Credibility in Context: An Analysis of Feature Distributions in Twitter

John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer
University of California
Santa Barbara, USA
Email: (jod, kang, gmeyer, holl)@cs.ucsb.edu

Sibel Adalı
Rensselaer Polytechnic Institute
Troy, New York, USA
Email: sibel@cs.rpi.edu

***Abstract*—Twitter is a major forum for rapid dissemination of user-provided content in real time. As such, a large proportion of the information it contains is not particularly relevant to many users and in fact is perceived as unwanted 'noise' by many. There has been increased research interest in predicting whether tweets are relevant, newsworthy or credible, using a variety of models and methods. In this paper, we focus on an analysis that highlights the utility of the individual features in Twitter such as hashtags, retweets and mentions for predicting credibility. We first describe a context-based evaluation of the utility of a set of features for predicting manually provided credibility assessments on a corpus of microblog tweets. This is followed by an evaluation of the distribution/presence of each feature across 8 diverse crawls of tweet data. Last, an analysis of feature distribution across dyadic pairs of tweets and retweet chains of various lengths is described. Our results show that the best indicators of credibility include URLs, mentions, retweets and tweet length and that features occur more prominently in data describing emergency and unrest situations.**

I. Introduction

Studying of information credibility on Twitter has become a popular topic. Can one determine whether information found on Twitter is credible or not? Which specific features are most relevant to identifying credibility of content [1]. In parallel with these studies are those that try to determine whether features obtained from Twitter activity can be used to determine who the experts are [2], which party will win the election [3], [4] or who the influential people are [4], [2]. Some of this work is now coming under some scrutiny [1] as to which degree they can be used to make generalized conclusions about Twitter vs. the specific data set that they are derived from.

When studying credibility for example, it is important to consider both the type of data that was used in analysis as well as the methods used to generate ground truth. For example, a post-hoc analysis may not reveal the true credibility of the data at the time it was processed. Another debatable aspect of such studies is the specific features chosen for study. Why are the studied features the correct features? The types of features available for study are large and they can be studied in various ways: linear regression of individual features vs network aggregation algorithms. More importantly, different features may end up being important in specific contexts. Given the question: "which features are best for determining content credibility", the most appropriate answer might be: "it depends". Overall, we believe that instead of trying to determine which features determine credibility, studies must concentrate on specific circumstances that determine when specific features are useful in determining credibility. Our paper is a first step towards in studying this problem.

To motivate this problem, let us consider the problem of information credibility. Suppose a Twitter study show that four most important factors that determine information credibility are: presence of URLs, the number of days the poster has been on Twitter, the number of followers she has and the presence of sentiments expressed. This raises a number of questions. Suppose we would like to study credibility in a specific scenario, i.e. an earthquake in Turkey. The user population in our study may be very different than those present in this scenario. Some features may be culturally dependent: in this case, everyone may have emotional responses regardless of whether the content is credible or not. The presence of URLs may not be helpful at all as the information is being generated very quickly to even post to a site. Furthermore, posters in the field may not be long time Twitter users or those with followers, just people trying to get the word out. In fact, most long term power users may be far removed from the actual event and the word becomes heresay by the time it reaches them. Overall, these features may not be useful at all to determine credibility in this scenario. This does not mean that the study was not useful for understanding credibility in general, but it does not provide a mechanism for customizing these features to a specific scenario.

As a first step towards solving this problem, we would like to propose a method to study credibility related features along a number of dimensions. The main focus of this initial study is to understand how the features in our study are distributed across a range of contexts, and whether the can or should be grouped into a single category. This will subsequently enable us to understand which features tend to provide similar or

Class	*Description*	*# of Contexts*
Diverse Topics	Diverse topics in Twiter; eg: #Romney #Facebook	8 different topics (see Table II)
Credibility	Manually provided assessments of tweets	Credible or non credible
Chain length	Mined retweet chains and classified based on length	Long or short
Dyadic pairs	Mined interpersonal interaction and classified	Dyadic or not dyadic

TABLE I
CONTEXT CLASSES IN OUR EVALUATION.

Set Name	*Core Tweeters*	*Core Tweets*	F_o *and* F_e (overlapped)	F_o *and* F_e (distinct)
Libya	37K	126K	94M	28M
Superbowl	191K	227K	N/A	N/A
Romney	226K	705K	N/A	N/A
Facebook	433K	217K	62M	37M
EnoughIsEnough	85K	129K	13M	4M
Egypt	49K	217K	73M	36M
Earthquake	67K	131K	15M	5M

TABLE II
OVERVIEW OF 7 TOPIC-SPECIFIC DATA COLLECTIONS MINED FROM THE TWITTER STREAMING API.

Set Name	*Tweets*
Libya	#Libya: Muammar #Gaddafi's base taken http://t.co/UKvSn7Jk #drumit
Superbowl	RT @mashable: The Giants may have won the #SuperBowl, but Madonna won the Google search competition - http://t.co/YRqErdkg
Romney	#Romney outlines economic plan - cut taxes across the board to boost economy, but won't add to the deficit. #GOP2012 http://t.co/AjbDWLKy
Love	You deserve better friends #love you.
Facebook	What pisses you off more: #Facebook changing every 2 minutes, or people complaining about #Facebook changing every two minutes?
Enoughisenough	i Mean come on Now #EnoughIsEnough
Egypt	#Egypt #July8 Egypt Mubarak-era minister jailed for corruption Albany Times Union http://t.co/5ucF7kDA #Feb17
Earthquake	A light intensity #earthquake, of magnitude 4.3 on the Richter Scale, occurred here at 8.51 p.m.

TABLE III
EXAMPLES OF 8 TOPIC-SPECIFIC TWEETS FROM THE CRAWLED SETS.

distinct types of information. We can use these categories to simplify the feature space and study the contribution of each category to the ground truth on credibility. Gaining a better understanding of feature usage and distibrutions can aid in the design of future models that attempt to predict elements of credibility. In this work we will also segment our data into disjoint groups that change important feature categories (for example, retweet chains, credibility assessments and dyadic interactions) significantly. In other words, we would like to find groups that behave very differently than the norm in Twitter. What are these groups? What determines them? We check whether these subgroups exhibit very different behavior in terms of feature distribution and why. By understanding which factors can impact credibility models the most, we hope to inspire a new set of studies that target closer study of these factors.

II. DATA GATHERING

The goal of this research is an examination of both the distribution and predictive utility of features across varying contexts in Twitter. Accordingly, we must first gather a diverse set of data from which to form these "contexts". Of course, there are infinite choices for context, so for our analyses we focus on a carefully chosen subset, which is described in TableI. In this section we first describe a collection of 8 corpuses spanning very diverse "topic contexts" on Twitter. This is followed by a discussion of a study to collect human assessments of credibility of individual tweets, which forms the basis of a "credibility context". Next, we discuss a path-based analysis of feature distribution based on information flow through retweeting. The last context in Table I discusses a comparison of features that exist in dyadic pairs against an intrinsic set of tweets.

A. Topic-Specific Collections

To gather a diverse sample of user and tweet data, a crawler was initiated for the topics in Table II. The table describes the properties of each collected set, including details on Friend to Follower ratios for some cases. Table III provides a set of sample tweets from each topic as an illustration of content diversity. Since Twitter has restricted API calls, and a complex network structure, a tailored algorithm was developed to traverse the space of users and tweets. Figure 1 provides a pseudocode description of the crawling process, which centered around a target topic tag, but also incorporates their other tweets, and information about their friends and followers.

B. Credibility Annotations

To create the classes in our "credibility" context, human-provided assessments on groups of tweets were sourced from an online evaluation. Tweet data was presented to Amazon Mechanical Turk users who were paid a nominal amount for their participation. In total 236 participants took part, and were asked to rate tweets on a Likert scale of 1-5 indicating their impression of the credibility of the tweet content. Participants also had an option to select "can't" answer. Overall participants were 39% female and 61% male, varying in age from 19 to 56 (median 28). Each participants ability to rate was tested using a 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. Some post-hoc information was collected from participants after the study. For example, participants were generally familiar with the Twitter domain (4/5) rating on average. In total, 6369 individual credibility assessments were collected, predominantly on the Libya topic collection from Table II. Participants were encouraged to

```
1: procedure CRAWL(topicsList, tweetsList)
2:     store = ∅
3:     for all topic ∈ topicsList do
4:         store ← topic
5:         topicTag = topic.getTopicTag()
6:         for all tweets ∈ tweetsList do
7:             if tweetsContains(topicTag) then
8:                 store ← getRelevantTweets()
9:             end if
10:        end for
11:        for all tweets ∈ store.getTweets() do
12:            store ← getUsers()
13:        end for
14:        for all users ∈ store.getUsers() do
15:            store ← getTweets()
16:        end for
17:        for all users ∈ store.getUsers() do
18:            store ← getFollowers()
19:            store ← getFollowing()
20:        end for
21:    end for
22: end procedure
```

Fig. 1. Data crawling procedure used for gathering 8 datasets from the twitter social graph, consisting of over 20 million users and 200 million tweets in total.

provide feedback comments. From analysis of the comments it was evident that the presence of provenance features such as URLs provided a sense of credibility. However, not all users agreed with this sentiment, and one user commented that "social network activity should have no bearing on credibility". Overall the two phases of the study (one in late 2011, and one in early 2012) took less than 24 hours to reach the target number of participants using Amazon Turk, which was much faster than previous similar studies performed by the authors in the past.

C. Retweet Chains

To analyze the influence of retweet behavior on feature distribution, the Libya dataset from Table II was crawled for all retweet chains. Twitter supports two mechanisms for identifying retweets: the "@rt" keyword in tweet text, and the retweet metadata from the API. For our purposes, the API provided easier access to the chain since it also provides the from and to identifiers, so this method was used. In total 2535 chains were computed, ranging from 3 to 15 hops in length with an average length of 3.3. We classified the chains into two contexts shown in Table I, a) long chains, having greater than 5 propagations, and short chains having less than 4.

D. Dyadic Pairs

To assess if there is an impact on interpersonal behavior on feature distribution in Twitter, a context was generated to represent pairwise interaction between Twitter users. This context used the "@" mention tag to isolate a group of tweets that had been part of conversations involving at least two messages between pairs of users. As shown in Table I, this context contains two simple classes, a) dyadic, and b) not dyadic.

Name	*% Present*	*Average score*	*Class*
Age	100.00	610.64	Social
listed_count	100.00	11.82	Social
status_count	100.00	554.49	Social
status_rt_count	100.00	10.17	Social
favourites_count	100.00	57.96	Social
followers	100.00	295.15	Social
followings	100.00	315.03	Social
fofe_ratio	100.00	5.81	Social
char	100.00	120.55	Content
word	100.00	18.69	Content
question	7.95	0.10	Content
excl	10.10	0.15	Content
uppercase	10.23	11.27	Content
pronoun	92.84	4.22	Content
smile	42.24	0.02	Content
frown	1.81	0.43	Content
url	14.17	0.42	Content
retweet	8.71	0.74	Content
sentiment_pos	71.51	1.53	Content
sentiment_neg	59.07	1.23	Content
sentiment	74.20	0.29	Content
num_hashtag	42.09	0.83	Content
num_mention	19.25	0.25	Content
tweet_type	100.00	1.10	Content
ellipsis	2.11	0.29	Content
news	5.13	2.03	Content
average balance of conversation	100.00	0.32	Behavioral
average number of friends in timeline	100.00	2086.28	Behavioral
average spacing between statuses in seconds in timeline	100.00	21959.07	Behavioral
average text length in timeline	100.00	104.52	Behavioral
average general response time	100.00	3.27	Behavioral
average number of messages per conversation	100.00	4.34	Behavioral
average trust value in conversation	100.00	0.10	Behavioral
fraction of statuses in timeline that are retweets	100.00	0.55	Behavioral

TABLE IV
THE SET OF TWITTER FEATURES ANALYZED IN OUR EVALUATION. FEATURES ARE GROUPED INTO THREE CLASSES: A) SOCIAL, B) CONTENT-BASED AND C) BEHAVIORAL. EACH IS A REPRESENTATIVE SUBSET FOR THE LARGER FEATURE SETS IN EACH CLASS.

III. SOCIAL, CONTENT-BASED AND BEHAVIORAL FEATURES

There are a wide range of behavioral, psychological, network-based, social and content-based features in Twitter that could potentially be used to predict credibility of a person or a piece of information in different contexts. For the purpose of this study, we limit our analysis to the feature set described in Table III. Many studies examine Twitter feature by classifying them into groups. For example, [5], [6], [7] all distinguish between social and content-based groups of features. For example, the number of followers a user has is a prominent "social" feature, whereas the use of URLs in a tweet is a a prominent content-based feature. While we do consider these groups independently, there is an additional classification

which we must consider for the analysis in this paper. *Ubiquity* of a feature in a specific context has a strong influence on its ability to predict useful information in different contexts. Features in Twitter have very diverse usages across different contexts, but some features are omnipresent in the data. This set of features aligns reasonably well with the "social model" feature set from [5], but has a different set of properties. The set of always-present features (shown in column 4 of Table III) have some distinguishing properties: for example, they are generally are not binary but have a defined value range, are generally far more complex to compute than other features, although they may be available from pre-computed metadata, and they are generally harder to "fake" than content, demographic or other profile-based other features.

Table III presents the list of Twitter features that we focus on in this study. Each feature is classified into one of three classes. The social class is a representative subset of features that deal with the properties of users in the microblog. For example, demographic information such as gender, profile information such as number of days the account has been active, and social features such as the ratio of friends to followers. This subset of features was described in [5] as part of a credibility prediction model. The second class of features focuses on content only. In addition to standard text features such as punctuation, character and words this class also includes some richer features that can be mined from tweet content. For example, positive or negative sentiment factors, which are computed though comparison with lexicons of keywords. The set also includes a news feature which is computed by comparing content with a popular news archive. In general the set of content-based features do not involve interactions with other users. The final set in Table III are a set of behavioral features. Again, this is a representative subset of a far larger class of features that focus on the dynamics of information flow in the microblog. These features model a far more complex information space than the content-based set. For example, by analyzing conversational aspects of the system (messages, mentions, balance etc) and information flow factors such as a user's propagation energy. The behavioral feature set is a small subset from a study by Adali in [7].

IV. Feature Analysis

Now that we have described our data collection and annotation processes, and outlined a set of target features for analysis....

A. Credible v/s Non-Credible Context

Our first analysis of feature distribution focuses on two simple credibility contexts derived from the credibility assessment study described earlier. Figure 2 shows the mean distribution of features per tweet across credible and non-credible classes. From the 1 to 5 Likert scale in the assessment study, tweets with scores of 1 or 2 were considered as not credible for this evaluation, and tweets with ratings of 4 or 5 were considered as credible. Those tweets with a rating of 3 were discarded to reduce the possibility of ambiguity. Fig 2 shows that

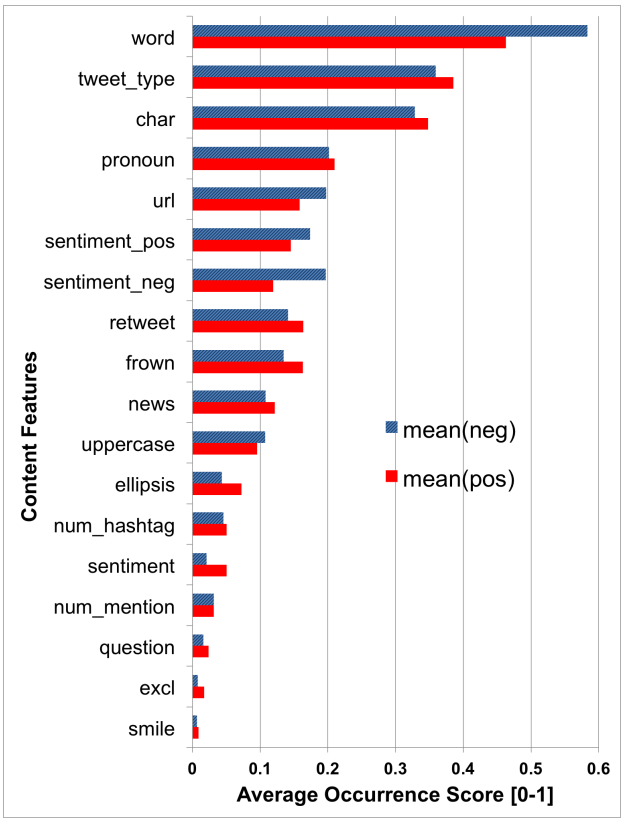

Fig. 2. Analysis of the distribution of selected content-based features based on two credibility contexts. The total score feature occurrences for the credible context was 0.145, compared with 0.148 for the non-credible case.

there is a mixed distribution of feature occurrences across the two classes. For those features that occurred less frequently (question, exclamation, smile etc.) they tended to occur more frequently in the positive class. Interestingly for the sentiment features, which were computed using correlations to positive and negative lexicons, the negative credibility context exhibited a higher relative distribution in both (although there was a 50% relative reduction in the positive sentiment case. This seems to imply that positive sentiment is not necessarily associated with credibility. The "news" feature, computed using correlations with data from the Twibes[1] website tended to occur more frequently in the positive credibility context. The total score feature occurrences for the credible context was 0.145, compared with 0.148 for the non-credible case, which implies that the simple presence of features does not necessarily increase credibility. Perhaps then, the particular usage of individual features or combinations of features has a stronger bearing on the credibility judgement.

B. Cross-corpus Analysis

Figure 3 presents an overview of the results of our cross-topic analysis of feature distribution. Using the topic collection from Table II, the content-based features from Table III were computed for each. The graph shows the average occurrence of

[1] www.twibes.com

a particular feature per tweet on the x-axis, with the exception of the "word" and "char" features which are computed relative to the maximum value. It is clear from the graph that there is a high variance for feature occurrence across the different topics. An interesting insight is illustrated in Figure 5, which shows the distributions of all features across the independent topics. There is a significant difference ($p < 0.05$) between the occurrences of special features across the different sets. Looking closer, the topics that cover emergency and crisis situations, namely, #Earthquake, #Egypt and #Libya (our initial crawl was performed during the 2011 political crises in North Africa), show a significant increase in the number of features occurring in the tweets. The #Libya set scores 0.13 while the best non-crisis topic (#superbowl) scored 0.9, which is a relative increase in feature usage of 44% the crisis situation. Conversely, run-of-the-mill topics such as #Facebook (0.6) and #Love (0.7) score lowest in terms of feature usage. This result has interesting implications, which may seem counter-intuative: it is perhaps easier to make automated feature-based predictions of behavior during crisis, simply because there are significantly more data points available to work with. Of course, this does not hold for all features, for example looking back to Figure 3 we can see that there are significantly less "hashtag" features than with the #superbowl topic.

C. Feature Distribution in Retweet Chains

Our third context class looks at one of the behavioral aspects of the twitter space. In general, content gets retweeted because it is of interest to a larger number of people. However, this may not necessarily mean that the content itself is credible. Consider the case of receiving a spam email from an otherwise credible source, or a case where a comic YouTube video or other such meme is propagated around in the usual "viral" manner. To further explore these questions a third context was established to examine feature distribution across a set of tweets that were found at the ends of long chains, and those that had been propagated, but only in short chains. To recap from our data crawling discussion, max chain length was 15, chains with 6 or more hops were considered "long", chains with 1 or 2 hops were considered "short". As a baseline comparison, a randomly chosen set of non-propagated tweets was added to this context. Figure 6 shows the results of the feature distribution analysis for this context. Probably the most notable result is the prominence of the URL feature in the longer chains, occurring in 50% of the long chain context, indicating that tweets with provenance links to other information do tend to get propagated more often. Similarly, longer tweets in terms of words and characters tend to appear more frequently in longer chains.

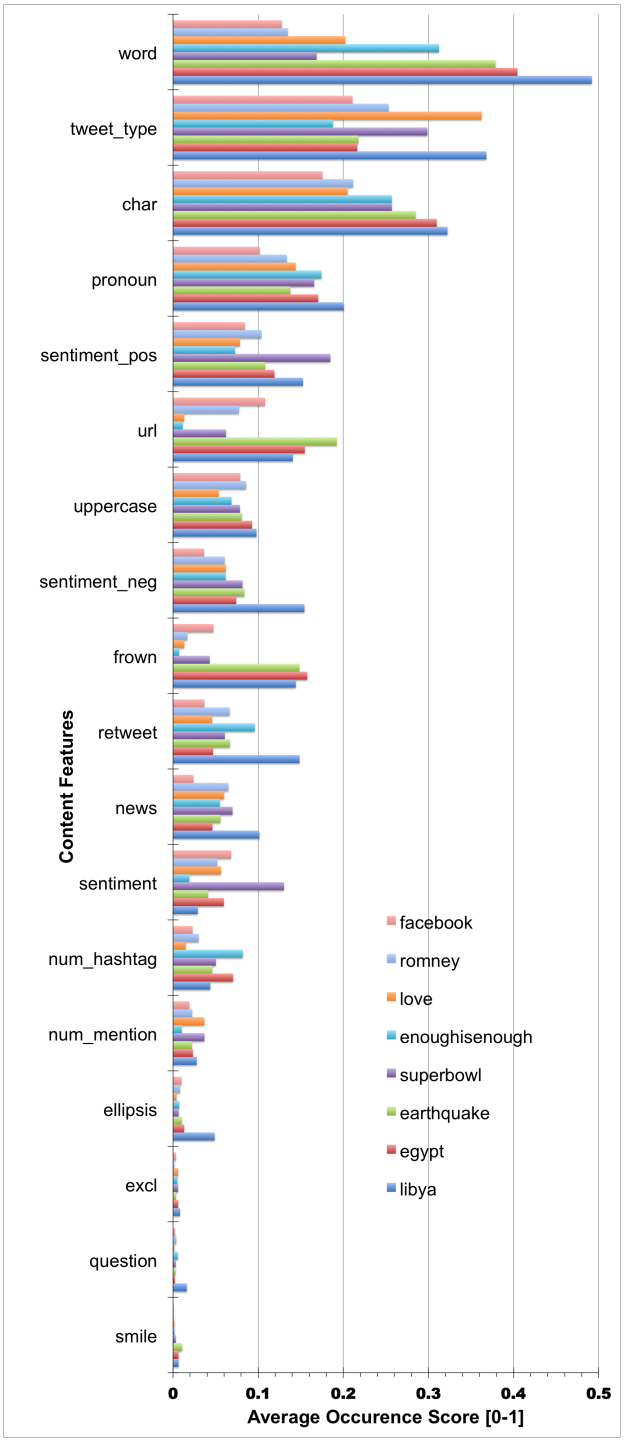

Fig. 3. Analysis of the distribution of selected content-based features across a diverse set of Twitter topics.

D. Feature Distribution in Dyadic Pairs

The final context analyzed in this research was also a behavioral context. Dyadic pairs of tweets –that is, tweets that come from conversations between two users where more than two messages have been exchanged using the '@mention' or '@reply' symbols, were isolated and compared against a standard set of tweets. The comparison again looked at feature distribution across the two contexts. Figure 7 shows the distributions for this context. In this case the variance in feature distribution is not as large as in the other contexts. The character and word features show that dyadic pairs tend to have more words, but shorter words than standard tweets. There are

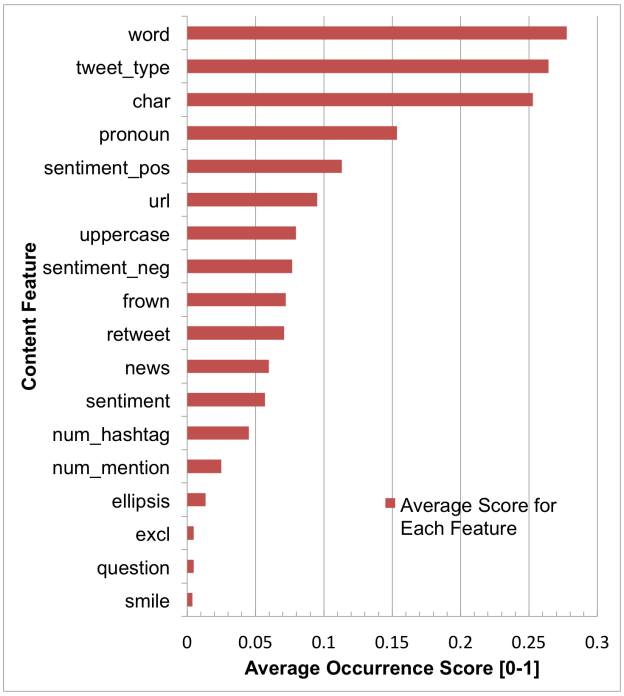

Fig. 4. A plot of the average occurrence of each feature per tweet across all collected data.

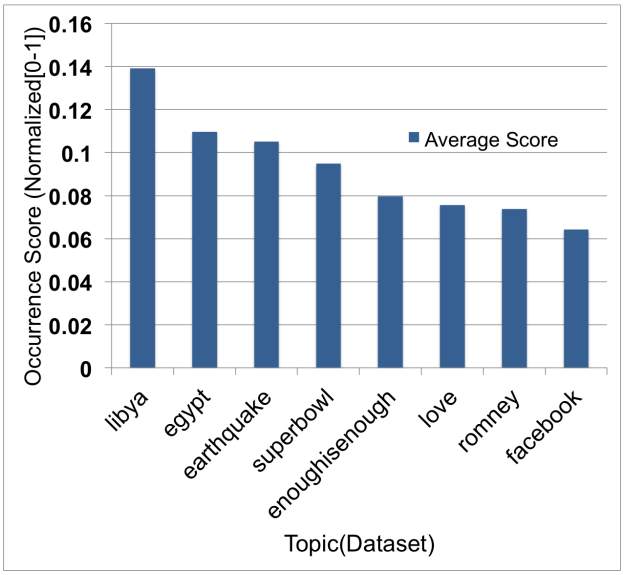

Fig. 5. Overview of the distribution of all features per topic.

significantly more uppercase letters, exclamation marks and negative sentiments in dyadic pairs than in standard tweets, and there are less hashtags and question marks. There is no significant difference in the overall distribution of features across the dyadic and non-dyadic groups.

V. Background

In this paper we have described an in-depth analysis of the distribution of features across a variety of contexts in the Twitter space. The motivation behind this work is that a better

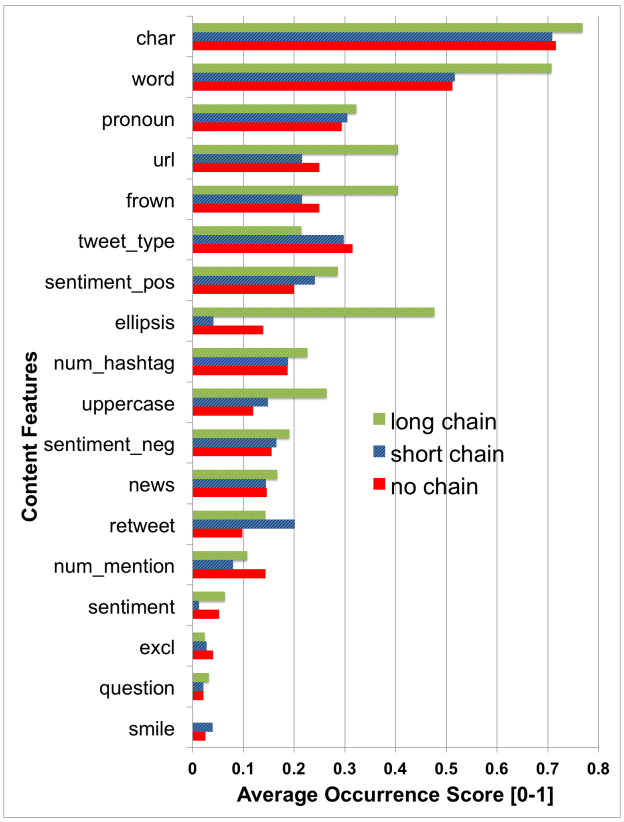

Fig. 6. Distribution of predictive features in three contexts formed around length of retweet chains. a) Non-retweeted content b) Short chains (1-3 hops) c) Long chains (>4 hops).

insight into the behavior of specific features and combinations of features can help to inform and inspire better algorithms to find relevant and credible information in the microblog. A large amount of research effort has focused on modeling trust and credibility of content producers. We now present a discussion and analysis of a representative sample of the most relevant material. Our analysis can be categorized into two parts: research in the general area of trust and credibility mining on the web, and research in the microblog domain. The discussion also touches on possible application areas for improved credibility modeling algorithms, for example, improving social search and designing better personalized information filtering systems such as recommenders.

a) 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 [8] and Milgram's famous small worlds experiment [9], trust has been shown to play an important role in social dynamics of a network. With the advent of social web API's, researchers now have orders of magnitude more data to work with, and accordingly, they can experiment with, and evaluate new concepts far more easily. This is evident across a variety of fields, for example, social web search [10], semantic web [11] [12], online auctions [13] [14] [15], personality and behavior prediction [16] [17], political predictions [18] and

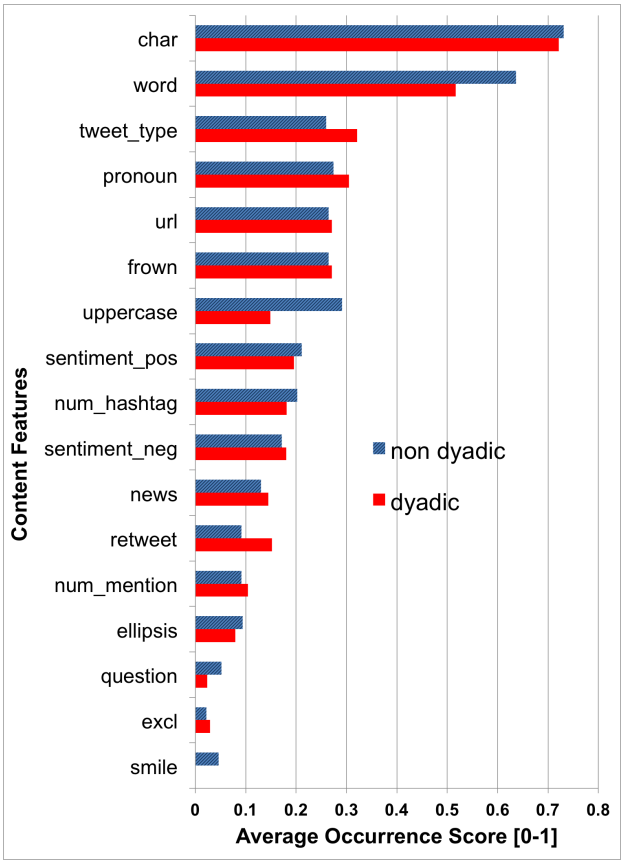

Fig. 7. Distribution of predictive features dyadic pairs. Tweets were selected for this group if they occurred in a pairwise conversation between two users in which more than two messages were exchanged, as measured by the mention and retweet metadata in Twitter API.

many others.

b) Credibility on Twitter: Twitter has a unique combination of text content and underlying social link structure, in addition to a variety of dynamic or ad-hoc structures, making it ideal for the study of information credibility. Some approaches, for example [19] rely on content classifiers or the social network individually, while others harness information from both sources. Canini et al. [20] 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 identifying 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 [20] that both network structure and topical content of a tweet have a bearing on perceived credibility. Adali et al. [7] describe a behavioral approach to analyzing credibility features in the Twitter domain. Our research incorporates features from their model as a representative set for the broad class of possible behavioral measurements in the network.

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, and other high-impact situations [21][22][6][23][24]. For example, Thomson et al. [24] model the credibility of different tweet sources during the Fukushima Daiichi nuclear disaster in Japan. They found that proximity to the crisis seemed to moderate an increased tendency to share information from highly credible sources, which is further evidence for our earlier argument that credibility models in Twitter need to account for and adapt to changes in context. Castillo et. al. [21] describe a 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 [21], and we add a brief comparison of findings in our result analysis. Mendoza et. al [6] 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. An analysis of Follower / Following relations in our crawled twitter data yields a very similar pattern to their result.

In addition to computing a credibility score based on some set of Twitter features, it is important to consider the end-user's *perception* of credibility. Morris et al. [25] performed a study to address users perceptions of the credibility of individual tweets in a variety of contexts, for example, from socially connected and unconnected sources. From the results, Morris et al. derive a set of design recommendations for the visual representation of social search results.

c) Applications of Credibility Models: Credibility models [26] have been shown to play an important role in the social web, from the self-regulating e-commerce systems such as eBay's platform to the process of content prediction (e.g: Amazon, Netflix etc.). They can be applied in social filtering to augment user similarity metrics in the recommendation process. [26]. They have also been shown to increase robustness of prediction algorithms in cases where bad (malicious / erroneous) ratings exist [27][28]. Credibility models have been applied in a variety of ways on Twitter data, as information filters [29], crime detectors [30] and in intelligent social tagging tools [31]. Models that incorporate explicit distrust have recently been shown to produce better predictions, for example, Victor et al. [32] 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 [33]. In this paper, we are not propagating credibility values around the

network, or computing direct interpersonal trust at the dyadic level, however, the authors believe that distrust metrics can potentially improve credibility predictions in Twitter.

VI. Discussion and Future Work

This paper has presented an analysis of the distribution of the salient features in Twitter that can be used to find interesting, newsworthy [6] and credible [5] information. Our analysis focused on feature distributions in four distinct contexts: diverse topics; credibility levels; retweet chains and dyadic interactions. Through our analysis of distributions we have shown that in general (across 8 data sets), feature usage tends to increase in emergency situations or situations of unrest. We believe that while some features may serve as good predictors of credible information, their usefulness can vary greatly with context, both in terms of the occurrence of a particular feature, and the manner in which it is used. Due to the size and rapid evolution of microblogs such as Twitter, it is exceedingly challenging to fully understand the subtle links between feature presence/usage and truly credible information. A follow up research will focus on combining predictive ability of different features with both distribution and particular usages to gain a more in-depth knowledge of the complex interactions that occur in the Twitter space, and to provide insight for future credibility prediction models.

Acknowledgment

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; 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, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter." in *WWW*, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, Eds. ACM, 2011, pp. 675–684. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2011.html#CastilloMP11

[2] Q. V. Liao, C. Wagner, P. Pirolli, and W.-T. Fu, "Understanding experts' and novices' expertise judgment of twitter users," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2461–2464. [Online]. Available: http://doi.acm.org/10.1145/2207676.2208410

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

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

[5] B. Kang, J. O'Donovan, and T. Hollerer, "Modeling topic specific credibility on twitter," in *IUI*, C. Duarte, L. Carriço, J. A. Jorge, S. L. Oviatt, and D. Gonçalves, Eds. ACM, 2012, pp. 179–188.

[6] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?" in *1st Workshop on Social Media Analytics (SOMA '10)*. ACM Press, Jul. 2010. [Online]. Available: http://chato.cl/papers/mendoza_poblete_castillo_2010_twitter_terremoto.pdf

[7] S. Adali, F. Sisenda, and M. Magdon-Ismail, "Actions speak as loud as words: predicting relationships from social behavior data," in *WWW*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, Eds. ACM, 2012, pp. 689–698.

[8] D. Pool and M. Kochen, "Contacts and influence," *Social Networks*, vol. 1, no. 1, pp. 5–51, 1978. [Online]. Available: not-available

[9] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 61–67, May 1967.

[10] K. McNally, M. P. O'Mahony, B. Smyth, M. Coyle, and P. Briggs, "Towards a reputation-based model of social web search," in *Proceedings of the 15th international conference on Intelligent user interfaces*, ser. IUI '10. New York, NY, USA: ACM, 2010, pp. 179–188. [Online]. Available: http://doi.acm.org/10.1145/1719970.1719996

[11] J. Golbeck, *Computing with Social Trust*. Springer Publishing Company, Incorporated, 2010.

[12] T. G. F. W. Haifeng Zhao, William Kallander, "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.

[13] D. Houser and J. Wooders, "Reputation in auctions: Theory, and evidence from ebay," *Journal of Economics and Management Strategy*, vol. 15, no. 2, pp. 353–369, 2006.

[14] P. Resnick and R. Zeckhauser, "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.

[15] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod, "Extracting and visualizing trust relationships from online auction feedback comments," in *IJCAI*, 2007, pp. 2826–2831.

[16] M. E. K. T. Jennifer Golbeck, Cristina Robles, "Predicting personality from twitter," in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.

[17] S. Adali, R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B. K. Szymanski, W. A. Wallace, and G. T. Williams, "Measuring behavioral trust in social networks," in *ISI*, 2010, pp. 150–152.

[18] J. Golbeck and D. L. Hansen, "Computing political preference among twitter followers," in *CHI*, 2011, pp. 1105–1108.

[19] Y. Suzuki, "A credibility assessment for message streams on microblogs," in *3PGCIC*, 2010, pp. 527–530.

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

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

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

[23] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 591–600.

[24] H. S. F. L. Y. L. R. H. R. I. Z. W. Robert Thomson, Naoya It, "Trusting tweets: The fukushima disaster and information source credibility on twitter," in *Proceedings of the 9th International ISCRAM Conference Vancouver, Canada, April 2012*, 2012.

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

[26] J. O'Donovan and B. Smyth, "Trust in recommender systems," in *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*. ACM Press, 2005, pp. 167–174.

[27] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, "Identifying attack models for secure recommendation," in *Beyond Personalisation Workshop at the International Conferece on Intelligent User Interfaces*. San Deigo, USA.: ACM Press, 2005, pp. 347–361.

[28] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness," *ACM Trans. Inter. Tech.*, vol. 7, no. 4, p. 23, 2007.

[29] S. Ravikumar, R. Balakrishnan, and S. Kambhampati, "Ranking tweets considering trust and relevance," *CoRR*, vol. abs/1204.0156, 2012.

[30] X. Wang, M. Gerber, and D. Brown, "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral - Cultural Modeling and Prediction*, ser. Lecture Notes in Computer Science, S. Yang, A. Greenberg, and M. Endsley, Eds. Springer Berlin / Heidelberg, 2012, vol. 7227, pp. 231–238.

[31] I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi, "In tags we trust: Trust modeling in social tagging of multimedia content," *Signal Processing Magazine, IEEE*, vol. 29, no. 2, pp. 98 –107, march 2012.

[32] P. Victor, C. Cornelis, M. D. Cock, and E. Herrera-Viedma, "Practical aggregation operators for gradual trust and distrust," *Fuzzy Sets and Systems*, vol. 184, no. 1, pp. 126–147, 2011.

[33] T. DuBois, J. Golbeck, and A. Srinivasan, "Predicting trust and distrust in social networks," in *Proceedings of the third IEEE International Conference on Social Computing (SocialCom)*, 2011.