Figure 1: Screenshot of *TweetCred* Chrome extension built and deployed for displaying credibility of tweets to users in real-time within their Twitter timeline.

In our previous work on the problem of assessing credibility, we analyzed Twitter data in a post-hoc setup [8]. We showed a proof of concept algorithm which took manually annotated tweets, and then used automated techniques to rank previously unseen tweets by credibility. We also used insights from the analysis of fake content in previous crisis events, reported in [9, 10], to create a novel system for credibility assessment in real-time. Our model for credibility ranking in this paper is based on a much more exhaustive and comprehensive set of features than our previous work. Also, the feature sets had to be modified according to the constraint of limited data in real-time. To the best our knowledge, this is the first research work that has produced a prototype for the credibility assessment problem that was deployed and evaluated by Twitter users. *TweetCred* takes a direct stream of tweets as input and computes the credibility for each of the tweets on a scale of 1 (low credibility) to 7 (high credibility).

The main contributions of this work are:

- We developed a semi-supervised ranking model using SVM-rank for assessing credibility based on learning data obtained from 6 high impact crisis events of 2013. An extensive set of 45 features was used to determine the credibility score for each of the tweets.

- We developed and deployed a real time system, *TweetCred*, in the form of a Chrome extension, Web application, and REST API. *TweetCred* was installed and used by 717 Twitter users within a span of three weeks, and used by them to compute the credibility of more than 1.1 million unique tweets.

- We evaluated the real-time performance of *TweetCred*, observing that 84% of the credibility scores were displayed for the corresponding tweets within 6 seconds. For 43% of the 936 tweets for which system received feedback, users agreed with the credibility score computed by the system. For a further 25% of tweets, their disagreement was of 2 points or less (on the 7-point scale).

This paper is organized as follows: Section 2 describes the literature review of work done around this domain; Section 3 gives our methodology in detail and in Section 4 we discuss the credibility ranking techniques and performance of our proposed solution. Section 5 describes the implementation details, usage analysis and performance evaluation of *TweetCred*. Finally, in the last section we provide the discussion of the results, their impact, and future work.

2. LITERATURE REVIEW

Researchers have attempted to solve the problem of trust and credibility on Online Social Media (OSM) using various techniques. There has been work done in identifying and filtering spam, phishing and other kinds of malicious contents from OSM data.

Trust/Credibility Assessment. In this section, we discuss some of the research work done to assess, characterize, analyze and compute trust and credibility of content on online social media. The first work discussed is Truthy,[2] which was developed by Ratkiewicz et al. [18] 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. In their work, they presented certain cases of abusive behavior by Twitter users. Truthy is a live web service built upon the above work. Supervised classification has been applied by researchers to detect credible and incredible content in OSM. Castillo et al. [3] showed that automated classification techniques can be used to detect news topics from conversational topics and assessed their credibility based on various Twitter features. They achieved a precision and recall of 70-80% using decision-tree based algorithm. They evaluated their results with respect to data annotated by humans as ground truth. The feature sets used in their work included message (tweet content), user, topic and propagation based features. They made some interesting observations, such as: tweets which do not include URLs tend to be related to non-credible news; tweets which include negative sentiment words are related to credible news.

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

Now we discuss research work that has been done focused on determining the credibility of the users in OSM. Canini et al. [2] analyzed usage of automated ranking strategies to measure credibility of sources of information on Twitter for any given topic. The authors define a credible information source as one which has trust and domain expertise associated with it. They observed that content and network structure act as prominent features for effective credibility based ranking of users on Twitter.

Some researchers focused their study of trustworthy or credible information during particular events which had high impact. Gupta et al. [7] 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 were with low Twitter reputation (less number of followers). This highlights the difficulty in measuring credibility of information and the need to develop automated mechanisms to assess credibility of information on Twitter. The authors in a follow up study applied machine learning algorithms (SVM-rank) and information retrieval techniques (relevance feedback) to assess credibility of content on Twitter [8]. They analyzed fourteen high impact events of 2011; their results showed that 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.

Another, similar work was done by Xia et al. [22] on tweets generated during the England riots of 2011. They used a supervised method of Bayesian Network to predict the credibility of tweets in emergency situations. They proposed and evaluated a two step methodology: in the first step they used a modified sequential K-means algorithm to detect an emergency situation; in the second step, a Bayesian Network structure learning algorithm was used to judge the information credibility. Donovan et al. [16] focused their work on finding indicators of credibility during different situations (8 separate event tweets were considered). Their results showed that the best indicators of credibility were URLs, mentions, retweets and tweet length. Also, they observed that the presence and effectiveness of these features increased a lot during emergency events.

A different methodology, than the above papers was followed by Morris et al. [15]. They conducted a survey to understand users' perceptions regarding credibility of content on Twitter. They asked about 200 participants to mark what they consider are indicators of credibility of content and users on Twitter. They found that the prominent features based on which users judge credibility are features visible at a glance, for example, username and picture of a user. By their experiments they showed that users are poor judges of credibility based only on content and are often biased by other information like username. Also, they highlighted that there exists a disparity between features a user considers relevant to credibility and those used by search engines.

Yang et al. [25] analyzed credibility perceptions of users on two micro-blogging websites: Twitter in the USA and Weibo in China. They found that location and network overlap features had the most influence in determining the credibility perceptions of users. They examined cultural differences and found that Chinese users were more sensitive to the context of an event, with their credibility perceptions changing according to context changes. Ghosh et al. [6] identified topic-based experts on Twitter using features obtained from user-created list, relying on the wisdom of Twitter's crowds.

Extracting Situational Awareness from Twitter. Work has been done to extract situational awareness information from the vast amount of data posted on Twitter during real-world events. Vieweg et al. [21] analyzed the Twitter logs for the Oklahoma Grass fires (April 2009) and the Red River Floods (March and April 2009) looking for situational awareness content. They developed an automated framework to enhance situational awareness during emergency situations, extracting location and location-referencing information from users' tweets. Verma et al. [20] used natural language processing techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness. Corvey et al. [4] also adopted a computational linguistics approach, analyzing the importance of linguistic and behavioral annotations. They considered data from four events: Hurricane Gustav in 2008, the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods, and the 2010 Haiti Earthquake. They concluded that users used a specific vocabulary to convey tactical information on Twitter, as evidenced by the accuracy achieved using bag-of-words model for situational awareness tweets classification.

Inflammatory and hate speech. Over recent years OSM has also been used to spread hate or inflammatory content. Such content if propagated during crisis situations can have major adverse implications. There have been few research works which have analyzed the hate content on YouTube and Twitter OSM. Sureka et al. [19] used semi-automated techniques to discover content on YouTube that spread hate. They discovered videos and users propagating hate, as well as hidden virtual communities, using data-mining and social network analysis techniques. The precision they achieved using bootstrapping techniques was 88% for the task of detecting users that spread hate. Xiang et al. [23] applied machine learning and topic modeling techniques to detect offensive content on Twitter. They achieved a true positive rate of approximately 75%, outperforming

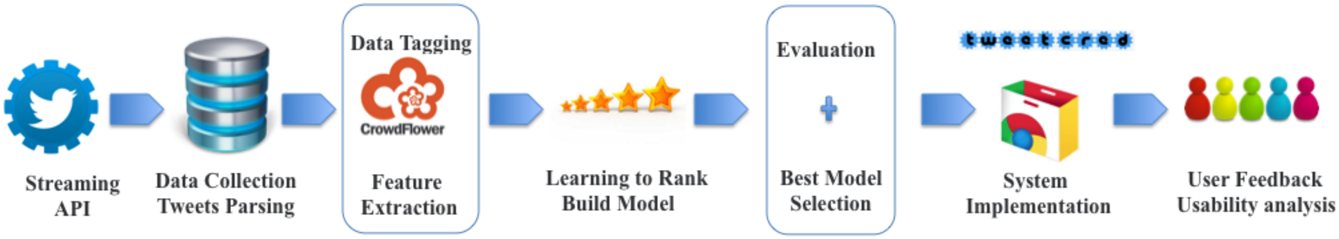


Figure 2: Diagram depicting the operation of *TweetCred* and the methodology followed in this research work.

keyword-based techniques. The authors used a seed lexicon of offensive words, and then applied Latent Dirichlet Allocation (LDA) models for topic discovery. One interesting finding of their work was that there are several words that are not offensive individually, but only when used in combination with other words.

To the best our knowledge, the work presented in this paper is the first research work that describes the creation and deployment of a practical system for credibility on Twitter, including the evaluation of such system with real users.

3. METHODOLOGY

At the core of our system is the capability of ranking tweets by credibility in real time. We propose, implement and evaluate algorithms for determining a credibility score for each tweet, taking into account variables from the tweet itself and from its author. For our study, we first collected data from Twitter for six prominent events of 2013, and then we extracted features from the collected tweets. Figure 2 depicts the methodology we followed.

After creating a model for credibility assessment, we invited users to test our model by downloading and installing a browser extension that seamlessly incorporates our credibility inferences into a users' Twitter experience.

3.1 Data Collection

We collected data from Twitter's streaming API.[3] We had a 24×7 data collection pipeline, which automatically collects data from Twitter for a set of pre-specified keywords. For this research work we considered six crisis events from different parts of the world during 2013. These events affected a large population and generated a high volume of content in Twitter. The events considered, and the corresponding number of tweets for each one, are listed in Table 1.

3.2 Data Labeling

Table 1: Summary statistics for the studied datasets.

Event	Tweets	Users
Boston Marathon Blasts	7,888,374	3,677,531
Typhoon Haiyan / Yolanda	671,918	368,269
Cyclone Phailin	76,136	34,776
Washington Navy yard shootings	484,609	257,682
Polar vortex cold wave	143,959	116,141
Oklahoma Tornadoes	809,154	542,049
Total tweets	10,074,150	4,996,448

In order to create ground truth for building our model for credibility assessment, we obtained human labels for around 500 tweets selected uniformly at random per event. The annotations were obtained through crowdsourcing provider CrowdFlower.[4] We selected only annotators living in the United States and for each task collected answers from three different annotators, keeping the majority among the options chosen by them.

The annotation proceeded in two steps. In the first step, we asked users if the tweet contained information about the event to which it corresponded, with the following options:

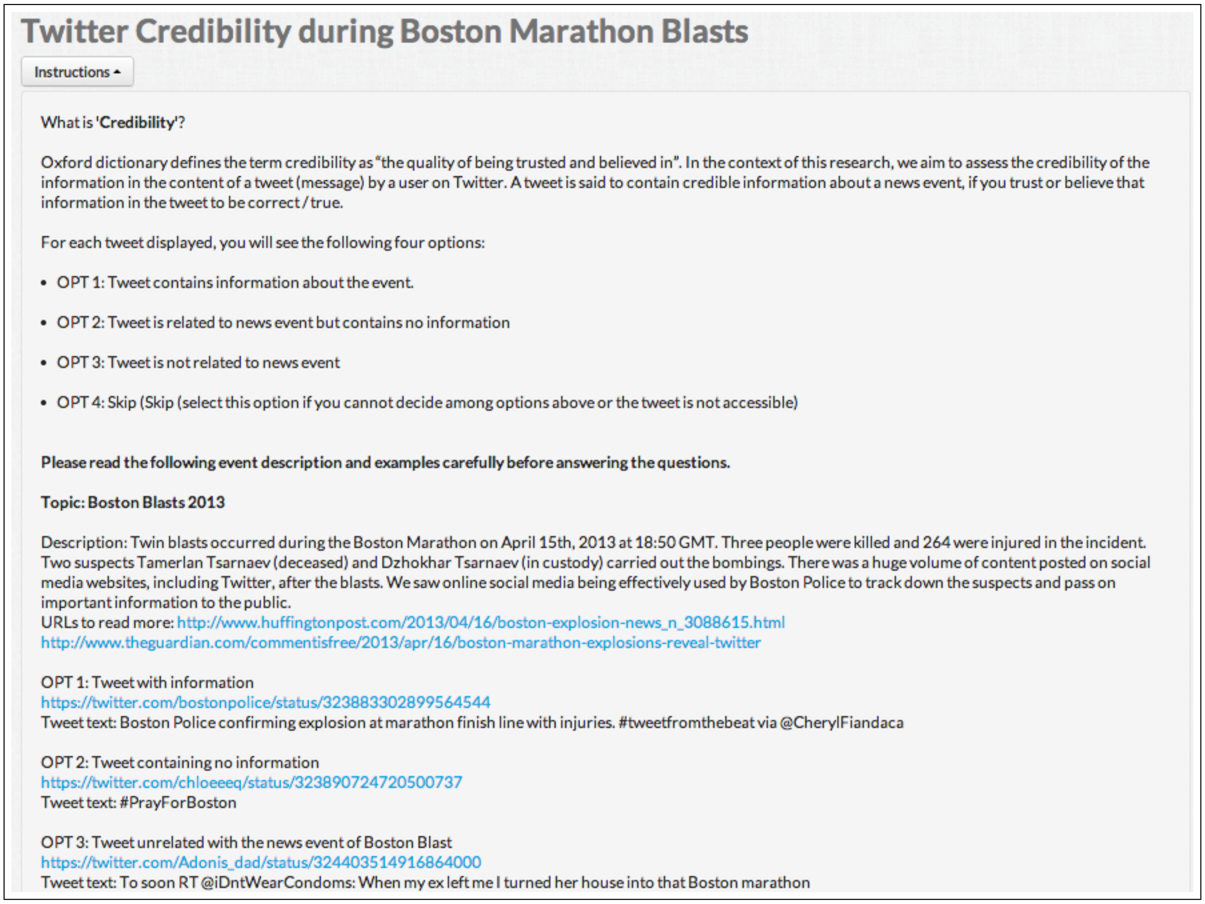
- The tweet contains information about the event.
- The tweet is related to the event, but contains no information.
- The tweet is not related to the event.
- Skip tweet.

Along with the tweets for each event, we provided a brief description of the event and links from where users can read more about it. We also showed users a definition of credibility and example tweets for each option in the annotation, as shown in Figure 3.

In the second step, we selected those tweets that were marked as informative (45% of the original tweets), and annotated them with respect to the credibility of the information conveyed by it. We asked workers to score each tweet according to its credibility with the following options:

[3]https://dev.twitter.com/docs/api/streaming

[4]http://www.crowdflower.com/



Twitter Credibility during Boston Marathon Blasts

Instructions

What is '**Credibility**'?

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.

For each tweet displayed, you will see the following four options:

- OPT 1: Tweet contains information about the event.
- OPT 2: Tweet is related to news event but contains no information
- OPT 3: Tweet is not related to news event
- OPT 4: Skip (Skip (select this option if you cannot decide among options above or the tweet is not accessible)

Please read the following event description and examples carefully before answering the questions.

Topic: Boston Blasts 2013

Description: Twin blasts occurred during the Boston Marathon on April 15th, 2013 at 18:50 GMT. Three people were killed and 264 were injured in the incident. Two suspects Tamerlan Tsarnaev (deceased) and Dzhokhar Tsarnaev (in custody) carried out the bombings. There was a huge volume of content posted on social media websites, including Twitter, after the blasts. We saw online social media being effectively used by Boston Police to track down the suspects and pass on important information to the public.
URLs to read more: http://www.huffingtonpost.com/2013/04/16/boston-explosion-news_n_3088615.html
http://www.theguardian.com/commentisfree/2013/apr/16/boston-marathon-explosions-reveal-twitter

OPT 1: Tweet with information
https://twitter.com/bostonpolice/status/323883302899564544
Tweet text: Boston Police confirming explosion at marathon finish line with injuries. #tweetfromthebeat via @CherylFiandaca

OPT 2: Tweet containing no information
https://twitter.com/chloeeeq/status/323890724720500737
Tweet text: #PrayForBoston

OPT 3: Tweet unrelated with the news event of Boston Blast
https://twitter.com/Adonis_dad/status/324403514916864000
Tweet text: To soon RT @iDntWearCondoms: When my ex left me I turned her house into that Boston marathon

Figure 3: Screenshot of the first annotation task done on crowd-sourcing provider CrowdFlower.

- Definitely credible
- Seems credible
- Definitely incredible
- I can't decide

Table 2 gives the distribution of the annotations. There were about 23% of tweets that contained definitely credible information about an event; and about 6% information that the users definitely did not trust.

Table 2: Distribution of labels over tweets.

Label	Percentage	
	2013 events	[8]
Definitively credible	23%	17%
Seems credible	16%	} 13%
Definitively incredible	6%	
Not informative	40%	56%
Not related to the event	15%	14%

For comparison, we also include in Table 2 the results of our previous work [8], based on 14 events from 2011. We observe that the distributions are not exactly equal, but similar. Though, we observe that non-informative content for an event has decreased from 2011 to 2013 by 16%, and the content that people trust on Twitter has increased by 5% in 2013.

4. CREDIBILITY RANKING ANALYSIS

Our aim is to develop a model for ranking tweets by credibility. We adopt a supervised learning to rank approach in three steps. First, we perform feature extraction from the tweets. Second, we test different learning schemes to develop models for credibility ranking. Third, we implement and deploy *TweetCred*, a real-time solution to measure credibility of tweets, and analyze its usage, performance and accuracy.

4.1 Feature Extraction

The first important step in data analysis for supervised learning algorithm is generating feature vectors from the data points. Since our work is aimed at building a real time system, the features we employ are restricted to those that can be derived from a single tweet. This excludes features from a group of tweets (as in e.g. [3]) as well as user-related features from past tweets. A tweet as downloaded from Twitter's API contain a se-

Table 3: Features used by the credibility model.

Tweet Meta-data Features: Number of seconds since the tweet, Source of tweet (mobile / web/ etc), Tweet contains geo-coordinates
Tweet Content Features: Number of characters, Number of words, Number of URLs, Number of hashtags, Number of unique characters, Presence of stock symbol, Presence of happy smiley, Presence of sad smiley, Tweet contains 'via', Presence of colon symbol
User based Features: Number of followers, friends, time since the user if on Twitter, etc.
Network Features Number of retweets, Number of mentions, Tweet is a reply, Tweet is a retweet
Linguistic Features: Presence of swear words, Presence of negative emotion words, Presence of positive emotion words, Presence of pronouns, Mention of self words in tweet (I, my, mine)
External Resource Features: WOT score for the URL, Ratio of likes / dislikes for a YouTube video

ries of fields [5] in addition to the text of the message. For instance, it includes meta-data such as posting date as well as information about its author at the time of posting (e.g. his/her number of followers). For tweets containing URLs, we enriched this data with information about that specific URL such as Web of Trust reputation (WOT) score for a domain. [6] The features we used can be divided into several groups, as shown in Table 3. In total, we used 45 features.

4.2 Learning to Rank Tweets

We tested and evaluated multiple learning-to-rank algorithms to learn a model that ranks tweets by credibility. We experimented with various methods that are typically used for information retrieval tasks: Coordinate Ascent [14], AdaRank [24], RankBoost [5] and SVM-rank [12]. We used two popular toolkits for ranking, RankLib[7] and SVM-rank.[8]

Coordinate Ascent is a standard technique of optimizing multi-variate optimization functions. It considers one dimension at a time and optimizes for the same. SVM-rank is pair-wise ranking technique that uses SVM (Support Vector Machines). It changes the input data, provided as a ranked list, into a set of ordered pairs. The (binary) class label for every pair is the order in which the elements of the pair should be ranked. At testing time, the classifier also predicts the ordering for an input pair. AdaRank trains the model by minimizing a loss function directly defined on the performance measures. It applies a boosting technique in ranking methods. Unlike other models like SVM-rank and RankBoost which are loosely dependent on performance measures, AdaRank directly enhances them in its training process. RankBoost is a boosting algorithm based on the AdaRank algorithm. It also, runs for many iterations or rounds and uses boosting techniques to combine weak rankings using the ranking features.

The two most important factors for a real-time system are correctness and response time, hence, we measured the effectiveness of rank prediction and time taken to compute the model for credibility ranking. We compared the methods based on two evaluation metrics, NDCG (Normalized Discounted Cumulative Gain) and execution times. For evaluating the relevance ranking results, we first used the standard metric of NDCG [11]. NDCG is preferred over MAP (Mean Average Precision), since it captures data with multiple grades. 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 (5=Credible, 4=Maybe credible, 3= Incredible, 2=Relevant but no information, 1=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 NDCGs of the queries in the test set.

Feature vectors for all the tweets annotated for the events were given as input to the ranking algorithms as training dataset. The ranking algorithm first learns a model for credibility assessment and then tests the results on the testing dataset. We applied 4-fold cross validation to our results. Table 4 shows the results obtained for the credibility ranking. We observe that AdaRank and Coordinate Ascent perform best in terms of $NDCG@n$ among all the algorithms in ranking the tweets correctly for their credibility; SVM-rank is a close second. The table also presents the learning and ranking times for each of the methods. The ranking time of all methods was nearly one second, but the learning time for SVM-rank was, as expected, much shorter than for any of the other methods. Considering these results, we implemented our system using SVM-rank.

For the above ranking task, we have considered only data collected for the six events of 2013 for this research work. We then analyzed if we can consider the data annotated in our 2012 study for fourteen events [8]. For

[5] https://dev.twitter.com/docs/api/1.1/get/search/tweets

[6] The WOT reputation system computes website reputations using ratings received from users and information from third-party sources. The API returns a reputations, categories, and third-party blacklist information for web URLs. https://www.mywot.com/

[7] http://sourceforge.net/p/lemur/wiki/RankLib/

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

Table 4: Evaluation of various ranking algorithms in terms of normalized discounted cumulative gain (NDCG) and execution times. Boldface values in each row indicate the best results.

	AdaRank	Coord. Ascent	RankBoost	SVM-rank
NDCG@25	**0.6773**	0.5358	0.6736	0.3951
NDCG@50	**0.6861**	0.5194	0.6825	0.4919
NDCG@75	0.6949	**0.7521**	0.6890	0.6188
NDCG@100	0.6669	**0.7607**	0.6826	0.7219
Time (learn+rank)	35-40 secs	1 min	35-40 secs	9-10 secs
Time (rank)	1 sec	1 sec	1 sec	1 sec

checking the same, we trained the ranking model using SVM-rank on 2011 events data and tested on 2013 events data. Table 5 shows the results of this experiment. We observe that for the given feature vectors, the SVM-rank gives good results when trained and tested on the same year dataset, when trained on 2011 and tested on 2013 dataset, we observe there is a drastic drop in the accuracy. This can be attributed to various factors like evolution of Twitter and its usage during large scale events over past few years.

Table 5: Performance of SVM-rank algorithm in credibility ranking of tweets using 2011 and 2013 data. We observe a significant drop in NDCG when training on data from one year and testing on data from a different year.

Training	NDCG @25	NDCG @50	NDCG @100	Testing
2011 events	0.4765	0.5966	0.7359	2011 events
2013 events	0.3951	0.4919	0.7219	2013 events
2011 events	0.3743	0.3693	0.3783	2013 events

Table 6 shows the top 10 features of the models for credibility ranking built for 2011 events [8] and 2013 events [this paper]. For both sets, we observe that both tweet- (e.g. number of characters in a tweet, presence of URL in tweet) and user-based (e.g. ratio of friends / followers, user location) features are important. The fact that many of the top features are different for both set of events, explains why the 2011 data should not be used to predict real-time credibility now. It also highlights that there is temporal evolution in the landscape of credibility prediction models. Hence, whatever system or model we build in this work, will require to be updated and re-trained in the future.

5. IMPLEMENTATION

In order to measure the effectiveness of above techniques and models in a large scale scenario, we developed *TweetCred* a real-time platform to measure the credibility of content on Twitter. *TweetCred* platform described herein consists of a Chrome extension, Web application, Twitter data acquisition module and credibility score computation module. Clients (Chrome extension, Web application) interface with credibility score computation module on the web server over RESTful HTTP APIs. We used credibility ranking model trained in the previous section using SVM-rank method as the backend for *TweetCred* system. When a new tweet comes in real-time, the rank of the tweet is predicted according to the pre-learnt model of SVM-rank, and displayed to the user on a scale of 1 (low credibility) to 7 (high credibility). For distinction between the ratings from 1 to 7, we defined the threshold values based on our training and testing values of our experiment described in previous section. In the initial pilot study, conducted for *TweetCred* we used the *Likert Scale* of score 1 - 5 for showing credibility for a tweet. [9] But, the users' found it difficult to differentiate between a high credibility score of 4 and a low credibility score of 2, as the difference in values seemed very less. They were more comfortable with a slightly larger scale of 1 - 7 ranking.

Table 6: Top 10 features obtained using SVM-rank for ranking tweets according to their credibility. We observe that many of the top features are different for both scenarios.

2013	2011
Tweet contains *via*	Presence of $ symbol
No. of characters in tweet	Tweet contains URL
Unique characters in tweet	User has location in profile
No. of words in tweet	User has URL in profile
User has location in profile	No. of characters in tweet
Number of retweets	No. of words in tweet
Age of tweet	Unique characters in tweet
Tweet contains URL	Friends / Followers
Statuses / Followers	Favorites / Statuses
Friends / Followers	User is verified

5.1 Design and Technology Details

In order to ensure that a user obtains credibility of tweets within the Twitter ecosystem, i.e. without logging into another application we developed the *TweetCred* Chrome Extension, which would display credibility score of each tweet embedded in the Twitter webpage. Figure 4 shows the basic architecture of the system. The flow of information in *TweetCred* is as follows: A user logs on to his Twitter account on *twitter.com* website, once the tweets starts loading on the webpage, the chrome extension passes the IDs of tweets displayed on the page to our web sever on which the

[9]http://www.clemson.edu/centers-institutes/tourism/documents/sample-scales.pdf

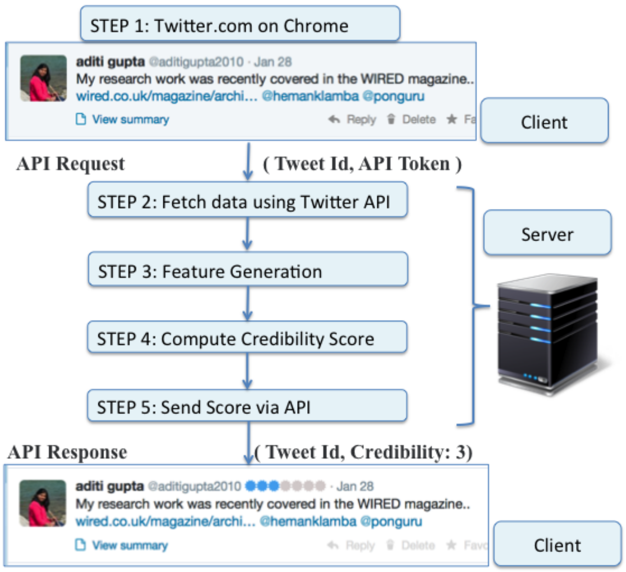


Figure 4: Data flow steps of the *TweetCred* extension and API.

credibility score computation module is hosted. We do not scrape the tweet or user information from the raw HTML of web page and merely pass the tweet IDs to web server. From the server an API request is made to *twitter.com* to fetch the complete *JSON* object of an individual tweet. Once the complete data for the tweet is obtained, the feature vectors are generated for the tweet, and then the credibility is computed using the prediction model of SVM-rank. The credibility score (between 1 - 7) computed using the threshold values, is now sent back to the user's browser via HTTP API, where it is displayed alongside each tweet. Figure 1 shows the credibility score of tweets as shown to the users on their Twitter timeline.

For the first iteration of *TweetCred*, Chrome extension was the ubiquitous choice, since, it enjoys the maximum user base among various Web Browsers.[10] In order to minimize computation load on the web browser, heavy computations were offloaded to the web server, hence the browser extension had a minimalistic memory and CPU footprint. This design ensures that the system is scalable and would not result in any performance bottleneck on client's web browser. All feature extraction and credibility computation scripts were written in *Python* with *MySQL* as a database back-end. The RESTful APIs were implemented using *PHP*. The hardware for backend was a mid-range server (Intel Xeon E5-2640 2.50GHz, 8GB RDIMM).

User feedback. To evaluate the performance of *TweetCred*, a feedback mechanism was added to the user interface. When end users were shown the credibility score for a tweet, they were given the option to provide feedback to the system, indicating if they agree or disagree with the credibility score for each tweet. Figures 5(a) and 5(b) show the two options given to the user upon hovering over the displayed credibility score. In case the user disagreed with the credibility rating, s/he was asked to provide what s/he considered should be the credibility rating as shown in Figure 5(c). The feedback provided by the user is sent over a separate REST API endpoint and recorded in the database.

5.2 Performance and Accuracy Evaluation

We uploaded *TweetCred* on Chrome Web Store, [11] and advertised its presence via OSM and blogs. We analyzed the deployment and usage activity of *TweetCred* from April 27th, 2014 to May 17th, 2014. *TweetCred* is a live system used by Twitter users, for analysis and statistics in this paper we consider data logged for only above mentioned *three* weeks. *TweetCred* was mostly used with the Chrome extension and few users explored and evaluated the browser-based version of the system. 717 unique Twitter accounts used *TweetCred* from 601 browser installations from Chrome web store–since the same browser can be used with more than one Twitter account. Table 7 presents a summary of usage statistics for *TweetCred*.

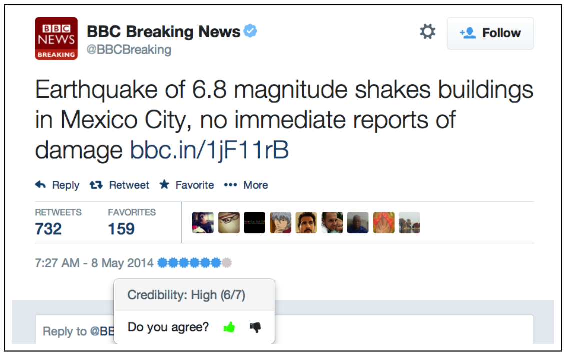
Table 7: Summary statistics for the usage of *TweetCred*.

Date of launch of *TweetCred*	27 Apr, 2014
Credibility score requests for all tweets	1,339,079
Credibility score requests for unique tweets	1,108,015
Credibility score requests for tweets (Chrome extension)	1,330,218
Credibility score requests for tweets (Browser version)	8,858
Downloads from Chrome store	601
Unique Twitter users	717
Feedback was given for tweets	936
Unique users who gave feedback	166
Unique tweets which received feedback	926

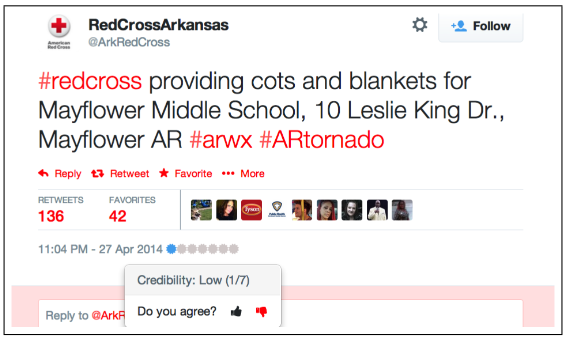
In total 1,339,079 API requests for the credibility score of a tweet were made on 1,108,015 unique tweets. Credibility scores were cached for 15 minutes, meaning that if a user requests the score of a tweet whose score was requested less than 15 minutes ago, the previously-computed score was re-used. After this period of time, cached credibility scores were discarded and computed again if needed, to account for changes in tweet or user features such as the number of followers, retweets, favorites and replies. In order to evaluate the performance

[10] http://www.w3schools.com/browsers/browsers_stats.asp

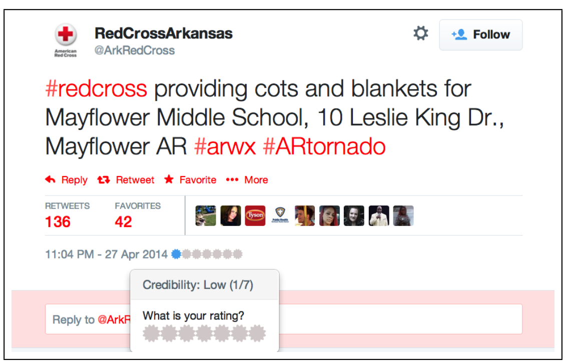
[11] http://bit.ly/tweetcredchrome



(a) A tweet from BBC's official account rated with high credibility (6 out of 7), showing agree/disagree buttons for feedback.



(b) A tweet from Red Cross's official account rated with low credibility (1 out of 7), showing agree/disagree buttons for feedback.



(c) A tweet from Red Cross's official account rated with low credibility (1 out of 7), showing user rating buttons for feedback.

Figure 5: Users can provide feedback by clicking on the "thumbs up" or "thumbs down" icons. Additionally, they can suggest what they would consider to be the correct level of credibility.

and usability of *TweetCred* we analyzed users' feedback, server logs and usage statistics.

Users who installed *TweetCred* are a diverse sample of Twitter users. We looked at their characteristics including the distribution of number of tweets evaluated and number of followers of users. We observed highly-skewed distributions as expected. For instance, one user used *TweetCred* to evaluate more than 50,000 tweets, while the majority of users evaluated less than 1,000 tweets. In terms of number of followers, the most followed user among those who installed *TweetCred* had 1.4 million followers.

5.2.1 Response Time

We analyzed the response time of the browser extension, measured as the elapsed time from the moment in which a request is sent to our system to the moment in which the resulting credibility score is returned by the server to the extension. Figure 6 shows the CDF of response times for all 1.1 million API requests. From the figure we can observe that for 84% of the users the response time was less than 6 seconds, while for 99% of the users the response time was under 10 seconds.

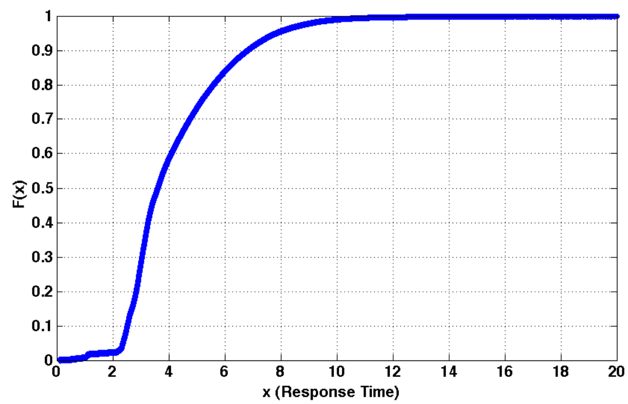


Figure 6: CDF of response time of *TweetCred*. For 84% of the users, response time was less than 6 seconds and for 99% of the users, the response time was under 10 seconds.

In addition to individual response time for API requests, it is also essential that under high load conditions, the response time of the system is still under acceptable limits. We plotted the average response time for all requests and the number of requests (load) sent to the credibility computation system per hour. Figure 7 shows that even during considerable load (more than 8,000 requests per hour), the average response time of the system remained under 8 seconds. There is a gradual increase in the response time every a few hours as the backend database becomes larger, but the response time drops again drops when the database is auto-flushed after a few hours.

5.2.2 User Feedback

We received feedback from users of our system in two ways, firstly, the users could give their feedback on each tweet for which a credibility score was computed. Secondly, we asked our users to fill a usability survey on our website. Out of 1.1 *million* tweets for which the

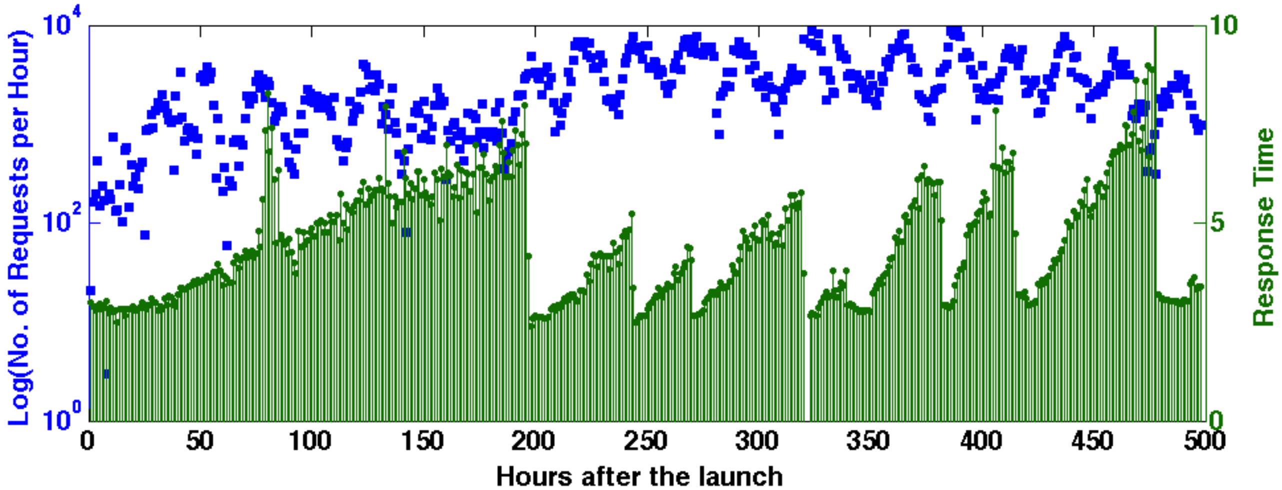


Figure 7: Number of requests per hour to *TweetCred* system and average response time per hour.

credibility scored was computed by *TweetCred*, for 936 of them we received feedback from our users. Users had the option of either agreeing or disagreeing with our score. In case they disagreed, they were asked to mark the correct score according to them. Table 8 shows the break-down of the received feedback. We observed that for 43% of tweets for which user's provided feedback agreed with the credibility score given by *TweetCred*, while 57% disagreed–we expect this to be the result of self-selection bias due to cognitive dissonance: users are moved to react when they see something that does not match their expectations. In addition to 43% for which they agreed, a further 25% of tweets, their disagreement was of 2 points or less (on the 7-point scale). Figure 8 shows the number of tweets per user for which *TweetCred* feedback was received.

Table 8: Feedback given by users of *TweetCred* on specific tweets ($n = 936$).

Agreed with score	42.95%
Disagreed: score should have been higher	46.26%
Disagreed: score should have been lower	10.79%
Disagreed by 1 point	10.04%
Disagreed by 2 points	15.17%
Disagreed by 3 points	11.86%
Disagreed by 4 points	8.65%
Disagreed by 5 points	5.77%
Disagreed by 6 points	5.56%

For the 57% tweets for which users disagreed with our score, for 46% of the tweets the users felt that credibility score should have been higher than the one given by *TweetCred*, while for approximately 11% thought it should have been lower. We think that one of the reason why users felt that credibility score given by *TweetCred* was less, is because a user often trusts other users on Twitter, because of their real-world or past online interactions. Such local friendships and trust relationships are not captured by a generalized model built for entire Twitter space.

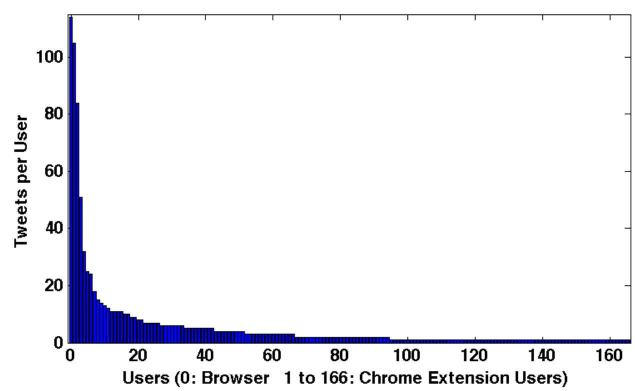


Figure 8: Distribution for number of tweets per user for which we received feedback.

Usability Survey for *TweetCred*. We conducted an online survey to assess the usability of the *TweetCred* browser extension. An unobtrusive link to the survey appeared on the right corner of Chrome's address bar when users visited Twitter.[12] The survey link was accessible only to those users who had installed the extension, this was done to ensure that only actual users of the system gave their feedback. A total of 52 users participated. The survey contained the standard 10 questions of the *System Usability Scale* (SUS) [1]. In addition to SUS questions, we also added questions about users' demographics such as gender, age, etc. We ob-

[12] `http://twitdigest.iiitd.edu.in/TweetCred/feedback.html`

tained an overall SUS score of 70 for *TweetCred*, which is considered above average from a system's usability perspective.[13] In the survey, 78% of the users found *TweetCred* easy to use (strongly agree / agree); 22% of the users thought there were inconsistencies in the system (strongly agree / agree); and about 80% of the users said that they may like to use *TweetCred* in their daily life. Some of the comments we received about *TweetCred* in the survey as well as from tweets were:

- "I plan on using this to monitor public safety situations on behalf of the City of [withheld]'s Office of Emergency Management."
- "Very clever idea but Twitter's strength is simplicity - I found this a distraction for daily use."
- "It's been good using #TweetCred & will stick around with it, thanks!"
- "It's unclear what the 3, 4 or 5 point rating mean on opinions / jokes, versus factual statements."

5.2.3 Credibility Rating by *TweetCred*

The credibility score was computed by *TweetCred* for about 1.1 million tweets. Figure 9 shows the distribution of scores. In addition to showing the distribution for all analyzed tweets, we also used keywords to select tweets corresponding to three crisis events that occurred during our experiment timeline: crisis in Ukraine (3,637 tweets), Oklahoma/Arkansas tornadoes (1,362 tweets) and an earthquake in Mexico (1,476 tweets).

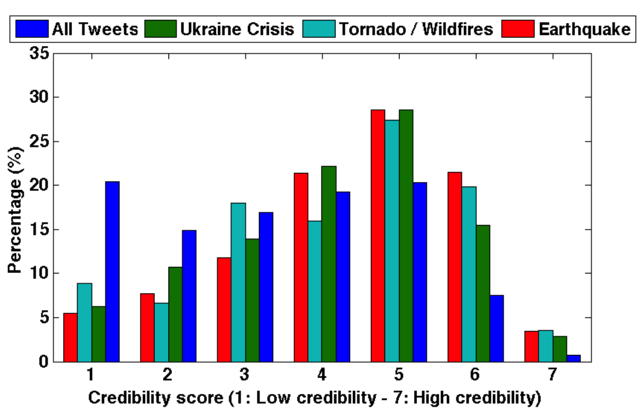


Figure 9: Distribution of credibility scores (1=low, 7=high) as given by *TweetCred*. We observe that during crisis events there are more tweets with high credibility than during non-crisis times.

Figure 9 shows that among all tweets scored by *TweetCred*, about 8% were marked with high credibility scores (6 or 7), while during crisis events more than 20% obtained these scores. Similarly, we observed a higher percentage of tweets getting low credibility for general tweets as compared to crisis tweets. These observations indicate that a crisis may generate a larger volume of credible information-rich content in Twitter, an interesting phenomenon that merits further study.

6. DISCUSSION

We have described the research, development, and evaluation of *TweetCred*, a real-time web-based system to automatically evaluate the credibility of content on Twitter. The system provides a credibility rating from 1 (low credibility) to 7 (high credibility) for each tweet on a user's Twitter timeline. The score is computed using a supervised automated ranking algorithm that determines the credibility of a tweet based on more than 45 features. All features can be computed online for single tweets. They include the tweets content, characteristics of its author, and information about external URLs. The system is trained on human labels obtained using crowd-sourcing. We obtained useful insights on how credibility evaluation models evolve over time and the features which indicate credibility change with time.

Our live deployment of *TweetCred* spanned three weeks, in which more than 717 unique Twitter users used our system. The system achieved a response time under 6 seconds for 84% of the users. They used *TweetCred* to compute credibility ratings for more than 1.1 *million* unique tweets and gave back feedback for about 936 Tweets. For about 43% of the tweets, the users agreed with the credibility score computed by *TweetCred*. For a further 25% of tweets, their disagreement was of 2 points or less (on the 7-point scale). Around 46% users thought the credibility scores should have been higher than that given by *TweetCred*, and 11% thought it should have been lower. Many of the users felt that the credibility score was low because, the model for credibility ranking developed in this work is a generalized model, it does not take into account, the real-world or online relationships of an user. In future, we would like to make *TweetCred* customizable for each user, in which the user can train the system according to him.

TweetCred stirred a wide debate on Twitter regarding the problem and solutions for the credibility assessment problem on Twitter. Our work was covered in many news websites and blogs such as Washington Post,[14] the New Yorker,[15] and the Daily Dot[16] among others, generating debates in these platforms also.

Future work. Some of the insights we obtained from our live experiment will help us build a more robust *TweetCred* in the next iterations. Some of the proposed enhancements we aim to introduce include:

[13] `http://www.measuringusability.com/sus.php`

[14] `http://wapo.st/1pWE0Wd`

[15] `http://newyorker.com/online/blogs/elements/2014/05/can-tweetcred-solve-twitters-credibility-problem.html`

[16] `http://www.dailydot.com/technology/tweetcred-chrome-extension-addon-plugin/`

- The meaning of *information credibility* is not clear for all users, particularly when applied to non-newsworthy content, which is frequent in Twitter. In these cases, and in cases where there is little or no content in the tweet, we should output a special symbol / outcome (e.g. "not enough information").
- More research is needed to find the most effective method of displaying the credibility score to users. We could use less levels (e.g. three instead of seven), or show only a warning next to the low-credibility items, or highlight the high-credibility ones.
- We have not yet reached a plateau in terms of ranking accuracy, which means that more training data should increase the effectiveness of our model. Moving to an online learning model in which we learn from user's feedback would also be an important step.
- *TweetCred* works currently only with the Chrome browser; we are developing a version that is compatible also with Mozilla Firefox.

TweetCred is the first practical system for credibility on Twitter. It acted as a catalyst in stirring up a debate and consciousness among Internet users regarding this issue, and has achieved to obtain partial success in solving the information credibility problem in social media. This research paper provided us with useful insights on how to make it a more robust and usable system in future.

Acknowledgments. We would like to thank Nilaksh Das and Mayank Gupta in helping us with the web development of *TweetCred* system. We would like to express our sincerest thanks to all members of Precog, Cybersecurity Education and Research Centre at IIIT-Delhi, and Qatar Computing Research Institute for their continued feedback and support.[17] We would like to thank the Government of India for funding this project.

7. REFERENCES

[1] J. Brooke. SUS: A quick and dirty usability scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, editors, *Usability evaluation in industry*. Taylor and Francis, London, 1996.

[2] Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom*, 2011.

[3] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on Twitter. In *Proc. WWW*, pages 675–684. ACM, 2011.

[4] William J. Corvey, Sudha Verma, Sarah Vieweg, Martha Palmer, and James H. Martin. Foundations of a multilayer annotation framework for twitter communications during crisis events. In *Proc. International Conference on Language Resources and Evaluation (LREC'12)*, may 2012.

[5] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, 2003.

[6] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proc. SIGIR*, 2012.

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

[8] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proc. 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8. ACM, 2012.

[9] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter. In *Proc. Eighth IEEE APWG eCrime Research Summit (eCRS)*, page 12. IEEE, 2013.

[10] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proc. WWW companion*, pages 729–736. International World Wide Web Conferences Steering Committee, 2013.

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

[12] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. KDD*, pages 133–142. ACM, 2002.

[13] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: can we trust what we rt? In *Proc. First Workshop on Social Media Analytics*, SOMA '10, pages 71–79. ACM, 2010.

[14] Donald Metzler and W Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[15] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proc. CSCW*. ACM, 2012.

[17] http://precog.iiitd.edu.in/
http://cerc.iiitd.ac.in/
http://www.qcri.org.qa/

[16] J. O'Donovan, B. Kang, G. Meyer, T. Hllerer, and S. Adali. Credibility in context: An analysis of feature distributions in twitter. *ASE/IEEE International Conference on Social Computing, SocialCom*, 2012.

[17] Onook Oh, Manish Agrawal, and H. Raghav Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, March 2011.

[18] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proc. WWW*, 2011.

[19] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. Mining YouTube to Discover Hate Videos, Users and Hidden Communities. *Accepted at Sixth Asia Information Retrieval Societies Conference*, 2010.

[20] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *Proc. of ICWSM*, 2011.

[21] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proc. SIGCHI*, CHI '10, pages 1079–1088. ACM, 2010.

[22] Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information credibility on twitter in emergency situation. In *Proc. Pacific Asia conference on Intelligence and Security Informatics*, PAISI'12, 2012.

[23] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proc. CIKM*, pages 1980–1984, New York, NY, USA, 2012. ACM.

[24] Jun Xu and Hang Li. Adarank: A boosting algorithm for information retrieval. In *Proc. SIGIR*, pages 391–398, New York, NY, USA, 2007. ACM.

[25] Jiang Yang, Scott Counts, Meredith Ringel Morris, and Aaron Hoff. Microblog credibility perceptions: Comparing the usa and china. In *Proc. CSCW*, pages 575–586, 2013.