Increasing the Veracity of Event Detection on Social Media Networks Through User Trust Modeling

Conference Paper · October 2014

DOI: 10.1109/BigData.2014.7004286

CITATIONS
51

READS
713

4 authors, including:

Conrad Tucker
Carnegie Mellon University
167 PUBLICATIONS **3,160** CITATIONS
SEE PROFILE

Sven G. Bilén
Pennsylvania State University
316 PUBLICATIONS **2,834** CITATIONS
SEE PROFILE

Increasing the Veracity of Event Detection on Social Media Networks Through User Trust Modeling

Todd Bodnar
Center for Infectious
Disease Dynamics
Pennsylvania State University
University Park, Pennsylvania 16802
Email: ToddBodnar@gmail.com

Conrad Tucker
Engineering Design, Industrial Engineering
Computer Science and Engineering
Pennsylvania State University
University Park, Pennsylvania 16802
Email: ctucker4@psu.edu

Kenneth Hopkinson
Electrical and Computer Engineering
Air Force Institute of Technology
Wright-Patterson Air Force Base
Fairborn, Ohio 45433
Email: Kenneth.Hopkinson@afit.edu

Sven G. Bilén
Engineering Design, Electrical Engineering
Pennsylvania State University
University Park, Pennsylvania 16802
Email: SBilen@engr.psu.edu

***Abstract*—With the success and ubiquity of large scale, social media networks comes the challenge of assessing the veracity of information shared across them that inform individuals about emerging real-world events and trends. We propose a veracity-assessment model for information dissemination on social media networks that combines natural language processing and machine learning algorithms to mine textual content generated by each user. Large scale social media networks (such as Twitter and Facebook) are considered digital communication platforms, in which information can be quickly and easily exchanged, thereby expanding the breadth of knowledge across the globe. In this paper, four case studies spanning multiple geographic regions, threat scenarios and time frames are investigated, in order to demonstrate how real-world events impact the manner in which information/misinformation is communicated and spread through a social media network. Our results show that metadata associated with each user can provide significant insight on the social media network's users' tendency to accurately discuss a topic.**

I. Introduction

Society generates more than 2.5 quintillion (10^{18}) bytes of data each day [1]. A significant amount of this data is generated through mobile and social media services such as Twitter, Facebook, and Google that process anywhere between 12 terabytes (10^{12}) to 20 petabytes (10^{15}) of data each day. These online user-propelled networks are now being referred to as digitized word of mouth networks [2], and have been successful in shaping everything from global politics [3] to financial markets [4]. Big data has emerged as a critical research topic in the information age and is typically defined as high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [5]. While varying definitions exist throughout the data mining community, the converging characteristics of big data are *volume*, *velocity*, *variety*, *veracity*, and *value* [6] with *volume* being the ever increasing size and magnitude of data; *variety* being the multiple forms and characteristics of data; *velocity* being the speed at which data are being generated, *veracity* referring to the accuracy and truthfulness of the data as well as the ability of the data to predict trends; and *value* being the cost–benefit of what the data provide [7].

Social media networks such as Twitter exhibit characteristics of "Big Data" and present challenges and opportunities for knowledge discovery relating to threat detection. For example, there are more than 500 million tweets generated by Twitter each day (*volume*) [8] corresponding to roughly one message every 173 microseconds (*velocity*). These tweets contain a combination of textual content, geo-location data, video, etc. (*variety*). The authors have assessed the *veracity* of social media networks in a range of applications ranging from predicting product demand [9] to medical diagnoses [10]. Such insights can be critical to companies looking to minimize product failures; hence, the reason why many companies continue to invest in customer relations management via social media (*value*) [11], [12]. However, the objective of discovering threat events in social media is not as trivial as "searching all messages, *m* that contain the threat word(s) **X**", due to several challenges:

- The heterogeneity in the manner in which individuals communicate. For example, one individual may express a threat event in a more direct manner "An explosion just went off at location X," while another individual may express a threat event in a more indirect manner, such as: "What's wrong with humanity, why can't we all just get along?"
- The ability of individuals expressing threat information in a social network to accurately decipher real versus fake events. For an example, a user saying

The views expressed in this document are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

"Another earthquake in San Francisco today" could either be real or fake.

- The speed at which information (fake or real) spreads through a social network and influences others within the network. For example, a hoax about a bombing at the White House was able to spread enough to cause a 13-point drop in the stock market within 5 minutes.

If each user in a Social Media Network is assumed to represent a classifier that takes various information and classifies an event as real or a hoax, we find that they are biased towards believing events are happening. We define the *trustworthiness* of a user by comparing his or her classification of an event to the actual value of the event. Just like machine learning based classifiers, not every user will necessarily perform well. For example, humans spread chain letters that falsely claim that bad things will happen to the receiver if not spread further. However, previous work assumes that each user's classifications are accurate or that each user has the same accuracy. Here, we question that assumption by considering events where the population, as a whole, was not immediately aware of whether or not the event was really happening or just a hoax. By analyzing how users, who were correctly skeptical during a hoax or correctly accepting when the event was real, differ from other users that did not accurately predict whether or not the event was real, we are able to generate trust profiles that can serve as a guide for analyzing events in which it is not yet certain whether or not the event is real.

The paper is structured as follows. In Section II we review previous work about trust management. In Section III we describe two traditional trust metrics and introduce a novel metric based on a user's profile information. In Section IV we apply these metrics to security-related events and compare their results. In Section V, we compare the differences between event detection on the original data stream and one that uses our trust metric as a filter. In Section VI we remark on the results from our case study and present conclusions.

II. Related Work

A. Mining Social Networks

Social network platforms have been the basis for novel information collection in a large range of domains. For example, Google searches have been used to track disease rates [13], Facebook data has been used to study survival rates of relationships [14], and Twitter has been used to detect FIFA World Cup goals [15]. These platforms have matured at the same time as big data. Indeed, since these systems generate terabytes of data each day [1], analysis requires modern processes such as scalable databases or a MapReduce methodology. However, for more specific analysis to be performed, the data must be filtered. This proves a challenge. In some applications, such as spam filtering, there is an advisory that is constantly attempting to find ways to bypass filters [16], [17]. In other applications, the relatively small size of an individual message provides complications as whether or not the user was being sarcastic can be hard to discern even by a human rater [18].

B. Event and Anomaly Detection

Anomaly detection aims to find irregular patterns that do not conform to normal patterns in data. Anomaly detection techniques have been proposed to address many applications, ranging from fraud detection [19], [20], [21], [22] (e.g., credit card fraud detection) to network intrusion [23], [24], [25], [26]. The importance of anomalies (good or bad) is their ability to provide high information throughput about an emerging event that deviates from that which is expected. From a data mining perspective, anomalies are data points that have a certain deviation from a distribution of known points. However, quantifying the distance of deviation that classifies a data point as anomalous is non-trivial. Hodge and Austin [27] provide an extensive survey of machine learning and statistical techniques on anomaly detection. In statistics, the classification of variation into common cause and special cause variation emphasizes the fact that deviations, while important, should not immediately be classified as anomalous that warrant action, as one could end up aiming to resolve deviations (i.e., common cause) that are inherent to the system itself. In an effort to categorize anomalies based on their inherent properties, Chandola et al. [28] propose three types of anomalies: Point Anomalies (an individual data instance that deviates from the rest of the data instance); Contextual Anomalies (a data instance that deviates; given a specific context but not otherwise); and Collective Anomalies (a collection of related data instances is anomalous with respect to the entire data set, but not on its own).

While extensive research has been performed on anomaly detection, the techniques proposed in the literature are not well suited for anomaly detection in large scale social media networks. Unlike traditional networks such as a banking network infrastructure that serve to perform a well-defined task (i.e., enable the secure and reliable management of finances), social media networks do not have a well defined objective. For example, Twitter can be used as a news source [29], a platform for humor and comedy [30], a venue to discuss politics [31], etc. Anomalous patterns in how humans communicate (e.g., sarcasm), also make anomaly detection challenging in large scale social media networks [32]. Thom et al. [33] have proposed a spatiotemporal anomaly detection model for mining Twitter messages. However, their approach requires term-specific instances of Twitter messages and input from a domain expert to determine what is considered a threshold for an anomalous event. Anantharam et al. [34] propose a method to detect anomalies in Twitter messages, based on the URLs that messages reference. However, many social media network messages contain only plain text and may not necessarily contain hashtag, URL, or other uniquely indefinable features.

Given the fact that social media networks are not context specific and contain unstructured, free-flowing textual content, existing methods are not well suited for anomalous events that occur beyond a specific time frame or context (e.g., terrorist event). We propose a method that can quantify the veracity of information spread through the network that is not context dependent, thereby enabling a wide variation of anomalies (real events occurring, hoaxes being propagated, etc.) to be detected in large sale social networks.

C. Gossip Protocols on Distributed Systems

Social networks can be seen as large-scale distributed systems and there are many different kinds of trust models and mechanisms that have been built for such systems [35],

[36]. One challenge in such large-scale systems is how to keep information flow coordinated with low overhead. Gossip protocols are often employed in such situations, ranging from straightforward uniform techniques [37] to more sophisticated stratification of gossip targets and probabilities [38], [39]. By building on top of these or other types of scalable communication mechanisms, reputation-based trust can be used to protect key assets, including cyberphysical critical infrastructure protection systems [40]. This same genre of technique can also be applied to the types of social media networks discussed in this article.

D. Alternative Trust Metrics

Previous work on filtering out messages have often focused on spam filters [17], which usually focus on the textual content of messages to determine if an event is spam or not spam (commonly referred to as "ham"). While these systems have been fairly successful, focusing on textual content will not work for our application. Spam is less context dependent: an email about purchasing Viagra from questionable sources will always be spam, but whether or not a message about an ongoing terrorist attack is accurate is dependent on whether or not there *is* an ongoing terrorist attack. Before proposing our own solution to this problem, we discuss two previously derived solutions.

It has been argued that individuals closer to an event are more likely to accurately report on the event [41]. Intuitively, this makes sense. If an individual is able to directly observe an event, he or she should be able to accurately infer if it is real or not. However, in most cases, this will be a very small subset of any event, unless the event is observable from several geographical regions at once (e.g., a lunar eclipse). On the other hand, individuals farther away from an event likely heard about the event through news sources, which may perform verification before reporting on an event. However, the demand for real time reporting of events may lower the veracity of these news sources [42].

Although social media has been heralded for the speed at which events can be reported, there are limits to how quickly these reports can be trusted. That is, an event detection system, that determines that there is an event the moment the first message is received, will be less accurate than one that patiently waits for more data to be collected. For example, Sakaki et al. [43] describe a method for reducing false reports of earthquakes by waiting based on the number of users reporting the event. Similarly, it is likely that a message that is posted later may be more accurate because the person reporting it has more time to gather information about the event before posting it. There is a trade off, however, as messages that are posted later are less valuable to an early warning system.

III. Methods

Here, we develop a system to determine the trustworthiness of a user's posts based on his or her social network profile (see Figure 1). That is, given that a user is discussing that an event is real or a hoax, how likely is it that the event is *actually* real or a hoax. This is done by a two step process. First, the obtained user profile data is *transformed* into a feature vector. Then, a *classifier* is applied to the feature vector to determine whether or not the user's posts are to be trusted. To avoid issues associated with arbitrarily choosing data transformation and classification methods, we describe variations on several key stages of the process and combinatorially build a model in each point of parameter space. These models are then evaluated and compared against each other. The strongest performing model is then chosen as our trust metric.

A. Data Transformation

We begin by collecting each user's profile data. The metadata associated with each user contains numeric and textual data. In the case of Twitter user profiles, this data includes the number of friends and followers they have, the number of messages a user has posted, the number of tweets a user has favored, the number of the user's tweets that other users have favored, the user's website (stored as a binary variable, does he or she have one?), and the user's description text. Variables related to a user's behavior are often log distributed [44]. Since many classifiers assume that variables are normally distributed, we consider transforming the numeric variables (number of friends, followers, tweets, favorites, and favored) by $\log(a+1)$ where a is the attribute's value and the addition of one is to avoid taking the log of zero.

Unlike numeric data, textual data must be transformed into a form that can be processed by a machine-learning algorithm. First, the text is converted into lower case and tokenized based on non-alphanumeric characters. We either stem these tokenized words with Iterated Levins stemming [45] or we do not stem the words. We then consider generating n-grams (where $n = 1, 2, 3$). We define an n-gram as a set of at most n concurrent words. For example, the 2-grams generated from the phrase "White House party" would be "White", "House", "party", "White_House", "House_party", and "White_House_party". We then generate an n-gram occurrence vector where entry i is 1 if the description field contains n-gram i and zero otherwise The total list of n-grams was generated from the set of all user's descriptions in the training set.

This list of n-grams can be excessively large, so we reduce it by one of two methods. First, we considered using only the 100, 500, or 1000 most common n-grams. Second, we considered dimensionality reduction. This was done by either ranking the n-grams based on information gain, performing principal component analysis, or performing latent semantic analysis. In these cases, we retained the 10, 50, or 100 most descriptive variables. By iterating through all combinations of text value generation, we generate 72 different methods for transforming textual data into binary data, which, when combined with the user's other profile information, are each tested as the input for our classification algorithm.

B. User Trust Classification

We consider 11 different classifiers—implemented through Weka [46] version 3.7.10—to apply to our datasets: Zero-Rule (guessing the mode), One-Rule, J48 (a C4.5 implementation), random forests, random trees, AdaBoost, support vector machines with a linear kernel, support vector machines with a Gaussian kernel, naïve Bayes classifiers, logistic regression, and multi-layer perceptrons. An in depth discussion of these

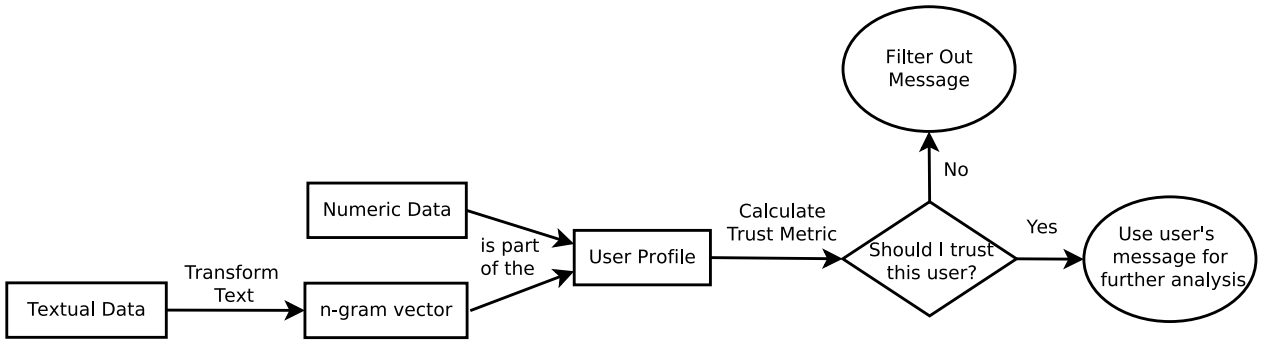

Fig. 1. Graphical representation of our system for filtering out untrustworthy users.

classifiers is available at [47]. We select three of the four events as the training set and test on the fourth. To avoid biasing the classifiers toward events with more users, a random sample of 100 users are taken from each of the three training sets, resulting in a total of 300 instances to train on. Each classifier is trained on each combination of input transforms and tested on each of the four datasets. To account for randomly selected subsets of training sets, each test is repeated multiple times.

We then determine the classifier data transformation combination that performs the best over all four training sets. Since these datasets can be very biased towards being accurate or inaccurate predictions of an event's truthfulness, we measure performance based on area under the ROC curve instead of simple accuracy. The end result of this process is a machine-learning algorithm that can serve as a filter to remove messages that come from users that may not be trustworthy.

C. Training Set Collection

We begin with a set of messages about an event of interest from which we want to determine their trustworthiness. These messages are hand rated by the authors based on whether it was likely that the user, who posted the message, believed that the event was *real* or that the event was a *hoax*. Examples of a real and hoax tweet for each event are given in Table I. However, one cannot assume that all social media users are accurate in classifying the event, necessitating the development of a filtering system to determine which messages are more trustworthy. Indeed, as shown later, users often report an event as real even if it is not happening. We define the user as being *accurate* if and only if he or she sends a message claiming an event is real or fake *and* that claim matches the real outcome of an event.

IV. Case Study

In this case study, we consider four security related events as a case study to evaluate the three trust metrics described above.

A. Event Selection

In this paper, we consider four security related events in the United States, two of which were hoaxes. The two *real* events were the bombing of the 2012 Boston Marathon[1] and the shooting in a movie theater during the midnight premier of the Batman movie, *The Dark Knight Rises*,[2] also in 2012. The two *hoax* events chosen were the alleged bombing of the White House in 2013, as reported by a news agency's hacked Twitter account, and a bomb threat at Harvard[3] during finals week in 2013. These four events had extensive news coverage and impact. For example, reports of explosions at the White House resulted in a 13-point drop in the stock market[4] despite being a hoax. It is believed that this drop was caused by automated investment systems that used Twitter streams in their analysis, which apparently did not use a trust system such as the one proposed here.

These events were paired with tweets collected during the two-hour time span surrounding the event.[5] Tweets for the Boston Marathon bombing were selected by searching for tweets containing the word "Boston", ignoring case. The Batman Shooting tweets were selected by searching for tweets containing the string "kill", "shoot", "bomb", or "Aurora". Note that the shooting did not involve bombs, but the search term is included because of initial media reports of a bombing. The White House explosion hoax tweets were selected by tweets that contained "Obama", "Whitehouse", or "White House". Finally, tweets involving the Harvard bomb hoax were selected by searching for tweets containing the word "Harvard". These search terms were the initial filtering process and attempted to collect as many relevant tweets as possible (high recall) with little concern to falsely selecting irrelevant tweets (low precision). These messages were then filtered by hand to remove irrelevant tweets. We find that, in all four events, the number of messages about the event actually happening is greater than the number of skeptical messages, even during events that were hoaxes (see Figure 3).

B. Comparing Metrics

To determine the relative performance of our model, we consider three alternative models to compare our model against. First, we define a baseline model that assumes users should be trusted completely. This is equivalent to not filtering out any untrustworthy individuals from the dataset. Second,

[1] http://bigstory.ap.org/article/two-explosions-boston-marathon-finish-line-0

[2] http://www.denverpost.com/ci_21124893/12-shot-dead-58-wounded-aurora-movie-theater

[3] http://www.bostonmagazine.com/news/blog/2013/12/16/harvard-university-warns-of-explosives-twitter/

[4] http://buzz.money.cnn.com/2013/04/23/ap-tweet-fake-white-house/

[5] Due to the large reaction, tweets related to the Boston bombing were limited to a one-hour time span.

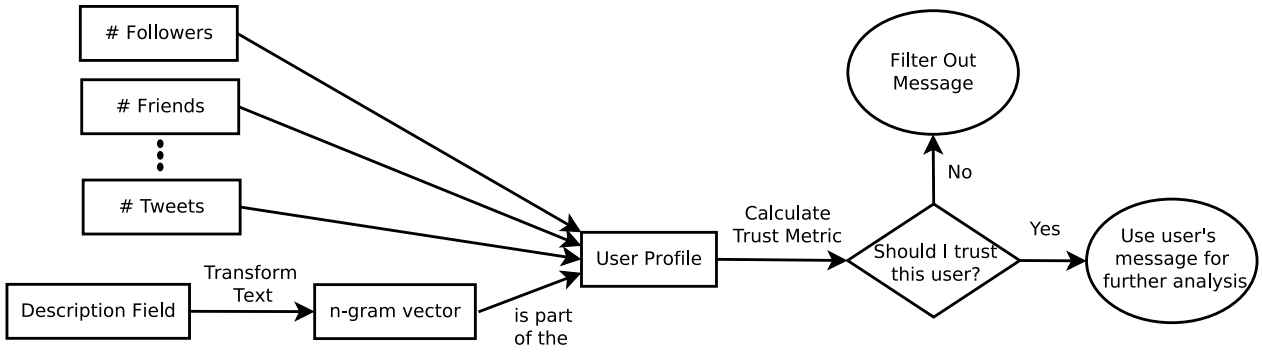

Fig. 2. An implementation of our system for filtering out untrustworthy users applied to Twitter users.

TABLE I. EXAMPLE OF MESSAGES FROM USERS THAT APPEAR TO BELIEVE AN EVENT IS *REAL* OR A *HOAX* BASED ON THEIR TWEETS.

Event	Belief	Example Tweet
Batman	**Real**	I'm listening to "Aurora Police and Fire" using the Scanner Radio app on my Android phone. #crazy #bombing #shooting
	Hoax	Woow there was noo shootings tonight ??????
Boston	**Real**	Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened?
	Hoax	I bet you the Boston Marathon explosion was another false-flag op by you criminal govt! These bastards are sick!
White House	Real	"@AP: Breaking: Two Explosions in the White House and Barack Obama is injured" OMG!!!!! PRAY FOR PEACE!
	Hoax	This is why "password1" is a bad password. RT @AP: Breaking: Two Explosions in the White House and Barack Obama is injured
Harvard	Real	BREAKING: explosions at Harvard. Science Center, Thayer, Sever, and Emerson.
	Hoax	My best guess is some Harvard College freshman really didn't want to take an exam today. With luck they'll identify the threat maker

we consider geographical distance from the event. Third, we account for the time between an event and when a message is posted. Note that these two other metrics cannot be incorporated into our model as our model is based solely on metadata provided with the message and not external variables such as the event's epicenter. Finally, we discuss the accuracy of our model on the four case-study events.

1) Baseline Metric: The baseline model is trivial to implement, but it is included for completeness. Even for the baseline model, we assume that some form of filtering out irrelevant messages is included. The accuracy of the baseline is the percentage of users in each database that accurately infer the event's truth. This can be exceptionally high, for example 99.8% of user's correctly inferred that the Boston bombing was real, or exceptionally low, for example 24.6% of user's correctly inferred that the Harvard bomb hoax was a hoax. Measuring area under the ROC curve corrects for this bias, and the baseline model only has an area of 0.5.

2) Distance-Based Metric: We determine trust levels based on distance by considering tweets with geographical information and determine their distance from the epicenter of the event that they are about. That is, we calculate the *distance*, in degrees, of the shortest path between the message. We then build a logistic regression model

$$p(\text{Message is Accurate}) = \left(1 + e^{-(\beta_0 + \beta_1(\text{distance}))}\right)^{-1} \quad (1)$$

TABLE II. EFFECTS OF TIME ON MESSAGE ACCURACY.

Event	β_0	β_1	p	80% Accuracy (min.)
Batman	1.54921	0.01512	0.241	NA
Boston	8.26294	0.03131	0.283	NA
White House	-0.96155	0.06612	1.02×10^{-5}	6.42
Harvard	-3.1949	0.014216	0.03425	204.50

where β_0 and β_1 are weights determined by fitting the data. We calculate the distance between the place where the tweet was posted and either the Century 16 multiplex in Aurora, the finish line of the Boston Marathon, the White House, or the center of the Harvard campus by the shortest arc that connects the two points. However, we fail to find a statistically significant relationship between the distance between the two points and whether it accurately measures the event ($p > 0.087$ in all cases), with Šidák correction for multiple tests applied. Thus, we ignore geographical knowledge in our model.

3) Time-Based Metric: As with the case of distance, we build a logistic regression model based on the number of minutes after an event has occurred to predict whether or not the message is accurate. Since we are studying events that, by their nature, are not announced beforehand, we do not need to consider the trustworthiness of messages that occur *before* the event occurs. Indeed, in this case study, we do not find any messages before the event discussing it, as expected.

The two *real* events (the Batman shooting and the Boston bombing) were consistently accurate regardless of time, thus messages about these two events could be considered accurate at least 80% of the time regardless of post time, by definition. We find a significant effect of time on the trustworthiness of a message (see Table II): after 6.42 minutes, tweets about the White House bombing were correctly determining that the event was fake 80% of the time, a speed which may be faster than the rate that a news agency would adapt. Finally, in the case of the Harvard bomb threat, our model predicts that messages were accurate after 3 hours 34.5 minutes. Future work will explore the causes for the differences in times to determine that the event is a hoax.

4) Proposed Trust Metric: We trained a total of 31,680 models on the user's profile information. We then selected the model that has the highest area under the ROC curve when averaged between the 3 repeats and 4 datasets. The model selected is the random tree classifier where only 1-grams are used, the numeric data is log transformed, iterated Levins stemming is applied, and the 50 unigrams with the highest

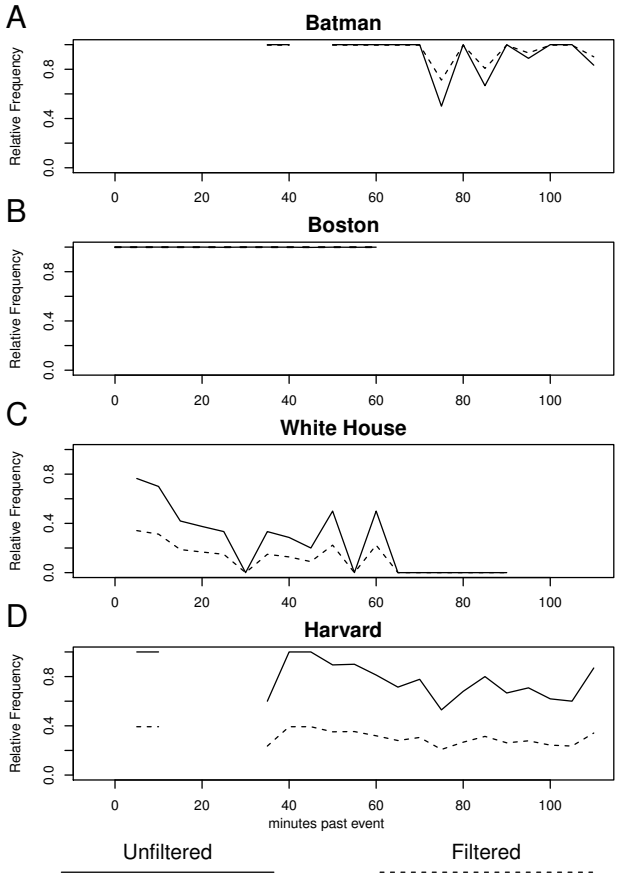

Fig. 3. Normalized rates of tweeting for the four events. Note that the two events that are real (Batman and Boston) maintain high levels of relative frequency of tweets saying the event is real, while the two events that are not real (White House and Harvard) do not. Graphs are aligned based on when the event occurred.

TABLE III. COMPARISON OF THE ABILITY FOR THE THREE METRICS TO DETERMINE TRUSTWORTHINESS.

Metric	Mean Accuracy	Mean Area Under ROC curve
Baseline	67.80%	0.50
Proximity	60.64%	0.50
Time	55.68%	0.53
Profiles	75.41%	0.73

information gain are retained. This classifier has a mean area under the ROC curve of 73.14% and mean accuracy of 75.41% on the test data. A table of results for of all 1,584 combinations (of classifiers, text transformation, and numeric transformation) is not included due to space constraints.

We compare the area under the ROC curve for each of these three metrics (see Table III). Since the distribution of *real* and *hoax* tweets can be very biased, accuracy should not be used as a measure of performance, but it is included for completeness. As with the profile-based metric, the other two metrics were trained on three events and tested on the remaining. That is, the unfiltered dataset. We observe that our trust metric out preforms the other two metrics by between 38% and 46%.

C. Detecting Security Related Events

We observe the proportion of tweets in each dataset that are from a user that believes the event is real for each 5-minute window for each event (see Figure 3, empty areas in the graph indicate no relevant discussion at that time.) The proportion of tweets from a user that believes the event is happening for the two real events (Batman and the Boston bombing) are on average of 92.59% and 99.94%, respectively (see Figure 3.a. and 3.b., respectively). This is higher than the two *hoax* events have lower proportions of *real* tweets, with an average of 25.95% of the White House explosion tweets (see Figure 3.c.) and an average of 78.74% of the Harvard tweets claiming the event is true (see Figure 3.d.) and being true. Note that, while the mean is relatively low for the entire duration for the White House explosion tweets, the majority (76.47%) of tweets in the first five minutes after the event considered the event to be true.

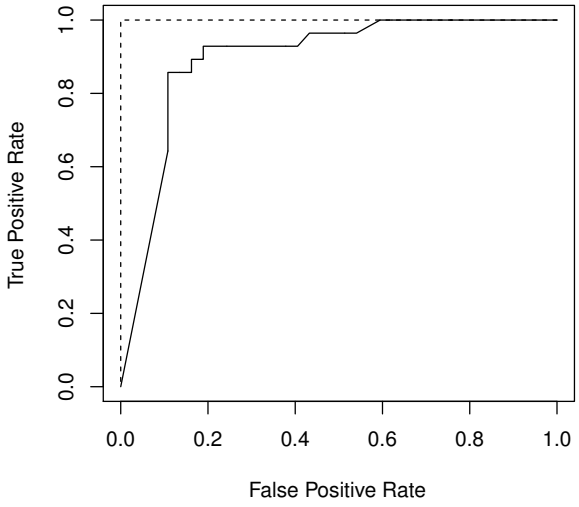

Fig. 4. Event detection accuracy for unfiltered (solid line) and filtered (dashed line) messages.

The classifier, determined in the previous section, is then applied to the data sets. This removes users from the data that are classified as untrustworthy, along with their tweets. This results in a dampening of messages that claim an event is happening during hoax events. The mean proportion of *real* tweets in the White House event is reduced from 25.95% to 8.09%. Similarly, the mean proportion for the Harvard bomb hoax was reduced from 78.74% to 30.91%. Additionally, the evidence from Twitter for the two real events increased slightly from 99.93% to 99.95% for the Boston bombing and 92.59% to 95.38% for the Batman shooting.

Now we consider detecting events through a simple, rule-based algorithm. As discussed above, this system is based on normalized frequencies of tweets claiming an event to be happening. Here, we define an event as occurring when the rate of messages is over a cut-off, x. A higher value of x results in more false positives but a lower value of x results in more false negatives. The trade-off between these two attributes can be determined by an analysis of the differing costs between these two errors. Hence, we can plot the ROC curve for unfiltered and filtered messages (see Figure 4), and note an increase in

performance over the unfiltered data.

V. Discussion and Conclusion

In this paper, we proposed a novel solution to improving the trust of social media streams that is generalizable to multiple events. We tested our machine learning algorithms on events that were independent of the training sets—compared to, for example, combing all messages into one dataset and performing k-fold cross validation—we can be relatively confident that our system would perform similarly well on new events.

Traditionally, researchers have assumed that an individual's messages should be trusted as accurate. In this paper, we propose a novel solution to improving the trust of social media streams. Trust measures based on either spatial or temporal proximity to an event require external knowledge about the event before trust levels can be inferred. That is, with location-based or time-based trust models, one would have to know when or where the event which is discussed in the message either algorithmically—a problem that has been shown [48] to be difficult to accurately perform—or through manual methods, which will be relatively slow. Alternatively, this system is based solely on metadata which is included with each message as it is posted. Since our system does not require secondary sources to determine the trust of the user (either by pinpointing the event's location or time in the other two cases), real-time event detection is feasible. Although, as shown above, our system out-performs trust models based on time or distance even if we allow for perfect knowledge about the event's time and location. However, note that the initial influx of messages for each event (see Figure 3) have some lag from the event's occurrence, limiting the potential for absolute-real-time event detection. The exact length of this lag may depend on several factors. For example, the Boston Marathon was being covered by the press and other people at the time of the bombing while news about the Batman shooting, which did not occur during a news worth event (a movie showing), took longer to disseminate.

Our system is 75% accurate, resulting in some user's being misclassified as trustworthy and noise still getting through the filter. This is to be expected, however, as it is unlikely that even the most trustworthy individual will be accurate 100% of the time. Furthermore, our system scales better than a human based classifier as the models are trained *a priori*. Our system does show that some users have traits related to their social network that allow us to judge their trustworthiness, opening the door to future work that could improve on trust classification schemes based on social network profiles. Indeed, this system could be extended to a wide range of social network trust applications such as determining a "tabloid" rating for news articles or predicting how trustworthy a job applicant would be.

References

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97–107, 2014.

[2] C. Dellarocas, "The digitization of word of mouth: Promise and challenges of online feedback mechanisms," *Management Science*, vol. 49, no. 10, pp. 1407–1424, 2003.

[3] H. H. Khondker, "Role of the new media in the Arab Spring," *Globalizations*, vol. 8, no. 5, pp. 675–679, 2011.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[5] Y. Demchenko, C. Ngo, and P. Membrey, "Architecture framework and components for the big data ecosystem," *Journal of System and Network Engineering*, pp. 1–31, 2013.

[6] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," *Semantic Web*, vol. 4, no. 3, pp. 233–235, 2013.

[7] E. G. Ularu, F. C. Puican, A. Apostu, and M. Velicanu, "Perspectives on big data and big data analytics," *Database Systems Journal*, vol. 3, no. 4, pp. 3–14, 2012.

[8] R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr., C. Wilson, A. Mislove, and V. Almeida, "Of pins and tweets: Investigating how users behave across image-and text-based social networks," 2014.

[9] S. Tuarob and C. S. Tucker, "Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data," in *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2013.

[10] 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*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 651–656.

[11] D. L. Hoffman and M. Fodor, "Can you measure the ROI of your social media marketing?" *Sloan Management Review*, vol. 52, no. 1, 2010.

[12] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," *Business Horizons*, vol. 52, no. 4, pp. 357–365, 2009.

[13] H. A. Carneiro and E. Mylonakis, "Google Trends: A WebBased Tool for RealTime Surveillance of Disease Outbreaks," *Clinical Infectious Diseases*, vol. 49, no. 10, pp. 1557–1564, Nov. 2009.

[14] B. State. (2014) Flings or lifetimes? the duration of facebook relationships. [Online]. Available: https://www.facebook.com/notes/facebook-data-science/flings-or-lifetimes-the-duration-of-facebook-relationships/10152060513428859

[15] L. Cipriani. (2014) Goal! Detecting the most important World Cup moments. [Online]. Available: https://blog.twitter.com/2014/goal-detecting-the-most-important-world-cup-moments

[16] D. Mazieres and M. F. Kaashoek, "The design, implementation and operation of an email pseudonym server," in *Proceedings of the 5th ACM Conference on Computer and Communications Security*. ACM, 1998, pp. 27–36.

[17] B. Krause, C. Schmitz, A. Hotho, and G. Stumme, "The anti-social tagger: Detecting spam in social bookmarking systems," in *AIRWeb '08: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*. ACM, Apr. 2008.

[18] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," *IEEE Symposium on Visual Analytics Science and Technology*, pp. 115–122, 2010.

[19] E. Aleskerov, B. Freisleben, and B. Rao, "Cardwatch: A neural network based database mining system for credit card fraud detection," in *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, Mar 1997, pp. 220–226.

[20] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, Jan. 1997. [Online]. Available: http://dx.doi.org/10.1023/A:1009700419189

[21] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Networking, Sensing and Control, 2004 IEEE International Conference on*, vol. 2, 2004, pp. 749–754 Vol.2.

[22] C. Phua, V. C. S. Lee, K. Smith-Miles, and R. W. Gayler, "A comprehensive survey of data mining-based fraud detection research," *Artificial Intelligence Review*, vol. 24, 2005.

[23] C. Krügel, T. Toth, and E. Kirda, "Service specific anomaly detection for network intrusion detection," in *Proceedings of the 2002 ACM Symposium on Applied Computing*, ser. SAC '02. New York, NY, USA: ACM, 2002, pp. 201–208. [Online]. Available: http://doi.acm.org/10.1145/508791.508835

[24] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, "A comparative study of anomaly detection schemes in network intrusion detection," in *In Proceedings of the Third SIAM International Conference on Data Mining*, 2003.

[25] K. Leung and C. Leckie, "Unsupervised anomaly detection in network intrusion detection using clusters," in *Proceedings of the Twenty-eighth Australasian Conference on Computer Science - Volume 38*, ser. ACSC '05. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2005, pp. 333–342. [Online]. Available: http://dl.acm.org/citation.cfm?id=1082161.1082198

[26] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection," in *Recent Advances in Intrusion Detection*, ser. Lecture Notes in Computer Science, E. Jonsson, A. Valdes, and M. Almgren, Eds. Springer Berlin Heidelberg, 2004, vol. 3224, pp. 203–222. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30143-1_11

[27] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9

[28] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[29] A. Hermida, "Twittering the news: The emergence of ambient journalism," *Journalism Practice*, vol. 4, no. 3, pp. 297–308, 2010.

[30] A. E. Holton and S. C. Lewis, "Journalists, social media, and the use of humor on Twitter," *Electronic Journal of Communication*, vol. 21, no. 1/2, 2011.

[31] B. L. Nacos, "Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public," *Political Science Quarterly*, vol. 128, no. 1, pp. 178–179, 2013. [Online]. Available: http://dx.doi.org/10.1002/polq.12021

[32] R. González-Ibáñez, S. Muresan, and N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 581–586. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002850

[33] D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," in *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, Feb 2012, pp. 41–48.

[34] P. Anantharam, K. Thirunarayan, and A. Sheth, "Topical anomaly detection from Twitter stream," in *Proceedings of the 4th Annual ACM Web Science Conference*, ser. WebSci '12. New York, NY, USA: ACM, 2012, pp. 11–14. [Online]. Available: http://doi.acm.org/10.1145/2380718.2380720

[35] W. Sherchan, S. Nepal, and C. Paris, "A survey of trust in social networks," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 47:1–47:33, Aug. 2013. [Online]. Available: http://doi.acm.org/10.1145/2501654.2501661

[36] T. Grandison and M. Sloman, "A survey of trust in internet applications," *Commun. Surveys Tuts.*, vol. 3, no. 4, pp. 2–16, Oct. 2000. [Online]. Available: http://dx.doi.org/10.1109/COMST.2000.5340804

[37] K. P. Birman, M. Hayden, O. Ozkasap, Z. Xiao, M. Budiu, and Y. Minsky, "Bimodal multicast," *ACM Trans. Comput. Syst.*, vol. 17, no. 2, pp. 41–88, May 1999. [Online]. Available: http://doi.acm.org/10.1145/312203.312207

[38] J. Fadul, K. Hopkinson, T. Andel, and C. Sheffield, "A trust-management toolkit for smart-grid protection systems," *Power Delivery, IEEE Transactions on*, vol. 29, no. 4, pp. 1768–1779, Aug 2014.

[39] K. Hopkinson, K. Jenkins, K. Birman, J. Thorp, G. Toussaint, and M. Parashar, "Adaptive gravitational gossip: A gossip-based communication protocol with user-selectable rates," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 12, pp. 1830–1843, 2009.

[40] E. Rivire and S. Voulgaris, "Gossip-based networking for internet-scale distributed systems," in *E-Technologies: Transformation in a Connected World*, ser. Lecture Notes in Business Information Processing, G. Babin, K. Stanoevska-Slabeva, and P. Kropf, Eds. Springer Berlin Heidelberg, 2011, vol. 78, pp. 253–284. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20862-1_18

[41] A. Croitoru, A. Crooks, J. Radzikowski, and A. Stefanidis, "Geosocial gauge: A system prototype for knowledge discovery from social media," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 12, pp. 2483–2508, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1080/13658816.2013.825724

[42] S. R. Maier, "Accuracy matters: A cross-market assessment of newspaper error and credibility," *Journalism & Mass Communication Quarterly*, vol. 82, no. 3, pp. 533–551, 2005.

[43] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[44] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772751

[45] J. B. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, no. 11, pp. 22–31, 1968.

[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[47] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, third edition*. Morgan Kaufmann, 2011.

[48] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter analytics support for situational awareness," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, 2011, pp. 181–190.