Modeling Topic Specific Credibility in Twitter

Byungkyu Kang, John O'Donovan,
Tobias Höllerer
Department of Computer Science
University of California Santa Barbara
{bkang, jod, holl}@cs.ucsb.edu

ABSTRACT
This paper presents and evaluates three computational models for recommending credible topic-specific information in Twitter. The first model focuses on credibility at the user level, harnessing various dynamics of information flow in the underlying social graph to compute a credibility rating. The second model applies a content-based strategy to compute a finer-grained credibility score for individual tweets. Lastly, we discuss a third model which combines facets from both models in a hybrid method, using both averaging and filtering hybrid strategies. To evaluate our novel credibility models, we perform an evaluation on 7 topic specific data sets mined from the Twitter streaming API, with specific focus on a data set of 37K users who tweeted about the topic "Libya". Results show that the social model outperfoms hybrid and content-based prediction models in terms of predictive accuracy over a set of manually collected credibility ratings on the "Libya" dataset.

Author Keywords
Credibility, Trust, Microblogs, Data Mining, Social Networking

ACM Classification Keywords
H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms
Experimentation

INTRODUCTION

Twitter's micro-blogging service is a free forum for the provision of all sorts of information, bringing together many millions of users, from individuals to large corporations. Twitter enables uploads and dissemination of short messages called "tweets" from an increasingly diverse number of client devices such as cellphones and tablets. For example, Apple's recent release of iOS 5 supports "tweeting" natively from millions of iOS 5 enabled devices[1], for both text and image content. All of these mechanisms make it easy for people to post facts, opinions, statements or any other arbitrary tweet in this melting pot of information. The primary focus and contribution of this paper is on an evaluation and comparison of three novel approaches for predicting credible information for specific topics on Twitter –an important challenge given the abundance of useless information in the forum. We first model social credibility, then focus on content-based (tweet-only) credibility, and lastly on a hybrid of features from both approaches. This aims to maximize the available window of information from the social web site upon which we can base a credibility assessment. The paper focuses on two core questions: 1) how well can we assess credibility in Twitter using our proposed models? 2) how do social, content-based and hybrid models perform at itentifying credible information? Analysis of our methods is performed on a range of metrics, from credibility-based predictions of simple features from available metadata, to prediction on thousands of tweets with manually labeled credibility assessments. Our evaluations also consider the contextual nature of credibility on Twitter, from the perspective that credibility can vary with perceived context. Results from a user study assessing credibility on a set of tweets with varying credibility indicators (context) are presented and discussed. We will begin by describing the Twitter domain in terms of social links and information flow, and define what we mean by "credibility" in this context, then follow with a discussion of relevant research in this area.

Information Flow

Propagation of information on Twitter is largely based on a "friends and followers" model, making it a suitable forum for the spread of news and other trends. Tweets and profiles are public and anyone can see a user's posts. Propagation of content generally occurs via posts from people you follow that are streamed proactively to your personalized twitter feed, and conversely your posts are streamed to your followers' feeds. Text-based search is also supported, along with a ranked list of "trending topics", which are tweets containing a topic keyword labeled with "#", such as "*#LondonRiots*" for instance. Trending topics are ranked both globally and by geographic area. Users can reference others in their tweets through the "mention" tag "@", as in "thanks @bob2011". Interpersonal communication is facilitated through a "reply" function, where a user can click a button and reply to the author of any tweet. Information flow is further enabled through

[1] http://blog.twitter.com/2011/10/twitter-and-ios-5-sharing-made-simple.html

a "retweet" mechanism, whereby a user can click a button and post another user's tweet to his/her own profile, and therefore broadcast it to his/her followers. "Credible" tweets contain factual, newsworthy or otherwise useful information, for example, "Libya names new armed forces chief aje.me/AcOXz4". Conversly, less credible tweets contain false, malicious or simply no useful information, for example, "I am Dajjal Mahmood commander w/ NATO in Libya. Hold this secret help secure 2m dinars I recover in the house of Gaddafi", or highly personal information such as "I like mexican food!!!". With more than 200 million users contributing up to 5000 tweets per second through the above methods, a thriving and highly dynamic social network exists, along with an abundance of meta-data and text-based content, which we believe is a suitable domain for a hybrid (social and content-based) prediction strategy.

Defining Credibility.

The classic problem of the social web: information overload, abounds in Twitter. This paper examines several strategies for distinguishing "credible" information (and information sources) about a topic of interest from the abundance of useless, false or nonsensical information available through the Twitter services. For the purpose of our discussion, we define two types of "credibility" in the context of a target topic of interest:

Definition 1. *Tweet-Level Credibility: A degree of believability that can be assigned to a tweet about a target topic, i.e: an indication that the tweet contains believable information.*

Definition 2. *Social Credibility: The expected believability imparted on a user as a result of their standing in the social network, based on any and all available metadata.*

Tweet-level credibility is similar to Castillo's definition in [4], with the addition of the topic level constraint. Tweet level credibility can also be summed/propagated to the user level by averaging over a profile of tweets. Conversely, a user's social credibility is attached to all of his/her tweets.

BACKGROUND

The primary contribution of this paper is the design and analysis of credibility models for microblogs. Since the proliferation of the social web as a platform for user provided content, a large amount of research effort has focused on modeling trust and credibility of content producers. In this discussion of related work we focus on a handful of the most relevant works. Our analysis falls into three broad categories: research in the general area of trust and credibility on the web; research in the microblog domain; and finally an overview of relevant works from the field of recommender systems, particularly those works that have guided the models and methods presented in this paper.

Credibility and Trust on the Web

Research on trust and credibility in a social context has been popular for many decades, from Kochen & Poole's experiments [20] and Milgram's famous small worlds experiment [16], trust has been shown to play an important role in social dynamics of a network. With social web API's, researchers now have many orders of magnitude more data at our fingertips, and we can experiment and evaluate new concepts far more easily. This is evident across a variety of fields, for example, social web search [13], semantic web [6] [8], online auctions [10] [22] [19], personality and behavior prediction [11] [1], political predictions [7] and many others.

Credibility on Twitter

Scale, network complexity and rich content make twitter an ideal forum for research on trust and credibility. Some approaches, for example [23] rely on content classifiers or the social network individually, while others harness information from both sources. Canini et al. [3] present a good example of the latter, to source credible information in Twitter. As with the methods in this paper, they concentrate on topic-specific credibility, defining a ranking strategy for users based on their relevance and expertise within a target topic. Based on user evaluations they conclude that there is "a great potential for automatically identiftying and ranking credible users for any given topic". Canini et al. also evaluate the effect of context variance on perceived credibility. Later in this paper, we provide a brief overview of a similar study performed on our data, correlating with the findings in [3] that both network structure and topical content of a tweet have a bearing on perceived credibility.

Twitter has been studied extensively from a media perspective as a news distribution mechanism, both for regular news and for emergency situations such as natural disasters for example [4][15][12]. Castillo et. al. [4] describe a very recent study of information credibility, with a particular focus on news content, which they define as a statistically mined topic based on word co-occurrence from crawled "bursts" (short peaks in tweeting about specific topics). They define a complex set of features over messages, topics, propagation and users, which trained a classifier that predicted at the 70-80% level for precision/recall against manually labeled credibility data. While the three models presented in this paper differ, our evaluation mechanism is similar to that in [4], and we add a brief comparison of findings in our result analysis. Mendoza et. al [15] also evaluate trust in news dissemination on Twitter, focusing on the Chilean earthquake of 2010. They statistically evaluate data from the emergency situation and show that rumors can be successfully detected using aggregate analysis of Tweets. Our evaluation of Follower / Following relations from our crawled data (shown in Figures 5 and 6 yields a very similar pattern to their result.

Credibility in Recommendation

Recommender systems have been the focus of research attention for many years, and reputation metrics (such as credibility) [18] have been shown to play an important role in the process of content prediction. They can be applied in social filtering to augment user similarity metrics in the recommendation process. [18]. They have also been shown to increase robustness of prediction algorithms in cases where bad (malicious / erroneous) ratings exist [2][17]. Models that include explicit distrust have recently been shown to produce better

predictions, for example, Victor et. al [24] highlight the advantage of combining trust and distrust metrics to compute predictions over multiple network paths, while a recent study by Golbeck shows that distrust metrics can be used to predict hidden trust edges in a network with very high accuracy [5]. In this paper, we are not propagating credibility values around the network, or computing direct interpersonal trust at the diadic level, however, the authors believe that distrust metrics can potentially improve credibility predictions in Twitter.

MODELING CREDIBILITY

Traditional recommendation strategies such as content-based [14] or collaborative filtering [9] [21] typically compute a *personalized* set of recommendations for a target user based on some derivation from that user's profile of item preferences. An important distinction between these techniques and the approaches presented here is that personalization is only performed at the topic level in our algorithms. That is, personalized preferences for a single target user are not considered, only those for a target group who are interested in a specific topic. While we believe that traditional personalization does play an important role for predicting credible content, the focus here is on predicting credible information within a target group centered around a topic of interest, and moreover on the prediction of credible content for target users beyond that group, where preference information is commonly inaccessible.

Given these goals and constraints, we now present three computational models for assessing credibility of information within a specific microblog topic. We begin with by defining nomenclature for the domain:

Definition 3. *The Twitter domain can be represented as a quintuple* (U, F_o, F_e, T, X)*, where* F_o *and* F_e *are two* $U \times U$ *matrices representing binary mappings* $f \in F_o, F_e \mapsto 0, 1$ *between users in* U *(termed "follower" and "following" groups, respectively).* T *is the set of tweets, distributed over* U*, and* X *is the set of topics in* T*.*

By this definition, Twitter is rich in both text content and social network links. Research in recommender systems has long argued the benefits of combining content based and collaborative approaches to recommendation to maximize information gain in the prediction process [14] [9] [21]. For example, while content-based methods tend to predict narrowly, in that they must match a text description of an item already in a target user's profile, collaborative techniques have the potential to provide more serendipitous predictions since they are based on subjective opinions of groups of similar users.

Since our domain is rich in both text content and network links, we propose the following three approaches for identifying credible information, borrowing from the content and collaborative synergies identified by the recommender system community.

1. *Social Model*: A weighted combination of positive credibility indicators from the underlying social network.
2. *Content Model*: A probabilistic language-based approach identifying patterns of terms and other tweet properties that tend to lead to positive feedback such as retweeting and/or credible user ratings.
3. *Hybrid Model*: A combination of the above, firstly by simple weighting, and secondly through cascading / filtering of output.

Social Model

Complex network structure and feed-based information flow make dissemination of information on Twitter highly dynamic and ephemeral in nature. Accordingly, the task of detecting factors of credibility is inherently difficult. Consider internet pioneer Tim Berners-Lee for example: Tim follows 83 people, but has almost 54,000 followers, compared to the global average of 126 followers. Former US House speaker Newt Gingrich has over 1.5 million followers, however a recent report[2] revealed that only 8% of these are real people, and the remainder are automatically generated profiles used to create a false impression of popularity. Another example outlier is the user "@Twitter", which has 6.4 million followers, since it shares useful news about the forum itself. Furthermore, direct marketing companies commonly follow users to collect information solely for target advertising, and don't behave as "regular" followers. There are also a large number of fake profiles that have become very popular, for instance CNET's analysis[3] revealed dozens of popular profiles for talk show host Stephen Colbert, all but one of which are fake. Our social model attempts to mitigate these problems by weighting a diverse range of *positive credibility indicators* within a target topic.

We first consider the "retweet" as an indication of credibility. Equation 1 gives a value for credibility based on the deviation of a user $u \in U$'s retweet rate RT_u from the average retweet rate $\overline{RT_x}$ in a target topic $x \in X$. In practise, values from the following equations are mapped to a log-log scale to handle large outliers in the data. Notation has been left out for simplicity.

$$Cred_{RT}(u, x) = |RT_u - \overline{RT_x}| \quad (1)$$

Keeping with retweet analysis, Equation 2 considers retweet rate but factors in usage rate and number of followers F_o, in other words, a utility metric from the potential number of retweets.

$$Utility_{RT}(u, x) = |\frac{RT_{u,x} \times F_o(u)}{t_{u,x}} - \frac{\overline{RT_x} \times \overline{F_{o,x}}}{t_x}. \quad (2)$$

Retweet metrics function over both the content of a collection of tweets and the underlying network. We believe that the network topology itself can also provide insights into credibility of a user. Equation 3 computes a social credibility score as

[2] http://gawker.com/5826645

[3] http://news.cnet.com/8301-17939_109-10218926-2.html

the deviation in the number of user u's followers from the mean number of followers in the domain, again normalized by number of tweets.

$$Cred_{social}(u) = |\frac{F_o(u)}{t_u} - \frac{\overline{F_o}}{t}| \quad (3)$$

Assuming that a "follow" request is usually an indication of credibility, we can now also weight Equation 3 by factoring in the ratio of friends to followers as a deviation from the norm for a given topic. For example, an information gathering agent for a direct marketing company is likely to follow many profiles, but have few followers. Equation 4 describes the social balance of a user u as the ratio of follower (F_o) to following (F_e) group size.

$$Balance_{social}(u) = |\frac{F_o(u)}{F_e(u)} - \frac{\overline{F_o}}{\overline{F_e}}| \quad (4)$$

There are cases where the opposite is true however, for example, a popularity-hungry politician may pay to have automated agents create accounts and follow his profile, but these profiles are not likely to have strong social connectivity, and can be discounted by other filters in this model, such as Equation 2 for example.

We also consider social connections within a given topic as a positive indication of credibility, both in the F_o and F_e groups. Consider a user who has tweeted frequently about a topic, lets say "#IUI2012". If that user has few or no followers with associations to that topic, this should raise suspicion about that user's credibility in the topic. Our findings indicate that network data is frequently too sparse within a specific topic for this metric to yield useful results, but we include it in the model because it leverages social connections in a potentially useful way.

$$Cred_{social}(u, x) = |\frac{F_o(u, x)}{t_{u,x}} - \frac{\overline{F_{o,x}}}{t_x}| \quad (5)$$

The final metric in our social credibility model addresses the focus of a target user within a given topic space as a function of their global profile. For example, many people have set up Twitter accounts solely for business or research purposes, and thereby have a more constrained number of topics that they tweet about, potentially indicating an increased level of credibility, since the likelihood of recurring topics is higher. A practical nugget being that the authors of this paper have tweeted about "#IUI2012" many times, and it appears as a peak in their personalized topic-distribution graphs. Equation 7 computes this metric as the sum of the tweets for a user u on topic x as a percentage of their total number of tweets t_u.

$$Focus(u, x) = |\frac{\sum_{t \in T} t_{u,x}}{\sum_{t \in T} t_u}| \quad (6)$$

Weighting Scheme

Now that we have described an array of potential credibility indicators in the microblog domain, we incorporate them into a single predictive mechanism that can be used to make inferences about credible sources. There are a variety of utility functions that can be applied to train the weighting scheme, for example, prediction of manually labeled "ground-truth" data, prediction of empirical "retweet" data. We will discuss these in detail in our evaluation section. For the purpose of the discussion here, we provide the simple weighted combination below:

$$C_u = \alpha(Focus(u, x) + \beta(Balance(u) \times Cred_{social}(u)) + \gamma(Utility_{RT}(u, x) \times Cred_{RT}(u, x)) \quad (7)$$

Content-based Model

We have described how the social provenance of a piece of information can have a bearing on its credibility. However, credibility can be assigned both to the information source, and to the information itself in an intrinsic way. Accordingly, our second credibility model focuses in on tweet content, isolated from the underlying social network. Following this discussion, we will describe integration strategies that harness strengths of both approaches.

We begin by representing all tweets in our topic-specific data sets as a set of salient credibility indicators (12 numeric and 7 binary). Approximately half of the these features are taken from a 2011 study by Castillo et al. [4]. However, their work defined a much larger set that incorporated features for Tweet, Topic, User and Propagation scopes, in a multiple topic setting. In this model, we are interested in tweet content only, for a single topic, and hence other features are not included.

Numeric Indicators:

1. *Positive Sentiment Factor*: Number of positive words (matching our lexicon)
2. *Negative Sentiment Factor*: Number of negative words
3. *Sentiment Polarity*: Sum of sentiment words with intensifier weighting (x2) ('very', 'extremely' etc)
4. *Number of intensifiers*: 'very', 'extremely' etc., based on our lexicon.
5. *Number of swearwords*: Simple count, based on lexicon.
6. *Number of popular topic-specifc terms*: Simple count, based on lexicon.
7. *Number of Uppercase Chars*: Simple Count
8. *Number of Urls*: Simple Count
9. *Number of Topics*: Number of topics '#' (All have at least 1)
10. *Number of Mentions*: Number of user's mentioned with '@'
11. *Length of Tweet (Chars)*: simple count.
12. *Length of Tweet (Words)*: simple count.

Binary Indicators:

1. *Is Only Urls*: No text, only links.
2. *Is a Retweet*: From metadata

3. *Has a Question Mark*: '?' or contains any of Who/What/Where/Why/When/How

4. *Has an Exclamation Mark*: '!'

5. *Has multiple Questions/Exclamations*: '??' '???' '!!' '!!!' etc.

6. *Has a positive emoticon*: :) :-) ;-) ;)

7. *Has a negative emoticon*: :(:-(;-(;(

To evaluate the utility of this model for predicting credible/useful information, we train a range of classifiers using 5000 manually annotated tweets from a user evaluation. Details and results of this analysis are presented in the evaluation section.

Hybrid Model

So far we have focused our discussion on credibility indicators at the user level and the tweet level individually. A logical progression is to combine aspects from both methods to maximize the information upon which we can base credibility decisions. Both models contain many variables, making it infeasible to conduct an exhaustive analysis of possible hybrid strategies in this paper. As a representative sample, we now present four novel methods for combining facets from the earlier models that aim to improve their ability to predict credible information and credible information sources. Since our earlier models compute credibility at different levels of granularity (user and information level), so also do the following hybrid strategies.

Content-based Ranking

This strategy predicts credibility at both the user and tweet levels. The hybrid algorithm first performs a filtering step based on the user level (social) credibility score from model 1, passing profiles with a credibility score above a threshold S_{min} to the second model. The content based model extract features from each tweet and computes a credibility score which is used to re-rank tweets from the set of credible users. $u \in U \; where \; S_u < S_{min}$.

Weighted Combination

This simple combination of output from the two earlier models predicts at the user level only. Credibility scores from the content-based model are aggregated over each $u \in U_t$. The resultant user-level score C_u is combined with the social credibility S_u using a harmonic mean weighting strategy to minimize outlier values: $C_{weighted} = \frac{2}{C_u + S_u}$.

Feature Combination

This strategy computes a credibility at the tweet level, and is designed to use all available data to generate a prediction. Feature lists from both the social model, the content based model, and a collection of other user metadata from the Twitter API are used to train a J-48 decision tree to generate a prediction model.

Content-boosted Social Credibility

The final hybrid method predicts at the user level, incorporating the aggregated content-based score for a user into the social model. This approach is similar to the Weighted Combination with the exception that the content-based credibility factor is considered at the same level as the $Cred_{RT}$ and $Cred_{social}$ scores from Equations 1 and 2 respectively.

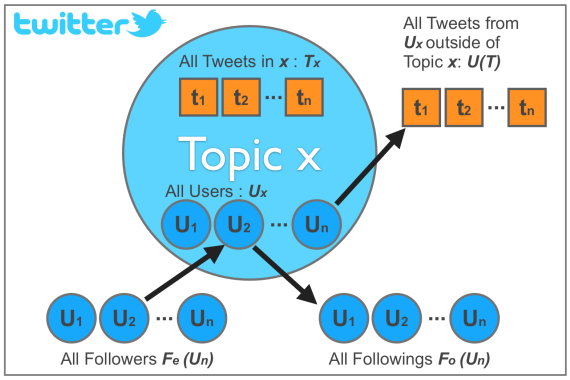


Figure 1. Illustration of the scope of crawled data for each of 7 current topics.

Set Name	*Core Tweeters*	*Core Tweets*	F_o *and* F_e (overlapped)	F_o *and* F_e (distinct)
Libya	37K	126K	94M	28M
Facebook	433K	217K	62M	37M
Obama	162K	358K	24M	5M
Japanquake	67K	131K	25M	4M
LondonRiots	26K	52K	30M	4M
Hurricane	32K	116K	35M	5M
Egypt	49K	217K	73M	36M

Table 1. Overview of 7 topic-specific data collections mined from the Twitter streaming API.

EVALUATION SETUP

Ultimately, we would like to incorporate the models presented in this paper into a real world system for recommending credible information and information sources from a microblog to collect rich preference feedback. Meanwhile, we evaluate our methods using 7 topic-specific data sets from Twitter. Our evaluation is organized as follows: Firstly we describe the data collection process and provide an overview of each topic-specific collection. Next, a brief statistical analysis of the data is presented to highlight core trends across each set, with specific focus on our larger data set on the topic "#Libya". Following this, we describe an online user study of 150 participants wherein ground truth credibility assessments were collected on the Libya data set. Lastly, we describe a set of predictive accuracy experiments performed for each of the three models and provide a detailed discussion and comparative analysis of the results.

Data Gathering and Analysis

We ran a python-based crawler[4] for 8 weeks from a cluster of 12 machines using 14 different Twitter authentications. Data was crawled from the Twitter streaming API[5] and stored in a

[4]The authors have made this crawler and data sets publicly available at www.wigis.net/twitterdata

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

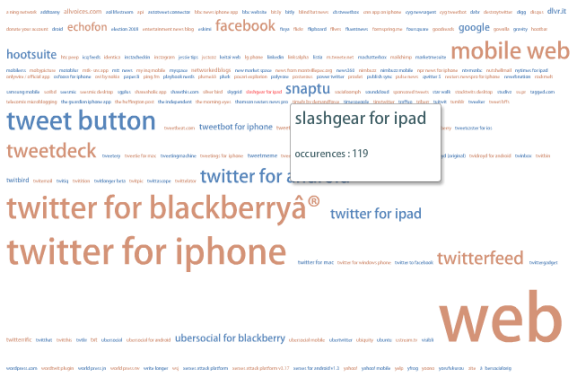


Figure 2. Word cloud showing origin of tweets in the Libya data set.

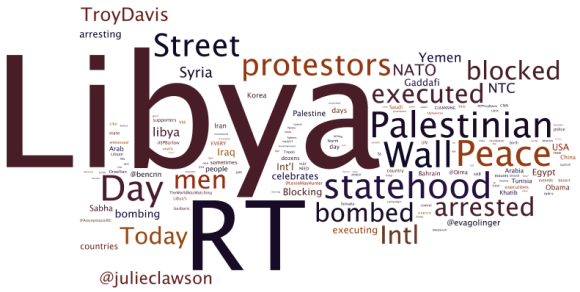


Figure 3. Word cloud showing distribution of popular terms in the Libya data set.

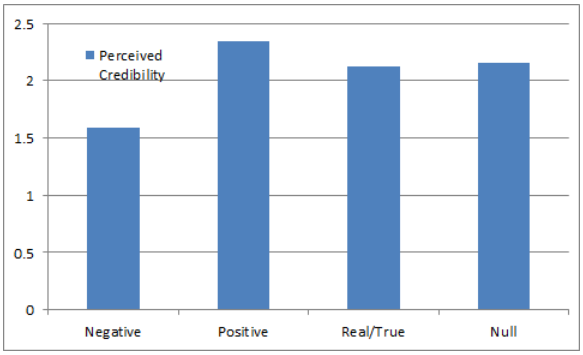


Figure 4. Plot showing perceived credibility in each of four Twitter contexts (Negative, Positive, True, Null).

relational database. Our choice of topics was guided by two factors: firstly we required popular topics that would create a flurry of activity in the network, yet we were also interested in topics which would yield significant interconnections in the underlying social networks of the users within the topics. Figure 1 is an illustration of the crawling algorithm. Starting at a topic keyword such as '#Libya' for instance, all tweets containing the tag were collected. Additionally, all users who tweeted with this tag were crawled and their tweets (with or without) the topic tag were stored, along with all available metadata from the API. For each of these "core users" information on their Follower and Following groups was also collected. We did not store the tweets of these users, only metadata such as number of following and followers, number of retweets etc. The main challenge in this process was rate limitation and 403-503 range error codes returned by the API, giving an approximate download rate of 350 queries per hour with approximately 200 tweets per query. As shown in Table 1, the second order social groups are exponentially larger, ranging in size from 25 to 94 million profiles across the 7 collections. Core users collections ranged from 26k to 433k. A small world effect is evident in these data sets, since there is very significant overlap in the F_o and F_e groups, as shown by the last two columns in Table 1. A visual overview of originating devices for the tweets in or Libya data set is given in Figure 2, and distribution of popular terms is shown in Figure 3.

User Study: Credibility and Context

An online study was set up for two purposes: Firstly to collect ground truth credibility assessments of the crawled tweets from real users, for the purpose of evaluating our prediction models. However, we were also interested in analyzing the effects of Twitter context on perceived credibility, so we performed a within subjects study, varying source context (dependent variable), with the goal of examining the effect on perceived credibility rating (independent variable). 145 participants took the online study, which lasted about 10 minutes. Participants were 39% female, 61% male, varying in age from 19 to 56 with an average age of 26 (median 28). Participants were generally familiar with Twitter (4 out of 5 rating on average). In total, credibility ratings on 5025 tweets were collected, excluding those that were rated as "can not answer". Participants were all asked to leave comments and insights in a feedback form. Comments varied widely, but there was a consensus that information provenance (i.e: links to other sources) produced a sense of credibility. One participant reported the opposite: "that social network activity should have no bearing on credibility". Each participant was shown four groups of 10 tweets and asked to provide a credibility rating for each tweet on a scale of 1 to 5, with an option to click a "can't answer" response. Following from the analysis in [4], the general statement "likely to be credible" was removed. Group 1 contained just the raw tweet, with no context information about the tweeter. Group 2 contained statistically poor credibility features for the tweeter (few followers, few retweets), while Group 3 contained statistically "good" context (many followers, many retweets). The final group was shown the true context, i.e: the user's real number of retweets and followers. It is important to note that only ratings from the "true" context were used in the evaluation of our prediction models. Figure 4 shows the average perceived credibility rating for each group. A one-way ANOVA showed a significant difference between each the groups ($p = 0.039$). Results indicate that context does have an effect on perceived credibility, with the positive context achieving an average rating of 2.34, and the negative context a score of 1.58. This reveals a relative rating shift of 0.76 or 26% between the extreme contexts.

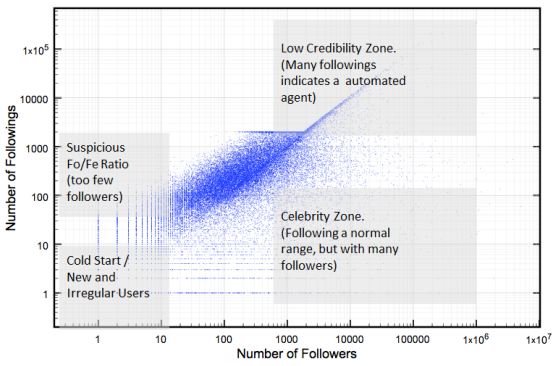
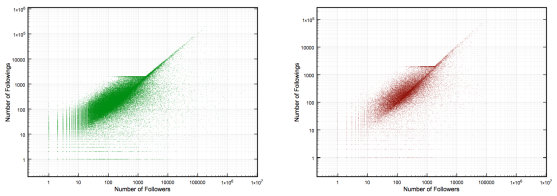


Figure 5. Plot showing number of followers to number of following profiles for the Libya data set. Areas of particular interested are shaded in grey and labeled accordingly.

EVALUATION

Now that we have presented our models for predicting credibility, and our ground truth collection process, we must assess and compare performance of each model. Given our available resources, substantial credibility assessment data could only be collected on the Libya data set, so we focus on that set for most of the following evaluation, namely, where comparisons against user provided assessments are performed.



(a) Japan Quake (b) Hurricane Irene

(c) Enough is Enough (d) Facebook

Figure 6. Friend to Follower patterns across four of our topic-specific data sets. All other sets exhibited a similar distribution.

Data Analysis

Before we describe our evaluation of predictive accuracy, we first take a broader statistical view of the collected data sets to gain insights about trends, clusters and any interesting anomalies about the data sets. A detailed comparison and discussion of every feature in our models is not feasible here, so we focus in on a selection of features that influence the models most (F_o, F_e, links and retweets), based on a best-first feature analysis performed using WEKA[6]. Figure 8 shows a comparative analysis of a selection of features from both social and content-based models. In this figure, lighter (red) nodes indicate positive credibility and darker (blue) nodes indicate negative credibility assessments from the user study. Clusters appear in some of the scatter plots, indicating that the feature does have some bearing on assessed credibility. For example, looking at the features for "char" and "word", it is clear that longer tweets tend to be assigned more credibility than shorter ones. Number of tweets (status-count) and number of listings (listed-count) also align well with reported credibility.

Friend to Follower Analysis

Figure 5 shows a comparison of the number of followers F_o to the number of following users F_e over the 37,000 core users in our Libya data set. The shaded areas in the resulting distribution reveal areas that have a potentially negative impact on credibility. For example, at a threshold where users are following more than approximately 5000 others, the data appears to form a straight line. However, on the log-log scale, this is evidence of the long tail of a power law distribution. This is marked as low credibility zone since the group size is abnormally high for a human user, and we must therefore assume that the profiles are based on automated agents or "bots". Vice-versa, small sized F_o and F_e groups indicate new or irregular users. This is a low credibility group because we do not have sufficient social information (and by correlation, content information) to perform a reasonable assessment of credibility. This is analogous to the cold start problem in the recommender system literature [21]. Groups along the other extremities of this graph are also interesting: those with very few followers but larger following groups are (again by correlation) likely to tweet less and be leaf nodes in retweet chains, while conversely, the "celebrity" group (high F_o and low F_e tend to be higher in retweet chains, and have many retweets. The latter two groupings do not necessarily bear on credibility, but the other shaded areas of Figure 5 do indicate negative credibility. Accordingly, the "balance" component of our social credibility model, shown in Equation 4 is weighted to penalize these groups. Figure 6 shows similar distributions of follower to following groups across the JapanQuake (67k users), Hurricane Irene(32k users), Enough is Enough(65k users), and Facebook (433k users) topics. We found similar distributions for all of the other sets in Table 1.

Followers and Credibilty

Figure 7 shows an analysis of the average credibility reported for tweets in the user study compared with number of followers. Binning of followers was applied to highlight the distribution. There is a significant correlation between reported credibility and number of followers up to a network size of approximately 1500 followers, after which, reported credibility drops off steeply. This result aligns well with our earlier analysis of follower to following groups. The reported drop in credibility past this threshold is likely due to the "bot" effect shown in Figures 5 and 6.

[6]www.cs.waikato.ac.nz/ml/weka/

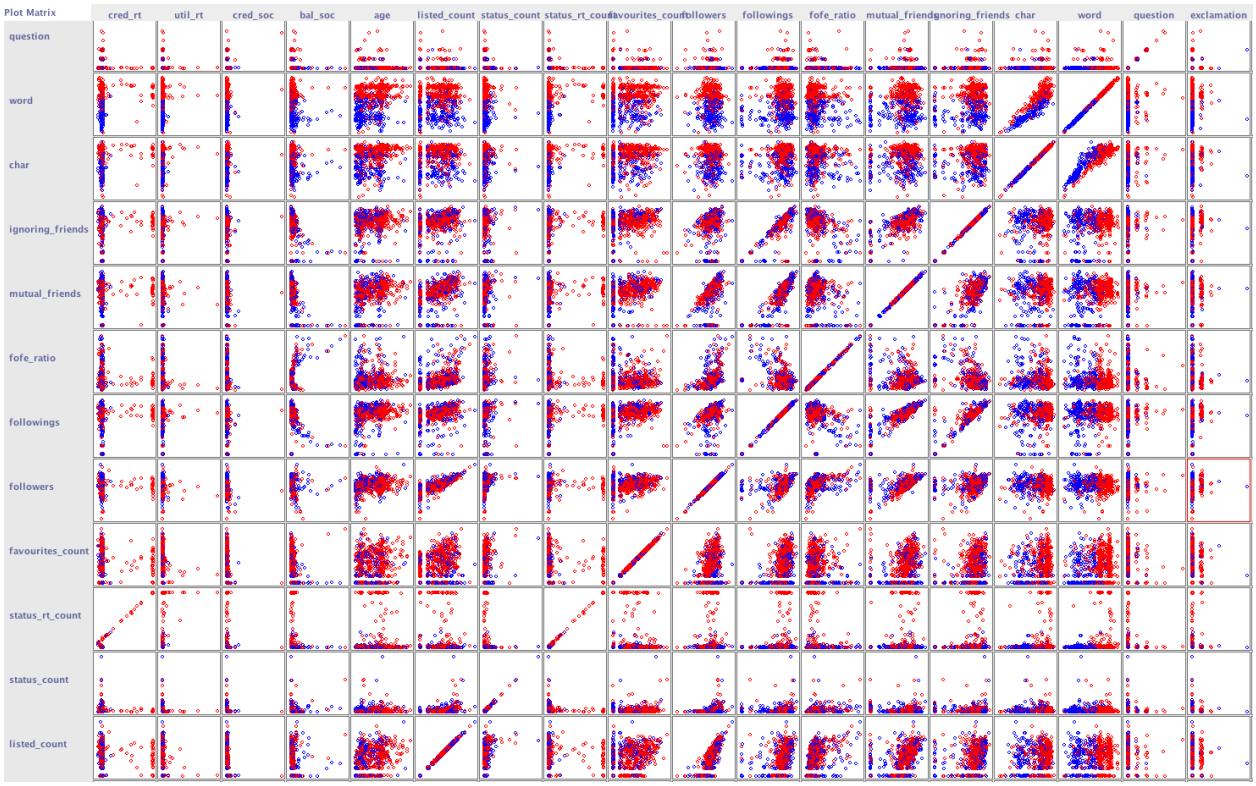


Figure 8. Comparisons of each feature used in computing the social credibility model. In this plot, lighter (red) areas indicate high credibility and darker (blue) areas indicate low credibility assesments.

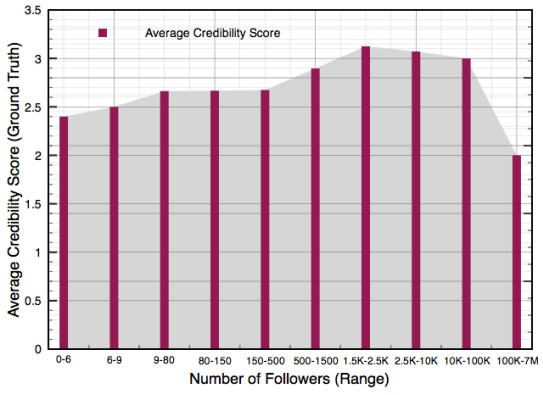


Figure 7. Average credibility rating from the web survey versus number of followers for the tweet authors (binned).

Retweets, Links and Other Credibility Indicators

The hybrid approaches proposed in this paper rely on a variety of different features from the social network, tweets and from user metadata. Once again a discussion of the benefits and merits of all features tested is not possible here, as a sample, we now discuss interesting findings from our analysis. The distribution graph in Figure 8 shows an overview of a subset of features. Links (presence of urls in tweets) were found to be a very positive of credibility and were more frequently used in older profiles. Users who provided links frequently tended to be listed more often, and added to other users' "favorite" groups. Presence of links in tweets was found to correspond to sentiment (presence of positive or negative sentiment keywords) in an interesting way: if the sentiment metric was polarized, either negatively or positively, links were far more common, and users tended to assess credibility more highly. Tweets containing links were also found to be retweeted more frequently. Retweets were generally reported as credible in our study, and were also more frequent in older user profiles. Interestingly, emoticons (both positive and negative) were found to be an indicator of retweeting. Additionally, longer tweets were retweeted more frequently than shorter ones.

Predicting Credibility

Treating each hybrid strategy independently, a total of 6 credibility prediction strategies were tested. Each strategy was

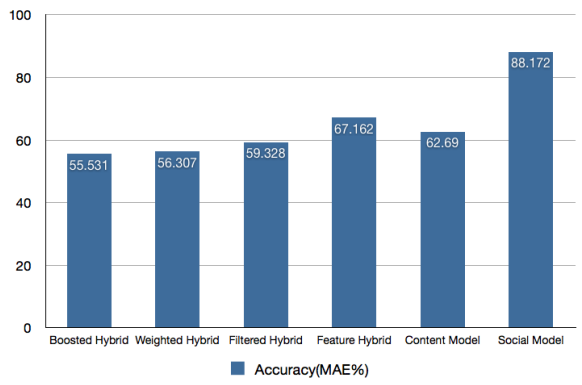


Figure 9. Comparison of predictive accuracy over the manually collected credibility ratings for each model.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         902               88.172  %
Incorrectly Classified Instances       121               11.828  %
Kappa statistic                          0.8222
Mean absolute error                      0.0771
Root mean squared error                  0.2162
Relative absolute error                 22.8584 %
Root relative squared error             52.6661 %
Total Number of Instances             1023

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.901     0.009      0.966     0.901     0.932      0.966    1.0
                 0.946     0.061      0.924     0.946     0.935      0.965    2.0
                 0.896     0.083      0.816     0.896     0.854      0.936    4.0
                 0.214     0.02       0.387     0.214     0.276      0.876    5.0
Weighted Avg.    0.882     0.054      0.872     0.882     0.875      0.952

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 200  21   1   0 |   a = 1.0
   5 424  17   2 |   b = 2.0
   2  12 266  17 |   c = 4.0
   0   2  42  12 |   d = 5.0
```

Figure 10. Statistical results from a J48 tree learner for the best performing credibility prediction strategy (Social Model)

represented as a set of weighted features and loaded as an input file to WEKA machine learning toolkit. Our goal was to accurately predict the user-provided credibility scores from the online study. Preliminary experiments were performed using Bayesian classifiers (and a range of others) to learn a model based on the features of each prediction strategy. For the full experiment a J48 tree-based learning algorithm was used, firstly since it performed well in preliminary tests, and secondly to allow for comparison of results with Castillo et. al's similar evaluation in [4]. Predictions were run on a training set of 591 tweets with annotated credibility scores. A 10-fold cross-validation was applied, and training sets were separate from test sets. The algorithm attempted to classify each test instance into one of two credibility classes. To clarify, we note that all predictions were made at the tweet level, that is, if a strategy (such as the standalone social model) predicts credibility at the user level, the evaluation applied this approach to predicting credibility of a tweet by that user. Class instances were evenly distributed in the training sets. For each strategy, the mean absolute error between the predicted rating and the user provided rating was recorded.

Figure 9 shows the results of this evaluation for each strategy. The content-based and hybrid models performed reasonably at the prediction task, but were far outperformed by the social model, which achieved an accuracy of 88.17%, an improvement of 11% over the feature hybrid which was the next best performer (statistical analysis shown in Figure 10. The content-based approach scored an accuracy of 63% while the hybrid approaches ranged from 56% to 67%. Our initial expectations were that the simple rule of "more features, better prediction" would apply across this study, but in this case our findings have indicated otherwise, since the social model outperform the hybrid and content-based methods significantly. The relatively poor performance of the content based model (67%) can perhaps be attributed to the fact that tweet text is short and does not always contain sufficient information to make a credibility judgement. The feature-hybrid method exhibited a small improvement in accuracy ($\tilde{1}0\%$) over the next best hybrid strategy, which was the filtered approach. An overview of the statistical output from the J48 learner process is provided in Figure 10 for our best performing method, showing a correct classification of 902 of the 1023 instances, yielding 88.172% accuracy. The content-based (and therfore, hybrid) approaches rely on tweet text, whereas our social model relies on rich interconnections in the twitter network, including dynamic information flow metrics (retweets). Our findings indicate that the underlying network and dynamics of infomation flow are better indicators of credibility than text content.

Castillo et. al's examination of credibility in Twitter produced similar accuracy scores to the above (8% less accurate than our best performing social model result), with "precision and recall in the range of 70-80%". Several key differences make it infeasable to perform a fair comparison of classification accuracy however. For example, [4] analyses groups of "newsworthy" tweets, whereas our analysis focuses on "credible" individual tweets or users, as per our earlier definition. Furthermore, our analysis are focused in a topic-specific domain consisting of a different set of users and tweets.

DISCUSSION AND FUTURE WORK

The study of credibility models presented here is my no means exhaustive, and we believe that there are still better prediction strategies to be found. It appears from our evaluations that while stand alone social or content-based approaches fair reasonably well at predicting user provided credibility ratings, they are outperformed by hybrid methods which combine features from both, ultimately basing credibility assumptions on a larger window of information. Our evaluation answers the research question posed in the introduction, that credible tweets can be automatically detected with high accuracy (88% for our social model). Accurate detection of credible information in Twitter has many practical implications. For example, automatic filtering/ranking of twitter feeds based on credibility, spam detection, automatic recommendation of credible infomration [18] and identification of key players in infomration dissemination, which can be useful in assessing/predicting situations of social unrest such as the "occupy" movements and the London riots for instance.

There are a significant number of possible next-steps for this research. We believe that improvements could be made by

incorporating interpersonal factors such as the credibility that exists between users, also known as a trust relation. Such metrics can be incorporated both at the positive and negative levels (distrust), and have been shown to be useful for finding credible information in microblog domain. Furthermore, we are interested in evaluating the mechanisms presented here in a real world system, to elicit significant user feedback on the credibility of information in a real-world information consumer context, as opposed to the simple user survey approach presented here. This includes considering the role of interfaces and interactions that communicate credibility to, and elicit credibility data from real users.

CONCLUSION

As with most interactions on the Social Web, the window of information upon which we can make credibility judgement on Twitter is limited. As this forum becomes more popular, it becomes increasingly important to investigate new models for assessing credibility of the information it distributes. This paper presented three computational models for assessing such credibility, using social, content-based and hybrid strategies. The models were evaluated on 6 collections of tweets about current topics, including the associated social network information for each tweeter, as provided by the Twitter streaming API. An automated analysis of the predictive ability of each model was performed, predicting on both empirical "retweet" data, and on a collection of manually assessed tweets collected in an online user survey. Results showed that the social model outperformed both content-based and hybrid models, achieving a predictive accuracy of 88.17%, compared with 62% and 69% for content-based and the next best performing hybrid (weighted strategy) respectively.

ACKNOWLEDGEMENTS

The authors would like to thank Cha Lee and Sibel Adali for their input on analysis methods and credibility indicators for microblogs, respectively. This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; by NSF grant IIS-1058132; by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number FA8650-10-C-7060; and by the U.S. Army Research Laboratory under MURI grant No. W911NF-09-1-0553; 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 ARL, AFRL, IARPA, 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. Adali, S., Escriva, R., Goldberg, M. K., Hayvanovych, M., Magdon-Ismail, M., Szymanski, B. K., Wallace, W. A., and Williams, G. T. Measuring behavioral trust in social networks. In *ISI* (2010), 150–152.
2. Burke, R., Mobasher, B., Zabicki, R., and Bhaumik, R. Identifying attack models for secure recommendation. In *Beyond Personalisation Workshop at the International Conferece on Intelligent User Interfaces*, ACM Press (San Deigo, USA., 2005), 347–361.
3. Canini, K. R., Suh, B., and Pirolli, P. L. 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).
4. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *WWW* (2011), 675–684.
5. DuBois, T., Golbeck, J., and Srinivasan, A. Predicting trust and distrust in social networks. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)* (2011).
6. Golbeck, J. *Computing with Social Trust*. Springer Publishing Company, Incorporated, 2010.
7. Golbeck, J., and Hansen, D. L. Computing political preference among twitter followers. In *CHI* (2011), 1105–1108.
8. Haifeng Zhao, William Kallander, T. G. F. W. Read what you trust: An open wiki model enhanced by social context. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)* (2011).
9. Herlocker, J. L., Konstan, J. A., and Riedl, J. Explaining collaborative filtering recommendations. In *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work* (2000), 241–250.
10. Houser, D., and Wooders, J. Reputation in auctions: Theory, and evidence from ebay. *Journal of Economics and Management Strategy 15*, 2 (2006), 353–369.
11. Jennifer Golbeck, Cristina Robles, M. E. K. T. Predicting personality from twitter. In *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)* (2011).
12. Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, ACM (New York, NY, USA, 2010), 591–600.
13. McNally, K., O'Mahony, M. P., Smyth, B., Coyle, M., and Briggs, P. Towards a reputation-based model of social web search. In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, ACM (New York, NY, USA, 2010), 179–188.
14. Melville, P., Mooney, R., and Nagarajan, R. Content-boosted collaborative filtering. In *In Proceedings of the Eighteenth National Conference on Artificial Intelligence* (2002).
15. Mendoza, M., Poblete, B., and Castillo, C. Twitter Under Crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA '10)*, ACM Press (July 2010).
16. Milgram, S. The small world problem. *Psychology Today 1* (May 1967), 61–67.
17. Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Inter. Tech. 7*, 4 (2007), 23.
18. O'Donovan, J., and Smyth, B. Trust in recommender systems. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, ACM Press (2005), 167–174.
19. O'Donovan, J., Smyth, B., Evrim, V., and McLeod, D. Extracting and visualizing trust relationships from online auction feedback comments. In *IJCAI* (2007), 2826–2831.
20. Pool, D., and Kochen, M. Contacts and influence. *Social Networks 1*, 1 (1978), 5–51.
21. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work* (1994), 175–186.
22. Resnick, P., and Zeckhauser, R. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-Commerce. Volume 11 of Advances in Applied Microeconomics.* (December 2002).
23. Suzuki, Y. A credibility assessment for message streams on microblogs. In *3PGCIC* (2010), 527–530.
24. Victor, P., Cornelis, C., Cock, M. D., and Herrera-Viedma, E. Practical aggregation operators for gradual trust and distrust. *Fuzzy Sets and Systems 184*, 1 (2011), 126–147.