS1 and FS1+FS2 (Paired T-test, t=7.47, p-value < 0.05). We conclude that both message and source based features help in predicting the rank of the tweets. Figure 8 (a) shows the cumulative gain using the two feature sets over baseline.

5.2.3 *Pseudo Relevance Feedback*

PRF aids in computing the best order among the top results obtained. We took the top 50 tweets as ranked by FS1+FS2 feature sets, and then extract 10 most frequent unigrams, after removing stop words, user-ids and URLs from the text. To re-rank the tweets, we use the similarity score between the tweet and the top 10 unigrams as given by the BM25 metric. A similarity score based on the above metric is computed for all frequent unigrams and the tweet. The top tweets are then reranked in the descending order of their similarity score. Figure 8 (b) shows that the results are enhanced considerably using the technique. Using the context (e.g. frequent n-grams from the event) in Twitter for ranking can be useful in increasing the effectiveness of ranking compared to not using the context.

5.2.4 *Type of Events*

When we evaluate the ranking outcomes with respect to the type of events, crisis and non-crisis events, the aim is to determine, if the situation of the event has an impact on the performance. Figure 8 (c) shows that NDCG values obtained for crisis events is higher than non-crisis events. The difference is statistically significant (Paired T-test, t=8.80, p-value < 0.05). The ranking algorithm is able to predict the credibility of tweets with a higher cumulative gain during crisis.

6. DISCUSSION

Enormous amount of content is getting generated on the online social media, in particular services like Facebook, Twitter, and YouTube. Twitter is a micro-blogging service which has become popular in the last few years and million of tweets are getting generated on a daily basis. Credibility of information on Twitter and other online social media is a big challenge in utilization of these services as news and information sharing platforms. In this paper, we considered fourteen large-scale events from all around the globe, and presented an automated techniques to rank the content on Twitter, according to the credibility of information contained in the tweet. We empirically showed that information diffusion (through retweets, and URLs) differs between crisis and non-crisis situations. Specifically, we observe that the percentage of tweets containing URLs in non-crisis situations is 20% greater than crisis situations and the number of retweets during crisis situations are more than non-crisis situation by 6%. Using the temporal information from the tweets, we saw that independent of it being crisis / non-crisis, all events plateau after 20 hours. From the human-

annotated data, we found that on average about 30% content on an event (crisis and non-crisis), provides information about the event and 17% of the content can be considered as credible information. To this best of our knowledge, this is the first work to study credibility of content on Twitter at the tweet level and exploring an automated ranking framework to predict rank of tweets according to their credibility. The ranking algorithm we presented used supervised machine learning and relevance feedback approach. We observe that both message and source based features help in predicting the rank of the tweets. We were able to achieve enhanced performance results, by reranking the top most tweets using the similarity of frequent unigrams and tweets. Thus, both context independent features (Twitter based) and context specific features (unigrams) aid in the ranking mechanism. Our results show that we can automate the extraction of credible information with high confidence from Twitter.

7. LIMITATIONS AND FUTURE WORK

In this paper, we considered only 14 events which occurred during the months of July-August, 2011. The Twitter dataset collected for the events does not contain the complete set of tweets posted on Twitter, we could collect only those number of tweets per topic as given by Streaming API of Twitter (the limitation is dependent on the bandwidth availability). In addition to the message and user properties, it would be interesting to study the effect of other features like network-based properties on ranking. The analysis in this paper depended on human annotation to establish the ground truth, we got 500 tweets per topic annotated, and the tweets may not be representative of all the tweets for the event.

Research work presented in this paper is part of a larger research project where we envision to create a real world application to aid in crisis management using the information explosion on online social media. We are drawing inspiration from the existing applications for creating the framework for our application – *Truthy*, *Trumorz* [10] and *zerozero88*. Our current framework will be collecting data continuously (in real-time) from Twitter for crisis events.

8. ACKNOWLEDGMENTS

We thank all members of PreCog research group at IIIT-Delhi for their valuable feedback and suggestions. Our special thanks to Niharika Sachdeva. We would like to acknowledge the support of Government of India for funding this research.

9. REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *CEAS*, 2010.

[2] C. Buckley, G. Salton, and J. Allan. Automatic retrieval with locality information using SMART. *NIST special publication*, (500207):59–72, 1993.

[3] K. R. Canini, B. Suh, and P. L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom*, 2011.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *WWW*, pages 675–684, 2011.

[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. CHI '10, pages 1185–1194, 2010.

[10] http://trumorz.com/

[6] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. CEAS 2011, pages 92–101, 2011.

[7] B. De Longueville, R. S. Smith, and G. Luraschi. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires, LBSN, 2009.

[8] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. WWW '10.

[9] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *COLING '10*.

[10] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security, 2010*.

[11] A. Gupta and P. Kumaraguru. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? IIIT, Delhi, Technical report, IIITD-TR-2011-005, 2011.

[12] A. L. Hughes and L. Palen. Twitter adoption and use in crisis twitter adoption and use in mass convergence and emergency events. In *ISCRAM*, 2010.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20:2002, 2002.

[14] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? WWW '10, pages 591–600, 2010.

[16] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, Mar. 1977.

[17] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *SOMA*. ACM Press, July 2010.

[18] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[19] O. Oh, M. Agrawal, and H. R. Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, 2011.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[21] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. WWW '11.

[22] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. WWW '10, 2010.

[23] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Nlp to the rescue? extracting "situational awareness" tweets during mass emergency. ICWSM, 2011.

[24] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI*, CHI '10, pages 1079–1088, 2010.

[25] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a Twitter network. *First Monday*, 15(1), Jan. 2010.

[26] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR'11*.