Understanding Information Credibility on Twitter

Sujoy Sikdar*, Byungkyu Kang†, John O'Donovan†, Tobias Höllerer†, Sibel Adalı*
*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA
Email: sikdas@rpi.edu, sibel@cs.rpi.edu
†Department of Computer Science, University of California, Santa Barbara, CA, USA
Email: {bkang, holl, jod}@cs.ucsb.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 defined. 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.**

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 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. 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]. 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 [2]. To which degree a credibility cue is used depends strongly on the decision making context [3]. 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.

User surveys for judging credibility are usually unbiased but 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. As surveys 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 [4], and credibility [5] 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 [6]. Overall, behavioral proxies

for credibility can be noisy, but they are also informed by the knowledge of the senders and the topic. 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 [7]. 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 with significantly different meanings. 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 individuals tweets with accuracy values of 0.93-0.95 for different ground truth definitions, highest shown in the literature. Furthermore, the prediction improves significantly over the baseline. This finding is true for both datasets, regardless of how the survey is conducted, which let's 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.

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], [8], [9]. A great deal of studies concentrate on creating feature based models. [1] studies the prediction of newsworthy topics that have credible information. Other work focuses on predicting influential users [10] or experts [11]. 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], [5], [11]–[14]. There is no uniform practice in conducting such surveys. Subjects are asked to judge a single tweet in some cases. Castillo et.al. [1] first determine newsworthy topics, then for these topics ask users to rate a sample of 10 tweets judging which topics tend to have credible messages. It is also not clear if newsworthiness is necessary for judging credibility. In prior work, there is little focus on the limitations of user surveys in determining content credibility or improvements over the baseline which is a topic we study in this paper.

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 [2]. 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 tends to agree that a retweet is a type of endorsement of a tweet for its credibility or interestingness [4], [5], [9], [15]. 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 that credibility studies often lack a clear study of confounding factors such as the level of penetration of Twitter in the public [16], the type of communities involved in the discussion such as the polarization around topics [15]. We show that our method for predicting credibility is effective in two different communities discussing the same topic, but from different perspectives.

III. Content and User Based Features

In this section, we briefly 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 [17], [18] 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 content based (see Table II) and user based information (see Table I).

Content based features evaluate the textual content alone, whether the text contains mentions, urls, specific type of words, sentiments expressed and so on as the content of the message

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 / sentiment_neg	positive/negative word count based on lexicon sourced from NLTK[1]
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, yn-question, 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 I. THE SET OF CONTENT-BASED TWITTER FEATURES ANALYZED IN OUR EVALUATION.

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 II. THE SET OF USER-BASED TWITTER FEATURES ANALYZED IN OUR EVALUATION.

is one of the most crucial factors in judging credibility. User based, social features try to assess the credibility or expertise of a person by their network, its size and level of activity. 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 [11], [17]. Often, they signal reputation which serves as a proxy for competence or expertise [11]. All the user features listed here are computed based on the statistical properties of behavior between pairs of individuals without considering message content. We do use meta-data to assess which messages are directed in conversation based features and compute all features using only the topic based collection, hence their computation does not create an additional cost. Propagation in our features is not retweet behavior, but pairs of messages that are likely to be a propagation based on statistical matching method [19]. If there are a lot of actions in which B appears to propagate from A, we can conclude that 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 a number of features summarizing the different aspects of chains. Details of behavioral features can be found in [17].

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

TABLE III. OVERVIEW OF THE DATA COLLECTIONS.

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 (Table III). 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

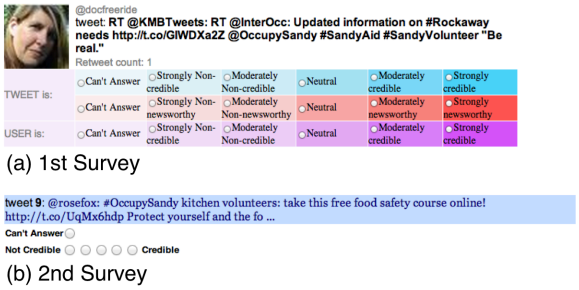

Fig. 1. Screen shot from the two MTurk tweet assessment surveys.

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 based on the the 2011 NIST Twitter dataset [2] containing 16 million representative tweets from 2011.

To analyze the tweets in terms of credibility we conducted two surveys. 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.

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`. 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 [20], [21]. 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.

To overcome possible issues related to the cues shown to the survey subjects, we conducted a second survey. 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. Furthermore, the users are forced to read the text of the messages that they are rating.

As mentioned in the introduction, surveys are particularly useful for evaluating the text content of messages, but the survey subjects are unlikely to be familiar with the survey topic or message senders. To overcome this problem, we compute two secondary measures of credibility based on the fact that the message was retweeted in the network. `RT` is the total number of retweets for a single tweet given by retweets of 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` or `RT` 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. 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 with high retweet representing higher credibilty.

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 construct a number of novel ground truth measures based on these measures as shown in Table IV with different levels of restrictiveness and different frames. For example, in `NT`, newsworthiness is the frame, where credibility is considered 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. This allows us to exclude ambiguous messages from our training set. 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`.

V. Results

A. Ground Truth Selection

We first look at the degree to which credibility judgements are correlated to each other in the two surveys (Table IV). 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`. However, `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,

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

The list of different ground truth measures used

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	2 or more times	0 times
NE	credible among newsworthy msgs	N & E	N & ¬ E
NT	credible among newsworthy msgs	N & T	N & ¬ T
RTE	credible among retweeted msgs	RT & E	RT & ¬ E
RTT	credible among retweeted msgs	RT & T	RT & ¬ T
TRT	credible & retweeted	T & RT	¬ T & ¬ RT
TRS	credible & retweeted	T & RS	¬ T & ¬ RS
NTRT	credible & retweeted among newsworthy msgs	N & T & RT	N & ¬ T & ¬ RT
NTRS	credible & retweeted among newsworthy msgs	N & T & RS	N & ¬ T & ¬ RS
rTRT	relaxed version of TRT	T & RT	¬ T ∨ ¬ RT
rTRS	relaxed version of TRS	T & RS	¬ T ∨ ¬ RS

Correlation of the various ground truth measures for the two datasets.

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 IV. DIFFERENT GROUND TRUTH MEASURES USED AND THEIR CORRELATION IN DIFFERENT DATASETS.

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.

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; 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 which is informed by the knowledge of the message topic and the sender. RT and RS are highly correlated with each other, despite measuring a slightly different behavior. This means that our sample (as in RS) is fairly representative of the actual retweet behavior. We only consider whether a message was retweeted or not, and disregard the actual number of retweets.

B. Predictability of Ground Truth

Table V 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 and the total number of features used in this section for each ground truth measure did not exceed 10. We also show the baseline accuracy (B) which is measured as the prediction accuracy (A) a classifier would achieve if it ignored all predictors and always predicted the majority class chosen at random. This also shows the class imbalance in the data. We also report the Kappa statistic (K) and the ROC Area achieved by the model. The Kappa statistic represents how much the classifier outperforms a random guess (ranges between -1, worst to 0, best). The ROC area represents the ability of the classifier to distinguish between the two classes (ranges between 0, worst to 1, 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.

	FR				OS			
GT	B	A	K	ROC	B	A	K	ROC
N	59.11	63.96	0.19	0.64	57.52	58.81	0.09	0.59
E	61.36	64.20	0.14	0.61	60.29	60.20	0.01	0.56
T	71.67	71.45	0.02	0.60	71.92	71.84	0	0.58
NE	87.03	87.03	0	0.51	82.08	82.08	0	0.53
NT	75.57	75.82	0.02	0.63	74.59	74.29	0	0.59
RT	64.59	86.97	0.72	0.92	64.44	**94.20**	0.88	0.97
RS	94.60	94.84	0.24	0.83	78.29	**94.89**	0.85	0.98
RTT	72.43	72.58	0.05	0.63	74.72	72.70	-0.03	0.56
RST	67.16	76.12	0.43	0.78	77.51	77.95	0.06	0.63
TRT	60.69	**93.24**	0.86	0.95	60.18	**94.84**	0.89	0.97
TRS	87.03	91.64	0.60	0.94	55.44	**95.26**	0.91	0.98
NTRT	67.40	**95.89**	0.90	0.94	70.0	**95.00**	0.88	0.96
NTRS	84.85	**93.18**	0.73	0.77	57.19	**95.41**	0.91	0.96
rTRT	72.48	**78.22**	0.42	0.88	71.79	**83.28**	0.59	0.90
rTRS	95.95	96.13	0.21	0.87	81.56	**89.30**	0.64	0.94

TABLE V. PREDICTION FOR THE VARIOUS GROUND TRUTH (GT) MEASURES FOR THE TWO DATASETS FOR BASELINE (B), ACCURACY (A), KAPPA (K) AND THE ROC AREA.

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). When predicting credibility of newsworthy tweets, we observe that they are also not much better than baseline either. 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 accuracy (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), hence

are stable ground truth values. Our results are superior to those reported in [1] despite the fact that our prediction is in single tweet level, not at the group level as in [1]. By combining these ground truth measures we are essentially finding agreement between two sets of raters, from the Twitterverse and from MTurk, which we show to be complementary credibility measures. Therefore, we are effectively reducing noise by combining two independent judgments which enables us to make better predictions. We see similarly improved performance when we attempt to predict credibility within 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). However, framing credibility within the context of newsworthiness is not necessary for achieving high accuracy in predictions. The classifier also performs significantly better than random, has high Kappa-statistic and has high ROC area values. The relaxed versions of the ground truth (`rTRT`, `rTRS`) bring more training data, but also more noise, resulting in slightly diminished accuracy. Our results indicate that by combining human annotation and retweet based judgements of credibility and by using classes that are well-separated logically, we can obtain many different meaningful and robust ground truth measures that are fairly inexpensive to compute.

VI. 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 and using carefully defined classes of credibility, 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 by framing credibility within the context of newsworthiness, the prediction accuracy remains unchanged. These findings are true in both datasets and the two different survey methods we study. Our message based on our findings is clear: any credibility study must carefully define and measure ground truth.

We note that it is possible messages satisfying these extended definitions of ground truth may 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. We also intend to 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. Furthermore, we also want to study how contextual factors such as network connectivity and uncertainty at the time of message impact the credibility models.

Acknowledgments. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. 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 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. 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.

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

[4] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *HICSS*, 2010, pp. 1–10.

[5] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *CSCW*. New York, NY, USA: ACM, 2012, pp. 441–450.

[6] Y. Suzuki, "A credibility assessment for message streams on microblogs," in *3PGCIC*. IEEE Computer Society, 2010.

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

[8] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *PSOSM*. New York, NY, USA: ACM, 2012, pp. 2:2–2:8.

[9] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: can we trust what we rt?" in *SOMA*, 2010, pp. 71–79.

[10] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *ICWSM*, 2010.

[11] K. Canini, B. Suh, and P. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *SocialCom*, 2011.

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

[13] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *CHI*, 2012, pp. 2451–2460.

[14] M. De Choudhury, N. Diakopoulos, and M. Naaman, "Unfolding the event landscape on twitter: classification and exploration of user categories," in *CSCW*, 2012, pp. 241–244.

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

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

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

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

[19] 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 *ISI*, 2010, pp. 150–152.

[20] 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.

[21] 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.