

Credibility Ranking of Tweets during High Impact Events

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

www.redpel.com
 +917620593389



ABSTRACT

Twitter has evolved from being a conversation or opinion sharing medium among friends into a platform to share and disseminate information about current events. Events in the real world create a corresponding spur of posts (tweets) on Twitter. Not all content posted on Twitter is trustworthy or useful in providing information about the event. In this paper, we analyzed the credibility of information in tweets corresponding to fourteen high impact news events of 2011 around the globe. From the data we analyzed, on average 30% of total tweets posted about an event contained situational information about the event while 14% was spam. Only 17% of the total tweets posted about the event contained situational awareness information that was credible. Using regression analysis, we identified the important content and sourced based features, which can predict the credibility of information in a tweet. Prominent content based features were number of unique characters, swear words, pronouns, and emoticons in a tweet, and user based features like the number of followers and length of username. We adopted a supervised machine learning and relevance feedback approach using the above features, to rank tweets according to their credibility score. The performance of our ranking algorithm significantly enhanced when we applied re-ranking strategy. Results show that extraction of credible information from Twitter can be automated with high confidence.

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

Credibility, Online social media, High impact events

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PSOSM '12, April 17 2012, Lyon, France
 Copyright 2012 ACM 978-1-4503-1236-3/12/04\$10.00.

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 [16]; secondly, the Internet has provided a platform for common people to share information and express their opinions. 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. Work presented in this paper primarily focuses on Twitter; 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 [16, 19].

One major difference between dissemination of news or information through traditional media and Twitter is that, Twitter is a crowd-sourced medium. In contrast to television or print or news websites where the source of information are few and known (i.e. credible), users on Twitter act like its sensors, and fill in the information gap about an event. Due to the anonymous and unmonitored nature of the Internet, 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 search results display tweets ordered from top to bottom, in reverse chronological order (i.e. the difference between the query and the time when the tweet was posted). When an event of a sizable magnitude and impact occurs, thousands of tweets are posted per hour.^{1 2} Due to the large amount of content generated on Twitter, it is hard to identify the tweets with credible information manually. In this research work, we propose an automated ranking scheme to present the user a ranked output of tweets

¹4.7 million tweets were posted after earthquake in Chile, 2010.

²Earthquake in Virginia (2011) generated more than 5,500 tweets per second.

according to the credibility of information in the tweet.

Though a large volume of content is posted on Twitter, not all information is trustworthy or useful in providing information about the event. The credibility and quality of information often plays a critical role during high impact events. Fake news and rumors also propagate along with genuine news [18]. Researchers have shown that role of Twitter during mass convergence and emergency events differs considerably than more general Twitter activity [12]. They showed that tweets during such events led to more information broadcasting and brokerage. This was the motivation for us to specifically consider some major events during 2011 for our analysis. Each of the events that we analyzed had thousands of tweets (minimum number of tweets for each event was 25,000 and the maximum number of tweets for one event was 542,685 posted about them from all around the globe. Figure 1 gives some sample tweets from our dataset for the event *Hurricane Irene*. All three tweets contain the words matching to the event and were posted while *Hurricane Irene* was the trending topic.³ 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 personal opinion of the user. Even though the bottom tweet contains related words, it includes a URL to an advertisement to sell a product, so it is a spam tweet with respect to the event. In this paper, information refers to the situational awareness information, that is information that leads to gain in the knowledge or update about details of the event, like the location, people affected, causes, etc. [20].



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 widespread of incredible information (like fake news or rumors). To the best of our knowledge, this is the first work to, use automated ranking techniques to assess credibility at the most atomic level of information on Twitter, i.e. at a tweet level. The main contributions of this paper are:

- We found that on average, 30% content about an event, provides situational awareness information about the event, while 14% was spam. In all, 17% of the content was found to be credible information by users.
- We performed linear logistic regression analysis on various Twitter based (content and user) features. Prominent content based features were number of unique

³Trending topics on Twitter are the current most talked about words or phrases on Twitter.

characters, swear words, pronouns and emoticons in a tweet; and user based features like the number of followers and length of username.

- We showed that automated algorithms using supervised machine learning and relevance feedback approach based on Twitter features can be effectively used in assessing credibility of information in tweets.

The rest of the paper is organized as follows: Section 2, describes the closely related work to this paper. 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 analysis performed. Section 5 discusses the credibility based relevance 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

Computer science research community has analyzed relevance of online social media, and in particular Twitter, as news disseminating agent, in the past. Kwak et al. showed the prominence of Twitter as a news media, they showed that 85% topics discussed on Twitter are related to news [16]. Their work highlighted the relationship between user specific parameters v/s the tweeting activity patterns, like analysis of the number of followers and followees v/s 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) [28]. They showed that Twitter users are relatively less interested in world news; still they are active in spreading news of important world events.

Researchers have highlighted that useful and actionable information can be extracted by mining Twitter data and activity during crisis events. Mendoza et al. used the data from 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity [18]. Their results showed that propagation of tweets related to rumors versus true news differed and could be used to develop automated classification solutions to identify correct information. 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. They 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 [24]. 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 [20]. 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 National ICT Australia Ltd. (NICTA) has been working on developing a focused search engine for Twitter and Facebook that can be used in humanitarian crisis situation.⁴ Hughes et al. in their work compared the properties of tweets and users during an emergency to normal situations [13]. They showed an increase in the use of URLs in tweets and a decrease in @-mentions 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; which helped in increasing situation awareness during emergency events [26]. Verma et al. used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [25].

Quality of Information on Twitter

Presence of spam, compromised accounts, malware, and phishing attacks are major concerns with respect to the quality of information on Twitter. Techniques to filter out spam / phishing on Twitter has been studied and various effective solutions have been proposed [1, 6, 10, 27]. Truthy⁵, was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a public stream of micro-blogging updates related to an event to detect political smears, astroturfing, misinformation, and other forms of social pollution [22]. 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 assessed their credibility based on various Twitter features [4]. They achieved a precision and recall of 70-80% using J48 decision tree classification algorithms. Canini et al. analyzed usage of automated ranking strategies to measure 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 credibility based ranking of users of Twitter. 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 difficulty in measuring credibility of information and 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 [21]. 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. Chen et al. built a tool called *zerozero88*,⁶ 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.

So far, the work done to assess credibility on Twitter, have explored credibility with respect to trending topics and users. Our work differs from that done by Castillo et. al [4] – their analysis was based on credibility of a trending topic (all tweets belonging to a topic were marked as credible or incredible) on Twitter, while we focus on assessing credibility at the level of tweets. This difference in approaches lends a significant impact in case of Twitter, since a topic (e.g. earthquake at a particular location) maybe credible, yet the tweets in that topic maybe of credible or incredible (e.g. Richter scale of the earthquake) in nature. Hence, credibility of a topic may not be a good indicator to judge the credibility of the content of the tweet. 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. We believe, our results can help users in making a decision on the credibility of the tweet.⁷

3. METHODOLOGY

In this section, we discuss the data collection setup, the process of selecting the events, and the coding scheme used to annotate the tweets. Figure 2 describes the methodology and analysis performed in the research presented in this paper.

3.1 Data Collection

We collected data from Twitter using the *Streaming API*.⁸ 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.⁹ 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

⁶<http://zerozero88.com/>

⁷We have already built an online portal precog.iiitd.edu.in/credtweet that applies our algorithms to show the credibility of the tweets. This portal is in the Alpha stage now; we plan to release the beta stage soon.

⁸<https://dev.twitter.com/docs/streaming-api>.

⁹<https://dev.twitter.com/docs/api/1/get/trends>

⁴<http://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/>

⁵<http://truthy.indiana.edu/>

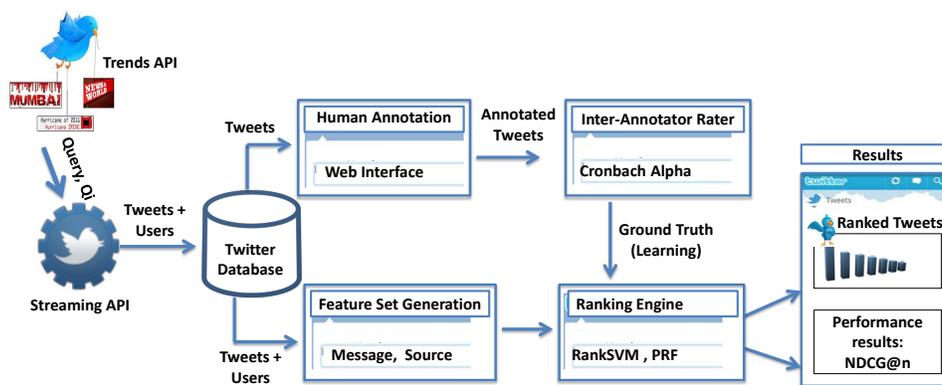


Figure 2: Describes the methodology and analysis performed in this paper.

until the time it remained as a trending topic. Castillo et al. also used a similar framework to collect tweets using current trending topics [4]. In our data collection, 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 12th July, 2011 to 30th 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 12th and August 30th, 2011. Each event had one or more trending topics associated with it, for example trending topics related to *Debt and downgrading crisis in the US* were *AAA to AA*, *S&P*, etc. For each event, we considered 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 number of tweets for each event, the corresponding trending topics for the event, and a short description of the event. We selected events covering various domains of news events like political, financial, natural hazards, terror strikes and entertainment news. 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, *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

trending topic on Twitter. For example, *Hurricane Irene* was a trending topic on Twitter for 150 hours.

3.3 Annotation Scheme

This section describes how we annotated 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 annotation 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 non-English tweets. For the purpose of annotation, we developed a web interface and we provided each annotator with an authenticating login and password.¹⁰

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 the news event, but contains no information
- Tweet is not related to news event
- Skip tweet

We provided the annotators with the definition of credibility¹¹ and then explained the above mentioned categories

¹⁰<http://precog.iiitd.edu.in/annotation/login1.php>

¹¹Oxford 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.

Events	Tweets	Trending Topics	Description
UK Riots	542,685	#ukriots, #londonriots, #prayforlondon	Riots in United Kingdom caused 5 deaths, 16 civilian and 186 police injuries
Libya Crisis	389,506	libya, tripoli	Rebels opposing Col. Qaddafi seized Zawiyah
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.
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
Hurricane Irene	90,237	Hurricane Irene, Tropical Storm Irene	Hurricane Irene in US caused 55 deaths and a damage of US \$10.1 billion
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.
Mumbai Blast, 2011	32,156	#mumbaiblast, Dadar, #needhelp	Three bomb blasts in Mumbai (India) on 13th July, 26 people died and 130 injured
JanLokPal Bill Agitation	182,692	Anna Hazare, #janlokalpal, #anna	An anti-corruption movement against the Government of India.
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
Google acquires Motorola Mobility	68,527	Google, Motorola Mobility	Google buying Motorola Mobility in a \$12.5bn cash deal, was a huge acquisition
News of the World Scandal	67,602	Rupert Murdoch, #murdoch	The News International phone hacking scandal exposed Rupert Murdoch
Abercrombie & Fitch stocks drop	54,763	Abercrombie & Fitch, A&F	Abercrombie & Fitch stocks drops 9% after a controversy
Muppets Bert and Ernie were gay	52,401	Bert and Ernie	Rumors circulated that muppet pair Ernie and Bert, are a gay couple
New Facebook Messenger	28,206	Facebook Messenger	Facebook launched a new messenger for mobile users

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. During our pilot study, we observed options ‘Seems Credible’ and ‘Seems Incredible’ were redundant, as both indicated that a tweet seemed both credible and incredible to the user. Hence, for the final annotation, we kept only one of the options. Each tweet for the events (500 tweets per event) was 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 7,000 (14 events * 500 tweets) tweets was 0.748. Alpha > 0.7 implies a high agreement between annotators [17]. We selected the majority score for a tweet (i.e. value given by at least 2 annotators) as the final scores for each tweet; 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 (around 80% from 7,000) tweets in our final annotated dataset.

4. ANALYSIS

We propose an automated ranking scheme to output of tweets ordered according to the credibility of information provided in them. We used a combination of supervised machine learning and relevance feedback approach to rank tweets. We analyzed the effectiveness of Twitter based features (message and source level) to rank tweets according to informa-

tion quality in the tweet. As a next step, we evaluated an enhancement to the above ranking technique by using pseudo feedback relevance re-ranking scheme. We used 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 [15]. We used the SVM^{Rank} implementation code from Cornell University.¹² SVM^{Rank} trains a Ranking SVM on the training set, and outputs the learned rule to a model file. Based on the learned model, the algorithm predicts a ranking score that are written to the output file. We performed four-fold cross validation of our results. The ground truth for the task was obtained from the human annotated tweet scores.

4.1 Types of Features

Two basic characteristics, the features of the message itself, and the properties of the user who posted the message characterize any post or tweet on Twitter. Table 3 presents features that are available in the message and the user. We consider following two types of features as input to the ranking algorithm:

- **Content or message level features:** The 140 characters posted by users contain data (e.g. words, URLs, hashtags) and meta-data (e.g. is tweet a reply or a retweet) related to it. We do not consider text semantic features here in our analysis.
- **Source or user level features:** The attributes of the user who posted the tweet. We consider properties

¹²http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

such as number of friends, followers and status messages of the user as part of this set.

Table 3: Content and source based features that were considered for ranking.

Message based features
Length of the tweet, number of words, number of unique characters, number of hashtags, number of retweets, number of swear language words, number of positive sentiment words, number 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

4.2 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) also known as Blind Relevance Feedback, is a prominent re-ranking 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 re-rank those documents according to a defined score. In our algorithm, we extracted most frequent unigrams from the top K tweets and used the text similarity between the most frequent unigrams and K tweets to re-rank them. The improvement achieved by re-ranking using PRF is highly dependent on the quality of top K results given by the ranking algorithm. We applied PRF to the best set of results obtained by previous analysis, that is the ranking results obtained using both message and source. We calculated the text similarity using the metric BM25 [23], 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 was represented by q_i . The BM25 metric is given by:

$$BM25(T, Q) = \sum_{q_i \in Q} \frac{IDF(q_i) \cdot tf(q_i, T) \cdot (k_1 + 1)}{tf(q_i, T) + k_1(1 - b + b \frac{Length(T)}{avglength})} \quad (1)$$

where $tf(q_i)$ is the frequency of occurrence of word q_i in Tweet T , $Length(T)$ denotes the length of T and $avglength$ represents average length of tweet in corpus. The variables k_1 and b are constants; we take their standard values as $k_1=1.2$ and $b=0.75$ in our case.¹³ The value of $IDF(q_i)$, Inverse Document Frequency for a query term q_i , is calculated as follows:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (2)$$

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

¹³<http://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html>

The algorithm (**Algorithm 1**) describes all the above steps of extracting top k ranked tweets. Function $ExtractFeatures(T)$ computes 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. $FreqLUnigrams(T_K)$ extracts the frequent L word unigrams from the top K tweets. $BM25$ method computes the similarity score between the top L unigrams and each tweet t_i in T .

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]
```

4.3 Evaluation metric

For evaluating the relevance ranking results, we used the standard metric of NDCG (Normalized Discounted Cumulative Gain) [14]. We chose this measure over other information retrieval measures like MAP (Mean Average Precision) as NDCG captures data with multiple grades. Given a rank-ordered vector V of results $\langle v_1, \dots, v_m \rangle$ 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:

$$DCG@n = \sum_{i=1}^n \frac{1}{\log_2(1+i)} (2^{label(v_i)} - 1) \quad (3)$$

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. EXPERIMENTAL RESULTS

For the human annotated data, on an average, 50% of tweets on an event are composed of tweets which were related to the event but provided no useful information about it. Such 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.

5.1 Regression Analysis

We performed logistic linear regression analysis, with respect to the features listed in Table 3 to estimate the good predictors for credibility of tweets. To perform the regression

analysis, we considered all tweets annotated as definitely and seems credible as the data points for the positive class (dependent variable = 1), and rest all as not credible class of tweets (dependent variable value = 0). As a result, we obtained the following features as strong indicators (p-value < 0.001) of credibility: number of characters present in tweet, number of unique characters present in tweet, presence of swear words, inclusion of pronouns and presence of sad / happy emoticons.

The number of unique characters present in tweet was positively correlated to credibility, this may be due to the fact that tweets with hashtags, @mentions and URLs contain more unique characters. Such tweets are also more informative and linked, and hence credible. Presence of swear words in tweets indicates that it contains the opinion / reaction of the user and would have less chances of providing information about the event. Tweets that contain information or are reporting facts about the event, are impersonal in nature, as a result we get a negative correlation of presence of pronouns in credible tweets. Low number of happy emoticons [:-), :) and high number of sad emoticons [:-(, :(] act as strong predictors of credibility. Some of the other important features (p-value < 0.01) were inclusion of a URL in the tweet, number of followers of the user who tweeted and presence of negative emotion words. Inclusion of URL in a tweet showed a strong positive correlation with credibility, as most URLs refer to pictures, videos, resources related to the event or news articles about the event.

By performing regression analysis for credibility, we were able to identify some prominent features based on the content and user of the tweet. We mostly got all tweet (content) based features to play an important role in determining credibility, but user based features like the number of followers also come out as a strong predictor. Hence, we can conclude that a ranking algorithm based on both the content of the tweets, and the user properties, would be effective in determining the credibility of information in the tweets.

5.2 Evaluation of Ranking

In this section, we evaluate the performance of Rank-SVM and PRF using the NDCG evaluation metric. We measured the ranking output's performance, for the top 25 ranked tweets in each of the analysis presented. For baseline evaluation, we considered time recency (most recent tweet on top), that is the order in which tweets are displayed by Twitter search. On an average for top 25 tweets, we achieved 0.37 NDCG for the recency metric.

5.2.1 Content and Source based Analysis

We performed ranking individually on content based features (tweet) and source level features (user). We observed that both set of features perform comparably. Using the combined set of features (tweet and user) there was a significant statistical improvement observed in the performance of ranking (Paired T-test, $t=7.47$, p-value < 0.05) than the performance of individual feature sets. We conclude that both message and source based features play a role in predicting the credibility based rank of the tweets. Figure 3 (a) shows the cumulative gain using the two feature sets over time recency. These results highlights the point that content based features are as important as the source based

features on Twitter with respect to credibility. Hence, it is not only important *who you are* when you tweet, but also the quality of *what you post*.

5.2.2 Relevance Feedback Re-ranking Analysis

For PRF analysis, we took the top 50 tweets obtained from Rank-SVM (tweet and user features), extracted the ten most frequent unigrams (after removing stop words, user-ids and URLs) from the tweets. We re-ranked the tweets, based on the similarity score based on BM25 metric, for the tweet and the top 10 unigrams. A similarity score based on this metric was computed for all frequent unigrams and the tweet. The top tweets were then re-ranked in the descending order of their similarity score. Figure 3 (b) shows that the performance of credibility is enhanced considerably using PRF (average 0.73 NDCG score). Using the context (e.g. frequent n-grams) in Twitter for ranking can be useful in increasing the effectiveness of credibility ranking.

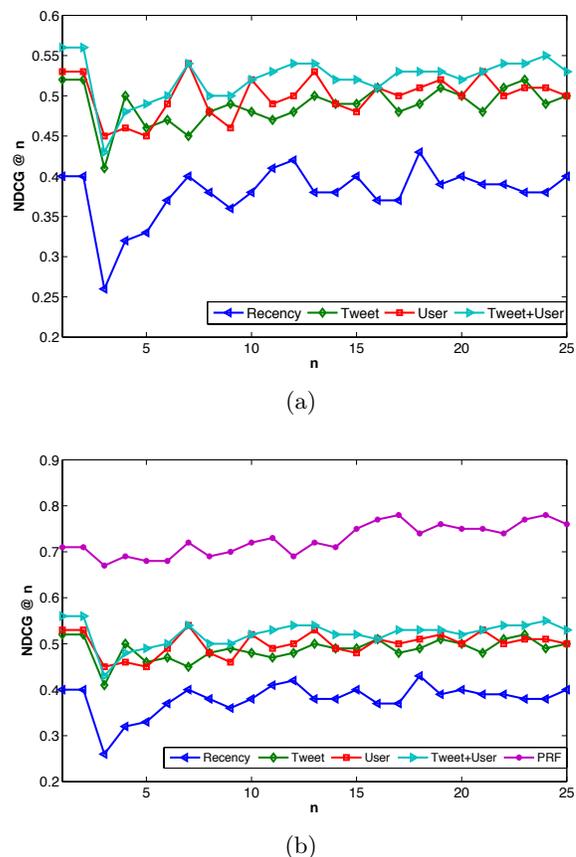


Figure 3: Performance Evaluation of ranking algorithm. (a) Recency results v/s Twitter features (tweet and user based); (b) Improved performance using PRF technique.

6. DISCUSSION

During high impact events, there is a sudden rise in activity over the Internet. People log on to social media websites to check for updates about the event and also to share information about the event with others. We considered, Twitter as our medium of information for this paper. In

recent years, Twitter has emerged as a news and information sharing platform during such events. Though a large volume of content is posted on Twitter, not all of the information is trustworthy or useful in providing information about the event. Credibility of information on Twitter is a big challenge in its utilization as news and information sharing platform. In particular, credibility of information during high impact events can be important. Researchers have shown that role of Twitter during mass convergence and emergency events differs considerably than regular Twitter activity. In this paper, we considered fourteen high-impact events from all around the globe and analyzed the tweets for these events for the credibility of information in them. By information here, we mean situation awareness information about an event. Situational awareness information is information that leads to gain in the knowledge or update about details of the event, like the location, people affected, causes, etc. We found that on average, 30% content about an event, provides situational awareness information about the event, while 14% was spam. Annotators found that only 17% of the total tweets contained situational awareness information that was credible. We applied linear logistic regression analysis to identify the prominent Twitter features (content and user based) which can help in assessing the credibility. Prominent content based features were number of unique characters, swear words, pronouns and emoticons in a tweet; and user based features like the number of followers and length of username. We evaluated an algorithm (using RankSVM and relevance feedback approach) to rank the tweets, according to the credibility of information contained in the tweet. We observed that the content based features were as important as the source based features on Twitter with respect to credibility. The results assert the fact that it is not only important *who you are* when you tweet, but also the quality of *what you post*. By applying the relevance feedback techniques based on most frequent n-unigrams, we were able to achieve considerably enhanced performance of ranking results. We show that both context independent features (Twitter based) and context specific features (unigrams) aid in the ranking mechanism. Results show that extraction of credible information from Twitter can be automated the with high confidence.

One of the limitation for the work presented here is that human annotation to establish the ground truth. We would like to develop self-learning mechanism and automatically adopting systems that do not require manual coding inputs.

7. 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.
- [6] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/Social: 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. Kolar, 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? IIT, Delhi, Technical report, IITD-TR-2011-005, 2011.
- [12] A. I. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the 2009 ISCRAM Conference*, 2009.
- [13] 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.
- [14] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20:2002, 2002.
- [15] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.
- [16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? *WWW '10*, 2010.
- [17] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, Mar. 1977.
- [18] M. Mendoza, B. Poblete, and C. Castillo. In *SOMA*, July.
- [19] B. O'Connor, R. Balasubramanian, 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.
- [20] 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.
- [21] 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.
- [22] 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*.
- [23] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. *IN*, 1999.
- [24] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *WWW '10*, 2010.
- [25] 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.
- [26] 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.
- [27] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a Twitter network. *First Monday*, 15(1), Jan. 2010.
- [28] 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*.