Information Credibility on Twitter

Carlos Castillo[1] Marcelo Mendoza[2,3] Barbara Poblete[2,4]

{chato,bpoblete}@yahoo-inc.com, marcelo.mendoza@usm.cl
[1]Yahoo! Research Barcelona, Spain
[2]Yahoo! Research Latin America, Chile
[3]Universidad Técnica Federico Santa María, Chile
[4]Department of Computer Science, University of Chile

ABSTRACT

We analyze the information credibility of news propagated through Twitter, a popular microblogging service. Previous research has shown that most of the messages posted on Twitter are truthful, but the service is also used to spread misinformation and false rumors, often unintentionally.

On this paper we focus on automatic methods for assessing the credibility of a given set of tweets. Specifically, we analyze microblog postings related to "trending" topics, and classify them as credible or not credible, based on features extracted from them. We use features from users' posting and re-posting ("*re-tweeting*") behavior, from the text of the posts, and from citations to external sources.

We evaluate our methods using a significant number of human assessments about the credibility of items on a recent sample of Twitter postings. Our results shows that there are measurable differences in the way messages propagate, that can be used to classify them automatically as credible or not credible, with precision and recall in the range of 70% to 80%.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Social Media Analytics, Social Media Credibility, Twitter

1. INTRODUCTION

Twitter is a micro-blogging service that counts with millions of users from all over the world. It allows users to post and exchange 140-character-long messages, which are also known as *tweets*. Twitter is used through a wide variety of clients, from which a large portion – 46% of active users[1] – correspond to mobile users. Tweets can be published by sending e-mails, sending SMS text-messages and directly from smartphones using a wide array of Web-based services. Therefore, Twitter facilitates real-time propagation of information to a large group of users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events.

For instance, in an emergency situation [32], some users generate information either by providing first-person observations or by bringing relevant knowledge from external sources into Twitter. In particular, information from official and reputable sources is considered valuable and actively sought and propagated. From this pool of information, other users synthesize and elaborate to produce derived interpretations in a continuous process.

This process can gather, filter, and propagate information very rapidly, but it may not be able to separate true information from false rumors. Indeed, in [19] we observed that immediately after the 2010 earthquake in Chile, when information from official sources was scarce, several rumors posted and re-posted on Twitter contributed to increase the sense of chaos and insecurity in the local population. However, we also observed that information which turned out to be false, was much more questioned than information which ended up being true. This seems to indicate that the social network somehow tends to favor valid information, over false rumors.

Social media credibility. The focus of our research is the *credibility* of information spread through social media networks. Over 20 years ago, Fogg and Tseng [10] described credibility as a *perceived quality* composed of *multiple dimensions*. In this paper we use credibility in the sense of believability: "*offering reasonable grounds for being believed*"[2]. We first ask users to state if they consider that a certain set of messages corresponds to a newsworthy event (as opposed to being only informal conversations). Next, for those messages considered as related to newsworthy events, we ask another group of users to state if they believe those messages are likely to be true or false.

Our main objective is to determine if we can automatically assess the level of credibility of content posted on Twitter. Our primary hypothesis is that there are signals available in the social media environment itself that enable users to assess information credibility. In this context we define *social media credibility* as the aspect of information credibility that can be assessed using only the information available in a social media platform.

[1] http://blog.twitter.com/2010/09/evolving-ecosystem.html

[2] http://www.merriam-webster.com/dictionary/credible

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

Contributions and paper organization. Our method is based on supervised learning, and the first step is to build a dataset for studying credibility on Twitter. We first extract a set of relevant discussion topics by studying bursts of activity. Then, each topic is labeled by a group of human assessors according to whether it corresponds to a newsworthy information/event or to informal conversation. After the dataset is created, each item of the former class is assessed on its level of credibility by another group of judges. This is described in Section 3.

Next, we extract relevant features from each labeled topic and use them to build a classifier that attempts to automatically determine if a topic corresponds to a newsworthy information/event, and then to automatically assess its level of credibility. This is described in Section 4. Finally, Section 5 presents our conclusions and directions for future work.

The next section outlines previous work related to our current research.

2. RELATED WORK

The literature on information credibility is extensive, so in this section our coverage of it is by no means complete. We just provide an outline of the research that is most closely related to ours.

Credibility of online news in traditional media and blogs. The perception of users with respect to the credibility of online news seems to be positive, in general. People trust the Internet as a news source as much as other media, with the exception of newspapers [8]. Therefore, and in part due to this, the Internet is the most important resource for news in the US among people under the age of 30, according to a survey in 2008 [23], and second only to television in the case of general audiences.

Among online news sites, blogs are considered less trustworthy than traditional news sites. A survey in 2005 showed that, even among young people, blogs are seen as significantly less trustworthy than traditional news sites [34]. An exception seem to be users with political interests, which rate the credibility of blogs sites high, particularly when they are themselves heavy blog users [14].

Twitter as a news media. While most messages on Twitter are conversation and chatter, people also use it to share relevant information and to report news [13, 22, 21]. Indeed, the majority of "trending topics" –keywords that experiment a sharp increase in frequency– can be considered "headline news or persistent news" [16].

The fact that Twitter echoes news stories from traditional media can be exploited to use Twitter, e.g. to track epidemics [17], detect news events [28], geolocate such events [27], and find controversial emerging controversial topics [24]. Recently Mathioudakis and Koudas [18] described an on-line monitoring system to perform trend detection over the Twitter stream. In this paper we assume that a system for trend detection exists (we use [18]) and focus on the issues related to labeling those trends or events.

Twitter has been used widely during emergency situations, such as wildfires [6], hurricanes [12], floods [32, 33, 31] and earthquakes [15, 7]. Journalists have hailed the immediacy of the service which allowed "to report breaking news quickly – in many cases, more rapidly than most mainstream media outlets" [25]. The correlation of the magnitude of real-world events and Twitter activity prompted researcher Markus Strohmaier to coin the term "Twicalli scale"[3].

Credibility of news on Twitter. In a recent user study, it was found that providing information to users about the estimated credibility of online content was very useful and valuable to them [30]. In absence of this external information, perceptions of credibility online are strongly influenced by style-related attributes, including visual design, which are not directly related to the content itself [9]. Users also may change their perception of credibility of a blog posting depending on the (supposed) gender of the author [3].

In this light the results of the experiment described in [29] are not surprising. In the experiment, the headline of a news item was presented to users in different ways, i.e. as posted in a traditional media website, as a blog, and as a post on Twitter. Users found the same news headline significantly less credible when presented on Twitter.

This distrust may not be completely ungrounded. Major search engines are starting to prominently display search results from the "real-time web" (blog and microblog postings), particularly for trending topics. This has attracted spammers that use Twitter to attract visitors to (typically) web pages offering products or services [4, 11, 36]. It has also increased the potential impact of orchestrated attacks that spread lies and misinformation. Twitter is currently being used as a tool for political propaganda [20].

Misinformation can also be spread unwillingly. For instance, on November 2010 the Twitter account of the presidential adviser for disaster management of Indonesia was hacked.[4] The hacker then used the account to post a false tsunami warning. On January 2011 rumors of a shooting in the Oxford Circus in London, spread rapidly through Twitter. A large collection of screenshots of those tweets can be found online.[5]

Recently, the Truthy[6] service from researchers at Indiana University, has started to collect, analyze and visualize the spread of tweets belonging to "trending topics". Features collected from the tweets are used to compute a *truthiness* score for a set of tweets [26]. Those sets with low truthiness score are more likely to be part of a campaign to deceive users. Instead, in our work we do not focus specifically on detecting willful deception, but look for factors that can be used to automatically approximate users' perceptions of credibility.

3. DATA COLLECTION

We focus on time-sensitive information, in particular on current news events. This section describes how we collected a set of messages related to news events from Twitter.

3.1 Automatic event detection

We use Twitter events detected by *Twitter Monitor* [18][7] during a 2-months period. Twitter Monitor is an on-line monitoring system which detects sharp increases ("bursts") in the frequency of sets of keywords found in messages.

[3] http://mstrohm.wordpress.com/2010/01/15/measuring-earthquakes-on-twitter-the-twicalli-scale/

[4] http://thejakartaglobe.com/home/government-disaster-\-advisors-twitter-hacked-used-to-send-tsunami-warning/408447

[5] http://www.exquisitetweets.com/collection/abscond/152

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

[7] http://www.twittermonitor.net/

For every burst detected, Twitter Monitor provides a keyword-based query. This query is of the form $(A \wedge B)$ where A is a conjunction of keywords or hashtags and B is a disjunction of them. For instance, $((cinco \wedge mayo) \wedge (mexican \vee party \vee celebrate))$ refers to the celebrations of "*cinco de mayo*" in Mexico. We collected all the tweets matching the query during a 2-day window centered on the peak of every burst. Each of these sub-sets of tweets corresponds to what we call a *topic*. We collected over 2,500 such topics. Some example topics are shown in Table 1.

Table 1: Example topics in April to July 2010. A tweet on a topic must contain all of the boldfaced words and at least one of the non-boldfaced ones.

Peak	Keywords
	News
22-Apr	**recycle**, **earth**, save, reduce, reuse, #earthday
3-May	**flood**, **nashville**, relief, setup, victims, pls
5-Jun	**notebook**, **movie**, makes, cry, watchin, story
13-Jun	**vuvuzelas**, **banned**, clamor, chiefs, fifa, silence
9-Jul	**sues**, **ntp**, tech, patents, apple, companies
	Conversation
17-Jun	**goodnight**, **bed**, dreams, tired, sweet, early
2-May	**hangover**, **woke**, goes, worst, drink, wake

In the table we have separated two broad types of topics: *news* and *conversation*, following the broad categories found in [13, 22]. The fact that conversation-type of messages can be bursty is a case of endogenous bursts of activity that occur this type of social system [5].

There are large variations on the number of tweets found in each topic. The distribution is shown in Figure 1. In our final dataset, we kept all the cases having at most 10,000 tweets, which corresponds to 99% of them.

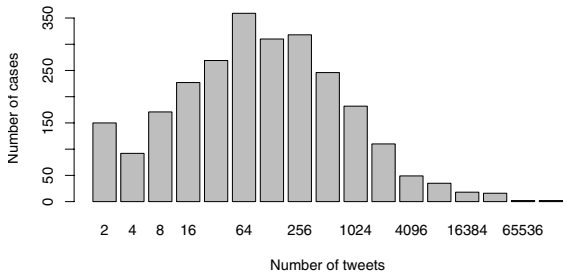

Figure 1: Distribution of tweets per topic.

3.2 Newsworthy topic assessments

Our first labeling round was intended to separate topics which spread information about a news event, from the cases which correspond to personal opinions and chat. In other words, we separate messages that are of potential interest to a broad set of people, from conversations that are of little importance outside a reduced circle of friends [2].

For this task we used Mechanical Turk[8], where we asked evaluators to assist us. We showed evaluators a sample of 10 tweets in each topic and the list of keywords provided by Twitter Monitor, and asked if most of the messages were spreading news about a specific event (labeled as class `NEWS`) or mostly comments or conversation (labeled as class `CHAT`). For each topic we also asked evaluators to provide a short descriptive sentence for the topic. The sentence allow us to discard answers without proper justification, reducing the amount of click spammers in the evaluation system.

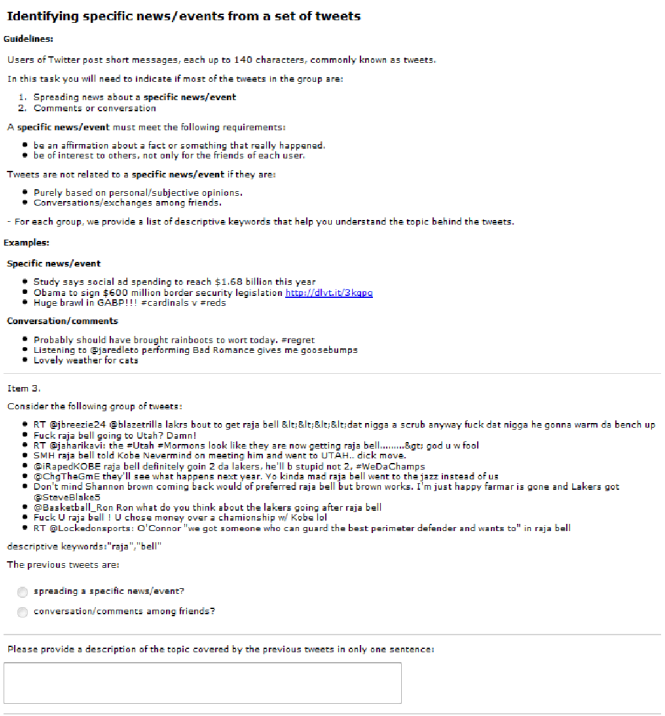

Figure 2: User interface for labeling newsworthy topics.

As shown in Figure 3.2, we provided guidelines and examples of each class. `NEWS` was described as statements about a fact or an actual event of interest to others, not only to the friends of the author of each message. `CHAT` was described as messages purely based on personal/subjective opinions and/or conversations/exchanges among friends.

Randomly we selected 383 topics from the Twitter Monitor collection to be evaluated using Mechanical Turk. We grouped topics at random, in sets of 3, for each task (called "human intelligence task" or *HIT* in Mechanical Turk jargon). During ten days evaluators were asked to assess HITs, and we asked for 7 different evaluators for each HIT. Evaluations that did not provide the short descriptive sentence were discarded.

A class label for a topic was assigned if 5 out of 7 evaluators agreed on the label. In another case we label the instance as `UNSURE`. Using this procedure, 35.6% of the topics (136 cases) were labeled as `UNSURE`, due to insufficient agreement. The percentage of cases labeled as `NEWS` was 29.5% (113 cases), and as `CHAT`, 34.9% (134 cases).

3.3 Credibility assessment

Next we focus on the credibility assessment task. To do this, we ran an event supervised classifier over the collection of 2,524 cases detected by Twitter Monitor. We will discuss details of this classifier in Section 4. Our classifier labels a total of 747 cases as `NEWS`. Using this collection of instances,

[8] `http://www.mturk.com`

we asked mechanical turk evaluators to indicate credibility levels for each case. For each one we provided a sample of 10 tweets followed by a short descriptive sentence that help them to understand the topic behind those tweets.

In this evaluation we considered four levels of credibility: (i) almost certainly true, (ii) likely to be false, (iii) almost certainly false, and (iv) "I can't decide". We asked also evaluators to provide a short sentence to justify their answers, and we discarded evaluations lacking that justification sentence. An example of this task is shown in Figure 3.3. We asked for 7 different assessments for each HIT. Labels for each topic were decided by majority, requiring agreement of at least 5 evaluators.

Figure 3: User interface for assessing credibility.

In a preliminary round of evaluation, almost all of the cases where labeled as "likely to be true", which turned out to be a very general statement and hence useless for our purposes. Hence, we removed the "likely to be true" option, forcing the evaluators to choose one of the others. The percentage of cases identified as "almost certainly true" was 41% (306 cases), "likely to be false" accounted for 31.8% (237 cases), "almost certainly false" accounted only for 8.6% (65 cases), while 18.6% (139 cases) were considered uncertain by evaluators, labeling these cases as "ambiguous".

4. AUTOMATIC CREDIBILITY ANALYSIS

On this section we discuss how, given a stream of messages associated to certain topics, we can automatically determine which topics are newsworthy, and then automatically assign to each newsworthy topic a credibility label.

4.1 Social media credibility

Our main hypothesis is that the level of credibility of information disseminated through social media can be estimated automatically. We believe that there are several factors that can be observed in the social media platform itself, and that are useful to asses information credibility. These factors include:

- the reactions that certain topics generate and the emotion conveyed by users discussing the topic: e.g. if they use opinion expressions that represent positive or negative sentiments about the topic;
- the level of certainty of users propagating the information: e.g. if they question the information that is given to them, or not;
- the external sources cited: e.g. if they cite a specific URL with the information they are propagating, and if that source is a popular domain or not;
- characteristics of the users that propagate the information, e.g. the number of followers that each user has in the platform.

We propose a set of features to characterize each topic in our collections. These include some features specific to the Twitter platform, but most are quite generic and can be applied to other environments. Many of the features follow previous works including [1, 2, 12, 26].

Our feature set is listed in Table 2. We identify four types of features depending on their scope: message-based features, user-based features, topic-based features, and propagation-based features.

Message-based features consider characteristics of messages, these features can be Twitter-independent or Twitter-dependent. Twitter-independent features include: the length of a message, whether or not the text contains exclamation or question marks and the number of positive/negative sentiment words in a message. Twitter-dependent features include features such as: if the tweet contains a hashtag, and if the message is a re-tweet.

User-based features consider characteristics of the users which post messages, such as: registration age, number of followers, number of followees ("friends" in Twitter), and the number of tweets the user has authored in the past.

Topic-based features are aggregates computed from the previous two feature sets; for example, the fraction of tweets that contain URLs, the fraction of tweets with hashtags and the fraction of sentiment positive and negative in a set.

Propagation-based features consider characteristics related to the propagation tree that can be built from the re-tweets of a message. These includes features such as the depth of the re-tweet tree, or the number of initial tweets of a topic (it has been observed that this influences the impact of a message, e.g. in [35]).

4.2 Automatically finding newsworthy topics

We trained a supervised classifier to determine if a set of tweets describes a newsworthy event. Labels given by Mechanical Turk evaluators were used to conduct the supervised training phase. We trained a classifier considering the three classes but performing a cost-sensitive learning process, increasing the relevance for the prediction of instances in the `NEWS` class. We considered a cost matrix into account during the training process ignoring costs at prediction time. We built a cost-sensitive tree, weighting training instances according to the relative cost of the two kinds of error, false positives and false negatives. The cost matrix weighted misclassifications containing the `NEWS` class as 1.0,

Table 2: Features can be grouped into four clases having as scope the Message, User, Topic, and Propagation respectively

Scope	Feature	Description
Msg.	LENGTH CHARACTERS	Length of the text of the tweet, in characters
	LENGTH WORDS	... in number of words
	CONTAINS QUESTION MARK	Contains a question mark '?'
	CONTAINS EXCLAMATION MARK	... an exclamation mark '!'
	CONTAINS MULTI QUEST OR EXCL.	... multiple question or exclamation marks
	CONTAINS EMOTICON SMILE	... a "smiling" emoticon e.g. `:-) ;-)` ...
	CONTAINS EMOTICON FROWN	... a "frowning" emoticon e.g. `:-( ;-(` ...
	CONTAINS PRONOUN FIRST | SECOND | THIRD	... a personal pronoun in 1st, 2nd, or 3rd person. (3 features)
	COUNT UPPERCASE LETTERS	Fraction of capital letters in the tweet
	NUMBER OF URLS	Number of URLs contained on a tweet
	CONTAINS POPULAR DOMAIN TOP 100	Contains a URL whose domain is one of the 100 most popular ones
	CONTAINS POPULAR DOMAIN TOP 1000	... one of the 1,000 most popular ones
	CONTAINS POPULAR DOMAIN TOP 10000	... one of the 10,000 most popular ones
	CONTAINS USER MENTION	Mentions a user: e.g. `@cnnbrk`
	CONTAINS HASHTAG	Includes a hashtag: e.g. `#followfriday`
	CONTAINS STOCK SYMBOL	... a stock symbol: e.g. `$APPL`
	IS RETWEET	Is a re-tweet: contains 'RT '
	DAY WEEKDAY	The day of the week in which this tweet was written
	SENTIMENT POSITIVE WORDS	The number of positive words in the text
	SENTIMENT NEGATIVE WORDS	... negative words in the text
	SENTIMENT SCORE	Sum of ± 0.5 for weak positive/negative words, ± 1.0 for strong ones
User	REGISTRATION AGE	The time passed since the author registered his/her account, in days
	STATUSES COUNT	The number of tweets at posting time
	COUNT FOLLOWERS	Number of people following this author at posting time
	COUNT FRIENDS	Number of people this author is following at posting time
	IS VERIFIED	1.0 iff the author has a 'verified' account
	HAS DESCRIPTION	... a non-empty 'bio' at posting time
	HAS URL	... a non-empty homepage URL at posting time
Topic	COUNT TWEETS	Number of tweets
	AVERAGE LENGTH	Average length of a tweet
	FRACTION TWEETS QUESTION MARK	The fraction of tweets containing a question mark '?'
	FRACTION TWEETS EXCLAMATION MARK	... an exclamation mark '!'
	FRACTION TWEETS MULTI QUEST OR EXCL.	... multiple question or exclamation marks
	FRACTION TWEETS EMOTICON SMILE | FROWN	... emoticons smiling or frowning (2 features)
	CONTAINS PRONOUN FIRST | SECOND | THIRD	... a personal pronoun in 1st, 2nd, or 3rd person. (3 features)
	FRACTION TWEETS 30PCT UPPERCASE	... more than 30\% of characters in uppercase
	FRACTION TWEETS URL	The fraction of tweets containing a URL
	FRACTION TWEETS USER MENTION	... user mentions
	FRACTION TWEETS HASHTAG	... hashtags
	FRACTION TWEETS STOCK SYMBOL	... stock symbols
	FRACTION RETWEETS	The fraction of tweets that are re-tweets
	AVERAGE SENTIMENT SCORE	The average sentiment score of tweets
	FRACTION SENTIMENT POSITIVE	The fraction of tweets with a positive score
	FRACTION SENTIMENT NEGATIVE	... with a negative score
	FRACTION POPULAR DOMAIN TOP 100	The fraction of tweets with a URL in one of the top-100 domains
	FRACTION POPULAR DOMAIN TOP 1000	... in one of the top-1,000 domains
	FRACTION POPULAR DOMAIN TOP 10000	... in one of the top-10,000 domains
	COUNT DISTINCT EXPANDED URLS	The number of distinct URLs found after expanding short URLs
	SHARE MOST FREQUENT EXPANDED URL	The fraction of occurrences of the most frequent expanded URL
	COUNT DISTINCT SEEMINGLY SHORTENED URLS	The number of distinct short URLs
	COUNT DISTINCT HASHTAGS	The number of distinct hashtags
	SHARE MOST FREQUENT HASHTAG	The fraction of occurrences of the most frequent hashtag
	COUNT DISTINCT USERS MENTIONED	The number of distinct users mentioned in the tweets
	SHARE MOST FREQUENT USER MENTIONED	The fraction of user mentions of the most frequently mentioned user
	COUNT DISTINCT AUTHORS	The number of distinct authors of tweets
	SHARE MOST FREQUENT AUTHOR	The fraction of tweets authored by the most frequent author
	AUTHOR AVERAGE REGISTRATION AGE	The average of AUTHOR REGISTRATION AGE
	AUTHOR AVERAGE STATUSES COUNT	The average of AUTHOR STATUSES COUNT
	AUTHOR AVERAGE COUNT FOLLOWERS	... of AUTHOR COUNT FOLLOWERS
	AUTHOR AVERAGE COUNT FRIENDS	... of AUTHOR COUNT FRIENDS
	AUTHOR FRACTION IS VERIFIED	The fraction of tweets from verified authors
	AUTHOR FRACTION HAS DESCRIPTION	... from authors with a description
	AUTHOR FRACTION HAS URL	... from authors with a homepage URL
Prop.	PROPAGATION INITIAL TWEETS	The degree of the root in a propagation tree
	PROPAGATION MAX SUBTREE	The total number of tweets in the largest sub-tree of the root, plus one
	PROPAGATION MAX | AVG DEGREE	The maximum and average degree of a node that is not the root (2 feat.)
	PROPAGATION MAX | AVG DEPTH	The depth of a propagation tree (0=empty tree, 1=only initial tweets, 2=only re-tweets of the root) and its per-node average (2 features)
	PROPAGATION MAX LEVEL	The max. size of a level in the propagation tree (except children of root)

and misclassifications involving only the `CHAT` and `UNSURE` classes as 0.5.

We also used a bootstrapping strategy over the training dataset. A random sample of the dataset was obtained using sampling with replacement considering a uniform distribution for the probability of extracting an instance across the three classes. A sample size was defined to determine the size of the output dataset. We perform bootstrapping over the dataset with a sample size percentage equals to 300%

Table 3: Summary for classification of newsworthy topics.

Correctly Classified Instances	89.121 %
Kappa statistic	0.8368
Mean absolute error	0.0806
Root mean squared error	0.2569
Relative absolute error	18.1388 %
Root relative squared error	54.4912 %

Table 4: Results for the classification of newsworthy topics.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
NEWS	0.927	0.039	0.922	0.927	0.924
CHAT	0.874	0.054	0.892	0.874	0.883
UNSURE	0.873	0.07	0.86	0.873	0.866
W. Avg.	0.891	0.054	0.891	0.891	0.891

and feature normalization. We perform also a 3-fold cross validation strategy.

We tried a number of learning schemes including SVM, decision trees, decision rules, and Bayes networks. Results across these techniques were comparable, being best results achieved by a J48 decision tree method. A summary of the results obtained using the J48 learning algorithm is shown in Table 3. The supervised classifier achieves an accuracy equal to 89 %. The Kappa statistic indicates that the predictability of our classifier is significantly better than a random predictor. The details of the evaluation per class are shown in Table 4.

As we can observe, the classifier obtains very good results for the prediction of `NEWS` instances, achieving the best TP rate and FP rate across the classes. An F-measure equivalent to a 92% illustrate that specially for this class the classifier obtains a good balance for the precision-recall tradeoff.

4.3 Feature analysis for the credibility task

Before performing the automatic assessment of credibility, we analyze the distribution of features values. To do this we perform a best-feature selection process over the 747 cases of the `NEWS` collection, according to the labels provided by the credibility task. We used a best-first selection method which starts with the empty set of attributes and searches forward. The method selected 15 features, listed in Table 5.

Table 5: Best features selected using a best first attribute selection strategy.

	Min	Max	Mean	StdDev
AVG REG AGE	1	1326	346	156
AVG STAT CNT	173	53841	6771	6627
AVG CNT FOLLOWERS	5	9425	842	946
AVG CNT FRIENDS	0	1430	479	332
FR HAS URL	0	1	0.616	0.221
AVG SENT SCORE	-2	1.75	-0.038	0.656
FR SENT POS	0	1	0.312	0.317
FR SENT NEG	0	1	0.307	0.347
CNT DIST SHORT URLS	0	4031	121	419
SHR MOST FREQ AU	0	1	0.161	0.238
FR TW USER MENTION	0	1	0.225	0.214
FR TW QUEST MARK	0	1	0.091	0.146
FR EMOT SMILE	0	0.25	0.012	0.028
FR PRON FIRST	0	1	0.176	0.211
MAX LEV SIZE	0	632	46	114

As Table 5 shows, the first four features consider characteristics of users such as how long they have been Twitter users, the number of tweets that they have written at the posting time, and the number of followers/friends that they have in the platform. The next ten features are aggregated features computed from the set of tweets of each news event. Notice that features based on sentiment analysis are very relevant for this collection. Other relevant features consider if the message includes a URL, a user mention, or a question mark. The last feature considers information extracted from the propagation tree that is built from the re-tweets.

To illustrate the discriminative capacity of these features we deploy box plots for each of them. In this analysis we distinguish between cases that correspond to the "almost certainly true" class (labeled as class A), and the "likely to be false" and "almost certainly false" (labeled as class B). We exclude from the analysis cases labeled as "ambiguous". The box plots are shown in Figure 4.

As Figure 4 shows, several features exhibit a significant difference between both classes. More active users tend to spread more credible information, as well as users with newer user accounts but with many followers and followees.

Sentiment based features are also very relevant for the credibility prediction task. Notice that in general tweets which exhibit sentiment terms are more related to non-credible information. In particular this is very related to the fraction of tweets with positive sentiments, as opposed to negative sentiments, which tend to be more related to credible information. Tweets which exhibit question marks or smiling emoticons tend also to be more related to non-credible information. Something similar occurs when a significant fraction of tweets mention a user. On the other hand, tweets having many re-tweets on one level of the propagation tree, are considered more credible.

4.4 Automatically assessing credibility

We trained a supervised classifier to predict credibility levels on Twitter events. To do this we focus the problem on the detection of news that are believed to be almost certainly true (class A), against the rest of news (class B), excluding topics labeled as "ambiguous". In total, 306 cases correspond to class A and 302 cases correspond to class B, achieving a data balance equivalent to 50.3 / 49.7. With this balanced output we can evaluate the predictability of the credibility data.

We tried a number of learning algorithms with best results achieved by a J48 decision tree. For the training/validation process we perform a 3-fold cross validation strategy. A summary of the classifier is shown in Table 6.

Table 6: Summary for the credibility classification.

Correctly Classified Instances	86.0119 %
Kappa statistic	0.7189
Mean absolute error	0.154
Root mean squared error	0.3608
Relative absolute error	30.8711 %
Root relative squared error	72.2466 %

As Table 6 shows, the supervised classifier achieves an accuracy of 86 %. The Kappa statistic indicates that the predictability of our classifier is significantly better than a random predictor. The details of the evaluation per class are shown in Table 7. The performance for both classes is similar. The F_1 is high, indicating a good balance bet-

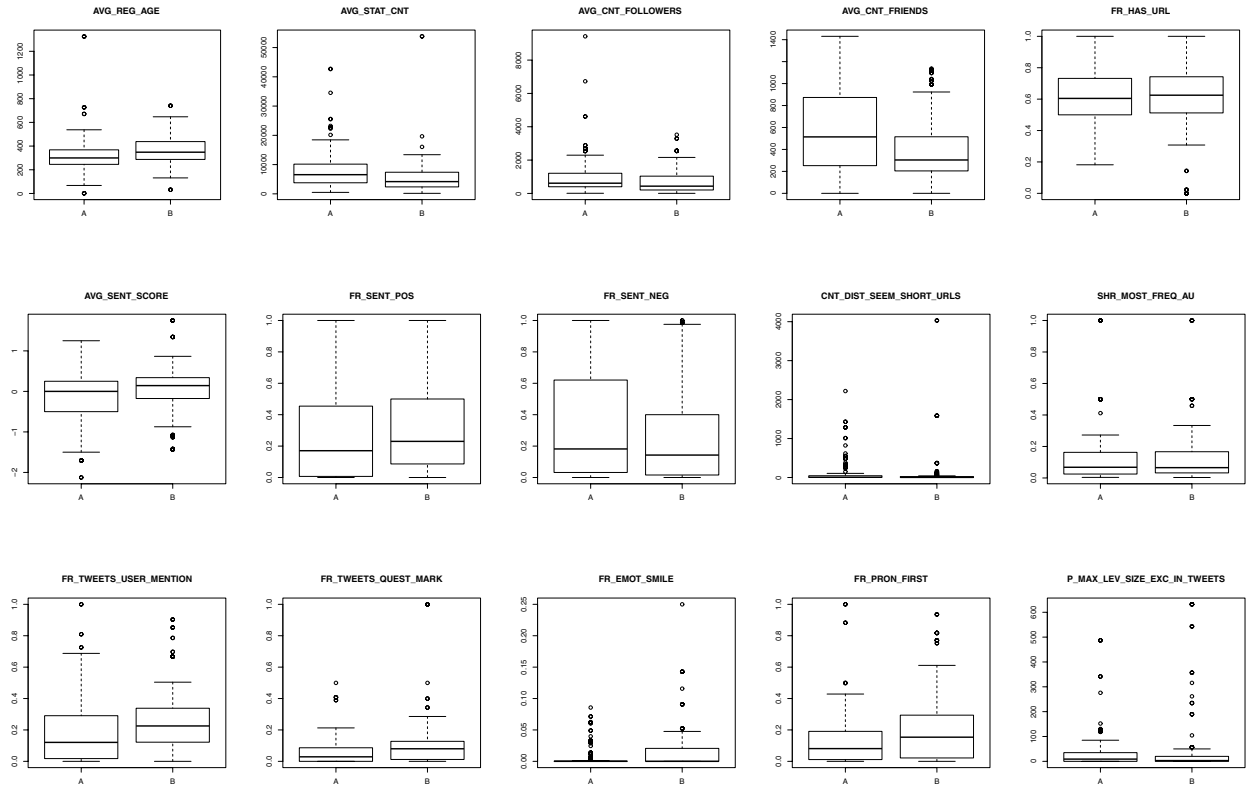

Figure 4: Box plots depicting the distribution for classes A ("true") and B ("false") of each of the top 15 features.

Table 7: Results for the credibility classification.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
A ("true")	0.825	0.108	0.874	0.825	0.849
B ("false")	0.892	0.175	0.849	0.892	0.87
W. Avg.	0.860	0.143	0.861	0.860	0.86

ween precision and recall. The last row of Table 7 shows the weighted averaged performance results calculated across both classes.

Best features. To illustrate the top features for this task, we analyze which features were the most important for the J48 decision tree, according to the GINI split criteria. The decision tree is shown in Figure 5. As the decision tree shows, the top features for this task were the following:

- Topic-based features: the fraction of tweets having an URL is the root of the tree. Sentiment-based features like fraction of negative sentiment or fraction of tweets with an exclamation mark correspond to the following relevant features, very close to the root. In particular we can observe two very simple classification rules, tweets which do not include URLs tend to be related to non-credible news. On the other hand, tweets which include negative sentiment terms are related to credible news. Something similar occurs when people use positive sentiment terms: a low fraction of tweets with

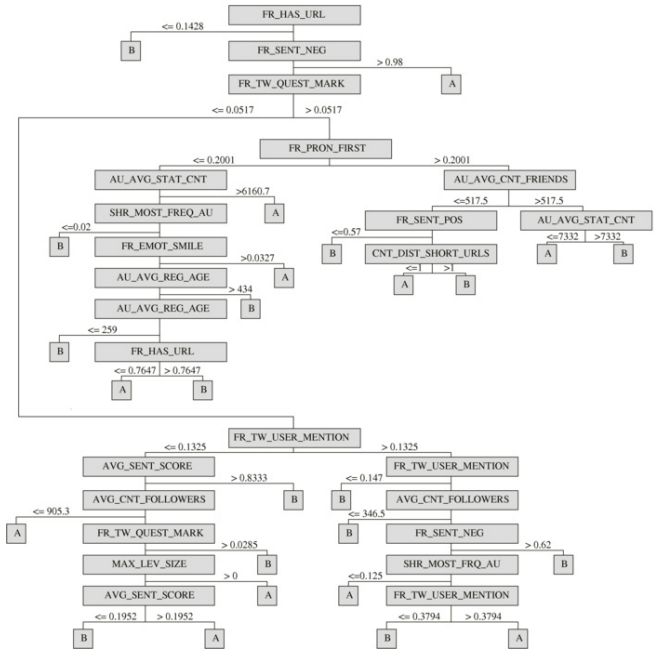

Figure 5: Decision tree built for the credibility classification. (A = "true", B = "false").

positive sentiment terms tend to be related to non-credible news.
- User-based features: these collection of features is very relevant for this task. Notice that low credible news are mostly propagated by users who have not written many messages in the past. The number of friends is also a feature that is very close to the root.
- Propagation-based features: the maximum level size of the RT tree is also a relevant feature for this task. Tweets with many re-tweets are related to credible news.

These results show that textual information is very relevant for this task. Opinions or subjective expressions describe people's sentiments or perceptions about a given topic or event. Opinions are also important for this task that allow to detect the community perception about the credibility of an event. On the other hand, user-based features are indicators of the reputation of the users. Messages propagated trough credible users (active users with a significant number of connections) are seen as highly credible. Thus, those users tend to propagate credible news suggesting that the Twitter community works like a social filter.

4.5 Credibility analysis at feature-level

In this section we study how specific subsets of features perform for the task of automatic assessment of credibility. To do this we train learning algorithms considering subsets of features. We consider 4 subsets of features grouped as follows:

- **Text subset**: considers characteristics of the text of the messages. This includes the average length of the tweets, the sentiment-based features, the features related to URLs, and those related to counting elements such as hashtags, user mentions, etc. This subset contains 20 features.
- **Network subset**: considers characteristics of the social network of users. This subset includes features related to the authors of messages, including their number of friends and their number of followers. This subset contains 7 features.
- **Propagation subset**: considers the propagation-based features plus the fraction of re-tweets and the total number of tweets. This subset contains 6 features.
- **Top-element subset**: considers the fraction of tweets that respectively contain the most frequent URL, hashtag, user mention, or author: 4 features in total.

We train a J48 decision tree with each subset feature as a training set. The instances en each group were splitted using a 3-fold cross validation strategy, as in the previous experiments.

Best features. In Table 8 we show with boldface the best results for each metric and class.

These results indicate that among the features, the propagation subset and the top-element subset are very relevant for assessing credibility. We observe that text- and author-based features are not enough by themselves for this task. Regarding non-credible news, high true positive rates are achieved using propagation features which indicate that graph patterns are very relevant to detect them. On the other hand, credible news are in general more difficult to detect. The top-element subset of features achieves the best results for this class indicating that social patterns measured through these features are very useful for this class.

Table 8: Experimental results obtained for the classification of credibility cases. The training step was conducted using four different subsets of features.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
Text subset					
A	0.636	0.152	0.808	0.636	0.712
B	0.848	0.364	0.700	0.848	0.767
W. Avg.	0.742	0.258	0.754	0.742	0.739
Network subset					
A	0.667	0.212	0.759	0.667	0.71
B	0.788	0.333	0.703	0.788	0.743
W. Avg.	0.727	0.273	0.731	0.727	0.726
Propagation subset					
A	0.606	**0.091**	**0.870**	0.606	0.714
B	**0.909**	0.394	0.698	**0.909**	0.789
W. Avg.	0.758	0.242	0.784	0.758	0.752
Top-element subset					
A	**0.727**	0.152	0.828	**0.727**	**0.774**
B	0.848	**0.273**	**0.757**	0.848	**0.800**
W. Avg.	0.788	0.212	0.792	0.788	**0.787**

To illustrate the dependence among these features according to the credibility prediction task, we calculate scatter plots for each feature pair considered in this phase. We show these plots in Figure 6.

As Figure 6 shows, most feature-pairs present low correlation, showing that the linear dependence between pairs of features is very weak. Something different occurs when sentiment-based features are analyzed, showing dependences among them. Regarding the class distribution, we can observe that every pair shows good separation properties, a fact that explains our results in credibility assessment.

5. CONCLUSIONS

Users online, lack the clues that they have in the real world to asses the credibility of the information to which they are exposed. This is even more evident in the case of inexperienced users, which can be easily mislead by unreliable information. As microblogging gains more significance as a valid news resource, in particular during emergency situations and important events, it becomes critical to provide tools to validate the credibility of online information.

On this paper, we have shown that for messages about time-sensitive topics, we can separate automatically newsworthy topics from other types of conversations. Among several other features, newsworthy topics tend to include URLs and to have deep propagation trees. We also show that we can assess automatically the level of social media credibility of newsworthy topics. Among several other features, credible news are propagated through authors that have previously written a large number of messages, originate at a single or a few users in the network, and have many re-posts.

For future work, we plan to extend the experiments to larger datasets, to partial datasets (e.g. only the first tweets posted on each topic), and to explore more deeply other factors that may lead users to declare a topic as credible. There are interesting open problems in this area, including studying the impact of the target pages pointed to by the URLs, or the impact of other factors of context that are displayed

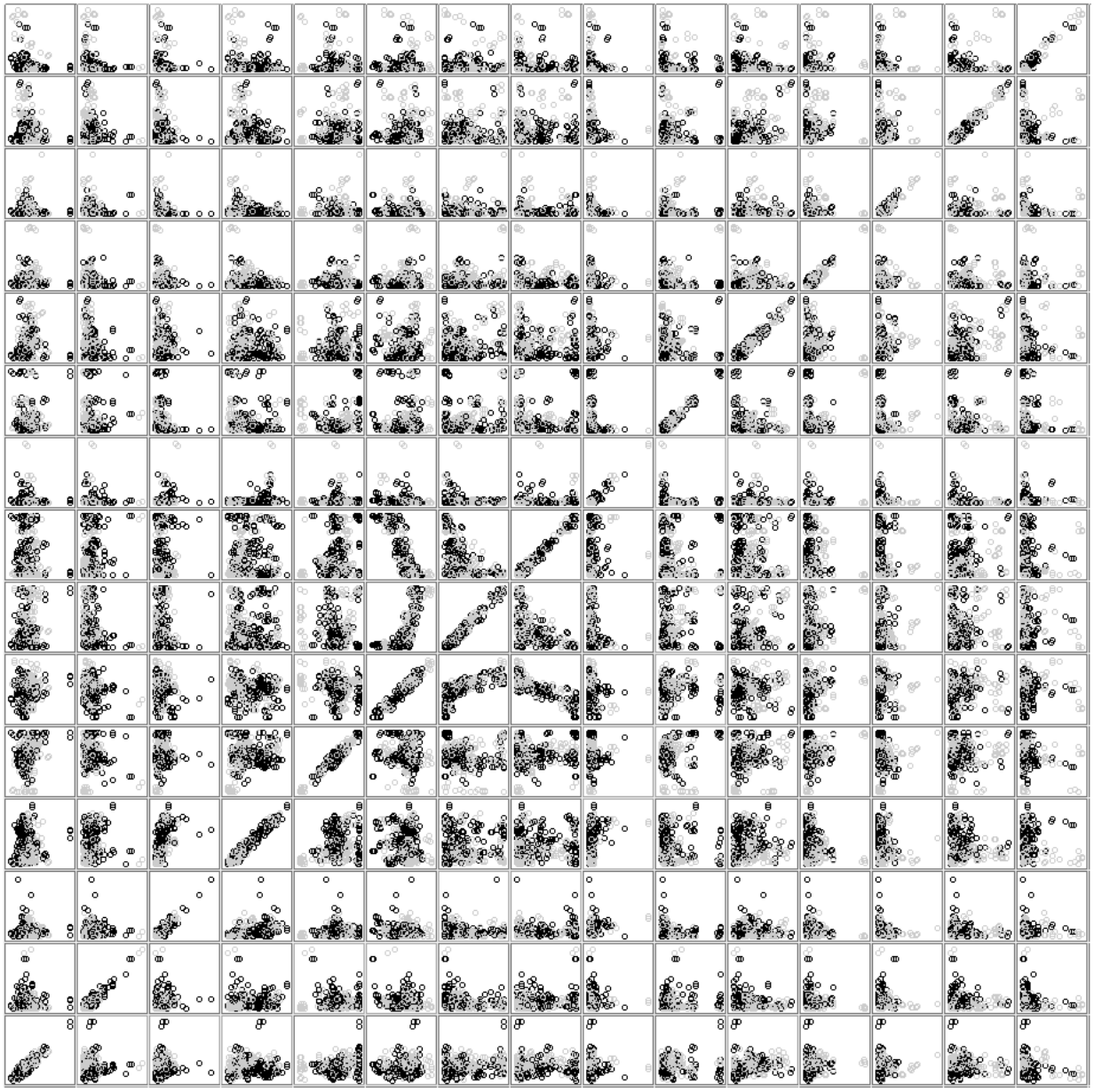

Figure 6: Scatter plots for features considered in the credibility prediction task. Black and gray points represent credible and non-credible information, respectively. each row represents a feature, from top to bottom: registration age, statuses count, number of followers, number of friends, tweets with URLs, sentiment score, positive sentiment, negative sentiment, shared URLs, shared author, tweets user mention, tweets with question marks, tweets with emoticon smiles, tweets with first pronoun, and max RT tree level size. The order in the columns goes from right to left.

in Twitter (e.g. the number of followers of each poster, the avatar used, etc.) on the assessments of credibility users do.

Acknowledgments. We would like to thank Michael Mathioudakis and Nick Koudas for lending us assistance to use the Twitter Monitor event stream. Carlos Castillo was partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (`http://cenitsocialmedia.es/`).

Key references: [18, 19]

6. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, New York, NY, USA, 2008. ACM.

[2] Alonso, Omar, Carson, Chad, Gerster, David, Ji, Xiang, and Nabar, Shubha. Detecting Uninteresting Content in Text Streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[3] C. L. Armstrong and M. J. Mcadams. Blogs of information: How gender cues and individual motivations influence

perceptions of credibility. *Journal of Computer-Mediated Communication*, 14(3):435–456, 2009.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[5] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008.

[6] 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. In *LBSN '09: Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 73–80, New York, NY, USA, 2009. ACM.

[7] P. S. Earle, M. Guy, C. Ostrum, S. Horvath, and R. A. Buckmaster. OMG Earthquake! Can Twitter improve earthquake response? *AGU Fall Meeting Abstracts*, pages B1697+, Dec. 2009.

[8] A. J. Flanagin and M. J. Metzger. Perceptions of internet information credibility. *Journalism and Mass Communication Quarterly*, 77(3):515–540, 2000.

[9] A. J. Flanagin and M. J. Metzger. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media Society*, 9(2):319–342, April 2007.

[10] B. J. Fogg and H. Tseng. The elements of computer credibility. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 80–87, New York, NY, USA, 1999. ACM.

[11] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *CCS '10: Proceedings of the 17th ACM conference on Computer and Communications Security*, CCS '10, pages 27–37, New York, NY, USA, October 2010. ACM.

[12] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. In *ISCRAM Conference*, May 2009.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[14] T. J. Johnson, B. K. Kaye, S. L. Bichard, and W. J. Wong. Every blog has its day: Politically-interested internet users' perceptions of blog credibility. *Journal of Computer-Mediated Communication*, 13(1), 2007.

[15] K. Kireyev, L. Palen, and K. Anderson. Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, December 2009.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *World Wide Web Conference*. ACM Press, 2010.

[17] V. Lampos, T. D. Bie, and N. Cristianini. Flu detector - tracking epidemics on twitter. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010)*, pages 599–602, Barcelona, Spain, 2010. Springer, Springer.

[18] M. Mathioudakis and N. Koudas. TwitterMonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pages 1155–1158. ACM, 2010.

[19] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *1st Workshop on Social Media Analytics (SOMA '10)*. ACM Press, July 2010.

[20] E. Mustafaraj and P. Metaxas. From obscurity to prominence in minutes: Political speech and real-time search. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, April 2010.

[21] M. Naaman, J. Boase, and C. H. Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM.

[22] Pear Analytics. Twitter study. http://www.pearanalytics.com/wp-content/uploads/2009/-08/Twitter-Study-August-2009.pdf, August 2009.

[23] Pew Research Center. Internet Overtakes Newspapers As News Outlet. http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-news-source 2008.

[24] A. M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1873–1876, New York, NY, USA, 2010. ACM.

[25] K. Poulsen. Firsthand reports from california wildfires pour through twitter. October 2007.

[26] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *arXiv*, Nov 2010.

[27] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, April 2010. ACM.

[28] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, New York, NY, USA, November 2009. ACM Press.

[29] M. Schmierbach and A. Oeldorf-Hirsch. A little bird told me, so i didn't believe it: Twitter, credibility, and issue perceptions. In *Proc. of annual meeting of the Association for Education in Journalism and Mass Communication*. AEJMC, August 2010.

[30] J. Schwarz and M. R. Morris. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *ACM Conference on Human Factors in Computing Systems (CHI)*. ACM Press, May 2011.

[31] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 241–250, New York, NY, USA, 2010. ACM.

[32] S. Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. In *Computer Supported Collaborative Work*, February 2010.

[33] S. Vieweg, A. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of ACM Conference on Computer Human Interaction (CHI)*, April 2010.

[34] C. R. W. Watch. Leap of faith: Using the internet despite the dangers. http://www.consumerwebwatch.org/pdfs/princeton.pdf, October 2005.

[35] D. J. Watts and J. Peretti. Viral Marketing for the Real World. *Harvard Business Review*, June 2007.

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