Cutting Through the Noise: Defining Ground Truth in Information Credibility on Twitter

Sujoy Sikdar
Dept. of Computer Science,
Rensselaer Polytechnic Institute
Email: sikdas@rpi.edu

Byungkyu Kang, John O'Donovan & Tobias Höllerer
Dept. of Computer Science,
U. of California, Santa Barbara
Email:
{bkang,jod,holl}@cs.ucsb.edu

Sibel Adalı
Dept. of Computer Science,
Rensselaer Polytechnic Institute
Email: sibel@cs.rpi.edu

ABSTRACT

Increased popularity of microblogs in recent years brings about a need for better mechanisms to extract credible or otherwise useful information from noisy and large data. While there are a great number of studies that introduce methods to find credible data, there is no accepted credibility benchmark. As a result, it is hard to compare different studies and generalize from their findings. In this paper, we argue for a methodology for making such studies more useful to the research community. First, the underlying ground truth values of credibility must be reliable. The specific constructs used to define credibility must be carefully identified. Secondly, the underlying network context must be quantified and documented. To illustrate these two points, we conduct a unique credibility study of two different data sets on the same topic, but with different network characteristics. We also conduct two different user surveys, and construct two additional indicators of credibility based on retweet behavior. Through a detailed statistical study, we first show that survey based methods can be extremely noisy and results may vary greatly from survey to survey. However, by combining such methods with retweet behavior, we can incorporate two signals that are noisy but uncorrelated, resulting in ground truth measures that can be predicted with high accuracy and are stable across different data sets and survey methods. Newsworthiness of tweets can be a useful frame for specific applications, but it is not necessary for achieving reliable credibility ground truth measurements. We also show that the underlying model for predicting credibility can differ depending on the underlying network context, which needs to be clearly identified and reported in credibility studies to improve their impact.

I INTRODUCTION

Increased popularity of microblogs in recent years brings about a need for better mechanisms to extract credible or otherwise useful information from noisy and large data. While there are a great number of studies that introduce methods to find credible data, there is no accepted credibility benchmark. As a result, it is hard to compare different studies and generalize from their findings. What are the desired properties of a credibility study? First of all, the exact definition of credibility must be made very clear by defining the underlying construct of credibility and the classes of credible and not credible messages. Methods to measure and obtain this ground truth must be justified. These methods must be robust, giving predictable results over repeated experiments. Obviously, the main purpose of a credibility study is to find models that can predict credibility with high accuracy and study the important features in such models. These models must also significantly improve on the baseline of random prediction especially in imbalanced prediction tasks. We will refer to a ground truth value as stable if it satisfies all these requirements. Our hypothesis in this paper is that credibility models trained using stable ground truth measures are portable to multiple data sets and studies. To increase portability further, credibility studies must also identify the underlying network context that shape how credibility is communicated. Without a stable definition of credibility, it is hard to judge to which degree a credibility model presents a novel scientific contribution.

Credibility is generally defined as the believability of information [1, 2]. People judge credibility based on many different constructs such as accuracy, objectivity, timeliness and reliability, and rely on different cues like source credibility, social prominence and domain knowledge [3]. To which degree a credibility cue is used depends strongly on the decision making context [4]. Since credibility judgements are subjective, researchers must pay careful attention to the way "ground truth" credibility data is collected. The methods for obtaining ground truth may vary considerably.

©ASE 2012

User surveys for judging credibility are usually unbiased, but unbiased and uninformed. Large-scale online user surveys such as those on Amazon's Mechanical Turk offer a direct way to measure credibility. As the raters of credibility tend not to know the message senders and do not have knowledge about the topic of the message, their ratings predominantly rely on whether the message text looks believable. The results of such studies can be extremely noisy at *a single tweet level* simply because the amount of information is too small to reliably assess credibility. Casual observations may not be as easy to classify as declarative statements. There is variation in how surveys are conducted, but the general expectation is that the survey results are unbiased except for the bias introduced by the cues presented to the raters such as the message sender's social network or the number of retweets for the message, and the way credibility is framed in the survey. Definitions given to the user or the other questions in the survey may be used to frame which specific credibility construct should be considered when judging the credibility of the message. The surveys are also typically uninformed about the topic. As they are performed post-hoc, they do not capture how credibility would have been judged at the time of the message based on the information that was available at that time.

In-network proxies for credibility are informed, but noisy and biased. In network behavior at the time of the message is a good proxy for credibility. For example, retweet behavior is often used as an endorsement for the quality and interestingness [5], and credibility [6] of the message. Given a retweet may mean many different things, it is a noisy factor. It can also be affected by other factors such as the trust for the sender if the sender is known personally, her reputation, information cascades and corroboration in the network. Note that this type of bias may actually improve the quality of credibility judgments obtained from observed behavior. In addition, the behavior reflects how credible the message was at the time it was sent, judged by people who have a stake in a given topic. It also takes into account the level of uncertainty in the network. Some messages may also be rumors that are later found to be false, but the social media network may actively stop rumors as well [7]. Overall, behavioral proxies for credibility can be noisy, but they are also informed by the knowledge of the senders and the topic.

Studies of credibility based on analysis of factors are incomplete if they do not introduce meaningful controls. It is hard to compare and use different studies due to lack of control variables in these studies. The network characteristics and network behavior differ greatly depending on the topic, who participates in the discussion and the level of uncertainty that exists at the time. There is little work that investigates what these controls might be. As a result of all these difficulties in determining ground truth and measuring it within the proper context, there are no benchmarks for credibility.

In this paper, we illustrate how to construct a stable ground truth value by carefully considering pros and cons of different ways to obtain it. To make this case, we conduct a unique study. We collect two data sets on the same topic, but from different perspectives. The first one is on Hurricane Sandy, collected during the storm. There is great uncertainty about what is happening and in fact there are even reported cases of misinformation being distributed [8]. The second data set is for the relief effort after the storm, from a period of lower uncertainty. Also, the network in the second data set is much more connected as it is initiated by people who have an existing social network. It is likely that people who know each other talk differently than those who talk to a general audience.

We construct different base ground truth values. We consider two different surveys in which subjects are shown different information about the same tweets. This allows us to test to which degree credibility judgments across different surveys are comparable and how the survey method influences the results. We also consider two different ways to quantify retweets, overall and at the time of the message, capturing the importance of the message at two different time granularities. We show that overall survey ratings for individual tweets are hard to predict and can vary greatly from survey to survey. Prediction of retweets may vary from data set to data set. Overall, user surveys and retweet behavior are noisy indicators of credibility, but they are uncorrelated and provide different type of information.

Then, we show that it is possible to construct multiple sophisticated ground truth values by combining individual base measures. We illustrate with examples how different definitions of credibility yield significantly different but valid sets of messages. We conduct a comprehensive study on the predictive accuracy of these ground truth values across both data sets. We show that it is possible to predict the credibility of individual tweets with accuracy values of 0.93-0.95 for different ground truth definitions, highest shown in the literature. Furthermore, the pre-

©ASE 2012

diction improves significantly over the baseline. This finding is true for both datasets, regardless of how the survey is conducted, which lets us conclude that our ground truth values are stable. We show that newsworthiness of tweets can be a useful frame for specific applications, but it is not needed for constructing ground truth values that can be predicted with high accuracy. Furthermore, our combined ground truth values are not only easier to predict, but also capture the best aspects of credible information: judged credible by survey subjects and found interesting/relevant within the network. We believe our method provides a step towards a more standardized approach to studying credibility.

Our findings provide compelling evidence that reliable and meaningful credibility measurements can be constructed by combining uncorrelated and noisy measurements. We illustrate how in the following sections.

II RELATED WORK

Due to the rising importance of social media sites, especially Twitter, in disseminating news and information, and organizing action in situations involving high uncertainty such as social movements and natural disasters, information credibility on Twitter has been studied extensively [1, 9, 10].

A great deal of studies concentrate on creating feature based models. [1] studies the prediction of newsworthy topics and newsworthy topics that have credible information. Whereas this work helps find topics or Twitter hashtags with credible information, our work aims to find the actual messages or tweets that have credible information regardless of the topic. Other work focuses on predicting influential users [11] or experts [12]. These studies tend to offer a set of network features and then report on the most predictive ones. An important aspect of such work is that they illustrate that it is possible to create highly predictive models. However, there is little work that discusses how the credibility ground truth underlying such work can be measured and what the pitfalls are. This is the topic we concentrate on.

Most work concentrating on creating feature based models of credibility rely on user surveys [1, 6, 12–15]. There is no uniform practice in conducting such surveys. What should be shown to the users? For example, it is well-known that many surface cues impact credibility judgments [16] and including such cues in surveys may bias the results. Given there are many constructs for credibility, should a specific construct of credibility be defined for the users before the survey? Overall, judging the credibility of a single tweet may be challenging given the short amount of text available. How good are users in assessing credibility of individual tweets? For example, work in [1] first asks whether a topic is newsworthy. Then within newsworthy topics, it samples a collection of 10 tweets, and asks the survey subjects whether the group of tweets is credible or not. While such an approach reduces the uncertainty involved in judging a single tweet, it does not provide credibility ratings at the single tweet level. In fact, this approach is useful for finding newsworthy topics in which there are credible tweets. In contrast, we look at predicting the credibility of individual tweets using features studied in prior work. Amazon Turk is a frequently used medium for user surveys. Prior work [17] has found the results from this medium acceptable for tasks involving a small number of annotations. This is the medium we use in this paper as well.

A further specialization of research in feature-based models examines topic-specific expertise in the network. Liao et al. [18] focus on judgement and perception of expertise from novice and expert perspective. Other researchers explore automated algorithms that model and predict expertise [19] and [2]. A common argument in these works is that since trust and credibility can vary across contexts, it is important to segment the network into different contexts or "topics" to achieve a more fine-grained understanding of credibility in the network. A key challenge in modeling topic-specific expertise in a complex network is deciding on the manner in which the topic data is collected [2]. Schall [20] highlight this by demonstrating a significant decay in topic similarity across hops in the friend-follower network on Twitter. Canini et al. [21] demonstrate how LDA 'topic modeling' algorithms can help identify expertise and credibility at a topic-specific granularity. Garcia Esparza [22] et al. argue that the Twitter follower network itself should be more topic-specific from the outset, and present a novel tool called CatStream to support simple categorization of information feeds on Twitter by expertise of the message originator. All of these researchers promote the concept of that segmentation of the network by expertise and/or topic helps support better information analysis. This paper presents a general methodology for collection of ground truth credibility information. We do not use a topic-specific approach, however, this is an avenue we would like to investigate in future work.

©ASE 2012

In prior work, there is little focus on the limitations of user surveys in determining content credibility which is a topic we study in this paper. Furthermore, in addition to showing that credibility of individual tweets can be predicted, we directly compare our findings to the baseline which we take as being the proportion of instances belonging to the majority class. Often in situations with great imbalance between predicted classes, the baseline may already be very high. Furthermore, we directly study how stable credibility ratings are by conducting two such surveys on the same messages. The results show that surveys alone are not very useful for constructing predictive models. We also show that it is possible to construct meaningful ground truth values that can be predicted with high accuracy by considering multiple constructs in conjunction. To our knowledge, this is the first such study in the literature.

The perceived credibility of the sources, especially determined by how knowledgeable the sources are on a topic, is frequently used to determine the credibility of a message [3]. Often rumors are suppressed by those who are knowledgeable about a given topic or user. Subjects of user surveys do not have access to such information. One clear measure of how other users view a message is their behavior towards it in the network. Retweeting is one such behavior. There is a growing body of work that aims to understand what is the underlying motivation behind retweeting, whether a retweet can be considered a type of endorsement of the credibility of a message. For example, a study of Microsoft employees who are Twitter users shows that a retweet is a significant factor in the perceived credibility of messages [6]. Retweet behavior is also studied directly in [5], which shows a variety of reasons of retweeting from amplifying tweets to new audiences to publicly agreeing with someone. Mustafaraj and Metaxas [23] show that in political tweets, people tend to retweet information that they directly agree with. Prior work has used retweets as a proxy of the sender's influence on the retweeter [11], and of credibility based on the number of retweets [7]. Work has suggested the depth of a retweet chain shows the relevance of the tweet to that community [10]. Overall, prior work tends to agree that a retweet is a type of endorsement of a tweet for its credibility or interestingness. However, there is no study that considers how retweet behavior might be used to complement ground truth assessments of credibility, which is what we study in this paper.

Another important problem is reporting on such credibility studies. Often such studies lack a clear study of confounding factors. The importance of such factors has been discussed in the case of political discourse, e.g. the level of penetration of Twitter in a specific area involved in the discussion [24], the type of communities involved in the discussion such as the polarization around the topics [23]. We show that our method for predicting credibility is effective in two different communities discussing the same topic, but from different perspectives. We also show that features predicted in both topics can be very different, requiring appropriate controls when reporting on feature based credibility models.

III CONTENT AND USER BASED FEATURES

In this section, we describe the various features used in our study. While some of the features are novel, the rest have been proposed in prior work by us [2,25] and others [1]. Our intention is to provide a set of features comparable with other studies of credibility. However, we remove features that are highly correlated with each other to increase the interpretability of the results. Note that our intention is not to provide a set of comprehensive features, but to give representative features that cover the frequently studied categories for user based and content based information. These include user's network, her behavior towards others, typical content of her messages, the properties of the message in question. Our features include most of the top rated features in prior work [1].

Content based features evaluate the textual content alone, whether the text contains mentions, urls, specific type of words, sentiments expressed and so on. Note that when judging credibility, the importance of the textual content cannot be disregarded [4]. For example, information that appears plausible is much more likely to be believed. Information that is familiar to the information consumer can be remembered quickly, and as a result may be judged more credible. These types of heuristics are often employed when judging credibility. In fact other cues such as the existence of cited sources (people or urls), whether it was retweeted or not, may be used to infer the authoritativeness of the message [3]. In fact, past work shows the importance of such features [1, 2]. However, the way authority is communicated differs based on who is speaking to whom. This well-studied notion in social sciences has not been properly studied in the credibility models. A person talking to public will use terms that are easily understood by everyone. A person talking to their social circle will

©ASE 2012

use words and expression that are known within the circle. This distinction between tacit vs. implicit knowledge is shown to be very relevant in many social situations [26]. In our study involving two separate datasets, we aim to test if this is indeed true in our case. As we will show, one of the data sets comes from a much more connected network for the same topic. The list of content based features are given in Table 1. Details on these features can be found in [2].

feature name	*description*
char/word	# chars/# words
question	# question marks
excl	exclamation marks
uppercase	# uppercases in text
pronoun	# pronouns (count by corpus)
smile	# smile emoticons
frown	# frown emoticons
url	# urls
retweet	RT in tweet text, 0: not retweeted, 1: retweeted once, 2: multiple times
sentiment_pos	positive/negative word count based on
sentiment_neg	lexicon sourced from NLTK[a]
sentiment	polarity (sentiment_pos - sentiment_neg)
num_hashtag	from entity metadata
num_mention	from entity metadata
ellipsis	counting ellipsis sign(. . .)
news	occurrence frequency of news sources
lex_diversity	proportion of unique words per tweet
dialog_act_type	category: statement, system, greet, emotion, ynquestion, whquestion, accept, bye, emphasis, continuer, reject, yanswer, nanswer, clarify, other
news_words	NLTK corpus of news article terms (sourced from Reuters), count of occurrence
chat_words	AOL messenger corpus, count of occurrence

Table 1: The set of content-based Twitter features analyzed in our evaluation.

[a]NLTK: Python Natural Language Processing Toolkit.

User based, social features try to assess the credibility or expertise of a person by the size of their network. The number of friends gives one access to diverse information, while number of followers allows them to distribute information widely. In addition, number of followers is an endorsement of the importance of a person in the network. There are many studies that elaborate on the importance of these features [12,25]. Often, they signal reputation which serves as a proxy for competence or expertise [12]. The age information, i.e. number of years one has been on Twitter impacts the network of a person as older users are likely to be more central in the overall Twitter network.

In addition, it is possible to develop user features based on how the user behaves in the network and even more importantly how the user's followers behave towards the user. Often behavior from both the sender's and receiver's perspective reveals more detailed information than the simple structural information. In prior work [25], we show that behavior differs towards friends versus acquaintances. All our behavioral features are computed based on the statistical properties of behavior between pairs of individuals without considering message content. We compute them using only the topic based collection, hence their computation does not create an additional cost.

feature name	*description*
u-friend/ u-follower	# friends/followers (log)
u-age	# years on Twitter
u-bal_soc	ratio of follower to friends
u-default_image	user has default image or not (0/1)
u-url/ u-mention	mean # urls /mentions in tweets
favorite-count	# tweets favorited
u-hashtag	mean # hashtags in tweets
u-length	mean text length in tweets
u-balance	mean balance of number of followers
u-conv-balance	mean balance of conversations
u-tweets/ u-favorite	# of tweets / tweets favorited
u-time	mean time between tweets
u-directed-ratio	# directed tweets/#broadcast tweets
u-retweet-ratio	# retweets/#tweets
u-prop-from	# users the user propagates from
u-prop-to	# users that propagate the user
u-convers-with	# users that converse with the user
u-propagated-tweets	# tweets propagated by other users
u-propagation-energy	amount of propagation energy spent on this user by others
u-worthiness	proportion of user's tweets found worthy of propagation by others
u-conv	mean # conversations
ss_length	avg length of chain-like behavior
ss_friends	ss_length * avg number of friends (log)
ss_followers	ss_length * avg number of followers (log)

Table 2: The set of user-based Twitter features analyzed in our evaluation.

Propagation type behavior by followers of a person combined with high number of followers, assortativity (balance of the number of friends and followers computed by entropy) are signals for asymmetric relationships that are signals of reputation [25]. Conversation type behavior and reciprocity of messages are a signal of friendship. For each pair, we seek at least one directed message in each direction. We then aggregate features for each user across all the friends

©ASE 2012

and followers to find the mean behavior for each user. Note that propagation in our features is not actual retweet behavior as we do not consider message content. It finds pairs of messages that statistically appear to be propagations [27]. Table 2 summarizes all the user features in our study.

For conversation based features, we look for a sequence of directed messages that appear close enough in time compared to the rest to be considered a unit. For propagation-like behavior, we consider the timing of the messages from a user A to a user B, and use a linear time maximum matching algorithm developed in [28] between B's incoming and outgoing messages satisfying a causality constraint with respect to time. A propagation-like behavior does not necessarily represent a retweet. If there are a lot of actions in which B appears to propagate from A, we can conclude that B receives a lot of messages from A and sends out a lot of messages. As a result, B is a good conduit. To further emphasize this concept, we compute chains of these behaviors and find the average length of such chains, which we will call social strength, `ss` for short. We also compute for each chain originating at node A, the average number of friends or followers along the chain multiplied by the length of the chain. We average these values and call it `ss_friends` (and similarly for followers). All `ss` features represent how well a node is as a conduit in the whole network. Details of behavioral features can be found in [25].

IV COLLECTION AND ANNOTATION OF TWITTER DATA

In this paper, we introduce a unique comparative study of two different data sets on the same topic, Hurricane Sandy. The first dataset `FR` was collected during Hurricane Sandy using keywords "#sandy" and "#frankenstorm", two keywords commonly used for the hurricane. The second data set `OS` was collected right after the Hurricane using keyword "#occupysandy". Occupy Sandy is a coordinated relief effort to distribute resources and volunteers to help neighborhoods and people affected by Hurricane Sandy. It has been started by those who have participated in Occupy Wall Street demonstrations in 2012. Our choice of these two topics reflects two different perspectives about the same broader event. During the hurricane, there is a great deal of uncertainty. The topic is also of great interest to a large group of people, many of whom may not know each other. The relief effort involves a more localized group of people who are likely to know each other to some degree. In fact, we tested the connectivity hypothesis. We collected samples of equal number of users from both datasets. We then computed the average number of connections to each other as friends in the sample. This value was 1.5 for `FR` and 6 for `OS`. Therefore, users in `OS` are much more connected to each other. As a result, both datasets offer us with a comparable study. They are on the same newsworthy topic. But, after controlling for topic, they represent two different contexts based on the level connectivity and uncertainty. We compare and contrast credibility measurements in these two different data sets.

We crawled both data sets using the Twitter Streaming API starting from Oct 29th, 2012 for two weeks. We applied keywords "#sandy" and "#frankenstorm" in order to generate the dataset `FR` during the storm and "#occupysandy" to generate dataset `OS` during the relief effort. The streaming API is not rate-limited, so it was possible to collect a large amount of tweets for our first data set `FR`. The second topic `OS` was far less popular (Table 3). From each dataset, we collected two basic samples of tweets, the first is a random set and the second is the set of tweets from users who had exchanged 2 or more messages with others in our collection. The survey tweets are a sample of 2,000 tweets each from each group with a total of 4,000 tweets. This allows us to have tweets from users with some social connectivity as well as random users.

In our samples, we excluded the users who are outliers, with more than 5K friends and 50K followers. These numbers were chosen as two standard deviations above the mean for typical Twitter users. We obtained these numbers by crawling the user info from the 2011 NIST Twitter dataset [1] containing 16 million representative tweets from 2011.

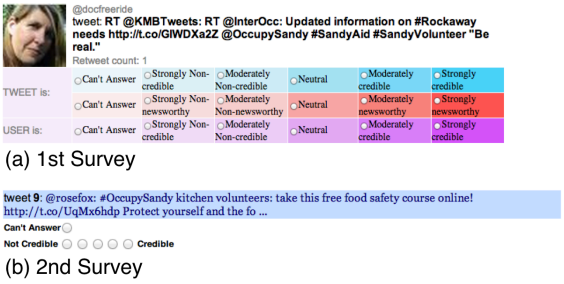


Figure 1: Screen shot from the two MTurk tweet assessment surveys.

[1] http://trec.nist.gov/data/tweets/

©ASE 2012

1 ANNOTATING TWITTER DATA

To analyze the tweets in terms of credibility we conducted two surveys. In survey 1 (Figure 1), we showed the users the message text, the source picture and retweet count, and sought three different types of annotations related to information credibility: the message is credible `E`, the message is newsworthy `N` and the user is credible `U`. Note that message credibility is extended with the additional source information, as a result, we will refer to this as `E`. In total 381 participants took part. Participants also had an option to select "can't answer". In all cases, assessments of 3 on the Likert scale and "can't answer" responses were discarded.

The existence of images in the survey `E` may impact the evaluation of credibility as faces are often used to identify whether a source is trustworthy or not [29–32]. In fact, facial evaluation is often much faster than the evaluation of text due to the dedicated processing of this signal in the brain. We expect that source credibility judgments can be impacted by this signal as there are only few other signals relating to the source. Furthermore, asking questions on newsworthiness of the tweet and the credibility of the user frame the message credibility judgment. We will test whether this frame had a noticeable impact in the next section.

Set Name	`FR`	`OS`
Seed Authors in Entire Collection	2,154,735	24,463
Seed Tweets in Entire Collection	3,801,395	60,671
Annotated Tweets in Survey `E`	8,728	6,503
Authors of Tweets in Survey `E`	7,974	3,239
Annotated Tweets in Survey `T`	3,471	3,639
Authors of Tweets in Survey `T`	2,654	1,657

Table 3: Overview of the two topic-specific data collections mined from Twitter.

To overcome possible issues related to the cues shown to the survey subjects, we conducted a second survey (Figure 1). This time participants were presented with a definition of credibility, given as: "The message states a true fact and/or is believable, regardless of whether it is a newsworthy item or a personal detail.", and shown only the textual content. We sought only a single ground truth `T`, whether the text is credible or not. In total 206 participants took part and at least 3 annotations were obtained for each tweet and the majority score was taken. If majority of the raters agreed on whether the message was credible or not, we used the corresponding label. Otherwise, this message was excluded. In this case, no information other than the text is available to judge credibility, but credibility is defined for the users as a construct more general than newsworthiness.

Both surveys were run using MTurk users. All assessments were given in 1-5 scale. Participants were presented with instructions, followed by a pre-survey questionnaire and a set of simple filtering questions to test for bots and other noise such as rapid tab-click behavior. Each participant's ability to rate was also tested using this set of pre-test questions. Those who did not answer the set reasonably were discarded, although this was unknown to them at the time of the study. Messages are then classified as credible (1) or not credible (-1) based on their score.

As mentioned in the introduction, surveys are particularly useful for evaluating the text content of messages. But, there are a number of limitations. The survey subjects are unlikely to be familiar with the survey topic and are more likely to use heuristics to evaluate the credibility of the message. Furthermore, as it is very unlikely that the survey subjects are familiar with the senders of the information, source credibility information will not be based on prior information regarding the source. Hence, the use of the survey is limited in measuring the credibility of the message as a function of the expertise and reliability of the source.

To overcome this problem, we compute a secondary measure of credibility based on the fact that the message was retweeted in the network. This means that others in the network endorsed the message in some way. We compute two ground truth values, `RT` is the total number of retweets for a single tweet. All tweets in our dataset get the same `RT` (for retweet total) value as the original message that they are a retweet of. However, these retweets may have happened before or after our collection. We also computed a measure of the number of retweets of messages during the time of the collection by finding a set of tweets that are retweets of the same message either by text similarity or by their metadata. Again all the messages in a retweet group are given the same value. We call this second measure `RS` (for retweet sample). This second value represents how the message was propagating during the event we were monitoring.

We assign tweets an `RS` value of 1 if the message appears more than twice in our sample and a value 0 if the message appears only once in our sample, disregarding the rest. Similarly, we assign an `RT` value of 1 if the message has a retweet count greater than one

©ASE 2012

Abbr	Description	Credible	Not credible
U	user credible or not (survey 1)	value 4,5	value 1,2
N	message newsworthy or not (survey 1)	value 4,5	value 1,2
E	message credible or not (survey 1)	value 4,5	value 1,2
T	message credible or not (survey 2)	value 4,5	value 1,2
RT	message retweeted or not	2 or more times	0 times
RS	message retweeted in the sample or not	more than 2 times	1 times
NE	credible among newsworthy messages	N & E	N & ¬ E
NT	credible among newsworthy messages	N & T	N & ¬ T
RTE	credible among retweeted messages	RT & E	RT & ¬ E
RTT	credible among retweeted messages	RT & T	RT & ¬ T
RSE	credible among retweeted messages	RS & E	RS & ¬ E
RST	credible among retweeted messages	RS & T	RS & ¬ T
ERT	credible and retweeted	E & RT	¬ E & ¬ RT
ERS	credible and retweeted	E & RS	¬ E & ¬ RS
TRT	credible and retweeted	T & RT	¬ T & ¬ RT
TRS	credible and retweeted	T & RS	¬ T & ¬ RS
NERT	credible and retweeted among newsworthy messages	N & E & RT	N & ¬ E & ¬ RT
NERS	credible and retweeted among newsworthy messages	N & E & RS	N & ¬ E & ¬ RS
NTRT	credible and retweeted among newsworthy messages	N & T & RT	N & ¬ T & ¬ RT
NTRS	credible and retweeted among newsworthy messages	N & T & RS	N & ¬ T & ¬ RS
rERT	relaxed version of ERT	E & RT	¬ E ∨ ¬ RT
rERS	relaxed version of ERS	E & RS	¬ E ∨ ¬ RS
rTRT	relaxed version of TRT	T & RT	¬ T ∨ ¬ RT
rTRS	relaxed version of TRS	T & RS	¬ T ∨ ¬ RS
rNERT	relaxed version of NTRT	N & E & RT	N & (¬ E ∨ ¬ RT)
rNERS	relaxed version of NTRS	N & E & RS	N & (¬ E ∨ ¬ RS)
rNTRT	relaxed version of NTRT	N & T & RT	N & (¬ T ∨ ¬ RT)
rNTRS	relaxed version of NTRS	N & T & RS	N & (¬ T ∨ ¬ RS)

Table 4: The list of different ground truth measures used

©ASE 2012

and an RT value of 0 if the message has never been retweeted. A benefit of this method is that while the survey is a post-hoc analysis, propagation looks at how credible the message was at the time it was traveling in the network. Information that was uncertain at the creation time may be known by the time the survey is conducted. The propagation information also incorporates how credible the source of the message was. The longer a message has traveled in the network, the more credible we consider it to be. Also, messages that are part of a long chain are likely to originate from users with higher credibility and reliability.

These measures of credibility serve as the basis of ground truth. However, we note that it is possible to augment ground truth judgments by combining complementary approaches. For example, E and RT provide different type of information about credibility. We expect both to be noisy indicators, but we also expect the noise to be uncorrelated. As a result, the combination ground truth that looks at tweets that are judged as credible and were also retweeted, is likely to be less noisy overall. Furthermore, these tweets constitute a more meaningful measure of ground truth, as credible messages that others in the network found useful and/or interesting. We will test in the next section various ways to construct ground truth and how well they can be predicted. To our knowledge, this unique approach has not been studied in any of the related work on predicting ground truth.

The list of ground truth measures tested in this paper are shown in Table 4. We consider multiple definitions of credible and not credible messages with different meaning and different levels of restrictiveness. We consider multiple criteria for framing credibility. For example, in NT, newsworthiness is the frame. The credibility is defined only for newsworthy messages. A message is considered credible if it is newsworthy **and** credible. A message is considered not credible if it is newsworthy **and** not credible. In RTE, retweets is the frame. We consider credibility only for those messages that are retweeted. The opposite class is defined by negating all the conditions for semantic clarity. For example, in NTRT, a newsworthy message that is not credible will be not credible with respect to T and not credible with respect to RT. We also test more relaxed versions of the opposite set using versions rTRT, rTRS, rNTRT, rNTRS. The relaxed version of ground truth measures are motivated by whether the two sets of raters could agree if the message was credible. The positive class corresponds to the case where both sets or raters agreed that the message was credible while the negative class corresponds to the case where either of the two sets of raters did not find the message credible.

U	N	E	T	NE	NT	RT	RS
1.00	0.48	0.55	0.06	0.47	0.42	0.06	0.06
0.48	1.00	0.55	0.06	0.83	0.81	0.05	0.05
0.55	0.55	1.00	0.04	0.64	0.49	0.06	0.36
0.06	0.06	0.04	1.00	0.07	0.29	0.04	0.03
0.47	0.83	0.64	0.07	1.00	0.84	0.04	0.04
0.42	0.81	0.49	0.29	0.84	1.00	0.09	0.09
0.06	0.05	0.06	0.04	0.04	0.09	1.00	0.89
0.06	0.05	0.04	0.03	0.04	0.09	0.89	1.00

(a) FR

U	N	E	T	NE	NT	RT	RS
1.00	0.42	0.42	0.03	0.45	0.37	0.05	0.05
0.42	1.00	0.41	0.04	0.82	0.74	0.10	0.11
0.42	0.41	1.00	0.01	0.57	0.34	0.06	0.07
0.03	0.04	0.01	1.00	0.04	0.28	0.07	0.08
0.45	0.82	0.57	0.04	1.00	0.76	0.12	0.13
0.37	0.74	0.34	0.28	0.76	1.00	0.15	0.16
0.05	0.10	0.06	0.07	0.12	0.15	1.00	0.91
0.05	0.11	0.07	0.08	0.13	0.16	0.91	1.00

(b) OS

Table 5: Correlation of the various ground truth measures for the two datasets.

V RESULTS

1 GROUND TRUTH SELECTION

In this section, we study the results from the different ground truth collection methods. As described in Section IV, we have conducted two different types of user surveys. In the first, users were shown tweet text as well as the image of the author and the retweet count for the tweet. They were also asked whether the tweet was newsworthy or not. In the second, only the tweet was shown. We even removed any RT in the beginning of the text to remove cues that the message was retweeted. As a result, participants were forced to read the messages and judge them on textual content alone. However, without any information regarding the source, the survey takers lacked a frequently used anchor for credibility. In the first survey, they had a chance to base their opinions on cues like author's user image, the RT in the text and actual retweet count.

We first look at the degree to which credibility judgements are correlated to each other in the two surveys (Table 5). High correlation would imply a stable way to obtain ground truth information. The first thing we notice is that E is not at all correlated with T. How-

©ASE 2012

ever, NE and NT are highly correlated (0.76-0.84). Assuming newsworthiness is a stable construct for surveys, we can conclude that credibility judgments are stable within newsworthy messages. But, without a specific construct, it is hard to get a stable survey response as subjects can use many different definitions. Note that in the second survey, we did not ask for newsworthiness directly, but used the ratings from survey 1. Our method is not directly comparable to survey in [1] that asks credibility for groups of tweets not individual messages and does not ask for newsworthiness of the message. However, the observation that asking for credibility of a single or multiple messages without a framing construct may provide noisy results is applicable in the general sense to any credibility study.

We also note that in the first survey, measures U, N, E are highly correlated with each other (0.4-0.55). This shows us that the source and information credibility are judged similarly. It is also likely that the existence of questions regarding the newsworthiness of a message had an impact on the framing of the judgments of credibility. For example, it is likely that newsworthy messages were more likely to be viewed as credible, and vice versa. However neither E or T are highly correlated with retweet based measures. Hence, we cannot conclude that showing number of retweets in E had a significant impact in credibility judgments. This leads us to conclude that RT and RS constitute an uncorrelated hence complementary measure of credibility on top of the user defined credibility measures. We also note that RT measure is informed by the knowledge of the message topic and sender that is not available to the survey subjects, and hence provides an independent type of credibility judgment.

RT and RS are highly correlated with each other, which means that our sample (as in RS) is fairly representative of the actual retweet behavior. This is also due to the fact that we only consider whether a message was retweeted or not, and disregard the actual number of retweets. We do note however that RT and RS ultimately measure a different behavior, at the time of message for RS, versus in the long run for RT.

2 PREDICTABILITY OF GROUND TRUTH

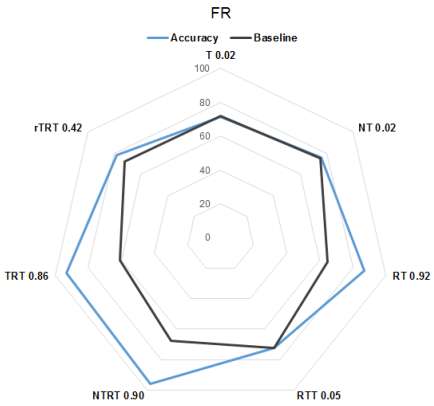
Table 6 shows the accuracy achieved by our model on the task of predicting different ground truth measures using 10-fold cross validation. For these tests, we chose the best features for each ground truth using all our features. The total number of features used in this section for each ground truth measure did not exceed 10. A best feature study is presented in the next section. We also show the baseline accuracy which is measured as the prediction accuracy a classifier would achieve if it ignored all predictors and always predicted the majority class. This also shows the class imbalance in the data. We also report the Kappa statistic and the ROC Area achieved by the model. The Kappa statistic represents how much the classifier outperforms a random guess (ranges between -1 for the worst, to 1 for the best). The ROC area represents the ability of the classifier to distinguish between the two classes (ranges between 0 for the worst and 1 for the best). Prediction rates were obtained using logistic regression which was chosen because it achieved the best results overall. We excluded all features that were computed based on the retweet counts while training our models to predict ground truth measures that include RT or RS.

In general, FR is a more noisy data set in which prediction is harder. For the task of predicting RT and RS, our model achieved prediction accuracies of 0.87 and 0.95 in FR and 0.94 and 0.95 in OS. Survey based measures (e.g. N, E, T) seem to be very noisy in comparison and prediction using our features is not very effective (not significantly better than baseline).

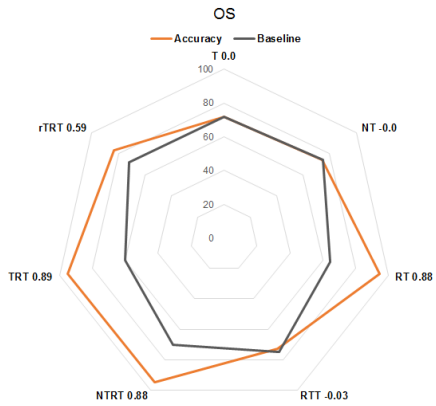
We then consider the problem of predicting credibility of newsworthy tweets, for which we had high correlation between NE and NT, and hence concluded that these were stable constructs. These are also not much better than baseline. Finally, we look at using retweet behavior as an anchor, and try to predict whether retweeted messages are credible or not (RTE, RTT, RSE, RST). Despite the fact that it is easy to predict whether a message is retweeted or not, it is not as easy to predict the credibility of such messages. We surmise that newsworthiness and relevance as measured by retweet rate are not good frames for obtaining reliable assessments of credibility or to make good predictions.

Our features yield significantly better predictive performance (0.92-0.95) at predicting the combined ground truth values (TRS, TRT) over both the FR and OS datasets (tweets that are credible and retweeted versus not credible and not retweeted). Similarly, ERT, ERS also show similar improvements, but the improvement over baseline is less significant in FR. We are able to build better models to predict TRT

©ASE 2012



(a) Dataset FR



(b) Dataset OS

Figure 2: Accuracy of prediction using different ground truth values. The values next to each ground truth value represents the Kappa value, representing how much the classifier outperforms the random guess (ranges between -1 worst, to 1 best).

Ground Truth	Baseline	Accuracy	Kappa	ROC Area
N	59.11	63.96	0.19	0.64
E	61.36	64.20	0.14	0.61
T	71.67	71.45	0.02	0.60
RT	64.59	86.97	0.72	0.92
RS	94.60	94.84	0.24	0.83
NE	87.03	87.03	0	0.51
NT	75.57	75.82	0.02	0.63
RTE	65.25	65.43	0.01	0.53
RTT	72.43	72.58	0.05	0.63
RSE	55.77	69.23	0.38	0.76
RST	67.16	76.12	0.43	0.78
ERT	52.87	**87.05**	0.74	0.92
ERS	92.73	92.98	0.24	0.84
TRT	60.69	**93.24**	0.86	0.95
TRS	87.03	91.64	0.60	0.94
NERT	78.93	93.77	0.80	0.90
NERS	75.00	84.09	0.56	0.89
NTRT	67.40	**95.89**	0.90	0.94
NTRS	84.85	**93.18**	0.73	0.77

(a) FR

Ground Truth	Baseline	Accuracy	Kappa	ROC Area
N	57.52	58.81	0.09	0.59
E	60.29	60.20	0.01	0.56
T	71.92	71.84	0	0.58
RT	64.44	**94.20**	0.88	0.97
RS	78.29	**94.89**	0.85	0.98
NE	82.08	82.08	0	0.53
NT	74.59	74.29	0	0.59
RTE	64.29	65.15	0.04	0.55
RTT	74.72	72.70	-0.03	0.56
RSE	66.10	66.83	0.08	0.60
RST	77.51	77.95	0.06	0.63
ERT	52.72	**93.84**	0.85	0.97
ERS	67.69	**93.25**	0.85	0.97
TRT	60.18	**94.84**	0.89	0.97
TRS	55.44	**95.26**	0.91	0.98
NERT	78.07	96.11	0	0.53
NERS	66.77	95.89	0.91	0.96
NTRT	70.0	**95.00**	0.88	0.96
NTRS	57.19	**95.41**	0.91	0.96

(b) OS

Table 6: Prediction for the various ground truth measures for the two datasets.

than ERT, but the difference is very small.

By combining these ground truth measures of credibility and retweets, we are essentially finding agreement between two sets of raters, from the Twitterverse and from MTurk. We had already concluded that these two were complementary credibility measures. Therefore, we are effectively reducing noise by combining two independent judgments which enables us to make better predictions. This new ground truth reflects all the desired properties of a credible message: it appears credible and it is propagated in the network. We see similarly improved performance when we attempt to predict credibility within

©ASE 2012

the context of newsworthiness. The combined ground truth measures (NTRS, NTRT) represent the same measures of credibility on newsworthy tweets where both sets of raters agree and our model achieves similarly high prediction accuracy at this task (0.93-0.96). The classifier also performs significantly better than random, has high Kappa-statistic and has high ROC area values.

It is interesting to note that the models thus built on our features perform equally well at predicting credibility over all tweets and over newsworthy tweets only. We conclude that placing credibility in the context of newsworthiness is not necessary to make good predictions of credibility if we can find the right construct of credibility and a robust ground truth measure. Ultimately, it is important to define the correct construct for judging credibility if we wish to get reliable results. Furthermore, these results appear stable across both datasets.

Dataset	Ground Truth	Baseline	Accuracy	Kappa	ROC
FR	rTRT	72.48	78.22	0.42	0.88
FR	rTRS	95.95	96.13	0.21	0.87
FR	rNTRT	68.94	79.42	0.53	0.88
FR	rNTRS	95.74	94.24	-0.02	0.88
OS	rTRT	71.79	83.28	0.59	0.90
OS	rTRS	81.56	89.30	0.64	0.94
OS	rNTRT	65.36	82.16	0.62	0.88
OS	rNTRS	75.03	88.38	0.71	0.92

Table 7: Prediction for the relaxation of the ground truth measures NTRT and NTRS for the two datasets.

Measures (rTRS, rTRT) and (rNTRT, rNTRS) represent a relaxation of these measures where the negative class comprises of messages where the two raters could not agree that the message was credible. Although this relaxation means we do not have as clear a separation between the classes, it does mean that we can make predictions and train over a larger proportion of tweets. We present the performance results in Table 7. We find that, despite the more noisy description, our model achieves relatively good prediction accuracy (0.78-0.89 and 0.79-0.88 respectively) at this task as well over both topic datasets. However, these measures are not too different than baseline for the more noisy data set FR.

We present examples of top and bottom most credible tweets as predicted by our classifier in Table 8. We trained a supervised classifier using a 66% split of a sample of the data using all of our features, and obtained classification results of whether the tweet belonged to the positive or negative class. Top and bottom tweets were chosen based on the confidence of the classifier that the instance belonged in the class predicted. As we can see, newsworthy and credible tweets resemble news items more closely. On the other hand, credible tweets for rTRS also include messages meant for exchanging information. There is a noticeable difference between top and bottom tweets. Top tweets contain more credible sources and more important information. Bottom tweets on the other hand include more conversational tweets in both cases, however bottom tweets for NTRS also include links to other sources and declarative statements more frequently when compared with rTRS. We also note a statement in the bottom tweets for TRS happens to be incorrect: the New York City Marathon was in fact cancelled in contrast with the speculation in this tweet.

Our findings and prediction results indicate that combining human annotation and retweet based judgements on credibility yields meaningful and robust ground truth measures. Many such measures can be constructed. We also find that it is possible to make reasonably high quality predictions of credibility across different datasets using features that are relatively inexpensive to compute. We note that we achieve higher accuracy in predicting credibility of individual messages than reported in [1].

3 BEST FEATURES IN DIFFERENT NETWORK CONTEXTS

In this section, we study the most predictive features for different ground truth values and compare the two different datasets corresponding to the two different network contexts. As discussed in the introduction, FR contains messages from a time of high uncertainty when compared to OS. Similarly, users in OS have higher percentage of social ties.

We use a heuristic based forward subset selection regression (FSS) to find a linear combination of the features that best predict the annotations in a given segment. FSS first finds the best single feature that approximates the given ground truth annotation. Then, it adds the next feature that minimizes the leave-one-out cross validation (LOO-CV) error until no improvements can be made to the LOO-CV error. This process typically produces a very sparse set of features and prevents over-fitting. We report only on

©ASE 2012

Source	Tweet Text
FR/NTRS	- RT @SportsCenter: Packers safety Charles Woodson said he's donating $100,000 to the Red Cross for assisting families hurt by Hurricane S ... - East Coast power outages from Hurricane Sandy reach 8.1 million: (Reuters) - East Coast electric companies say o... http://t.co/wVER1VsO - RT @cnnbrk: At least 50 U.S. deaths now linked to #Sandy – among a total of 118 worldwide. http://t.co/W3BSwLBL
OS/NTRS	- RT @OccupySandy: Today @520ClintonOS received so many donations! So many, UPS has agreed to donate a fleet of delivery trucks #mutualaid ... - RT @OccupySandy: Dozens of new volunteers lined up to get to work at the St. Jacobi #OccupySandy hub. http://t.co/k66ZSbF9 - RT @OccupyWallStNYC: Where do I find info to help #SandyVolunteer? http://t.co/HJEBBhdi Keepin' it Simple and REAL. #SandyAid #SandyHelp ...
FR/rTRS	- SANDY: Bloomberg - NYC put stickers on homes & buildings in SI & other places in NYC with different level colors to notify if they can enter - Long Islanders Use Facebook, Google Docs to Find Loved Ones Post-Sandy: Whether looking for a sk... http://t.co/BbUiFuUU via @mashable - RT @piersmorgan: I've changed my mind about this - Mayor Bloomberg should postpone the NYC marathon. Priority must be the #Sandy rescue ...
OS/rTRS	- RT @rhookinitiative: #RHISupports RT @shawncarrie: #ParkSlope needs volunteers to go to flooded areas. Meet at 8th & Garfield. #San ... - RT @AnthonyQuintano: RT @OccupySandy: Want to volunteer doing #SandyRelief in NYC? Start by filling our volunteer form: http://t.co/6CePyV2c - RT @OccupySandy: NEW: Want to do some #SandyAid in New Jersey? Check out our NJ info page: http://t.co/MLRAlm58 and map: http://t.co/35j ...

(a) Examples of top tweets

Source	Tweet Text
FR/NTRS	- RT @TimTebow: My thoughts &; prayers go out to everyone effected by Hurricane Sandy. Please be safe & help each other through thi ... Garden City resident: Sandy was a rude awakening http://t.co/wtQl3JK9 - @RZA ˜ @RedCross Donation Info Video ˜ @fema ˜ #Sandy ˜ http://t.co/hIIbnBUz ˜ to Donate $10 Text REDCROSS to 90999 _breakingDawn_XXX
OS/NTRS	- @OccupyWallSt @occupysandy ASIA ALMADEH woman 45 y lost her life by #death gas Crimes of the repressive regime in #Bahrain #HRW - @forwardretreat @520ClintonOS @OccupySandy Thanks for the suggestion. We are currently coordinating with @520ClintonOS - @ALAG_Aims check out the wedding.g registry. It has what they need. @OccupySandy
FR/rTRS	- The adventures of @Hanssie & her 2 friends trying to make it home after #Sandy #Getmehome http://t.co/CajGamps - The mayor is clearly not going to cancel it, so it is up to the runners to do the right thing. #sandy #nyc #marathon #volunteer - It's cold here tonight so glad for heating. I can't think how cold it is for people affected by #Sandy with no power/heating thinkin of U
OS/rTRS	- @_opXpress @AirOccupy @JemaNdunerkant Did they actually do this to anyone? #OccuChat #occupysandy - @OccupySandy thank you! Just making sure. :) - . @hey_haywood @OccupySandy The fun thing with Clothes Mountain is as soon as it's sorted and out the door, IT REAPPEARS from new donations!

(b) Examples of bottom tweets

Table 8: Examples of top and bottom tweets from a sample for various datasets and ground truth measures.

those features with significance at 1% in Figure 3.

There are many differences between the two data sets. There are also many differences between the ground truths as expected, illustrating that these are in fact different constructs. In almost all ground truth values, FR models employ a more diverse set of features than OS. Also, the features in FR tend to include reputational features like long chains (short chains in OS in contrast) and user properties that would imply that the user is a heavy Twitter user with the use of mentions, hashtags and picking of favorite tweets. There is also difference in the content based features from FR, including different punctuation and use of unique words. Overall, shorter messages are more credible in FR and longer messages in OS. These distinctions could be due to different reasons. Users in FR come from a more varied group, as anyone interested in the hurricane was likely participating in the discussion. In OS, a specific group of people organizing the effort was more active. These people knew each other and hence are likely to be more similar in the way they talk. One conclusion could be that there are a diverse number of features associated with retweeted messages in FR as it comes from a diverse set of users. As OS is from a more tight group of individuals, the messages have more normative features. One can also attribute this to the difference in the nature of the discussion: in FR which is on trying to assess the damage, while in OS in trying to organize others and give information. In OS, having more broadcasts than directed messages is significant, but not in FR. It is likely that most users in FR have this property and the feature is not distinctive.

If we look at similarities across all the tests, citing

©ASE 2012

news sources, not having mentions, use of ellipsis for explanation or emphasis are uniformly important for predicting credibility.

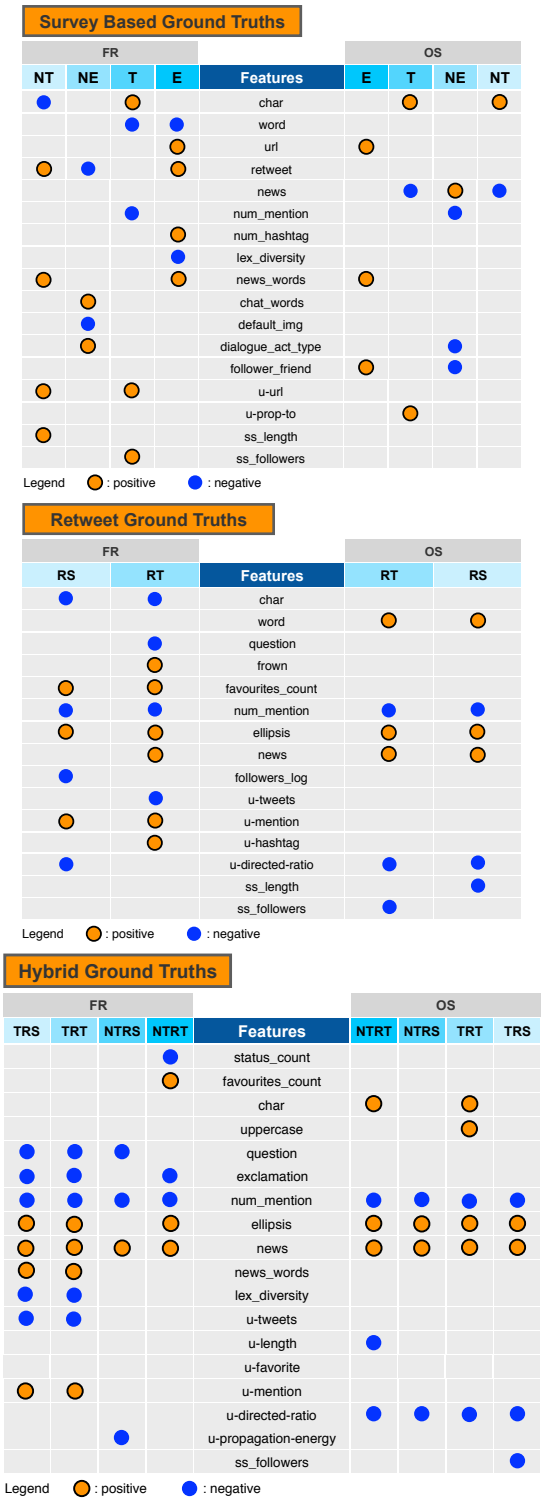


Figure 3: Best features for the different ground truths in each data set.

Overall, we can easily conclude that the best features for measuring credibility are highly dependent on the network context and the specific ground truth studied. This line of study also opens up many interesting new questions that can be asked by similar comparative study of different data sets and ground truth values. By seeing the differences between the best features, we can better understand when certain features are relevant and can make better informed choices on modeling credibility.

VI A GUIDE TO STUDYING INFORMATION CREDIBILITY

In this section, we summarize our findings into a number of action items that are applicable to all credibility studies, in particular to feature based models in microblogs. Any study on information credibility must satisfy as many of the following requirements as possible.

- Define the exact notion of credibility used in the study and describe the credibility definition as explicitly as possible.
- Use multiple credibility measures that are independent or have different biases as much as possible. Combine these measures in the ground truth construction.
  - Credibility constructs that incorporate assessment of textual credibility as well as source expertise on a topic are expected to be richer and more valuable constructs.
  - In network activity is a good proxy for measuring source expertise.
  - User surveys are a good way to obtain textual credibility, but users may lack topical expertise.
  - In network activity can also be a good proxy for measuring textual credibility, but may incorporate uncertainty arising from the lack of information at the time of message.
- Test credibility measures in multiple datasets to measure their correlation and prediction, to ensure that they are stable constructs.
- Define credible and not credible message classes carefully and unambiguously. Remove middle range of responses.

©ASE 2012

  - One method to do this is by removing middle range of responses from surveys or other measurements.
  - Another method to do this is by explicit construction in which the "not credible" class is the opposite of the "credible" class for every measure used in it. For example, if "credible" is defined as *A and B*, then "not credible" should be defined as *not A and not B*.

    Note that this is not the logical opposite, it throws away classes that are ambiguous, i.e. *A and not B* and *not A and B*.

- If a specific frame such as newsworthiness is appropriate for the given study, then define these classes within the given frame.

  Example: Given N is for newsworthy, credible with *N and A and B* and not credible with *N and not A and not B*.

- Design a survey to solicit user ratings that are appropriate for the given definition and frame of credibility.
  - Users should be given the correct definition of credibility.
  - Questions should be chosen carefully so that they do not influence how users perceive credibility.

- In feature based studies, do not forget that a single model is not likely to apply to all possible decision contexts. Identify the relevant contextual parameters for the study, measure and report them as much as possible.
  - Possible contextual parameters include the strength and type of ties between those involved in a conversation. People will likely use different words when talking to a stranger than to a friend, and when talking to an audience versus a single person.
  - Another contextual parameter is uncertainty and risk. In situations involving high risk (such as natural or other disasters) and in times of high uncertainty, people are likely to rely on the most trusted information sources and have limited resources for processing information.

We hope that future research will concentrate on standardized ground truth data sets developed by following these guidelines and be made available to the research community at large.

VII CONCLUSIONS AND FUTURE WORK

This paper described a novel method of constructing reliable and meaningful credibility ground truth values for microblogging sites like Twitter at the individual message level. We have shown that survey results can be noisy, affected by the specific framing of the questions and may differ greatly from survey to survey. Overall, it is hard to create prediction methods with high accuracy based on survey methods alone. Retweet behavior is easier to predict with network based features, but can differ from network to network. However, these two measures convey different and complementary information about credibility. We show that these two measures are uncorrelated in reality. Hence, by combining them, we are able to get ground truth values that are less noisy, can both be predicted with very high accuracy (0.93-0.95), and also capture the properties of the type of messages we would like to predict: credible text that has been endorsed as important by the network. We also show that while by framing credibility within the context of newsworthiness we do achieve high prediction accuracy, it does not necessarily result in a great increase in performance. The most important part is to choose a stable definition of credibility. In fact, we show multiple such definitions in this paper and also illustrate some that do not work very well. We show that it is possible to measure credibility for newsworthy messages as well as for general messages with high accuracy. These findings are true in both datasets and the two different survey methods we study. We have also shown that the best features for credibility differ based on the underlying network context. As a result, feature studies must carefully consider and report on the relevant contextual elements. Examples of such contextual elements are cultural norms for behavior, level of penetration of the social media site and the type of users. In this paper, we show that the social connectivity of individuals is likely an important contextual factor. Our message based on our findings is clear: any credibility study must carefully define and measure ground truth, and quantify the relevant contextual factors.

We note that it is possible for messages to satisfy these extended definitions of ground truth and still contain misinformation. To improve further on such a metric, we can consider more sophisticated measures such as the embeddedness of different sources in the network. Investigating the effectiveness of such expensive measures is future work. We also intend to

©ASE 2012

expand this study further by looking at how ground truth for other constructs such as expertise and interpersonal trust can be constructed using a combination of complementary methods. Additionally, we would like to expand our work towards understanding topic based expertise and credibility.

Acknowledgments. This research was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and by NSF grant IIS-1058132. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *WWW*, 2011, pp. 675–684.

[2] B. Kang, J. O'Donovan, and T. Hollerer, "Modeling topic specific credibility on twitter," in *Proceedings of IUI*, 2012, pp. 179–188.

[3] B. Hilligoss and S. Y. Rieh, "Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context," *Information Processing and Management*, vol. 44, pp. 1467–1484, 2008.

[4] S. Adalı, *Modeling Trust Context in Networks*. Springer Briefs, 2013.

[5] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, ser. HICSS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2010.412

[6] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW '12. New York, NY, USA: ACM, 2012, pp. 441–450. [Online]. Available: http://doi.acm.org/10.1145/2145204.2145274

[7] Y. Suzuki, "A credibility assessment for message streams on microblogs," in *Proceedings of the 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, ser. 3PGCIC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 527–530. [Online]. Available: http://dx.doi.org/10.1109/3PGCIC.2010.90

[8] J. Keller, "How truth and lies spread on twitter," october 31, 2012, BusinessWeek.

[9] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, ser. PSOSM '12. New York, NY, USA: ACM, 2012, pp. 2:2–2:8. [Online]. Available: http://doi.acm.org/10.1145/2185354.2185356

[10] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: can we trust what we rt?" in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 71–79. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964869

[11] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social Media*, 2010.

[12] K. Canini, B. Suh, and P. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011.

[13] C. Castillo, M. Mendoza, and B. Poblete, "Predicting Information Credibility in Time-Sensitive Social Media," *Internet Research*, 2013.

[14] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2451–2460. [Online]. Available: http://doi.acm.org/10.1145/2208276.2208409

[15] M. De Choudhury, N. Diakopoulos, and M. Naaman, "Unfolding the event landscape on

©ASE 2012

twitter: classification and exploration of user categories," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW '12. New York, NY, USA: ACM, 2012, pp. 241–244. [Online]. Available: http://doi.acm.org/10.1145/2145204.2145242

[16] P. Stavri, D. Freeman, and C. Burroughs, "Perception of quality and trustworthiness of internet resources by personal health information seekers," in *AMIA Annual Symposium Proceedings*, 2003.

[17] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613751

[18] Q. V. Liao, C. Wagner, P. Pirolli, and W.-T. Fu, "Understanding experts' and novices' expertise judgment of twitter users," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2461–2464. [Online]. Available: http://doi.acm.org/10.1145/2208276.2208410

[19] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 91–100.

[20] M. Schaal, J. O'Donovan, and B. Smyth, "An analysis of topical proximity in the twitter social graph." in *SocInfo*, 2012, pp. 232–245.

[21] K. R. Canini, B. Suh, and P. L. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.

[22] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth, "Catstream: categorising tweets for user profiling and stream filtering," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, ser. IUI '13. New York, NY, USA: ACM, 2013, pp. 25–36. [Online]. Available: http://doi.acm.org/10.1145/2449396.2449402

[23] E. Mustafaraj and P. T. Metaxas, "What edited retweets reveal about online political discourse," in *AAAI-11 Workshop on Analyzing Microtext*, 2011.

[24] D. Gayo-Avello, "I wanted to predict elections with twitter and all i got was this lousy paper," 2012.

[25] S. Adalı, M. Magdon-Ismail, and F. Sisenda, "Actions speak as loud as words: Predicting relationships from social behavior data," in *Proceedings of the WWW Conference*, 2012.

[26] D. Z. Levin and R. Cross, "The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer," *Academy of Management Journal*, vol. 50, no. 11, pp. 1477–1490, 2002.

[27] S. Adalı, R. Escriva, M. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B. K. Szymanski, W. A. Wallace, and G. M. Williams, "Measuring behavioral trust in social networks," in *Proc. International Conference on Intelligence and Security Informatics (ISI)*, 2010, pp. 150–152.

[28] J. Baumes, M. Goldberg, M. Hayvanovych, M. Magdon-Ismail, W. Wallace, and M. Zaki, "Finding hidden group structure in a stream of communications," *Intel. and Sec. Inform. (ISI)*, 2006.

[29] A. Todorov and N. N. Oosterhof, "Modeling social perception of faces," *IEEE Signal Processing Magazine*, vol. 117, 2011.

[30] V. Wout and Sanfey, "Friend or foe: the effect of implicit trustworthiness judgments in social decision-making," *Cognition*, vol. 108, no. 3, pp. 796–803, 2008.

[31] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall, "Inferences of competence from faces predict election outcomes," *Science*, vol. 308, pp. 1623–1626, 2005.

[32] L. J. Chang, B. B. Doll, M. van't Wout, M. J. Frank, and A. G. Sanfey, "Seeing is believing: Trustworthiness as a dynamic belief," *Cognitive Psychology*, vol. 61, no. 2, pp. 87–105, 2010.

©ASE 2012