The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1066-2243.htm

Received 22 May 2012
Revised 30 November 2012
Accepted 3 December 2012

Predicting information credibility in time-sensitive social media

Carlos Castillo
Qatar Computing Research Institute, Doha, Qatar
Marcelo Mendoza
Universidad Técnica Federico Santa María, Santiago, Chile, and
Barbara Poblete
Department of Computer Science, University of Chile, Santiago, Chile

Abstract

Purpose – Twitter is a popular microblogging service which has proven, in recent years, its potential for propagating news and information about developing events. The purpose of this paper is to focus on the analysis of information credibility on Twitter. The purpose of our research is to establish if an automatic discovery process of relevant and credible news events can be achieved.

Design/methodology/approach – The paper follows a supervised learning approach for the task of automatic classification of credible news events. A first classifier decides if an information cascade corresponds to a newsworthy event. Then a second classifier decides if this cascade can be considered credible or not. The paper undertakes this effort training over a significant amount of labeled data, obtained using crowdsourcing tools. The paper validates these classifiers under two settings: the first, a sample of automatically detected Twitter "trends" in English, and second, the paper tests how well this model transfers to Twitter topics in Spanish, automatically detected during a natural disaster.

Findings – There are measurable differences in the way microblog messages propagate. The paper shows that these differences are related to the newsworthiness and credibility of the information conveyed, and describes features that are effective for classifying information automatically as credible or not credible.

Originality/value – The paper first tests the approach under normal conditions, and then the paper extends the findings to a disaster management situation, where many news and rumors arise. Additionally, by analyzing the transfer of our classifiers across languages, the paper is able to look more deeply into which topic-features are more relevant for credibility assessment. To the best of our knowledge, this is the first paper that studies the power of prediction of social media for information credibility, considering model transfer into time-sensitive and language-sensitive contexts.

Keywords Information credibility, Online social networks, Model transfer, Time sensitiveness, Social media prediction

Paper type Research paper

1. Introduction

Microblogging is a well-established paradigm for interaction in online social networks. In a microblogging platform, users post short messages which are shown to their followers (or users which subscribe to them) in a real-time fashion. These short messages are known as microblog posts, status updates, or tweets – a term referring to

The authors 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/). Barbara Poblete was supported by FONDECYT grant 11121511 and Program U-INICIA VID 2012, grant U-INICIA 3/0612; University of Chile. Marcelo Mendoza was supported by FONDECYT grant 11121435.

Most of this work was done while the first-named author was at Yahoo! Research Barcelona.

Internet Research
Vol. 23 No. 5, 2013
pp. 560-588

1066-2243
DOI 10.1108/IntR-05-2012-0095

the popular microblogging platform, Twitter[1]. This communication format is ideal for posting from mobile internet devices, and indeed a majority of users (54 percent) access the service from their phones[2].

Microblog messages display a wide variety of content, as much as everything else on the web. Nevertheless, most messages correspond to either (Harcup and O'Neill, 2001): conversation items, which are valuable to the user and its immediate circle of friends (e.g. updates about personal activities or whereabouts, gossip, chat, etc.), or information or news items, which have the potential to be valuable to a broader community (e.g. announcements or comments about relevant and/or timely topics of general interest).

In this work we focus mostly on the credibility of newsworthy information propagated through Twitter. Besides helping to communicate relevant events on a day-to-day basis, microblogging can be particularly helpful during emergency and/or crisis situations. Under these circumstances, microblog messages are used to provide real-time information from the actual location where the crisis is unfolding. This information often spreads faster and to a wider audience than what traditional news media sources can achieve.

We have observed in our prior work (Mendoza *et al.*, 2010; Castillo *et al.*, 2011), that there is a correlation between how information propagates and the credibility that is given by the social network to it. Indeed, the reflection of real-time events on social media reveals propagation patterns that surprisingly has less variability the greater a news value is. Accordingly, we explore the nature of these patterns and the relation with newsworthiness and credibility. We believe that users can benefit from models which aid them in the process of discovering reliable information. Furthermore, given prior evidence, we believe that this can be achieved in an automatic way using features extracted from information cascades.

1.1 Roadmap and contributions

In Section 2 we outline previous work relevant to our research. The following sections present the main contributions of our work:

- In Section 3, we present a case study about information propagation during a natural disaster. We describe differences in how a sample of confirmed news and false rumors propagate in the aftermath of the 2010 Earthquake in Chile. This study provides valuable insight into credibility characteristics in microblogging.
- In Section 4 we present the process of creating a labeled data set of newsworthy events for credibility assessment. This procedure is performed using crowdsourcing tools. Once our data set is built we introduce a series of information cascade-based features for modeling information on Twitter.
- In Section 5 we create two automatic classifiers, trained on our labeled data sets. The first classifier detects newsworthy information cascades. The second classifier receives the newsworthy cascades and classifies them according to their credibility. We analyze the precision of these models with varying amounts of data and time elapsed before and after the event detection. In addition, we round up our work and return to the data set about the Chilean earthquake, validating the transferability of our system to another language (Spanish), using only language-independent features.

Finally, Section 6 summarizes our main findings and presents directions for future research.

1.2 Comparison with our prior work

Section 3 summarizes the main findings of a sample of cases from the Chilean earthquake of 2010 described in our paper "Twitter under crisis: can we trust what we RT?" (Mendoza *et al.*, 2010).

Sections 4 presents work from the article "Information credibility on Twitter" (Castillo *et al.*, 2011). Section 5 extends (Castillo *et al.*, 2011) by re-designing the learning scheme, introducing new experiments that enable early prediction by observing only the messages posted before the event is actually detected, and presenting model transfer results.

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.

2.1 Newsworthiness and credibility of online contents

The prediction of news values, often referred as newsworthiness, is a key brick in the process of news construction because can help to determine the impact of a news story to a given audience. In a seminal study in this topic, Galtung and Ruge (1965) showed that there are quantitative factors that are consistently applied by journalists for news priorization across different cultures and organizations. However, Harcup and O'Neill (2001) argue that the advent of social media suggests that news priorization is a two-way process where the audience plays a crucial role.

By 2010 the internet was the source of news for 61 percent of users (Pew Research Center, 2010) and kept growing. People trust the internet as a news source as much as other media, with the exception of newspapers (Flanagin and Metzger, 2000).

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 (Schwarz and Morris, 2011). Recently, Morris *et al.* (2012) showed that there are discrepancies between features people rate as relevant to determine credibility and those used by major search engines as Google and Bing.

Additionally, Schmierbach and Oeldorf-Hirsch (2010) conducted an experiment in which the headline of a news item was presented to users in different ways. Users found the same news headline significantly less credible when presented on Twitter. Twitter users also perceive the credibility of individuals differently in Twitter according to their behavior, in general believing more in users who post what appears to be personal information (Marwick and Boyd, 2011; Johnson, 2011).

2.2 Microblogging as news

While most messages on Twitter are conversation and chatter, people also use it to share relevant information and to report news (Java *et al.*, 2007; Pear Analytics, 2009; Naaman *et al.*, 2010). Twitter has been used to track epidemics (Lampos *et al.*, 2010), geolocate events (Sakaki *et al.*, 2010), and find emerging controversial topics (Popescu and Pennacchiotti, 2010).

Twitter has been used widely during emergency situations, such as wildfires (De Longueville *et al.*, 2009), hurricanes (Hughes and Palen, 2009), floods (Vieweg, 2010; Vieweg *et al.*, 2010; Starbird *et al.*, 2010), and earthquakes (Kireyev *et al.*, 2009; Earle *et al.*, 2009; Mendoza *et al.*, 2010). Journalists have hailed the immediacy of the service which allowed "to report breaking news more rapidly than most mainstream

media outlets" (Poulsen, 2007). The correlation of the magnitude of real-world events and Twitter activity prompted researcher Markus Strohmaier to coin the term "Twicalli scale"[3].

In order to detect news in Twitter, Sankaranarayanan *et al.* (2009) used a text-based naive Bayes classifier. A different approach to event detection is to look at changes in the frequencies of the keywords in tweets, as Twitter Monitor (Mathioudakis and Koudas, 2010) does. Indeed, the majority of "trending topics" – keywords that experiment a sharp increase in frequency – can be considered "headline news or persistent news" (Kwak *et al.*, 2010).

2.3 Spam and misinformation in Twitter

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 created a new "bubble" of web visibility (Gori and Witten, 2005) prompting people to use all sort of deceptive tactics to promote their contents. This has attracted spam (Benevenuto *et al.*, 2010; Grier *et al.*, 2010; Yardi *et al.*, 2010) as well as political propaganda (Mustafaraj and Metaxas, 2010).

An application to demonstrate how false rumors spread in social media was shown by the British newspaper *The Guardian*. They presented a web-based visualization showing the results of a manual classification of seven confirmed truths and seven false rumors related to recent riots in London (*The Guardian*, 2011).

2.4 Feature extraction from blogs and microblogs

Several signals of blog credibility described in Rubin and Liddy (2006) are implemented in Weerkamp and De Rijke (2008). The length of posts and the number of comments are found to be important features, a finding confirmed in (Wanas *et al.*, 2008), who also use the presence of emoticons and capitalization among the top features.

In Ulicny and Baclawski (2007), links to news sources, the presence of the full name of the author and his/her affiliation, unquoted contents and comments where good indicators of reputable ("tenured" blogs). Along this same line, Suh *et al.* (2010) shows, that tweets containing a URL, and tweets containing a hashtag, are more likely to be re-tweeted that those not containing these elements. In general, the presence of links seems to be consistently a positive signal for credibility (Fogg, 2002; Stewart and Zhang, 2003).

To establish credibility of messages in Twitter, Al-Eidan *et al.* (2010), Al-Khalifa and Al-Eidan (2011) use among other features the presence of links to authoritative/reputable news sources and whether the user is verified or not. They also look at the presence of URLs, user mentions, re-tweets, and hash tags. For the authors, they look at whether there is a location, a biography, and a web site, as well as the number of followers and friends and the "verified" status of the account.

2.5 Systems to find credible tweets

The Truthy[4] service collects, analyze, and visualize the spread of tweets belonging to political "trending topics." It also compute a score for sets of tweets (Ratkiewicz *et al.*, 2011) that reflects the probability that those tweets are deceptive (astroturfing). In contrast, in our work we do not focus on willful deception but on factors that can be used to automatically approximate users' perceptions of credibility.

A different system for identifying credible Twitter news in Arabic (Al-Eidan *et al.*, 2010; Al-Khalifa and Al-Eidan, 2011) uses as one of its main features the similarity of

the posts with contents in reputable news sources, following (Juffinger *et al.*, 2009), and the result of an heuristic from the "Twitter Grader" system[5]. In contrast with their work, instead of analyzing individual tweets, in our work our basic unit of information is a set of tweets. Recently, Metaxas and Mustafaraj (2012) has proposed a system that can maintain trails of trustworthiness propagated through a number of real-time streams, where the provenance, credibility, and independence of the multiple information sources can be provided to users.

3. Case study: twitter during a crisis event

The earthquake that hit Chile on February 2010 provided the initial motivation for this work. This section summarizes some observations presented in Mendoza *et al.* (2010) about the use of Twitter during this event. Specifically, we compare a sample of posts containing false rumors with a sample of posts containing confirmed news.

3.1 Event characterization

The earthquake occurred off the coast of the Maule region of Chile, on Saturday, February 27, 2010 at 06:34:14 UTC (03:34:14 local time). It reached a magnitude of 8.8 on the Richter scale and lasted for 90 seconds; as of May 2012 it is considered the eight stronger earthquake ever recorded in history[6]. A few minutes after the earthquake, a tsunami reached the Chilean shores. Nearly 560 people died and more than two million people were affected directly in some way.

In the hours and days after this earthquake, Twitter was used to post time-critical information about tsunami alerts, missing people, deceased people, available services, interrupted services, road conditions, functioning gas stations, among other emerging topics related to the catastrophe. After the earthquake, bloggers published first-hand accounts on how they used twitter during the emergency[7].

3.2 Twitter reaction

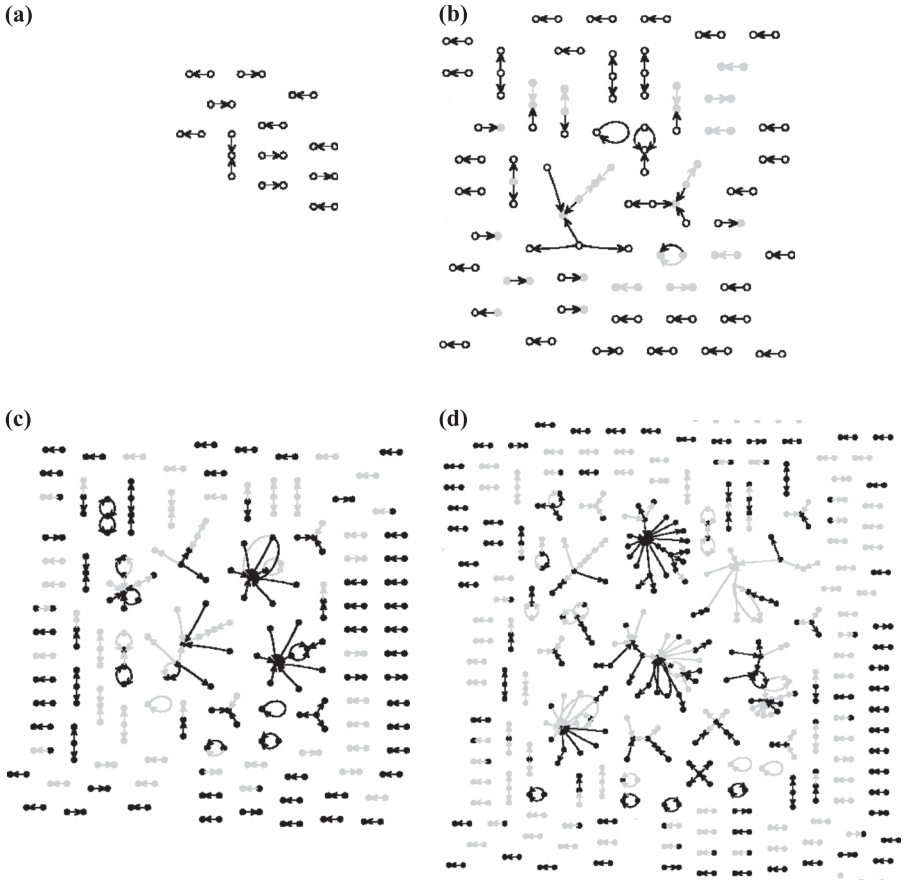
The earthquake reached the level of trending topic in Twitter a few hours after the event (e.g. #terremotochile).

To study how Twitter was used during the earthquake in Chile, we collected public tweets during the time window between February 27, 2010 and March 2, 2010. To determine the set of tweets related to the event, we used a filter-based heuristic approach. This was necessary because the data at our disposal from Twitter did not provide geographical information about its users and there were no IP addresses or reliable location information in general. Therefore, we focussed on the community that surrounded the topic of the earthquake.

We selected all tweets in the time zone of Santiago (GMT-04:00), plus tweets which included a set of keywords which characterized the event. This set of keywords was chosen by the authors after inspection of a large set of tweets. It included hashtags such as #terremotochile and the names of the towns, cities, and administrative regions affected. The set of keywords is listed in the supplementary material (see Section 6). This process selected 4,727,524 tweets with a 19.8 percent of them corresponded to replies to other tweets. The posts are related to 716,344 different users, which registered an average of 1,018 followers and 227 followees at the time of the earthquake.

3.3 Information propagation behavior

We illustrate the impact of the quake by measuring the re-tweet (RT) activity during the first hours. In Figure 1 we show the re-tweet graphs that emerge in the first hour



Notes: (a) 03:35-03:49; (b) 03:50-04:04; (c) 04:05 - 04:19; (d) 04:20-04:34. Gray edges indicate past re-tweets

Figure 1.
Trend propagation: tweets and re-tweets that include the term "earthquake" in the first hour post-quake

post-quake. In order to illustrate how the propagation process works over the Twitter social network we plot the graphs considering intervals of 15 minutes.

Figure 1 shows that tweets with the term "earthquake" are propagated through the social network. In fact, we observe that only 30 minutes after the quake some re-tweet graphs show interesting patterns. In some cases tweet propagation takes the form of a tree. This is the case of direct quoting of information. In other cases the propagation graph presents cycles, which indicates that the information is being commented and replied, as well as passed on. This last case involves bi-directional flows in the information dissemination process.

3.4 False rumor propagation

In this section we analyze the credibility of information on Twitter and how this information is spread through the social network. To achieve this task, we manually searched some relevant cases of valid news items, which were confirmed at some point by reliable sources. We refer to these cases as confirmed truths. Additionally, we

manually searched important cases of rumors which emerged during the crisis but were later confirmed to be false. We refer to these cases as false rumors. Our goal is to observe if users interact in a different manner when faced with these types of information. Each case studied was selected according to the following criteria:

(1) a significant volume of tweets is related to the case (close to 1,000 or more); and

(2) reliable external sources allow to verify if the claim is true or false.

The next step was to create a list of seven confirmed truths and seven false rumors. This list was obtained by manually analyzing samples of tweets and also using first-hand background knowledge of the crisis. A large majority of the news spread through twitter during this event were actually true, as confirmed in Section 5.4; so finding the confirmed truths was easier than finding the false rumors.

Examples. A true news item (confirmed truth) was the occurrence of a tsunami in the locations of Iloca and Duao. In fact this information was quickly informed through Twitter sources while government authorities ignored its existence initially, later on it was confirmed to be one of the most devastating aspects of this earthquake. On the other hand, a rumor that turned out to be false was the death of locally famous artist Ricardo Arjona.

In each case we collected between 42 and 700 unique tweets for classification (identical re-tweets were discarded for classification purposes). These tweets were retrieved by querying the collection using keywords related to each true or false case.

Manual classification of individual messages. The next step was to classify tweets into the following categories: affirms (propagates information confirming the item), denies (refutes the information item), questions (asks about the information item), and unrelated or unknown. We automatically propagated labels in such a way that all identical re-tweets of a tweet get the same label. The results of the classification are shown in Table I.

The classification results (see Table I) shows that a large percentage (>95 percent approximately) of tweets related to confirmed truths validate the information ("affirms" category label). The percentage of tweets that deny these true cases is very low. On the other hand, we observe that the number of tweets that deny information becomes much larger when the information corresponds to a false rumor. In fact, this category concentrates around 50 percent of tweets. There are also more tweets in the "questions" category in the case of false rumors.

The main conclusion we extract from this experiment is that false rumors tend to be questioned much more than confirmed truths, which is encouraging as it provides at least one feature that could be used for this automatic classifier. In Section 4, we introduce and test many other features. In Section 5 we return to work further on this data set.

4. Modeling newsworthy and credible information

The main focus of our research is on credibility of time-sensitive information, in particular on current news events. It is important to note that both goals, that of estimating the newsworthiness of a discussion topic, and that of determining its credibility, are equally important to us.

We describe the creation of our labeled data set, which contains newsworthy events and credible newsworthy events. We describe how this data set is built from emerging discussion topics on Twitter and their related messages. This section is divided into

	Tweets	RT%	Affirms	Denies	Questions
Confirmed truths					
The international airport of Santiago is closed	301	81	291	0	7
The *Viña del Mar International Song Festival* is canceled	261	57	256	0	3
Fire in the Chemistry Faculty at the University of Concepción	42	49	38	0	4
Navy acknowledges mistake informing about tsunami warning	135	30	124	4	6
Small aircraft with six people crashes near Concepción	129	82	125	0	4
Looting of supermarket in Concepción	160	44	149	0	2
Tsunami in Iloca and Duao towns	153	32	140	0	4
Total	1,181		1,123	4	30
Average	168.71		160.43	0.57	4.29
			97.1%	0.3%	2.6%
False rumors					
Death of artist Ricardo Arjona	50	37	24	12	8
Tsunami warning in Valparaiso	700	4	45	605	27
Large water tower broken in Rancagua	126	43	62	38	20
Cousin of football player Gary Medel is a victim	94	4	44	34	2
Looting in some districts in Santiago	250	37	218	2	20
"Huascar" vessel missing in Talcahuano	234	36	54	66	63
Villarrica volcano has become active	228	21	55	79	76
Total	1,682		502	836	216
Average	240.29		71.71	119.43	30.86
			32.3%	53.8%	13.9%

Note: All figures correspond to unique (non-duplicate) tweets

Table I.
Number of "affirms", "denies" and "questions" for each of the cases studied of confirmed truths and false rumors

three parts: emerging event detection, manual labeling, and feature extraction. On the next section we use this data for automatic classification purposes.

This section summarizes and extends our prior work (Castillo *et al.*, 2011).

4.1 Automatic event detection

As mentioned before, our basic research unit is an information cascade, which is composed of all of the messages which usually accompany newsworthy events. The detection of such sets of messages is not a part of this study; we obtain them from Twitter Monitor (Mathioudakis and Koudas, 2010) during a two months period.

Twitter Monitor is an online monitoring system[8] which detects sharp increases (bursts) in the frequency of sets of keywords found in messages. For every burst detected, Twitter Monitor provides a keyword-based query 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. The elements matching the query are a subset of those who caused the event detector to trigger. For details on how Twitter Monitor works please see reference Mathioudakis and Koudas (2010). We collected the tweets matching the query during a two-day window centered on the peak of every burst.

Each of these subsets corresponds to what we call a topic; some examples are shown in Table II. We have separated them in two broad types of topics: news and conversation (Java *et al.*, 2007; Pear Analytics, 2009). Contrary to what one might expect, conversation-type messages can be bursty, corresponding to endogenous bursts of activity (Crane and Sornette, 2008).

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

Note: Following Twitter Topic's algorithm, a tweet on a topic must contain all of the words in *italic* face and at least one of the other ones

Table II.
Sample topics extracted by Twitter Topic from April to July 2010

The topic extraction phase yields a large variation in the number of tweets found for each topic, as shown in Figure 2. For our particular data set, we decided to keep all of the topics which had at most 10,000 tweets, which corresponds to 99 percent of the cases and over 2,500 topics. Therefore, our initial topic data set contained approximately 1,873,000 messages.

4.2 Data set labeling process

Newsworthy topic labeling. Given that the scope of our work is credibility of newsworthy information, we first need to separate newsworthy topics. These are topics which are of interest to a broad set of people, as opposed to conversations/chat, which are of little importance outside a reduced circle of friends (Alonso *et al.*, 2010).

The manual data labeling process used the crowdsourcing tool Mechanical Turk[9]. For this task we presented evaluators a randomly selected sample of ten different tweets from a topic and the list of bursty keywords detected by Twitter Monitor for the topic. We asked if in general, most of the messages were spreading news about a specific event (labeled as class NEWS) or mostly comments or conversation (labeled as

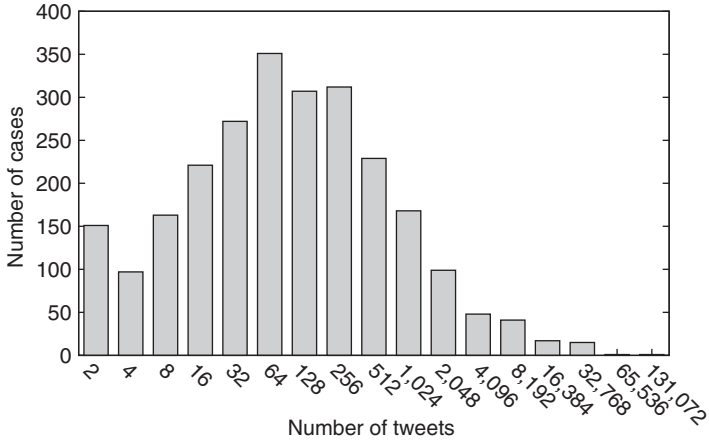


Figure 2.
Distribution of the number of unique tweets per topic

class CHAT). For each topic we also asked evaluators to provide a short summary sentence for the topic. Evaluators that did not provide justifications were discarded, to reduce the effect of click spammers in the evaluation system. A screenshot of the labeling interfaces, including examples and guidelines is shown in Figure 3. It should be noted that to simplify labeling task, we choose to present the evaluators with a sample of related tweets (on average each topic had thousands of tweets). Empirically, we perceived this information to be sufficient for determining the nature of the topic.

We selected uniformly at random 383 topics to be evaluated. In total, these topics involve 221,279 tweets. We required seven different evaluators for each topic, in total 2,681 unique assessments. Evaluations that did not provide the short summary sentence at this stage were discarded. A class label for a topic was assigned if at least five out of seven evaluators agreed on the label. Cases which did not meet this majority were labeled as UNSURE. Using this procedure, 35.6 percent of the topics (136 cases) were labeled as UNSURE, 29.5 percent as NEWS (113 cases), and 34.9 percent as CHAT (134 cases).

Credibility assessment. We focus next on the perceived credibility of the newsworthy topics. For the task of automatic credibility estimation, newsworthy topics are not known a priori, but must be found in the data stream. Hence, we use an automatic

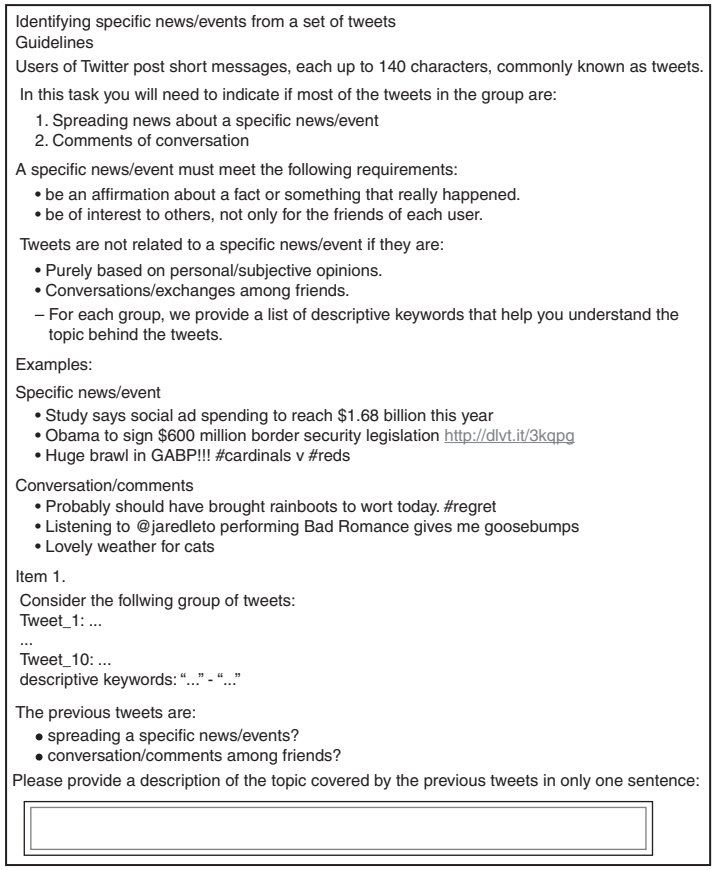


Figure 3.
User interface for labeling newsworthy topics, as seen by Mechanical Turk workers

classifier built using the data manually labeled as newsworthy in our data set (details of this model are given on Section 5.1). Therefore, we are able to expand to 747 the topics labeled as NEWS. We use this extended set of newsworthy topics as an input for our second labeling round, described next.

We asked evaluators to indicate a credibility level of the topic that was presented to them. Again we presented a sample of ten different tweets per topic, followed by the set of bursty keywords detected by Twitter Monitor for the topic.

We considered four levels of credibility: almost certainly true, likely to be false, almost certainly false, and "I can't decide"[10]. An example of the interface is shown in Figure 4.

As before, we discarded evaluations lacking a justification sentence and asked for a majority of five out of seven labels. This round produced the following results: 41 percent topics were considered "almost certainly true" (306 cases), "likely to be false" accounted for 31.8 percent (237 cases), "almost certainly false" accounted only for 8.6 percent (65 cases), while 18.6 percent (139 cases) were uncertain.

4.3 Feature extraction

The main hypothesis on which we base our work is that the level of credibility of information disseminated through social media can be estimated automatically.

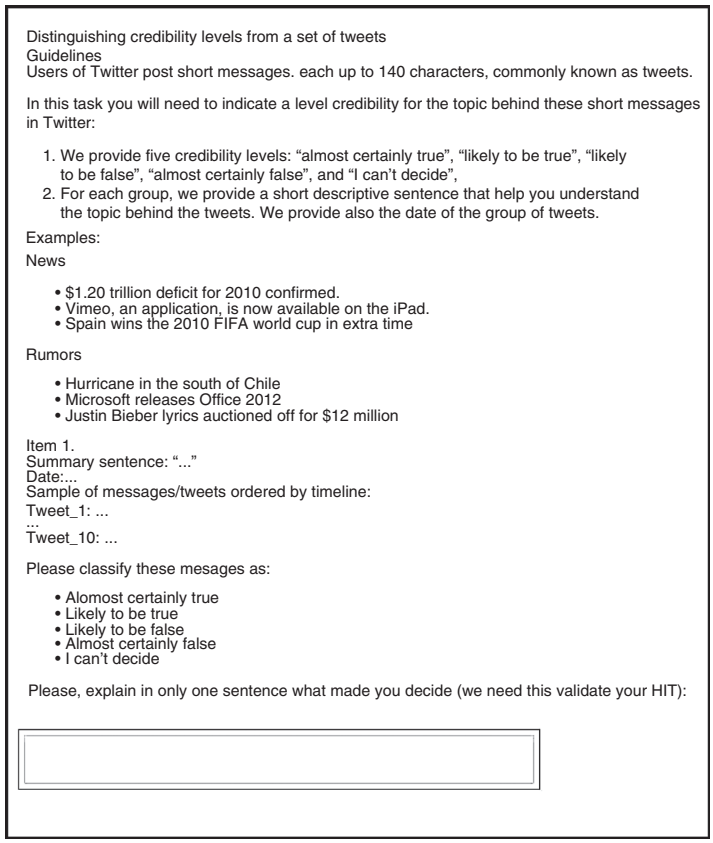


Figure 4.
User interface for assessing credibility, as seen by Mechanical Turk workers

We believe that there are several factors that can be observed in the social media platform itself that are useful to establish 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; and
- characteristics of the users that propagate the information, e.g. the number of followers that each user has in the platform.

Based on this reasoning, we propose an extended set of features (68 in total), many of which are based on previous works (Agichtein *et al.*, 2008; Alonso *et al.*, 2010; Hughes and Palen, 2009). We use these features to model topics and information cascades associated to them. Some of these features are specific to the Twitter platform, but most are quite generic and can be applied to other environments. We divide features into four main types described below, and detailed in Tables AI-AIV (see Section 6 for a detailed description of each feature).

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 Watts and Peretti, 2007).

5. Prediction model creation

In this section we present the methodology we use to create the models involved in automatic newsworthiness and credibility prediction for topics on Twitter. For this, we use the labels and features described in Section 4. It is important to recall, that we use information cascades as a research unit. Therefore, we train our classifiers only on the features which represent aggregated values for a topic, discarding the use of features calculated at user or message level.

5.1 Automatic discovery of newsworthy topics

We train a supervised classifiers to determine if a set of tweets describes a newsworthy event. For our supervised training phase we use the labeled topics obtained through

Mechanical Turk. These labels consider three classes (NEWS/CHAT/UNSURE) discussed in Section 4.2.

The effect of the UNSURE label. We study the impact of using instances labeled as UNSURE in the training process. For the interested reader, a more detailed account of experimental results are available as supplementary material (see Section 6). In summary, we compare classifier accuracy when training on different sets of labels, such as NEWS vs CHAT + UNSURE, and NEWS vs CHAT. Our experiments show that performance improves considerably if instances labeled as UNSURE are completely removed from our training data set. Therefore, we select only topics labeled as NEWS and CHAT for our training data.

Choice of a learning scheme. Next, we compare experimentally the performance of different learning schemes: Naive Bayes, Bayes Net, Logistic Regression, and Random Forest. The results of this comparison show that Bayes Net is the best learning scheme in this scenario. Therefore, we select it for our news classifier (although, Random Forest also performed well). A more detailed account of performance results is available as supplementary material (see Section 6).

Feature subsets. Once the details of our learning model have been determined, we study how features contribute in the prediction of newsworthy topics. We evaluate features by dividing them into the following subsets.

Text-only subset. Considers all of the features that are based on aggregated properties of the message text. This includes the average length of the tweets, sentiment-based features, features related to URLs, and those related to counting distinct elements (hashtags, user mentions, etc.) per topic. This subset contains 20 features.

User subset. Considers all of the features which represent the social network of users. This subset includes aggregated properties of message authors, including number of friends and number of followers. This subset contains the seven features described in Table AII.

Topic subset. Considers all of the topic features which include the fraction of tweets that contain one of these four elements (at topic level): most frequent URL, most frequent hashtag, most frequent user mention, and most frequent author.

Propagation-RT subset. Considers the propagation-based features plus the fraction of re-tweets and the total number of tweets. This subset contains the seven features described in Table AIV.

Table III shows results obtained for the Bayes networks-based classifier trained on each of these subsets. Noticeably, the text-only-based subset of features achieves the same performance as the classifier over the full feature set. On the other hand, the

Table III.
Results summary for different feature subsets over newsworthy detection using Bayes networks. RMS indicates root mean squared

	Text-only	User	Topic	Propagation-RT	All
Correctly Classified (%)	80.1	66.4	68.4	54.3	80.2
Incorrectly Classified (%)	19.8	33.6	31.6	45.8	19.8
Kappa statistic	0.59	0.31	0.39	0	0.6
Mean absolute error	0.19	0.42	0.4	0.5	0.19
RMS error	0.42	0.48	0.45	0.5	0.42
Relative error (%)	39.4	86.2	81.1	99.9	39.9
Relative RMS error (%)	84.3	95.1	90	100	85.5

propagation-based subset by itself achieves the lowest performance, with a κ-statistic = 0, which indicate a performance equal to that of a random predictor.

Feature selection. We test combinations of feature subsets, selecting arbitrary sets of features, with a best-first strategy and a CFS subset evaluation. We evaluated 684 subsets, and the best subset achieved a merit of 0.371, using eight features. The boxplots for these features are shown in Figure 5:

- fraction of authors in the topic that have written a self-description ("bio" in Twitter terms);
- count of distinct URLs;
- fraction of URLs pointing to domains in the top 100 most visited domains on the web;
- average length of the tweets;
- count of distinct user mentions;
- fraction of tweets containing a hashtag;

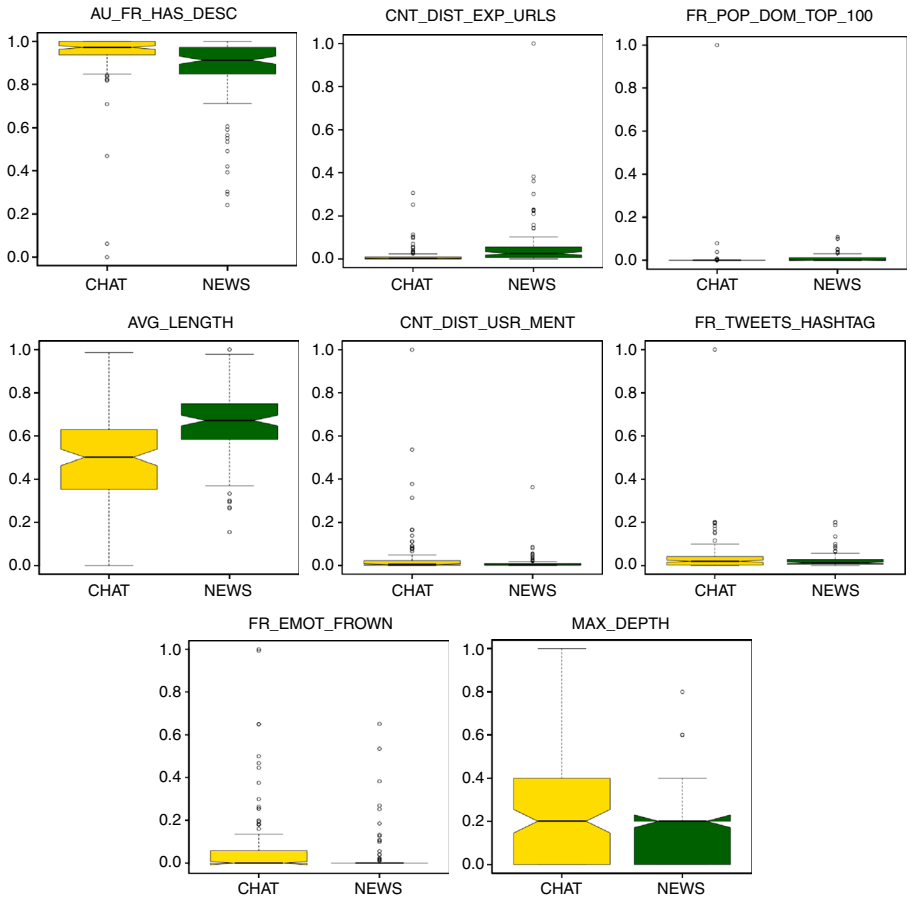


Figure 5.
Boxplots depicting the distributions of each of the eight features that best separate newsworthy and chat topics

- fraction of tweets containing a "frowning" emoticon; and
- maximum depth of propagation trees.

When analyzing in detail these features we observe that the feature "fraction of authors with a description" indicates that users with information in their profiles are more likely to propagate chat. Also, topics which have longer tweets tend to be more related to news topics. Intuitively, newsworthy topics share more URLs which belong to the top-100 most popular domains, and chat contains more frown emoticons. Also, news topics tend to have less hashtags than chat, and contain more URLs.

5.2 Credibility prediction

In this section we study how to automatically assign a credibility level (or score) to a topic deemed newsworthy by our previous classifier. Similarly to the newsworthy classification task, we divide this process into labeled data set selection, learning scheme analysis and feature space reduction.

Labeled data set selection. Our objective is to differentiate reliable news from those which are more doubtful. Therefore, we focus on topics labeled as almost certainly true, from now on assigned to the class CREDIBLE. The remaining credibility labels are joined in a class called NOT-CREDIBLE. We apply this criteria to the labeled data obtained in Section 4.2. Therefore, our problem is now reduced to that of binary classification. In total, 152 topic instances correspond to the class CREDIBLE and 136 to class NOT-CREDIBLE, achieving a class balance equivalent to 47.2/52.8 percent. In total, these topics involve 165,312 tweets.

Choice of a learning scheme. We study how different learning algorithms perform in our particular learning scenario. Similarly to Section 5.1. The data set used at this point is more reduced, as it is a subset of the one used for newsworthiness classification. Therefore we select a leave-one-out validation. We explore the performance of Bayesian methods, Logistic Regression, J48, Random Forest, and Meta Learning based on clustering. These former three methods achieve the best performance, shown in Table IV. Random Forest achieves high accuracy rates, although the accuracy of the Meta Learner is quite similar. Regarding error, the best results are obtained by the meta learner, with best accuracy/error rate. We observe that the predictability of the problem is very difficult, with very moderated κ-statistic values, indicating that improvements over a random predictor are limited. In addition, we analyze in detail how each method performed for each class, and the main conclusions are that misclassification of not-credible topics as credible is significant, but recall rates are quite acceptable.

Table IV. Results summary for credibility assessing using different learning algorithms

	Random forest	Logistic	Meta learning
Correctly Class. Instances (%)	61.8056	59.375	61.4583
Incorrectly Class. Instances (%)	38.1944	40.625	38.5417
Kappa statistic	0.226	0.1821	0.2174
Mean absolute error	0.4434	0.4599	0.3854
Root mean squared error	0.5015	0.5074	0.6208
Relative absolute error (%)	88.6465	91.9454	77.0537
Root relative squared error (%)	100.1069	101.2885	123.9281

Feature selection. We analyze feature selection to determine which features are the most useful for credibility prediction reducing dimensionality. As for the first classifier, we select subsets of features using a best-first strategy and a CFS subset evaluation method. We evaluate 684 subsets, with the best result obtaining a merit of 0.104, using 16 features, shown in Figure 6:

- the average number of tweets posted by authors of the tweets in the topic in the past;
- the average number of followees of authors posting these tweets;
- the fraction of tweets having a positive sentiment;
- the fraction of tweets having a negative sentiment;
- the fraction of tweets containing a URL that contain the most frequent URL;
- the fraction of tweets containing a URL;
- the fraction of URLs pointing to a domain among the top 10,000 most visited ones;

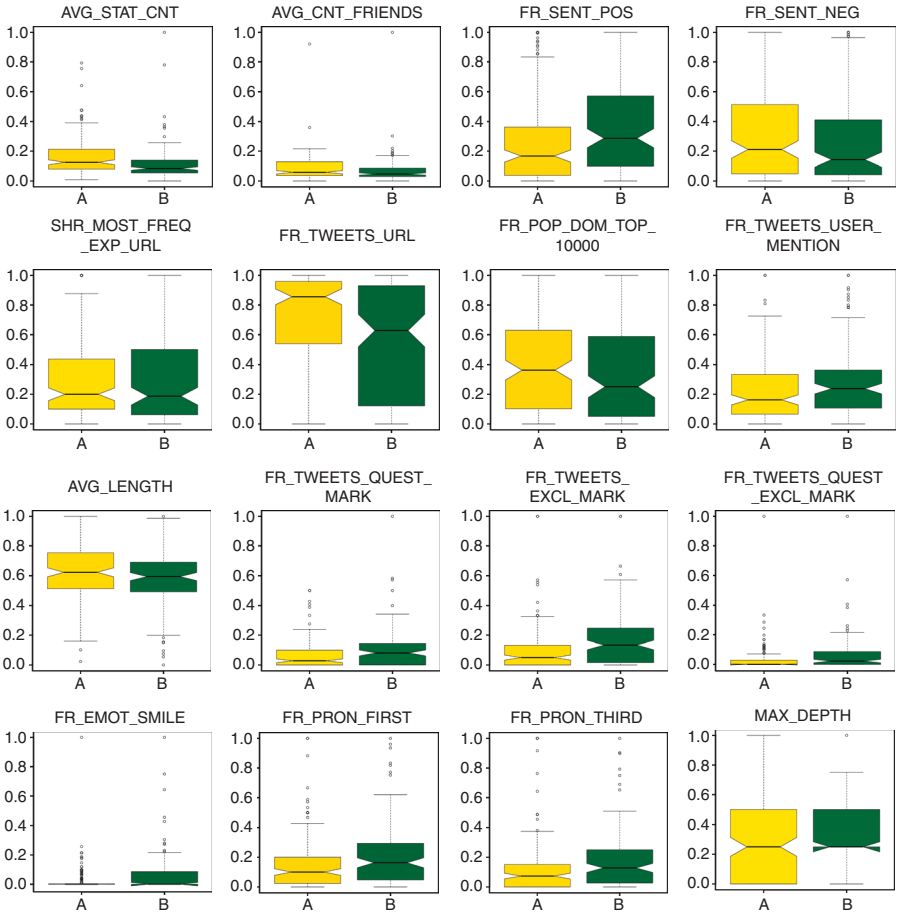


Figure 6.
Boxplots depicting the distributions of each of the 16 best features that best separate credible and non-credible tweets

- the fraction of tweets containing a user mention;
- the average length of the tweets;
- the fraction of tweets containing a question mark;
- the fraction of tweets containing an exclamation mark;
- the fraction of tweets containing a question or an exclamation mark;
- the fraction of tweets containing a "smiling" emoticons;
- the fraction of tweets containing a first-person pronoun;
- the fraction of tweets containing a third-person pronoun; and
- the maximum depth of the propagation trees.

Detailed analysis shows that several features display good properties for class separability. For example, users which spread credible tweets tend to register more friends. Note that credible tweets tend to include references to URLs which are included on the top-10,000 most visited domains on the web. In general, credible tweets tend to include more URLs, and are longer than non-credible tweets. Regarding polarity, non-credible tweets tend to concentrate more positive polarity scores, as opposite to credible tweets, which tend to express negative feelings. Some very interesting facts can be observed when we consider question and exclamation marks. People tend to concentrate question and exclamation marks on non-credible tweets, frequently using first and third-person pronouns.

We train using the top-3 classifiers for this task (Random Forest, Logistic, and Meta Learner) on the subset of 16 features, using a leave-one-out validation. Results are shown in Table V. It is important to note that even though we are using less features, there is no significant performance reduction. Moreover, the results of the logistic regression classifier actually improve achieving better results than classifiers over the full feature set. A possible explanation for the limited scope of these results is that binary separation between credible and not-credible topics is not truly realistic. In addition, since the logistic regression classifier is able to produce output scores, we observe that by using a threshold of 0.6, a 40 percent of the predictions achieves a 70.4 percent precision.

5.3 Early prediction of newsworthy and credible topics

We analyze model performance at the moment of topic detection. To simulate this situation we use the same topics in our previous data set, but only using tweets registered before the first activity burst. This experiment is intended to show if we can give early indicators of topic newsworthiness and credibility.

Table V. Results summary for credibility assessing using feature selection with different learning algorithms

	Random forest	Logistic	Meta learning
Correctly Class. Instances (%)	60.4167	62.3894	61.1111
Incorrectly Class. Instances (%)	39.5833	37.6106	38.8889
Kappa statistic	0.1972	0.2449	0.2088
Mean absolute error	0.4451	0.4243	0.3889
Root mean squared error	0.5024	0.5115	0.6236
Relative absolute error (%)	88.9936	81.3815	77.7479
Root relative squared error (%)	100.2795	98.1573	124.4841

As part of this analysis we compare how feature distributions vary in comparison to those of the complete data set. Similar distributions lead to similar classifier performance, also different distributions can lead to a decrease in performance. For this comparison we use a Kolmogorov-Smirnov test for each feature and we evaluate on the best feature sets selected for classification in the previous section. Results show that the distribution differs significantly only in the case of six features, four of them used by the newsworthy model and three used by the credibility model. Some interesting findings at this point are that tweets tend to include more URLs and hashtags and exclamation marks before the first burst. Also more users which have information in their profile participate before the burst, indicating greater incorporation of occasional users after the burst.

We study how the different models perform for the cases whose features were calculated using only the tweets before the peak. For the newsworthy detection phase we use the Bayesian model, which achieves the best performance results in the testing phase. Each case was labeled according to the Bayesian model and then only cases labeled as NEWS were analyzed by the credibility model. Thus, both models were evaluated considering error propagation, because they were assembled as a sequential filter bank. As a model for credibility assessment, we explore the performance of the logistic regression model, which allows for the inclusion of an output score threshold. We show a summary of these results in Tables VI and VII, respectively.

Table VI shows that the accuracy rate decreases in 9 percent, without a significant decrease in the κ-statistic, which indicates that the predictability of the newsworthy model is significantly better than a random predictor. Overall we note that the newsworthy model can detect news events by using only tweets registered before the first activity peak without significant performance loss. On the other hand, Table VII shows results for the Logistic Regression classifier with and without the output score threshold of 0.6. This classifier uses 43 percent of the original instances, which were the ones labeled as NEWS and when applying the threshold 19 percent of these instances are discarded. In this scenario almost 75 percent of the instances are correctly labeled, which increases to almost 80 percent when the threshold is applied. κ-statistics are quite significant, indicating that early prediction of credibility is possible.

Correctly Class. Instances (%)	71.4286
Incorrectly Class. Instances (%)	28.5714
Kappa statistic	0.4385
Mean absolute error	0.36
Root mean squared error	0.4591

Table VI. Result summary for early prediction of newsworthy topics

	Logistic	Logistic (Th = 0.6)
Correctly Class. Instances (%)	74.8503	79.9107
Incorrectly Class. Instances (%)	25.1497	20.0893
Kappa statistic	0.4681	0.5954
Mean absolute error	0.2599	0.2794
Root mean squared error	0.3929	0.3901

Table VII. Results summary for credibility classification with two different strategies for early credibility prediction

5.4 Credibility during a crisis event and model transfer

As a logical next step, we evaluate the transferability of our model for topics composed of tweets in a language different than English, in this case Spanish. In addition to this we also change the evaluation scenario, to a very different one, that of a crisis event.

We build the information cascades to create the evaluation data set for the Chilean earthquake. For this, we cluster tweets collected for the time frame described in Section 3. This data set is too large for a sequential clustering approach, therefore we segment our data into 24-hour time windows and use Mahout[11] with k-means. It should be noted that Twitter Monitor bursts were not available to us for this time period.

We apply the newsworthy topic classifier on the clustering output, which produced 1,566 clusters labeled as NEWS or CHAT. Most instances (93.8 percent) were labeled in this last category, and only 97 of the clusters were labeled as NEWS. Inspection of the derived clusters shows that many clusters contained combinations of news and chat tweets. This included noise, which was not the case when using information cascades derived from Twitter Monitor. Given the large bias in our class labels, to avoid problems in our evaluation, we create a ground truth using a stratified sampling process over clusters. We generate six cluster folds, each containing 130 clusters, where 30 clusters were sampled from those labeled as NEWS and others chosen at random with replacement. This method considers 562 clusters in total.

Folds are then manually labeled in three rounds. For each round a label is assigned to the cluster when there is an agreement of three human evaluators. It should be noted that crowdsourcing tools were not used in this process, since this task requires understanding of Spanish and some background on the crisis situation itself. Therefore, labeling was performed by a set of six evaluators composed of graduate students in Chile and the authors themselves.

The first labeling round is intended for establishing if the cluster is ON-TOPIC or OFF-TOPIC. Meaning that the cluster itself represents a coherent discussion topic (on-topic), and is not extremely noisy (off-topic). This round produces 224 ON-TOPIC instances. The second round is used to manually label on-topic clusters into the classes NEWS/CHAT/UNSURE. This process generates 52 properly labeled clusters: 34 labeled as NEWS, 15 as CHAT, and only three as UNSURE. Finally, for the instances labeled as NEWS, we separate clusters which spread confirmed truths from false rumors. This generates 27 CREDIBLE clusters, and only three NOT-CREDIBLE clusters.

Similarly to Section 5.3 we compare the feature distributions for the features extracted from the earthquake data set. We do this for instances labeled as ON-TOPIC, and compare them to the features from our original training data sets. Again we use the a Kolmogorov-Smirnov test for each pair of distributions. This comparison shows that features used for newsworthy classification differ significantly in four cases, but feature distribution in credibility classification are quite similar.

We study how our models perform in the Chilean earthquake situation. We perform the newsworthy detection phase over the set of cases labeled as ON-TOPIC. Then, we study credibility performance over the set of cases labeled as NEWS, considering error propagation between both models. For the newsworthy detection phase we use the Bayesian model. Credibility assessment was explored using the Logistic Regression based model without consider an output score threshold. Performance measures per class are in Table VIII.

Table VIII shows that the newsworthy detection phase displays similar performance to that of the original data set collection. Therefore, the newsworthy

model has good generalization properties for this news data set. In particular, the newsworthy model achieves high precision rate for NEWS cases. A good balance was achieved between precision and recall, reaching an *F*-measure equal to 81 percent for NEWS cases. For the newsworthy model it was more difficult to detect CHAT topics, achieving a similar recall rate as the one achieved for NEWS cases but with a lower precision. A ROC area equal to 0.807 indicates a good balance between false positives and true positives.

Regarding credibility assessment, CREDIBLE cases were detected with high precision and recall rates. For the credibility model it was hard to detect NOT-CREDIBLE cases, achieving only a precision rate equals to the 50 percent of the cases. Notice that the high false positive rate registered for the CREDIBLE class is due to the unbalance between the number of cases labeled as CREDIBLE and NOT-CREDIBLE, showing that in the labeled data set of the earthquake, the absence of false rumor cases was important. Notice that the 90 percent of the cases labeled as NEWS were labeled as CREDIBLE, being the 96 percent of these cases correctly labeled by our model. Regarding the tradeoff between false positives and true positives, the credibility models show good balance between both rates (ROC area = 0.824).

6. Conclusions

Our main conclusion is that newsworthy and credible information can be found in a microblogging platform using supervised learning methods over a set of appropriate features. In order to achieve this, we have described an approach in which initially the messages are collected into "topics," then an automatic classifier finds the newsworthy topics, and a second classifier finds the credible topics among the newsworthy ones.

For the creation of topics we used two existing methods: one based on frequency of keywords, and one based on clustering. Not surprisingly, the method based on frequency of keywords generates more newsworthy topics than the method based on clustering. However, it does not generate exclusively newsworthy topics.

We built a classifier for finding newsworthy topics, which achieved AUC of 0.86, and a classifier for finding credible topics, which achieved AUC of 0.64.

Next we attempted to find newsworthy and credible topics immediately at the moment in which they are detected (which means operating with less data). This classifier was worse at detecting newsworthy topics, with an AUC of 0.78 for the newsworthiness task, but more accurate at detecting credible topics among those found to be newsworthy, with AUC of 0.86 for the credibility task.

Finally, we tested if the classifier trained over the English topics detected based on frequency of keywords during a "normal" period of time, performed well over Spanish topics detected based on clustering during a time of crisis. We observed a good

	FP Rate	Precision	Recall	F-Measure	ROC area
Newsworthiness					
NEWS	0.266	0.866	0.764	0.811	0.807
CHAT	0.235	0.578	0.733	0.646	0.807
Weighted average	0.256	0.776	0.754	0.759	0.807
Credibility					
CREDIBLE	0.666	0.928	0.962	0.944	0.824
NOT-CREDIBLE	0.037	0.5	0.334	0.4	0.824
Weighted Average	0.603	0.885	0.899	0.889	0.824

Table VIII. Results for news and credibility prediction over the Chilean earthquake data collection

performance with AUC of 0.81 and 0.82 in the newsworthiness and credibility tasks, respectively. Overall, these results demonstrate the effectiveness of this approach.

Future work. In our experiments, we have considered Twitter as a closed system, without considering e.g. the pages pointed to by the tweets, or the content of other sources of information such as news media or blogs. Those sources may contradict or confirm what is being said in Twitter.

We have also largely ignored the past history of users. We consider shallow characteristics such as their number of followers, but we do not distinguish between those who have accumulated a good reputation in the past, those who have been spreading misinformation or spam in the past, etc. This is also important contextual information that can be exploited.

In general, this research can be extended both by considering more contextual factors when evaluating Twitter information, or by developing improvements or new methods that can be used to establish the credibility of user generated content.

Data availability and experimental details

The identifiers of the tweets that were processed and the labeled data used for training and testing are available upon request.

In addition, detailed plots and experimental results of our evaluation process, are available in the form of supplementary material. The suplementary material is available in the following link: http://www.dcc.uchile.cl/docs/2012/20121005_supplementary.pdf

Notes

1. http://twitter.com/
2. http://www.pewinternet.org/Reports/2011/Twitter-Update-2011/Main-Report.aspx
3. http://mstrohm.wordpress.com/2010/01/15/measuring-earthquakes-on-twitter-the-twicalli-scale/
4. http://truthy.indiana.edu/
5. http://graderblog.grader.com/twitter-grader-api/bid/19046/How-Does-Twitter-Grader-Calculate-Twitter-Rankings
6. http://en.wikipedia.org/wiki/List_of_earthquakes#Largest_earthquakes_by_magnitude
7. http://portalcesfam.com/index.php?option = com_content view = article id = 932:entrevista-en-diario-medico- catid = 88:informacion Itemid = 103 http://curvaspoliticas.blogspot.com/p/especial-terremoto.html (both in Spanish).
8. http://www.twittermonitor.net/
9. http://www.mturk.com
10. There is no option "likely to be true" as in a preliminary round of evaluation it was observed that it attracted almost all the responses from the evaluators.
11. Machine learning library which runs on Hadoop http://mahout.apache.org/

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G. (2008), "Finding high-quality content in social media", *WSDM'08: Proceedings of the International Conference on Web search and web data mining, ACM, New York, NY*, pp. 183-194.

Al-Eidan, R.M., Al-Khalifa, H.S. and Al-Salman, A.S. (2010), "Measuring the credibility of Arabic text content in twitter", *Proceedings of the 5th International Conference on Digital Information Management, Thunder Bay, ON, July 5-8*, pp. 285-291, available at: http://dx.doi.org/10.1109/ICDIM.2010.5664223

Al-Khalifa, H.S. and Al-Eidan, R.M. (2011), "An experimental system for measuring the credibility of news content in twitter", *International Journal of Web Information Systems*, Vol. 7 No. 2, pp. 130-151.

Alonso, O., Carson, C., Gerster, D., Ji, X. and Nabar, S. (2010), "Detecting uninteresting content in text streams", SIGIR Crowdsourcing for Search Evaluation Workshop, Geneva, July 23, available at: http://ir.ischool.utexas.edu/cse2010/slides/alonso-paper.pdf (accessed September 24, 2013).

Benevenuto, F., Magno, G., Rodrigues, T. and Almeida, V. (2010), "Detecting spammers on twitter", *Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, Redmond, WA, July 13-14*, available at: http://ceas.cc/2010/papers/Paper%2021.pdf (accessed September 24, 2013).

Castillo, C., Mendoza, M. and Poblete, B. (2011), "Information credibility on twitter", in Srinivasan, S. Ramamritham, K. Kumar, A. Ravindra, M.P. Bertino, E. Kumar, R. (Eds), *Proceedings of the 20th International Conference on World Wide Web, WWW'11*, ACM, New York, NY, pp. 675-684, available at: http://doi.acm.org/10.1145/1963405.1963500 (accessed September 24, 2013).

Crane, R. and Sornette, D. (2008), "Robust dynamic classes revealed by measuring the response function of a social system", *Proceedings of the National Academy of Sciences*, Vol. 105 No. 41, pp. 15649-15653, available at: http://dx.doi.org/10.1073/pnas.0803685105

De Longueville, B., Smith, R.S. and Luraschi, G. (2009), "'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 Zhou, X. and Xie, X. (Eds), *LBSN'09: Proceedings of the 2009 International Workshop on Location Based Social Networks*, ACM, New York, NY, pp. 73-80, available at: http://dx.doi.org/10.1145/1629890.1629907 (accessed September 24, 2013).

Earle, P.S., Guy, M., Ostrum, C., Horvath, S. and Buckmaster, R.A. (2009), "OMG Earthquake! Can Twitter improve earthquake response?", *AGU Fall Meeting Abstracts*, B1697 + .

Flanagin, A.J. and Metzger, M.J. (2000), "Perceptions of internet information credibility", *Journalism and Mass Communication Quarterly*, Vol. 77 No. 3, pp. 515-540, available at: www.comm.ucsb.edu/publications/flanagin/Flanagin-Metzger-2000-JMCQ.pdf

Fogg, B.J. (2002), "Stanford guidelines for web credibility", technical report, Stanford University, Stanford, CA, available at: http://credibility.stanford.edu/guidelines/

Galtung, J. and Ruge, M. (1965), "The structure of foreign news: the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers", *Journal of International Peace Research*, Vol. 2 No. 1, pp. 64-90, available at: http://dx.doi.org/10.1177/002234336500200104

Gori, M. and Witten, I. (2005), "The bubble of web visibility", *Commun. ACM*, Vol. 48 No. 3, pp. 115-117.

(The) Guardian (2011), "Reading the riots", *The Gaurdian*, December 12, 2007, available at: www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter (accessed September 24, 2013).

Grier, C., Thomas, K., Paxson, V. and Zhang, M. (2010), "@spam: the underground on 140 characters or less", in Al-Shaer, E., Keromytis, A.D. and Shmatikov, V. (Eds), *CCS'10: Proceedings of the 17th ACM conference on Computer and Communications Security, CCS'10*, ACM, New York, NY, pp. 27-37, available at: www.icir.org/vern/papers/ccs2010-twitter-spam.pdf (accessed September 24, 2013).

Harcup, T. and O'Neill, D. (2001), "What is news: Galtung and Ruge revisited", *Journal of Journalism Studies*, Vol. 2 No. 2, pp. 261-280, available at: http://dx.doi.org/10.1080/14616700118449

Hughes, A.L. and Palen, L. (2009), "Twitter adoption and use in mass convergence and emergency events", *ISCRAM Conference, Göteborg, May 10*, available at: www.slideshare.net/guest8c177f/twitter-adoption-and-use-in-mass-convergence-and-emergency-events (accessed September 24, 2013).

Java, A., Song, X., Finin, T. and Tseng, B. (2007), "Why we twitter: understanding microblogging usage and communities", *WebKDD/SNA-KDD'07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, ACM, New York, NY, pp. 56-65, available at: http://dx.doi.org/10.1145/1348549.1348556

Johnson, K. (2011), "The effect of Twitter posts on students' perceptions of instructor credibility", *Learning, Media and Technology*, Vol. 36 No. 1, pp. 21-38, available at: http://dx.doi.org/10.1080/17439884.2010.534798

Juffinger, A., Granitzer, M. and Lex, E. (2009), "Blog credibility ranking by exploiting verified content", *Proceedings of the 3rd Workshop on Information Credibility on the Web, WICOW'09*, ACM, New York, NY, pp. 51-58, available at: http://doi.acm.org/10.1145/1526993.1527005

Kireyev, K., Palen, L. and Anderson, K. (2009), "Applications of topics models to analysis of disaster-related twitter data", NIPS Workshop on Applications for Topic Models: Text and Beyond, Whistler, BC, December 11.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is twitter, a social network or a news media?", *World Wide Web Conference, Raleigh, NC, April 27*, ACM Press, available at: http://an.kaist.ac.kr/haewoon/papers/2010-www-twitter.pdf (accessed September 24, 2013).

Lampos, V., Bie, T.D. and Cristianini, N. (2010), "Flu detector – tracking epidemics on twitter", *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010)*, Springer, Barcelona, pp. 599-602, available at: http://springerlink.com/content/03437106w6438777/fulltext.pdf

Marwick, A. and Boyd, D. (2011), "To see and be seen: celebrity practice on twitter", *Convergence: The International Journal of Research into New Media Technologies*, Vol. 17 No. 2, pp. 139-158, available at: http://dx.doi.org/10.1177/1354856510394539

Mathioudakis, M. and Koudas, N. (2010), "Twitter monitor: trend detection over the twitter stream", *Proceedings of the 2010 International Conference on Management of Data, ACM*, pp. 1155-1158.

Mendoza, M., Poblete, B. and Castillo, C. (2010), "Twitter under crisis: can we trust what we RT?", in Melville, P., Leskovec, J. and Provost, F. (Eds), *1st Workshop on Social Media Analytics (SOMA'10)*, ACM Press, available at: http://chato.cl/papers/mendoza_poblete_castillo_2010_twitter_terremoto.pdf (accessed September 24, 2013).

Metaxas, P. and Mustafaraj, E. (2012), "Trails of trustworthiness in real-time streams", Design, Influence and Social Technologies Workshop of CSCW, Seattle, WA, February 11-15, available at: http://cs.wellesley.edu/pmetaxas/TrustTrails.pdf (accessed September 24, 2013).

Morris, M., Counts, S., Roseway, A., Hoff, A. and Schwarz, J. (2012), "Tweeting is believing? Understanding microblog credibility perceptions", in Poltrock, S.E., Simone, C., Grudin, J., Mark, G. and Riedl, J. (Eds), *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*, ACM, pp. 441-450, available at: http://doi.acm.org/10.1145/2145204.2145274 (accessed September 24, 2013).

Mustafaraj, E. and Metaxas, P. (2010), "From obscurity to prominence in minutes: political speech and real-time search", *Proceedings of the WebSci10: Extending the Frontiers of Society*

On-Line, Raleigh, NC, April 26-27, available at: http://journal.webscience.org/317/ (accessed September 24, 2013).

Naaman, M., Boase, J. and Lai, C.H. (2010), "Is it really about me? Message content in social awareness streams", in Quinn, K.I., Gutwin, C. and Tang, J.C. (Eds), *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ACM, New York, NY, pp. 189-192, available at: http://dx.doi.org/10.1145/1718918.1718953 (accessed September 24, 2013).

Pear Analytics (2009), "Twitter study", available at: www.pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf

Pew Research Center (2010), "The new news landscape: rise of the internet", available at: http://pewresearch.org/-pubs/-1508/-internet-cell-phone-users-news-social-experience

Popescu, A.M. and Pennacchiotti, M. (2010), "Detecting controversial events from twitter", in Huang, J., Koudas, N., Jones, G.J.F., Wu, X., Collins-Thompson K. and An, A. (Eds), *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 1873-1876, available at: http://dx.doi.org/10.1145/1871437.1871751 (accessed September 24, 2013).

Poulsen, K. (2007), "Firsthand reports from California wildfires pour through twitter", available at: www.wired.com/threatlevel/2007/10/firsthand-repor/

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Patil, S., Flammini, A. and Menczer, F. (2011), "Truthy: mapping the spread of astroturf in microblog streams", *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 249-252, available at: http://doi.acm.org/10.1145/1963192.1963301 (accessed September 24, 2013).

Rubin, V.L. and Liddy, E.D. (2006), "Assessing credibility of weblogs", *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 187-190.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), "Earthquake shakes twitter users: real-time event detection by social sensors", in Rappa, M., Jones, P., Freire, J. and Chakrabarti, S. (Eds), *Proceedings of the 19th International Conference on World Wide Web, WWW'10*, ACM, New York, NY, pp. 851-860, available at: http://dx.doi.org/10.1145/1772690.1772777 (accessed September 24, 2013).

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D. and Sperling, J. (2009), "TwitterStand: news in tweets", in Agrawal, D., Aref, W.G., Lu, C.-T., Mokbel, M.F., Scheuermann, P., Shahabi, C. and Wolfson, O. (Eds), *GIS'09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, New York, NY, pp. 42-51, available at: http://dx.doi.org/10.1145/1653771.1653781 (accessed September 24, 2013).

Schmierbach, M. and Oeldorf-Hirsch, A. (2010), "A little bird told me, so i didn't believe it: Twitter, credibility, and issue perceptions", *Proceedings of Annual Meeting of the Association for Education in Journalism and Mass Communication, AEJMC, Denver, CO, August 4-7*.

Schwarz, J. and Morris, M.R. (2011), "Augmenting web pages and search results to support credibility assessment", *ACM Conference on Human Factors in Computing Systems (CHI), ACM Press, Vancouver, BC, May 7-12*, (accessed September 24, 2013).

Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. (2010), "Chatter on the red: what hazards threat reveals about the social life of microblogged information", *CSCW'10: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ACM, New York, NY, pp. 241-250.

Stewart, K.J. and Zhang, Y. (2003), "Effects of hypertext links on trust transfer", *Proceedings of the 5th International Conference on Electronic Commerce, ICEC'03*, ACM, New York, NY, pp. 235-239, available at: http://doi.acm.org/10.1145/948005.948037

Suh, B., Hong, L., Pirolli, P. and Chi, E.H. (2010), "Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network", in Elmagarmid, A.K. and Agrawal D. (Eds), *Proceedings of the 2nd International Conference on Social Computing, Minneapolis, MN, August 20-22*, pp. 177-184, available at: http://dx.doi.org/10.1109/SocialCom.2010.33

Ulicny, B. and Baclawski, K. (2007), "New metrics for newsblog credibility", *Proceedings International Conference on Weblogs and Social Media, Boulder, CO, March 26-28*, available at: www.icwsm.org/papers/4--Ulicny-Baclawski.pdf (accessed September 24, 2013).

Vieweg, S. (2010), "Microblogged contributions to the emergency arena: discovery, interpretation and implications", Computer Supported Collaborative Work, available at: http://research.microsoft.com/en-us/um/redmond/groups/connect/CSCW_10/docs/p515.pdf

Vieweg, S., Hughes, A., Starbird, K. and Palen, L. (2010), "Microblogging during two natural hazards events: what twitter may contribute to situational awareness", *Proceedings of ACM Conference on Computer Human Interaction (CHI), Atlanta, GA, April 10-15*, available at: http://dl.acm.org/citation.cfm?doid = 1753326.1753486 (accessed September 24, 2013).

Wanas, N., El-Saban, M., Ashour, H. and Ammar, W. (2008), "Automatic scoring of online discussion posts", in Tanaka, K., Matsuyama, T., Lim, E.-P. and Jatowt, A. (Eds), *Proceeding of the 2nd ACM Workshop on Information Credibility on the Web, WICOW'08*, ACM, New York, NY, pp. 19-26, available at: http://doi.acm.org/10.1145/1458527.1458534 (accessed September 24, 2013).

Watts, D.J. and Peretti, J. (2007), "Viral marketing for the real world", *Harvard Business Review*, Vol. 85 No. 5.

Weerkamp, W. and De Rijke, M. (2008), "Credibility improves topical blog post retrieval", ACL-08: HLT, Association for Computational Linguistics, pp. 923-931, available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi = 10.1.1.164.1598.

Yardi, S., Romero, D., Schoenebeck, G. and Boyd, D. (2010), "Detecting spam in a Twitter network", *First Monday*, Vol. 15 No. 1, available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2793

Appendix

Feature	Description
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 1,000	… one of the 1,000 most popular ones
CONTAINS POPULAR DOMAIN TOP 10,000	… 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

Table AI. Message-level features

Feature	Description
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

Table AII. User-level features

Table AIII.
Topic-level features

Feature	Description	Abbreviation
COUNT TWEETS	Number of tweets	CNT_TWEETS
AVERAGE LENGTH	Average length of a tweet	AVG_LENGTH
FRACTION TWEETS QUESTION MARK	The fraction of tweets containing a question mark "?"	FR_TWEETS… …_QUEST_MARK
FRACTION TWEETS EXCLAMATION MARK	… an exclamation mark "!"	FR_TWEETS… …_EXCL_MARK
FRACTION TWEETS MULTI QUEST OR EXCL.	… multiple question or exclamation marks	FR_TWEETS… ._QUEST_EXCL_MARK
FRACTION TWEETS EMOTICON SMILE – FROWN	… emoticons smiling or frowning (2 features)	FR_EMOT_SMILE FR_EMOT_FROWN
CONTAINS PRONOUN FIRST – SECOND – THIRD	… a personal pronoun in 1st, 2nd, or 3rd person. (3 features)	FR_PRON_FIRST FR_PRON_SECOND FR_PRON_THIRD
FRACTION TWEETS 30PCT UPPERCASE	… more than 30\% of characters in uppercase	FR_30PCT_UP
FRACTION TWEETS URL	The fraction of tweets containing a URL	FR_TWEETS_URL
FRACTION TWEETS USER MENTION	… user mentions	FR_TWEETS_USR_MENT
FRACTION TWEETS HASHTAG	… hashtags	FR_TWEETS_HASHTAG
FRACTION TWEETS STOCK SYMBOL	… stock symbols	FR_TWEETS_ST
FRACTION RETWEETS	The fraction of tweets that are re-tweets	FR_RT
AVERAGE SENTIMENT SCORE	The average sentiment score of tweets	AVG_SENT_SCORE
FRACTION SENTIMENT POSITIVE	The fraction of tweets with a positive score	FR_SENT_POS
FRACTION SENTIMENT NEGATIVE	… with a negative score	FR_SENT_NEG
FRACTION POPULAR DOMAIN TOP 100	The fraction of tweets with a URL in one of the top-100 domains	FR_POP_DOM_TOP_100
FRACTION POPULAR DOMAIN TOP 1,000	… in one of the top-1,000 domains	FR_POP_DOM_TOP_1,000
FRACTION POPULAR DOMAIN TOP 10,000	… in one of the top-10,000 domains	FR_POP_DOM_TOP_10,000
COUNT DISTINCT EXPANDED URLS	The number of distinct URLs found after expanding short URLs	CNT_DIST_EXP_URLS

(*continued*)

Feature	Description	Abbreviation
SHARE MOST FREQUENT EXPANDED URL	The fraction of occurrences of the most frequent expanded URL	SHR_MOST_FREQ_EXP_URL
COUNT DISTINCT SEEMINGLY SHORTENED URLS	The number of distinct short URLs	CNT_DIST_SE_URLS
COUNT DISTINCT HASHTAGS	The number of distinct hashtags	CNT_DIST_HASHTAGS
SHARE MOST FREQUENT HASHTAG	The fraction of occurrences of the most frequent hashtag	SHR_MOST_FREQ_HASHTAG
COUNT DISTINCT USERS MENTIONED	The number of distinct users mentioned in the tweets	CNT_DIST_USR_MENT
SHARE MOST FREQUENT USER MENTIONED	The fraction of user mentions of the most frequently mentioned user	SHR_MOST_FREQ_USR_MENT
COUNT DISTINCT AUTHORS	The number of distinct authors of tweets	CNT_DIST_AU
SHARE MOST FREQUENT AUTHOR	The fraction of tweets authored by the most frequent author	SHR_MOST_FREQ_AU
AUTHOR AVERAGE REGISTRATION AGE	The average of AUTHOR REGISTRATION AGE	AVG_REG_AGE
AUTHOR AVERAGE STATUSES COUNT	The average of AUTHOR STATUSES COUNT	AVG_STAT_CNT
AUTHOR AVERAGE COUNT FOLLOWERS	… of AUTHOR COUNT FOLLOWERS	AVG_CNT_FOLLOWERS
AUTHOR AVERAGE COUNT FRIENDS	… of AUTHOR COUNT FRIENDS	AVG_CNT_FRIENDS
AUTHOR FRACTION IS VERIFIED	The fraction of tweets from verified authors	FR_IS_VER
AUTHOR FRACTION HAS DESCRIPTION	… from authors with a description	FR_HAS_DESC
AUTHOR FRACTION HAS URL	… from authors with a homepage URL	FR_HAS_URL

Table AIII.

Feature	Description	Abbreviation
PROPAGATION INITIAL TWEETS	The degree of the root in a propagation tree	INIT_TWEETS
PROPAGATION MAX SUBTREE	The total number of tweets in the largest sub-tree of the root, plus one	MAX_SUBTREE
PROPAGATION MAX – AVG DEGREE	The maximum and average degree of a node that is not the root (2 feat.)	MAX_DEGREE AVG_DEGREE
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)	MAX_DEPTH AVG_DEPTH
PROPAGATION MAX LEVEL	The max. size of a level in the propagation tree (except children of root)	MAX_LEVEL

Table AIV.
Propagation-level features

Corresponding author
Barbara Poblete can be contacted at: barbara@poblete.cl