2016 International Conference on Collaboration Technologies and Systems

Features for Ranking Tweets Based on Credibility and Newsworthiness

Jacob Ross, Krishnaprasad Thirunarayan
Kno.e.sis: Ohio Center of Excellence in Knowledge-enabled Computing
Department of Computer Science and Engineering
Wright State University
Dayton, Ohio 45435
{ross.138, t.k.prasad}@wright.edu

***Abstract*—We create a robust and general feature set for learning to rank tweets based on credibility and newsworthiness. In previous works, it has been demonstrated that when the training and testing data are from two distinct time periods, the ranker performs poorly. We improve upon this by creating a feature set that does not overfit a particular year or set of topics. This is critical for robust analysis of social media over time. In order to derive such features, we use the studies done on credibility perception of social media as well as the clues provided in past works in this domain. We also present new features that, to our knowledge, are more effective than the state of the art.**

***Keywords*—Twitter, Social Media, Credibility, Learning to Rank**

I. Introduction and Motivation

There is no denying that the popularity of social media has risen greatly over the past few years. Currently, there are 320 million monthly active users on the micro-blogging site, Twitter[1]. Twitter is also a global phenomenon: 77% of Twitter accounts are outside of the United States, and Twitter supports 33 languages.

Social media has been studied for the purposes of disaster response [1], [2], [3], [4], to analyze disease outbreak [5], to predict user location, [6], [7], to predict stock market trends [8], [9], and for many other domains. However, false information spreading on social media has serious ramifications. There is a need to automatically determine the credibility of such tweets, and many techniques already exist. Most approaches report high accuracy, but the training and testing datasets come from the same year and have the same topics [10], [11], [12], [13], [14], [15], [16], [17], [18]. We are constantly gaining new tweet data, and how people utilize Twitter and social media is constantly changing. However, the approaches that do train and test on datasets from two distinct years with distinct topics perform poorly [18], [19]. Gupta [19] takes a Learning to Rank approach in order to automatically determine tweet credibility. However, they show that while some features do accurately capture credibility for their 2011 dataset, they fail for the 2013 dataset. They also substantiate this by showing a decrease in ranker performance when trained on the 2011 dataset and tested on the 2013 dataset. We aim to create a feature set that will lead to improved ranker performance when the training and testing datasets are from two distinct years with different topics.

There are many instances of false information that involve high impact events being spread on social media. Gupta [20] reported on fake images that were spread on twitter during Hurricane Sandy. They discovered that 86% of the tweets containing fake images were retweets, implying that the vast majority of the fake images were not sent from the source of the fake images, rather, they were spread by users thinking the images were legitimate. Gupta later analyzed how false information was spread on Twitter during the Boston Marathon bombings in 2013 [21]. They found that 29% of the viral content generated regarding the Boston Marathon bombings were fake.

False information spread on Twitter can incite panic, but can also affect real world disaster relief efforts. Gupta [16] showed there were over 2000 fake tweets that claimed the need for blood donations during the 2011 bombing in Mumbai. In reality there was no shortage of blood, but the false tweets prompted citizens to arrive at hospitals with the intent of donating blood.

In Section II, we discuss previous works that leverages social media to solve real world problems, as well as the current approaches for automatically determining tweet credibility. In Section III, we describe the datasets we used to carry out our experiments. In Section IV, we discuss which features we select from previous work and the new features we derive for classification. In Section V, we show how well the classifier performs on the datasets with our new feature set. In Section VI, we discuss our results and the implications of our approach. In Section VII, we discuss worthwhile research endeavors in order to improve the state of the art for automatically determining tweet credibility.

[1]about.twitter.com/company

978-1-5090-2300-4/16 $31.00 © 2016 IEEE
DOI 10.1109/CTS.2016.21

II. Related Work

First, we discuss the research done on understanding credibility in terms of social media. Second, we discuss the current approaches for automatically determining tweet credibility.

A. Twitter and Credibility

Morris conducted a survey in order to understand the user's perception of credibility on Twitter [22]. They focused on users searching for content on Twitter based on topic and not on a specific author. Typically, a user will accept a known author's claims as fact. However, now that users tend to search based on topic, they cannot be guaranteed that the authors of the tweets for a particular topic are the ones they are familiar with. They make some interesting observations, such as, users do not have enough clues to accurately assess credibility on content alone; users will use other clues such as profile image and user name in order to help them assess credibility. In their experiments, Morris et al aim to uncover what features of a tweet or tweet source users utilize in order to judge its credibility.

A majority of the features that enabled the users to determine credibility were associated with the author of the tweet. The author based features can be grouped into three categories: *influence, topical expertise,* and *reputation*. Influence based features include follower, retweet, and mention counts. Topical expertise features are features an author has that indicate they are an expert on the topic the viewer is interested in. Morris gleaned such indicators through the author's homepage, the author's tweet history, outside webpages that are on topic that mention the author, and the author being in a location relevant to the topic. Reputation based features help indicate the familiarity a user has for the author of a tweet. Such features include whether the author is followed by the user, the author is someone the user has heard of before, or if the author's account has been verified by Twitter.

The content based features that revealed the most about a tweet's credibility were if the tweet contained a URL that leads to a reputable webpage, or if there are multiple tweets that make the same claim as the tweet in question, and use standard grammar. Users noted that the image an author uses as their profile picture affects how the credibility of their tweets are judged. Users were more willing to trust an author that had an image of themselves or of an image related to the topic they are interested in. Other profile images, such as cartoons or the default Twitter picture, indicated decreased credibility. Similarly, the structure of the author's username seemed to impact the user's credibility of the author's tweet.

Shariff [23] conducted a study to analyze how online social media users tend to judge or misjudge tweet credibility. They reveal that topics involving politics have the largest number of misjudged tweets. This can be attributed to the fact that a majority of the tweets involving politics are often questions and opinions. They note that the political tweets that are labeled correctly are usually linked to a known and reputable news source. Shariff also did an in-depth analysis of the tweets most users misjudged. 95% of those tweets were breaking news and political news. They show that tweets that lack a link to outside resources, such as a URL, are often difficult for users to judge.

B. Automatically Determining Tweet Credibility

Castillo [10] produced one of the earliest works on automatically predicting tweet newsworthiness and tweet credibility. Their data collection approach consisted of two phases. First, label and keep tweets that are deemed newsworthy. Second, label the newsworthy tweets with a credibility score. In order to obtain these annotations, Castillo utilized Amazon Mechanical Turk[2] and label tweets based on newsworthiness and credibility. They accept only those labeled tweets for which five out of seven users agree on the score. For the credibility labeling, users were asked to assign one of four scores: almost certainly true, likely to be false, almost certainly false, and cannot be decided.

By analyzing annotator comments, Castillo learned the key factors that annotators used to make their judgments. They note that some topics will evoke emotion out of the users posting about that topic. Typically, the overall emotion of a topic will match the emotion of the tweets for that topic. Annotators commented on how you can estimate the certainty a user has of the credibility of the information they are sharing. If the propagator has low confidence, they will typically question the information they are sharing. Annotators typically labeled tweets to have high credibility if the tweet cites an external link to a known source. The link can act as proof for the claim made in the tweet. Annotators also commented on characteristics of the tweet's author; annotators took things such as screen name, profile description, and user picture into account. The dataset for training and testing have features that fall into one of four categories: message based features, user and author based features, topic based features, and propagation based features.

Castillo evaluated various cost sensitive classifiers. The J48 classifier yielded the best results with 89% accuracy. For the credibility classification portion, they yielded 86% accuracy. Tweets that contain a URL tend to be credible, tweets with negative sentiment tend to be credible, and tweets with many retweets tend to be credible. A URL essentially acts as proof for a claim made in the tweet. If a tweet is retweeted many times, this indicates many users found the tweet useful. They also reveal the tendencies of non-credible tweets. Information deemed non-credible tends to be created by people who have a low tweet counts, and a small fraction of tweets with positive sentiment were deemed to be credible.

Yang [11] conducted a similar study to Castillo et al. in [10], except that they focus on Sina Weibo rather than Twitter. Sina Weibo is the most popular online social media outlet in China, and has nearly 8 times as many users as Twitter. Sina

[2]https://www.mturk.com/mturk/welcome

Weibo has a built-in rumor detection tool, so they were able to collect confirmed false rumors from Sina Weibo directly, and there is no need for a data annotation step. Although Twitter and Sina Weibo differ, there are many text and author based features that are relevant to both. Yang defined five categories for tweet features: content based features, client based features, account based features, propagation based features, and location based features.

Content based features are very similar to those in Castillo [10]. Examples of content based features include whether the message contains pictures, number of positive and negative emoticons, and whether the tweet contains a URL or not. The client based features describe what medium the user used in order to post their message, for example, if the user posted from a mobile client or from a web browser. Examples of account based features include whether or not the user has a verified account, whether or not the user has a description, gender, and number of friends. The gender feature is an example of a feature that is not readily available from the Twitter API, but is available from Sina Weibo. Location based features capture where the event being discussed has taken place. Propagation based features include features such as whether the message was re-tweeted and the number of comments for the message. In addition to these features, Yang proposed two new features: location of the event and the client program used to post the message. Overall, these new features were helpful, and shows that putting effort into creating new features can help increase classifier accuracy.

In [24], Gupta used a ranking SVM and pseudo relevance feedback (PRF) in order to rank tweets based on their credibility score. They address the problem that even though a topic in general can be credible, tweets on that topic can be non-credible and can potentially contain false information. A prime example of this are the false images that spread through Twitter during Hurricane Sandy. Hurricane Sandy was not a rumored event, however false information pertaining to it spread on Twitter rapidly and from otherwise credible sources [20]. Tweets were labeled with the following annotations: definitely credible, seems credible, definitely non-credible, related to a topic but has no information, and tweet is unrelated to topic.

Features are categorized as either content based features or source based features. They make some key observations about what features a tweet has that are correlated with credibility and newsworthiness. Tweets with a large number of unique characters tend to be credible. They attribute this to the fact that tweets with mentions, hashtags, and URLs will contain more unique characters. Tweets that contain swear words correlate with being relevant to a topic, but it is often a reaction to the topic and contains no information. As has been revealed in numerous previous works, they also discovered that a tweet with a URL tends to be credible. With their approach, they achieve a $NDCG@50$ score of 0.73 after applying PRF.

There are two related papers that highlight the problem we aim to solve. In [19], Gupta demonstrated that their approach sharply degrades if the training and testing datasets are from two distinct years. In [18], Boididou demonstrated a decrease in accuracy when the training and testing datasets are on two different topics.

In [19], Gupta extended their work from [24] in order implement a real time browser based system for ranking tweets based on credibility and newsworthiness called TweetCred. They gather tweets from six topics that occurred in 2013. Annotation is similar to their earlier paper [24]. First, tweets are labeled as either: containing information about the event, related to the event but has no information, and not related to the event. Then, the tweets deemed newsworthy were labeled as either: definitely credible, seems credible, definitely incredible, and cannot be determined.

This implies that a newsworthy tweet contains information about the event, but could have non-credible elements in it. In their ranking scheme, a tweet that is newsworthy but non-credible will rank higher than a tweet that is an opinion and contains no information about the event itself. Gupta defined five categories for tweet features: tweet meta-data, tweet content features, user based features, linguistic based features, and external resource features.

Tweet meta-data features include features such as number of seconds since the tweet was posted and the source of the tweet. Tweet content features include number of characters, number of words, and number of URLs. Linguistic based features include features such as presence of swear words, negative emotion words, and number of positive emotion words. External based features include the Web of Trust score for provided URLs and ratio of likes to dislikes to an attached YouTube video. For their purposes, the response time of their system was important, so they sacrificed ranker accuracy for faster training and testing times. Coordinate Ascent yielded the best NDCG@100 score of 0.7607. For their system, Gupta et al. chose SVMRank since it performed only slightly worse in terms on NDCG ($NDCG@100 = 0.719$) but was much faster than the Coordinate Ascent approach. They show that when training on the 2011 dataset and testing on the 2013 dataset, ranker performance degrades significantly. They achieve an NDCG@100 score of 0.3783 when the training and testing datasets are configured this way. This is problematic, because we will be constantly gaining new Twitter data, and how users utilize social media is constantly changing.

Boididou [18] exhibits a similar problem, in that their approach suffers when the training and testing datasets are on two distinct topics. When training and testing on their Hurricane Sandy dataset with 10-fold cross-validation, they achieve 80% accuracy. When training and testing on their Boston Marathon dataset with 10-fold cross-validation, they achieve 80% accuracy. However, when training on the Hurricane Sandy dataset and testing on the Boston Matathon dataset, they can only achieve 59% accuracy. This is also problematic, because

new events and topics will arise constantly, and training seems to overfit the data.

III. Dataset

A. Dataset Properties

Here we discuss the datasets used to carry out our experiments. We have two distinct sets of annotated tweets from 2011 and 2013. We obtained these datasets from Gupta to enable comparisons with their work. The 2011 dataset has been discussed in their first paper on ranking tweets [24] and the 2013 dataset has been discussed in their TweetCred paper [19].

Each dataset was labeled by human annotators as described in [24] and [19]. Each tweet was labeled with one of the following numerical annotations in order of most relevant to least relevant: (5) Tweet contains information and is credible, (4) Tweet contains information and seems credible, (3) Tweet contains information and is non-credible, (2) Tweet is relevant to the topic but contains no information, and (1) Tweet is spam.

B. 2011 Dataset

The 2011 Dataset contains 4028 tweets that spans 14 different topics. We explain these topics in Table I. In Table II we show the class distributions for the 2011 Dataset. The topics vary from disaster scenarios (i.e., the stage collapse at the Indiana State fair and the London Riots), to politics (i.e., the Anna Hazare protests in India), and topics involving entertainment and technology (i.e., Steve Jobs resigning and the Facebook messenger app.)

The 2013 dataset contains 2198 Tweets that spans 6 different topics. We explain the 2013 topics and class distributions in Tables I and III respectively. The 2013 topics contain mostly disaster scenario events.

IV. Feature Selection

A. Popular Features from Previous Works

We summarize the most popular features used in previous works. In total, we gather features from 17 different papers, some of which we have already discussed. We gather features from works that use classifiers to automatically predict credibility [10], [11], [12], [13], [14], [15], [16], [17], [18], features from works that use learning to rank algorithms [19], [24], and features gleaned from works that take hybrid or other approaches to quantify and model credibility [25], [26], [27], [28], [29]. We use the features that appear in 6 or more of these papers.

The most popular tweet based features are: tweet is a retweet (boolean), tweet length (integer), number of words (integer), number of user mentions (integer), number of hashtags (integer), number of URLs (integer), tweet has a URL (boolean), number of retweets (integer), has a happy emoticon (boolean), has a sad emoticon (boolean), and sentiment score (decimal).

The most popular author based features are number of followers (integer), number of friends (integer), number of tweets (integer), has description (integer), is verified (boolean), and friends to followers ratio (decimal).

We use these features as a starting point for our own feature set. For all sentiment based features, we use the lexicon from Hu [30] and we use the lexicon of curse words from Wang [31].

B. New Features

We create two features that aim to capture when the sentiment of a tweet matches the overall sentiment of the topic it is in, *differenceFromMeanPositive* and *differenceFromMeanNegative*. We hypothesize that tweets that have similar sentiment to the rest of the tweets in the topic will be credible. However, if a tweet does not have the same sentiment of the topic overall, this could mean the tweet is non-credible or not newsworthy.

For each topic, we calculate the average number of positive and negative words for the tweets in that topic. Then, for each tweet, compute the difference between the average number of positive words for that tweet's topic, and how many positive words appear in that tweet. We do the same process for negative words. In Table IV we show the average number of positive and negative words per tweet for each topic. In general, the high impact and negative topics (i.e., The Boston Marathon Bombings, Mumbai Blasts) tend to have more negative words per tweet than positive words. Topics that are less negative (i.e., Bert and Ernie Gay Marriage, the Facebook Messenger) tend to have more positive words. Here we show how these two features are calculated. We use the lexicon from Hu [30] to determine which words convey positive sentiment and which words convey negative sentiment.

The number of positive words p for a given tweet t is the summation of the number of words in that tweet that are in the positive word lexicon PW. For each word, return $\mathbb{1}$ if it is in PW.

$$p(t) = \sum_{w \in t} \mathbb{1}(w \in PW) \tag{1}$$

Similarly, we can compute the number of negative terms $n(t)$ as:

$$n(t) = \sum_{w \in t} \mathbb{1}(w \in NW) \tag{2}$$

The average number of positive words a^+ for a given topic T_i is the summation of all the positive words in all tweets t belonging to that topic divided by the number of tweets in that topic $|T|$

$$a^+(T_i) = \frac{1}{|T|} \sum_{t \in T_i} p(t) \tag{3}$$

TABLE I
TWEET TOPICS

Topic	Description
UK Riots (2011)	Riots take place in London; 5 people died and many more injured
Libya Rebels (2011)	Rebels take control on city in Libya fighting Qadaafi's forces
Virginia Earthquake (2011)	5.8 magnitude earthquake hits Virginia
Stocks Downgrade (2011)	S&P downgrades from AAA to AA-plus
Hurricane Irene (2011)	Hurricane causes $10.1 billion in damages, 55 deaths
Indiana Fair Stage Collapse (2011)	Stage collapses during performance at the Indiana State Fair, 5 dead and 40 injured
Mumbai Bombings (2011)	Three bombing take place in Mumbai. 26 people dead and 130 injured.
Anna Hazare Anti-Corruption (2011)	Anna Hazare's anti-corruption protests against the Government of India
Steve Jobs Resigns (2011)	Steve Jobs resigns as Apple's CEO.
Google Purchases Motorola (2011)	Google buys Motorola for $12.5 billion.
Rupert Murdoch Scandal (2011)	Phone hacking scandal involving Rupert Murdoch.
The Situation and Abercrombie and Fitch (2011)	Abercrombie and Fitch asks "The Situation" to stop wearing their clothing.
Bert and Ernie Gay Marriage	Rumors of Bert and Ernie from Sesame Street being a gay couple.
Facebook Messenger (2011)	Facebook launches their new, independent messenger app.
Boston Marathon Bombing (2013)	Bombs explode near the finish line of the 2013 Boston Marathon. 3 people are killed and 260 are injured.
Typhoon Haiyan (2013)	Record breaking typhoon near the Philippines that claimed 6,000 lives.
Cyclone Phailin (2013)	550,000 people evacuated due to tropical cyclone near India.
Washington Navy Yard Shooting (2013)	12 people shot and killed by gunman inside the Naval Sea Systems Command.
Polar Vortex (2013)	Mid-western United States hit with record low temperatures in the winter of 2013.
Oklahoma Tornadoes (2013)	24 dead and 212 injured in a series of tornadoes that hit Oklahoma

TABLE II
CLASS DISTRIBUTION FOR THE 2011 DATASET

Class	Number of Tweets
5 - Definitely Credible	656
4 - Seems Credible	602
3 - Definitely Non-Credible	113
2 - Relevant but not Newsworthy	2352
1 - Spam	305

TABLE III
CLASS DISTRIBUTION FOR THE 2013 DATASET

Class	Number of Tweets
5 - Definitely Credible	555
4 - Seems Credible	371
3 - Definitely Non-Credible	95
2 - Relevant but not Newsworthy	845
1 - Spam	332

Similarly, we can define the average number of negative words for a topic $a^-(T_i)$ as:

$$a^-(T_i) = \frac{1}{|T|} \sum_{t \in T_i} n(t) \tag{4}$$

For each tweet t, the *differenceFromMeanPositive* feature for that tweet is the average number of positive terms for the topic that tweet belongs to $a^+(T)$ minus the number of positive words for that tweet $p(t)$.

$$d_T^+(t) = a^+(T) - p(t) \tag{5}$$

Similarly, we can create the *differenceFromMeanNegative* feature $d_T^-(t)$ as:

$$d_T^-(t) = a^-(T) - n(t) \tag{6}$$

Morris et al. [22] discovered in their study on credibility perceptions that users perceive irregular grammar as a sign of non-credibility [22]. In order to capture this, we create two

TABLE IV
AVERAGE NUMBER OF POSITIVE AND NEGATIVE WORDS PER TWEET FOR EACH TOPIC

Topic	Positive	Negative
Anna Hazare Protests (2011)	.3714	.3
Virginia Earthquake(2011)	.0857	.2
Facebook Messenger (2011)	.3584	.0189
Google Buys Motorola (2011)	.1119	.0299
Hurricane Irene (2011)	.0945	.1417
State Fair Stage Collapse (2011)	.0833	.2121
Libya Rebels (2011)	.1075	.3118
London Riots (2011)	.16	.38
Mumbai Blasts (2011)	.1795	.8461
Rupert Murdoch Scandal (2011)	.0571	.3333
Stock Downgrade (2011)	.1034	.5747
Bert and Ernie Gay Marriage (2011)	.1379	.2241
Steve Jobs Resigns (2011)	.2439	.1382
The Situation and A&F (2011)	.2214	.1526
Boston Marathon Bombings (2013)	.1357	.5
Cyclone Phailin (2013)	.1899	.6783
Navy Yard Shooting (2013)	.0867	.5667
Oklahoma Tornadoes (2013)	.1912	.4645
Philippine Typhoon (2013)	.352	.504
Polar Vortex (2013)	.1939	.6182

new features, *ratioPunctuationNumWords* and *ratioPunctuationNumCharacters*. It is normal for users to not adhere to standard grammar rules when composing tweets due to the character limit on Twitter. Thus, we believe one way to capture anomalous grammar is if a user uses too many punctuation marks, or if they use too few punctuation marks. Given that there can only be a maximum of 140 characters per tweet, raw punctuation count will not reveal much. We want to compare the number of punctuation marks relative to how many words are in the tweet.

Here we describe three author based features. We hypothesize that people who have been on Twitter for a longer period of time will tend to have more followers, more friends, and

have produced more tweets. If people have an unusually high number of followers relative to their friends, then this could influence perceived credibility of this author. If a user has a high number of followers relative to the amount of time they have been active on Twitter, this could imply that many of their followers are bots. If a user produces a large number of tweets for the amount of time they have been active on Twitter, this can capture whether or not the user tends to tweet many spam or non-newsworthy tweets.

We also create three author based features that are based on the sentiment of the description of the author: *numNegativeWordsDescription*, *numPositiveWordsDescription* and *numCurseWordsDescription*. Sentiment based features for the text of tweet has been well explored in previous works. We simply map this feature to the optional descriptions each user on Twitter has. If the author does not have a description, each of these features are set to 0. These features have been used on the text of a tweet many times in the past, however, we want to see the usefulness of applying this feature to the optional description a Twitter user may set.

We create a feature, *hasPray* to detect whether or not the tweet has the substring " pray". We pick this word because it appears in tweets that are relevant to a topic, but do not contain any newsworthy information. Tweets that contain the " pray" substring are often people offering emotional support, and do not contribute to helping the reader understand the event. This feature ranks very highly for the 2013 dataset, but lowly in the 2011 dataset. This means that this feature exhibits the same problem the *hasStockSymbol* has on the results. It ranks highly for the 2013 dataset because all of the topics in the 2013 dataset are disaster scenarios in which many users will be offering emotional support, but not tweet anything newsworthy. However, the 2011 dataset is not entirely composed of disaster related topics, meaning the *hasPray* feature will be far less prominent.

C. Feature Ranking

We rank our entire feature set for each dataset with the LibSVM extension developed by Chen et al. that uses F-Score to rank the features [32].

In Table V, we list the top 20 features ranked in order from best to worst. There is significant overlap in the top 10 features across each dataset. One feature in particular, *hasStockSymbol* ranks very highly for the 2011 dataset, but is ranked extremely low in the 2013 dataset. This was also the case in the TweetCred paper by Gupta [19]. This is a sign that this particular feature overfits the 2011 dataset, and has no bearing for the 2013 dataset. In the next section, we will show the affects of removing this particular feature and how it affects ranker performance.

The top features for each dataset come as no surprise. The feature *numURLS* ranks as the best feature for each dataset. The feature *hasColonSymbol* is the text based feature where it indicates whether or not the tweet has a colon in it. Colons appear in tweets that re-tweets, and colons often precede a URL. The feature *numUniqueCharacters* correlates with high credibility because tweets that have hashtags and URLs are likely to have more unusual characters than just plain text. Two of our new features, *ratioPunctuationTweetLength* and *ratioPunctuationNumWords* appear in the top 10 features for each data set. The *hasPray* feature ranks 13th for the 2011 dataset and 5th for the 2013 dataset. The worst features are features that are not present in a majority of tweets or users, or, these features appear equally for all users and for all tweets. The feature *isVerified* ranks as the worst feature because most Twitter users do not have verified accounts.

TABLE V
OUR TOP 20 FEATURES BASED ON F-SCORE. OUR NEW FEATURES ARE DENOTED WITH AN ASTERISK.

Feature Ranking for 2011 Dataset	Feature Ranking for 2013 Dataset
numURLS (.3215)	numURLS (.3215)
hasStockSymbol(.2529)	hasColonSymbol(.2779)
hasColonSymbol(.2445)	numUniqueCharacters(.2034)
numUniqueCharacters(.1464)	numPunctuation(.1910)
numPunctuation(.1348)	**hasPray*(.1496)**
ratioPunctuationTweetLength*(.1006)	**ratioPunctuationTweetLength*(.1188)**
ratioPunctuationNumWords*(.0712)	**ratioPunctuationNumWords*(.0819)**
numSelfWords(.06)	numSelfWords(.0817)
hasVia(.031471)	tweetLength(.069)
isReply(.0237)	**numNegativeWordsDescription*(.045)**
numHashtags(.0237)	numNegativeWords(.0437)
numQuestionMarks(.0227)	**differenceFromMeanNegative*(.0391)**
hasPray*(.0177)	numStatuses(.0369)
numPositiveWords(.0158)	**ratioNumStatusesUserAge*(.0342)**
differenceFromMeanPositive*(.0141)	hasVia(.0304)
numPositiveWordsDescription*(.0135)	numExclamationMarks(.0245)
numStatuses(.0111)	numHashtags(.0226)
ratioNumStatusesUserAge*(.0099)	numQuestionMarks(.0214)
numCurseWords(.00838)	numWords(.0172)
tweetLength(.0078)	isReply(.0161)

V. RESULTS

Ranking SVM is described by Joachims [33] and we use the implementation that is an extension of LibSVM [34] by Lee [35]. We use Normalized Discounted Cumulative Gain (NDCG) [36] as our performance metric in order to compare results with the previous work [15]. For our purposes, the possible relevancy scores are 5 (definitely credible and newsworthy), 4 (seems credible and newsworthy), 3(non-credible and newsworthy), 2 (relevant to a topic, but contains no information), and 1 (spam).

We apply linear scaling to each feature so that each value falls between -1 and 1. We also randomly shuffle the data instances so that when we use cross validation, tweets from similar topics are not grouped together. We chose 4-fold cross validation in order to compare our results with that of Gupta in [19].

In Table VI, we show how well the ranking SVM algorithm performs on the 2011 dataset with 4-fold cross validation. For the 2011 dataset, ranker performance decreases when removing the *hasStockSymbol* feature. Earlier we discussed how this feature was deemed important for the 2011 dataset, but was deemed unimportant for the 2013 dataset. This is also reflected

TABLE VI
RESULTS OF OUR NEW FEATURE SET WHEN TRAINING AND TESTING ON THE 2011 DATASET WITH 4-FOLD CROSS VALIDATION. WE ALSO SHOW THE EFFECTS OF REMOVING THE STOCK SYMBOL FEATURE. NOTE THE DROP IN NDCG SCORE WHEN THE STOCK SYMBOL FEATURE IS DROPPED.

NDCG	Stock 1	Without Stock
NDCG@25	.9698	.6313
NDCG@50	.9337	.6433
NDCG@75	.855	.6404
NDCG@100	.8082	.6492

TABLE VII
RESULTS OF OUR NEW FEATURE SET WHEN TRAINING AND TESTING ON THE 2013 DATASET WITH 4-FOLD CROSS VALIDATION. NOTE NDCG SCORE IMPROVES MARGINALLY WHEN THE STOCK SYMBOL FEATURE IS DROPPED, EXCEPT FOR NDCG@100.

NDCG	Stock	Without Stock
NDCG@25	.7770	.7976
NDCG@50	.7391	.7489
NDCG@75	.7287	.7359
NDCG@100	.7147	.6492

in Tables VII and VIII. Table VII shows the results when training and testing with the 2013 dataset. Ranker performance increases slightly after removing the *hasStockSymbol* feature. Table VIII shows that ranker performance improves slightly after removing the *hasStockSymbol* feature that seems to overfit the 2011 dataset.

In the previous work from Gupta, they yield an NDCG score of 0.3783 when training on the 2011 dataset and testing on the 2013 dataset. With our updated feature set, we yield an NDCG score of .6998 on the same dataset, as shown in Table VIII.

It is important to note that the topics in 2013 are almost entirely disaster or otherwise high impact scenarios (the exception being the Polar Vortex event.) The 2011 dataset contains disaster scenarios as well as other less serious events, which may include topics where the stock symbol may arise. We want to see if we can train on the 2013 dataset, and test on the 2011 dataset with adequate performance since the 2011 dataset encompasses similar topics to the 2013 dataset, but not vice versa. In Table IX we show the results of training on the 2013 dataset that is almost entirely high impact and disaster scenarios, and testing on the 2011 dataset that has many topics that are not high impact or disaster scenarios.

TABLE VIII
RESULTS OF OUR NEW FEATURE SET TRAINING ON THE 2011 DATASET AND TESTING ON THE 2013 DATASET

NDCG	Stock	Without Stock
NDCG@25	.5784	.6286
NDCG@50	.6407	.6637
NDCG@75	.6342	.7033
NDCG@100	.6641	.6998

TABLE IX
RESULTS OF OUR NEW FEATURE SET WHEN TRAINING ON THE 2013 DATASET AND TESTING ON THE 2011 DATASET

NDCG	Score
NDCG@25	.7601
NDCG@50	.6923
NDCG@75	.6810
NDCG@100	.6688

VI. CONCLUSION

We have derived a set of features that are indicative of a tweet's credibility regardless of the time period and topic of that tweet. Our set of features were derived by combining popular and effective features from previous works, as well as deriving new features. Features can be broadly categorized as either user based features or tweet based features. We create sentiment based features that aim to capture when a tweet's sentiment is anomalous given the context of its topic. We create grammar based features that aim to capture when unusual grammar is used given that typical grammatical nature of a tweet does not adhere to formal grammar. We also apply common tweet based features, such as the number of curse words, the number of positive words, and the number of negative words to the optional description Twitter users may write about themselves. With our new feature set, we are able to train and test on datasets from two different time periods and improve upon the results from the state of the art.

VII. FUTURE WORK

One of the challenges of doing research in this domain is that there is no universally accepted definition of credibility. In the previous works, each approach had their own definition of credibility that the human annotators use. This means that people who use a different definition for credibility cannot meaningfully compare results with one another. We need to develop a standard definition for credibility so that we can meaningfully compare results.

In order to show the true robustness of our feature set, we need to evaluate our approach on larger and more general datasets. In Mitra, [37] they create an annotated tweet database that is much more general and diverse than the datasets used in this work. By evaluating our approach on this data, we can further quantify how general and robust our features are.

We also observe that these features can be exploited for malicious purposes. Malicious users who wish to pass on non-credible information via Twitter can go undetected by the automatic techniques discussed in this paper by simply injecting the features into tweets which usually indicate credibility. For instance, the body of the tweet can contain false information but contain a URL. In this paper and previous works, a URL usually indicates high credibility. In future work, we want to try and derive more robust and less vulnerable features that are not susceptible to exploitation.

ACKNOWLEDGMENT

We would like to thank Hemant Purohit, Wenbo Wang, and Pramod Anantharam for enlightening discussions and advice. We would also like to thank Carlos Castillo and Aditi Gupta for sharing their annotated tweet database.

We acknowledge partial support from the National Science Foundation (NSF) award: EAR 1520870: Hazard-SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

[1] A. Acar and Y. Muraki, "Twitter for crisis communication: lessons learned from Japan's tsunami disaster," *International Journal of Web Based Communities*, vol. 7, no. 3, pp. 392–402, 2011.

[2] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response," *Proc. of ISCRAM*, 2014.

[3] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of Twitter during a natural disaster: Case study of 2011 Thai Flood," in *2012 Proceedings of PICMET'12: Technology Management for Emerging Technologies (PICMET)*, pp. 2227–2232, IEEE, 2012.

[4] H. Purohit, A. Hampton, S. Bhatt, V. L. Shalin, A. P. Sheth, and J. M. Flach, "Identifying Seekers and Suppliers in Social Media Communities to Support Crisis Coordination," *Computer Supported Cooperative Work (CSCW)*, vol. 23, no. 4-6, pp. 513–545, 2014.

[5] T. Bodnar, V. C. Barclay, N. Ram, C. S. Tucker, and M. Salathé, "On the ground validation of online diagnosis with Twitter and medical records," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pp. 651–656, International World Wide Web Conferences Steering Committee, 2014.

[6] R. Krishnamurthy, P. Kapanipathi, A. P. Sheth, and K. Thirunarayan, "Knowledge enabled approach to predict the location of twitter users," in *The Semantic Web. Latest Advances and New Domains*, pp. 187–201, Springer, 2015.

[7] J. Mahmud, J. Nichols, and C. Drews, "Where Is This Tweet From? Inferring Home Locations of Twitter Users.," in *ICWSM*, 2012.

[8] T. Rao and S. Srivastava, "Analyzing stock market movements using Twitter sentiment analysis," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 119–123, IEEE Computer Society, 2012.

[9] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through Twitter; I hope it is not as bad as I fear," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[10] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," in *Proceedings of the 20th International Conference on World Wide Web*, pp. 675–684, ACM, 2011.

[11] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, p. 13, ACM, 2012.

[12] M. Gupta, P. Zhao, and J. Han, "Evaluating Event Credibility on Twitter," in *SDM*, pp. 153–164, SIAM, 2012.

[13] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on Twitter," in *Proceedings of the 2012 ACM international Conference on Intelligent User Interfaces*, pp. 179–188, ACM, 2012.

[14] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao, "Information credibility on Twitter in emergency situation," in *Intelligence and Security Informatics*, pp. 45–59, Springer, 2012.

[15] C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media," *Internet Research*, vol. 23, no. 5, pp. 560–588, 2013.

[16] A. Gupta and P. Kumaraguru, "Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? IIIT," tech. rep., Delhi, Technical report, IIITD-TR-2011-005, 2011.

[17] A. Gün and P. Karagöz, "A Hybrid Approach for Credibility Detection in Twitter," in *Hybrid Artificial Intelligence Systems*, pp. 515–526, Springer, 2014.

[18] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman, "Challenges of computational verification in social multimedia," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pp. 743–748, International World Wide Web Conferences Steering Committee, 2014.

[19] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter," *CoRR*, vol. abs/1405.5490, 2014. Accessed May 2014.

[20] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy," in *Proceedings of the 22nd International Conference on World Wide Web Companion*, pp. 729–736, International World Wide Web Conferences Steering Committee, 2013.

[21] A. Gupta, H. Lamba, and P. Kumaraguru, "$1.00 per rt# boston-marathon# prayforboston: Analyzing fake content on Twitter," in *eCrime Researchers Summit (eCRS), 2013*, pp. 1–12, IEEE, 2013.

[22] 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*, pp. 441–450, ACM, 2012.

[23] S. Shariff, X. Zhang, and M. Sanderson, "User Perception of Information Credibility of News on Twitter," in *Advances in Information Retrieval*, pp. 513–518, Springer, 2014.

[24] 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*, p. 2, ACM, 2012.

[25] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589–1599, Association for Computational Linguistics, 2011.

[26] Y. Ikegami, K. Kawai, Y. Namihira, and S. Tsuruta, "Topic and Opinion Classification Based Information Credibility Analysis on Twitter," in *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4676–4681, IEEE, 2013.

[27] Y. Suzuki, "A credibility assessment for message streams on microblogs," in *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2010 International Conference on*, pp. 527–530, IEEE, 2010.

[28] S. Ravikumar, R. Balakrishnan, and S. Kambhampati, "Ranking tweets considering trust and relevance," in *Proceedings of the Ninth International Workshop on Information Integration on the Web*, p. 4, ACM, 2012.

[29] R. Al-Eidan, H. Al-Khalifa, and A. Al-Salman, "Measuring the credibility of Arabic text content in Twitter," in *2010 Fifth International Conference on Digital Information Management (ICDIM)*, pp. 285–291, IEEE, 2010.

[30] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (New York, NY, USA), pp. 168–177, ACM, 2004.

[31] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Cursing in english on twitter," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, (New York, NY, USA), pp. 415–425, ACM, 2014.

[32] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, pp. 315–324, Springer, 2006.

[33] T. Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 217–226, ACM, 2006.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at your mom.

[35] C.-P. Lee and C.-J. Lin, "Large-scale linear ranksvm," *Neural Computation*, vol. 26, no. 4, pp. 781–817, 2014.

[36] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[37] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations," in *Proceedings of the 9th International Conference on Web and Social Media, Oxford, UK*, 2015.