A Topic-focused Trust Model for Twitter

Liang Zhao [1], Ting Hua[1], Chang-Tien Lu and Ing-Ray Chen
Department of Computer Science
Virginia Tech
{liangz8,tingh88,ctlu,irchen}@vt.edu

Abstract

Twitter is a crucial platform to get access to breaking news and timely information. However, due to questionable provenance, uncontrollable broadcasting, and unstructured languages in tweets, Twitter is hardly a trustworthy source of breaking news. In this paper, we propose a novel topic-focused trust model to assess trustworthiness of users and tweets in Twitter. Unlike traditional graph-based trust ranking approaches in the literature, our method is scalable and can consider heterogeneous contextual properties to rate topic-focused tweets and users. We demonstrate the effectiveness of our topic-focused trustworthiness estimation method with extensive experiments using real Twitter data in Latin America.

Keywords: trust management; social networks; Twitter; trustworthiness; credibility.

1. Introduction

As one of the most popular social messaging tools, Twitter is experiencing a tremendous growth. The number of users is over 200 million as of 2013, contributing over 200 million of tweets every day [1]. The posts in Twitter can be about any domain and any topic in the world, ranging from daily conversations to socially crucial issues. Thanks to the 140 character limitation of length, "timeliness" and "brevity" become the most distinguishing features of tweets. This empowers the freshness of the Twitter posts which usually beat traditional breaking news broadcasting media. Therefore, Twitter is becoming a promising information source to get the most timely knowledge and news around us [2]. Since different users may favor information of different topics, how to identify credible tweets belonging to the specific topics according to users' interests is of great importance. This paper is particularly concerned with the issue of how to treat Twitter as a news channel and use our proposed trust model to identify trustworthy tweets/users.

Despite the advantages of timeliness, Twitter suffers from the fact that it is hardly a trustworthy news resource. First, tweets are usually posted by individual users instead

[1]These two authors contributed equally to this work

of news authorities. The trustworthiness of tweets or users is hard to be ascertained. Second, the spread of posts or tweets in Twitter is through social networks instead of formal news broadcasting like traditional media. In Twitter, the trustworthiness of tweets/users can only be estimated through indirect means, such as the number of followers of a user or a tweet, and the number of retweets of a tweet. This is potentially problematic and can even foster the spread of rumors, because a malicious user can easily forge followers or retweets. Finally, the noisy nature of tweets (largely due to unstructured languages and abbreviations) further hinders accuracy of trustworthiness assessment. Tweets are often written in a casual style, without following standard grammatical rules. For example there is no verb in the tweet "Pretty bad day ioi waiting for it to go by already". New abbreviations and slangs are emerging each day, such as TMB (tweet me back) and abt (about). These noises make it difficult to understand tweets and to properly assess their trustworthiness.

Considering the social impact of information trustworthiness in Twitter, currently there is significant interest on trustworthiness evaluation of tweets or users [3, 4]. A thread of works focused on the evaluation of credibility of tweets by inspecting the contextual contents of tweets [5, 6, 7, 8, 9, 10]. Typically, key features indicating the quality/credibility, such as the length and the language style, are chosen as the features to train a classifier using tweets manually labeled as credible and incredible. Another thread of works focused on investigating the trustworthiness of the users by considering the underlying social network structure of Twitter through the numbers of followers and retweets and the social relationships between users [11, 12, 13, 14].

We observe a number of deficiencies in the works cited above and we aim to devise an effective trustworthiness estimation method to remove these deficiencies:

1. Most of current work focused on evaluating the credibility of general tweets. Credibility evaluation for topic-focused tweets of users' interest is of significantly practical use, yet hasn't been well studied. Supervised learning method is often applied to identify the tweets of specific domains; however, it is not scalable to manually label credible and incredible tweets for supervised learning. To build a training dataset for supervised learning, current technologies require extensive human effort to label tweets. Moreover, labeling of tweets in the training dataset must be updated periodically. There is a need to automatically rate tweets dynamically for scalability. In our work, we do not use supervised learning so there is no need building a training set. Instead, we automatically rate topic-focused tweets by means of a novel similarity-based trust evaluation mechanism.

2. Prior works treat tweets as independent of each other. Tweets are typically classified by a feature vector while the relationships between tweets are neglected. In Twitter, however, one must consider the relationships (e.g., replying, retweeting, authorship, and semantic context) among tweets as these are strong indicators to trustworthiness. For example, the tweets posted by the same untrustworthy user tend to be less trustworthy. In our work, we consider the social and contextual relationships between users/tweets for trustworthiness estimation dynamically by means of a novel iterative trust propagation algorithm.

3. Prior works are based on a social graph trust model [4] with which the credibility of a user is determined by its surrounding neighbors, e.g., how many social connections a user has. However, the social graph model is often constructed without considering the possibility that the edges in the graph can be artificially manufactured by a malicious user. One example is political astroturf, where political campaigns fake as spontaneous "grassroots" that are actually carried out by a malicious plotter or a conspiracy organization [15]. Our work is also based on social graphs. However, we do not use the social graph for directly inferring tweet trustworthiness. Rather, we rate topic-focused tweets by means of a novel similarity-based trust evaluation mechanism and then use the social and contextual relationships described by a social graph for trust propagation dynamically to achieve trust accuracy.

4. Prior works consider that trust is context independent, i.e., trust is deterministic in any situation and any context. However, in reality, trust is context dependent. A node may be trustworthy in one context, but not in another context. For example, a doctor is not as trustworthy when talking about laws, compared with medicine. In our work, we consider textual, spatial and temporal contextual features as we estimate trustworthiness of one user/tweet against another user/tweet.

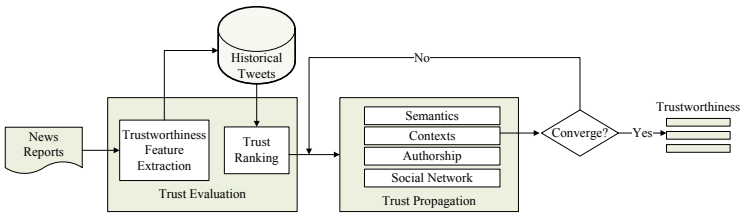


Figure 1: System Architecture

In this paper, we propose a novel method to estimate the user/data trustworthiness in Twitter. Our method first accurately identifies topic-focused trustworthy tweets, and then updates the user/data trustworthiness through iterative trust propagation. To address the scalability issue, we apply our similarity-based trust evaluation method with contextual heterogeneous properties to rate users/tweets against trustworthy users/tweets (say from authorities) without the need of human efforts in labeling credible tweets for supervised learning. As shown in Figure 1, our system consists of two main components: topic-focused similarity-based trust evaluation and trust propagation. The first module rates users/tweets against trustworthy users/tweets for the initial trustworthiness scores, and then the second module further propagates trustworthiness scores among tweets. Our contributions are as follows:

1. Untreated in the literature, we assess trustworthiness of users/tweets by a novel topic-focused trustworthiness estimation method. We propose a new

design notion of similarity-based trust evaluation by which a candidate tweet is considered trustworthy if it is non-conflictingly similar in contextual properties against trustworthy tweets or trustworthy news reports from broadcasting stations. Twitter data are noisy and pointless. However, we can "infer" trust from trustworthy news reports to noisy tweets if there is a sufficient context similarity between news reports and tweets, considering textual, spatial, and temporal contextual properties. Our method is scalable and can consider heterogeneous contextual properties to rate topic-focused tweets/users.

2. We propose a novel trust propagation algorithm which iteratively re-estimates the trustworthiness of users/tweets, by jointly considering their social and contextual relationships in a Twitter social graph. The theoretical proof of convergence is demonstrated.

3. We demonstrate the scalability of our topic-focused trustworthiness estimation method with raw tweet data (Latin America civil unrest tweets) without the need to manually label credible and incredible tweets in a training set for supervised learning.

2. Related Work

In this section we survey the state of the art in user and tweet trustworthiness assessment in Twitter. Existing approaches in general can be categorized into two types, namely, feature-based trust ranking [5, 6, 7, 8, 9, 11, 12, 16, 17, 18, 35, 36, 37], and social graph based trust ranking [10, 13, 14, 15, 19, 20, 21, 22, 38, 39]. We discuss them in Sections 2.1 and 2.2, respectively. In particular, we survey the subject area of tweet trustworthiness in [5, 6, 7, 8, 9, 10, 16]; user trustworthiness in [11, 12, 13, 14, 35, 36, 37]; rumor and misinformation propagation in [15, 21, 22, 38, 39]; supervised learning based on classification in [5, 6, 8, 35, 37]; and unsupervised learning based on clustering in [17, 18, 19, 20].

2.1. Feature-based Trust Ranking

Existing works in this category in general classify tweets related to a target topic based on credibility "features" of tweets and then apply supervised learning to classify if a tweet is credible. [8] provided a SVM-rank based system TweetCred to assign a credibility score to tweets in a user's timeline. [11] studied features that affected user perception. [16] identified eight features that cannot be automatically identified from tweets, but are perceived by users as important when judging information credibility. [6] used several credibility indicators and divided them into post-level (e.g., spelling, timeliness and document length) and blog-level (e.g., regularity, expertise, and comments). Based on these credibility indicators, they proposed a series of credibility ranking methods to find top credible tweets. They concluded that using the post-level indicators combined with comments and pronouns can provide the best performance. [9] conducted controlled experiments to study the impact of several tweet features (message topic, user name, and user image) on the user perception of tweet truthfulness. They showed that user judgments on tweet truthfulness are

biased, and often are based on heuristics (e.g., retweeting). [12] studied the impacts of several microblog features such as gender, name style, profile image, location, and degree of network overlap with the reader, on the credibility perception of users from different countries. They demonstrated that cultural differences can result in different perceptions on user credibility. For example, Chinese users are easier to trust pseudonymously authored tweets and have a strong dependency on microblogs as an information source. [7] studied three types of features: content relevance features (i.e., length and similarity), Twitter specific features (i.e., whether a tweet contains a URL link), and account authority features (i.e., the number of followers). They concluded that URL and time information contained in a tweet are the most effective features for tweet credibility. [35] evaluated user trustworthiness through a classifier trained by multiple features, such as the number of followers, friends, and tweet posts. Tweet posts by credible users are retained for analysis, while those written by untrustworthy users are discarded. [36] modeled user trust using the Analytic Hierarchy Process for measuring trust in a multi-criteria scenario (e.g., followers, retweets, and mentions). Their approach could integrate perception and sentiment of analysts into the problem solving process. [5] applied a supervised learning classification model to classify tweets as credible or not based on features extracted including message, user, topic, and propagation based features. [37] proposed an approach to train a classifier that, starting with some labeled data, identifies trustworthy users through profile/content/graph/neighbor features, propagates trust through the social network, and finally reuses the most trustworthy users to retrain the classifier. Given a set of human participants of unknown trustworthiness together with their sensory measurements, [17] [18] applied Bayesian reasoning and maximum likelihood estimates to determine the probability that a given measurement is true. Relative to the works cited above, our topic-focused trustworthiness estimation method is efficient and scalable, as it does not need to label credible and incredible tweets in a training set for supervised learning

2.2. Graph-based Trust Ranking

In contrast to feature-based trust ranking, graph based trust ranking infers trustworthiness information through social connections by means of a social graph. [10] proposed RAProp which combines two measures of trustworthiness of a tweet. One measure is the trustworthiness of the source of the tweet, which may be a user, a retweet or a webpage cited in the tweet. Another measure estimates tweet trustworthiness by analyzing the tweet content to discover the tweet's corroborating relationship with other tweets. [38] evaluated trust and distrust of users by implicit or explicit recommendations received from other users through user-to-user social connections. Based on social similarity between neighboring nodes, [39] explored the local structure of social networking by means of a graph pruning technique, and evaluated combined trust and distrust through a variation of Page-Rank Algorithm. [13] measured the credibility of social media users based on their online behavior. Users with similar behavior are clustered together and are assigned a similar credibility. However, they failed to give a clear picture about user behavior. To rank credibility of tweets on a topic, [14] proposed to build a social graph modeling web documents, tweets, and users. By connecting users who share similar contents, the social graph

is capable of linking tweets and web documents, filtering informal writing and noise, and inferring unseen relationships between users and tweets from explicit ones. [19] considered tweet trustworthiness as "believability that can be assigned to a tweet about a target topic" and provided three strategies for credibility computation: user-level, content-level, and hybrid. User-level strategies make use of dynamics of information flow from the underlying social network to compute credibility ratings for users. Content-level strategies identify topic patterns and tweet properties which can lead to positive feedback such as re-tweeting and/or credible user ratings. Hybrid strategies combine user-level and content-level strategies by using a weight, cascade or filter connector. Relative to the works cited above, our approach is also based on social graphs. However, we do not use the social graph for inferring tweet trustworthiness. Rather, we rate topic-focused tweets by means of a novel similarity-based trust evaluation mechanism and then use the social and contextual relationships described by a social graph for trust propagation dynamically to achieve trust accuracy. [20] ranked tweets through relevance to the query, aiming to identify latent spatial events based on the tweet graph built. Our work is different from [20] in that we intend to evaluate tweet credibility. [15, 21, 22] studied rumor propagation in social networks. [21] identified rumors relevant to Ebola outbreaks using dynamic query expansion. [15] studied astroturf political campaigns on microblogging platforms by using multiple centrally-controlled accounts to create the appearance of widespread support for a candidate or opinion. [22] proposed to identify rumors by examining the following three aspects of diffusion: temporal, structural, and linguistic. Different from the above cited works, our approach is to assign trustworthiness scores to tweets to differentiate trustworthy tweets from rumors.

3. Problem Formulation

In this section, we begin with a few key concepts, and then we formally define the problem we are solving and the two major tasks in our protocol design for solving the problem.

Definition 1. ***Twitter Collection****: A Twitter collection denoted by $\mathcal{C} = \{C_1, C_2, \ldots, C_T\}$ is a collection of time-ordered Twitter data separated by T time intervals, where $C_t \in \mathcal{C}$ represents the subcollection of the tth time interval. A Twitter subcollection C_t is captured by a Twitter social graph which we call a Twitter heterogeneous information network in this paper.*

Definition 2. ***Twitter Heterogeneous Information Network****: A Twitter heterogeneous information network (for describing a Twitter subcollection) is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V} = W \cup D \cup U$, where W, D and U denote the node sets of "words," "data" (i.e., tweets) and "users," respectively. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ stands for the undirected edge set.*

Figure 2 illustrates an instance of the Twitter heterogeneous information network. The ith word is denoted as W_i. Similarly, D_j and U_k stands for the jth tweet and the kth user. A tweet can contain geo-location information denoted as $l(D_i)$, and timestamp

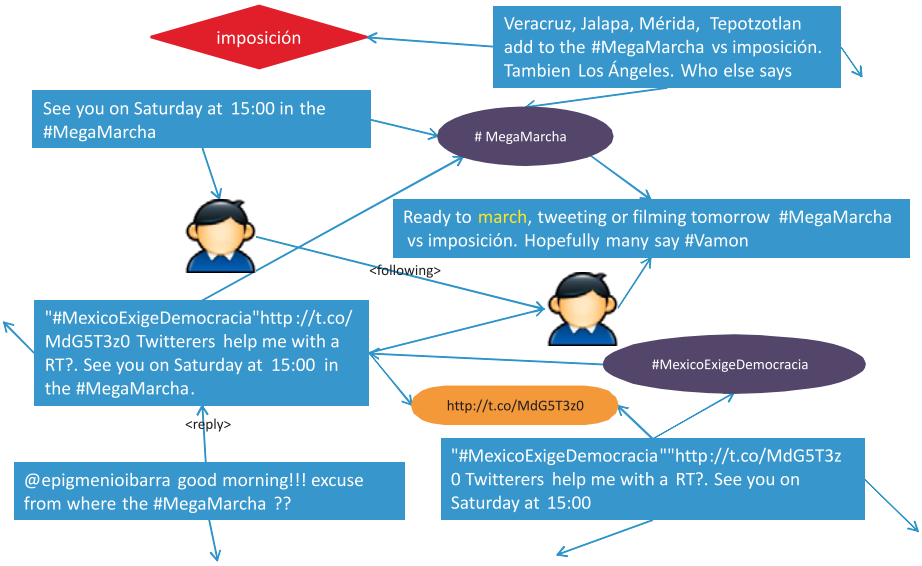


Figure 2: A Twitter heterogeneous information network describing heterogeneous entities (e.g., words, tweets, and users) and their relationships (e.g., "replying" between tweets, "containment" between tweets and words, "co-occurrence" between words, and "friendship" between users.)

information denoted as $t(D_i)$. The edge set $\mathcal{E}$ consists of the relationships among heterogeneous entities, such as "replying" between "tweets," "authorship" between a "user" and a "tweet," and "containment" between a "term" and a "tweet."

Definition 3. ***Trustworthiness Score****: The trustworthiness score denotes the degree that a tweet or a user is deemed credible. We will simply use "trustworthiness" to refer to "trustworthiness score" for short when the context is clear. Mathematically, we denote the trustworthiness of a tweet D_i as $R(D_i)$ while we denote the trustworthiness of a user U_j by $R(U_j)$.*

The trustworthiness of a tweet can be estimated by whether its content refers to things that really happened. The trustworthiness of a user can be estimated by the user's posts. Trustworthiness is context-dependent. In other words, it depends on specific topic domains, e.g., sports, laws, civil unrest. For example, a user who is a doctor is not as trustworthy when talking about laws as when this user is talking about medicine.

Definition 4. ***Event****: An event $x = (l(x), t(x))$ is a significant real-world thing which happened at location $l(x)$ and time $t(x)$. We define the set of events in the same topic domain p as domain $\mathbf{X}_p$.*

Authoritative news outlets are trustworthy sources for event reports. These event reports are not necessarily tweets but can be accesses through public media.

Definition 5. ***News Article****: A news article is a collection of event reports generated by authoritative news outlets for a particular event that happened in the past. The set of news articles in the system is expressed as* $\mathbf{A}$*. The set of news articles in domain p is expressed as* $\mathbf{A}_p$*. An article* $a_x \in \mathbf{A}_p$ *represents authoritative news reports about event x. Note that one event may be associated with multiple event reports, in which case we merge these event reports into one news article.*

Problem Definition: Given a set of news articles in a specific topic domain as well as a Twitter collection, the problem is to determine user/tweet trustworthiness in this topic domain. More specifically, given a set of events X_p, and a set of news articles $\mathbf{A}_p$ in domain p, as well as a Twitter collection $\mathcal{C}$, the problem of trustworthiness evaluation is to determine the trustworthiness of each tweet $R(D_i)$ where $D_i \in D$, and the trustworthiness of each user $R(U_i)$ where $U_i \in U$.

To solve this problem, we define the following two tasks:

Task 1: **Trust Evaluation.** Given a set of trustworthy news articles A_p in domain p, trust evaluation is to calculate the trustworthiness of each tweet $R(D_i)$ and the trustworthiness of each user $R(U_i)$ according to the degree of feature similarity between a tweet and the corresponding news article.

Task 2: **Trust Propagation**. Given the ranked tweets, trust propagation is to refine the trustworthiness of users/tweets based on the links defined in the Twitter heterogeneous information network for achieving trust accuracy.

4. Topic-focused similarity-based Trust Evaluation

In this section, we discuss in detail of our topic-focused similarity-based trust evaluation design. The basic idea is that news articles (see the definition in Section 3) are of high trustworthiness, so we can infer trustworthiness from news articles to Twitter users/data in the same topic domain when there is a non-conflictingly contextual similarity including textual, spatial and temporal features. As illustrated in Figure 1, topic-focused similarity-based trust evaluation has two main components, namely, *trustworthiness feature extraction* and *trustworthiness ranking*.

4.1. Trustworthiness Feature Extraction

The goal of *trustworthiness feature extraction* is to find the most trustworthy features that can identify a specific event in a topic domain. Although tweets and news articles are quite different in format, they are likely to share some semantic features when describing the same event. We represent these features as *domain words* and *event words*. *Domain words* are the most representative words for an event in a domain. For example, "protest" and "march" can be *domain words* for "civil unrest" events. *Event words* are words that can distinguish a particular event from other events in the same domain. For example in a news article describing the "dog protest" event, "YoSoyCan26" and "Zocalo" are *event words* that rarely appear in other "civil unrest" events. We identify these two types of words through *domain weight* and *event weight* defined as follows.

Definition 6. *A **domain weight** $C(W_i, p)$ quantifies the ability of a word W_i in representing the topic domain p. Given a news article set $\mathbf{A}_p$ in domain p, and an open-domain document set $\mathbf{A}$, $C(W_i, p)$ is computed as the product of the normalized term frequency $f(W_i, \mathbf{A}_p)$ of word W_i in set $\mathbf{A}_p$, and the inverse document frequency of W_i in set $\mathbf{A}$.*

$$C(W_i, p) = \frac{f(W_i, \mathbf{A}_p)}{\max\{f(W, \mathbf{A}_p) : W \in \mathbf{A}_p\}} \times \lg\left(\frac{|\mathbf{A}|}{|\{a \in \mathbf{A} : W_i \in a\}| + 1}\right) \tag{1}$$

A domain words set $\mathbf{W_p^{(d)}} \subset W$ is a word subset of words in $\mathbf{A}_p$, containing words with a relatively high domain weight. Initially, all words in $\mathbf{A}_p$ can be viewed as elements in the set of *domain words*. The first factor in the right side of Equation 1 is the term frequency of word W_i [23], and the second factor is the inverse document frequency of W_i in set $\mathbf{A}$ [24]. Following the popular setting of TF-IDF methods, the logarithmic is used to "dampen" the effect of inverse document frequency and therefore enhance the impact of term frequency. Based on the domain weight, we remove trivial words with domain weight smaller than a threshold η_d calculated by the MAD algorithm [25].

$$\delta_d = median(|f(W, \mathbf{A}_p) : \forall W \in \mathbf{A}_p|) \tag{2}$$

$$\eta_d = \delta_d + \alpha_d \times median(|f(W, \mathbf{A}_p) - \delta_d, \forall W \in \mathbf{A}_p\ |) \tag{3}$$

Definition 7. *An **event weight** $E(W_i, x)$ quantifies the ability of a word W_i in identifying event x. It is computed as the product of the term frequency of word W_i in the news article a_x for event x, and the inverse document frequency of W_i in the news article set $\mathbf{A}_p$ in domain p.*

$$E(W_i, x) = \frac{f(W_i, a_x)}{\max\{f(W, a_x) : W \in a_x\}} \times \lg\left(\frac{|\mathbf{A}_p|}{|\{a \in \mathbf{A}_p : W_i \in a\}| + 1}\right) \tag{4}$$

For each event x, an *event words* set $W_x^{(e)} \subset W$ is a subset of words in article a_x, containing words with a relatively high event weight. Similarly, to remove trivial words from the *event words* set, we again set a threshold δ_e computed by the MAD algorithm.

By now, we have obtained *domain words* and *event words* from news reports. Next, we use these words as queries to search Twitter data. Only tweets containing at least one *domain word* or one *event word* are retrieved and sent to the next module *trust ranking*.

4.2. Trust Ranking

In *trust ranking* module, we evaluate the trustworthiness of tweets. We consider three similarity features.

- **Textual Similarity.** We define textual similarity ϕ_{x,D_y} between event x and tweet D_y as the product of tweet words' domain weight sum and event weight sum, as follows:

$$\phi_{x,D_y} = \sum_{W_i \in (D_y \cap \mathbf{W_p^{(d)}})} C(W_i, p) \times \sum_{W_i \in (D_y \cap \mathbf{W_x^{(e)}})} E(W_i, x) \tag{5}$$

Only words in the *domain word* set $\mathbf{W_p^{(d)}}$ are considered when calculating the domain weight sum, and only words in the *event word* set $\mathbf{W_x^{(e)}}$ of event x are considered when computing the event weight sum. There are two reasons behind the formula. The first reason is that, the more *domain words* and *event words* one tweet contains, the more likely it will be event-related. Therefore, both the first and the second terms in Equation 5 are in the form of word weight sum. The second reason is that only tweets containing both *domain words* and *event words* are qualified as event-related. One tweet has many *domain words* but few *event words* may discuss other events in the same domain. Likewise, tweet with many *event words* (e.g., event location names) but few *domain words* may relate to events in other domains (e.g., something which also happened in the same location). To make a balance between *domain words* and *event words*, we multiply the domain weight sum with the event weight sum.

- **Spatial Similarity.** Spatial similarity between tweet D_y and event x is decided by two factors: 1) the distance between tweet location $l(D_y)$ and event occurrence location $l(x)$, and 2) the spatial influence scope of tweet D_y. The first factor is to relate them to the same location. The second factor is to further enhance the tweet trustworthiness with a high textual similarity score. Intuitively, within the same distance to the event occurrence location, tweet with a higher textual-similarity score is more likely to be event-related. Therefore, we model a tweet D_y's spatial influence to event x as a Gaussian distribution $\varphi_{x,y} = \mathrm{N}(x|l_{D_y}, \sum_{x,y})$, centered at tweet D_y's location $l(D_y)$, with influence scope $\sum_{x,y} = \begin{pmatrix} \phi_{x,y} & 0 \\ 0 & \phi_{x,y} \end{pmatrix}$, where $\phi_{x,y}$ is the textual similarity defined in Equation (5).

- **Temporal Similarity.** After a burst of tweets upon the occurrence of a particular event, the number of event-related tweets usually decreases as a Poisson process ([26]). In other words, the possibility of tweet D_y being related with event x decreases as time goes by, i.e., it also decreases following a Poisson process. Therefore, we model temporal similarity between tweet D_y and event x by an exponential distribution.

$$\rho_{x,D_y} = \lambda e^{-\lambda|t(x)-t(D_y)|} \tag{6}$$

  where $t(x)$ is the occurrence time of event x and $t(D_y)$ is the timestamp of tweet D_y.

By integrating the textual, spatial, and temporal similarity scores, we rank the trustworthiness of tweet D_y for event x, ψ_{x,D_y}, by the following function:

$$\psi_{x,D_y} = \phi_{x,D_y} \cdot \varphi_{x,D_y} \cdot \rho_{x,D_y} \tag{7}$$

In general the trustworthiness of each tweet D, $R_0(D)$, is given by:

$$R_0(D) = \{\psi_{x,D_y}, x \in X_p, D_y \in D\} \tag{8}$$

For a tweet D_y, we choose event x^* that maximizes ψ_{x,D_y} as its most correlated event:

$$x^* = g(D_y) = \arg\max_{x \in \mathbf{X}_p} \psi_{x,D_y} \tag{9}$$

5. Trust Propagation

5.1. Design Principle

Twitter is a social network in which multiple entities exist along with heterogeneous relationships. When evaluating the trustworthiness of an entity in Twitter, the impact from all its neighbor entities in the Twitter heterogeneous information network must be taken into consideration. For example, a user's trustworthiness can be inferred by looking at the trustworthiness of its posts, and also be influenced by the trustworthiness of his/her friends. The trustworthiness of a tweet can be heavily influenced by the trustworthiness of its author, and can also have relevance to the trustworthiness of other tweets that are semantically or contextually similar to it. Our design principle for trust propagation comprises the following four rules:

- Rule 1: If two tweets have a similar (non-conflicting) semantic features, then it is likely they have a similar degree of trustworthiness.
- Rule 2: If two tweets have a similar (non-conflicting) conversational context, then they are likely to have a similar degree of trustworthiness.
- Rule 3: If a tweet is trustworthy, then the user who posted it is likely to be trustworthy, and other tweets authored by this user are also likely to be trustworthy.
- Rule 4: If a user's friends are trustworthy, then it is likely this user is trustworthy.

5.2. Trust Propagation Based on Semantic or Contextual Relationship, Authorship, or Friendship

We follow the four rules discussed in Section 5.1 to execute trust propagation.

- **Trust Propagation based on tweet semantic relationship:** Based on Rule 1, we do trust propagation from one tweet to another based on their semantic relationship as follows:

$$R_1(D) = \Omega_1 \cdot B \cdot H \cdot B^T \cdot R_0(D) \tag{10}$$

where $R_0(D)$ is the existing tweet trustworthiness inferred by Equation 8. $R_1(D)$ is the inferred tweet trustworthiness based on semantic relationships. B is the tweet semantic relationship matrix between all tweet nodes and word nodes, so $B * B^T$ denotes the word based semantic relationship among all tweet nodes. To reflect the different weights of words on a domain, we define H as:

$$H = diag(H_1, H_2, \cdots, H_{|W|}),\ H_i = C(W_i, p) \tag{11}$$

where $C(W_i, p)$ is the domain weight of word W_i on domain p as defined in Equation 1. Ω_1 is used to normalize the matrix $B \cdot H \cdot B^T$ by column.

- **Trust Propagation based on tweet contextual relationship:** Based on Rule 2, we do trust propagation from one tweet to another based on their conversational contextual (i.e., replying) relationship as follows:

$$R_2(D) = \Omega_2 \cdot G \cdot R_0(D) \tag{12}$$

where $R_2(D)$ is the inferred tweet trustworthiness based on contextual relationships. G is the tweet replying relationship adjacency matrix where $G_{ij} = 1$ means that tweet i and tweet j have replying conversational relationships, and $G_{ij} = 0$, otherwise. G is column-normalized by Ω_2.

- **Trust Propagation based on tweet authorship:** Based on Rule 3, we align the credibility of the tweets posted by the same author as follows:

$$R_3(D) = \Omega_3 \cdot E \cdot E^T \cdot R_0(D) \tag{13}$$

where E is the adjacency matrix between all the tweet nodes and the user nodes, and thus the $E * E^T$ represents whether any two tweets share the same author. $R_3(D)$ is the inferred trustworthiness based on authorship relationships. Ω_3 is utilized to normalize the matrix $E \cdot E^T$ by column.

- **Trust Propagation based on friendship:** Based on Rule 4, we infer the credibility of a tweet based on the author's credibility in the social network as follows:

$$R_4(D) = \Omega_4 \cdot E \cdot F \cdot E^T \cdot R_0(D) \tag{14}$$

where F is the adjacency matrix among all the user nodes denoting their friendships. $R_4(D)$ is the inferred trustworthiness based on user friendships. Ω_4 is used to normalize the matrix $E \cdot F \cdot E^T$ by column.

By combining all the above considerations with appropriate weights, we can get the trust propagation calculated as:

$$R(D) = \lambda_1 \cdot R_1(D) + \lambda_2 \cdot R_2(D) + \lambda_3 \cdot R_3(D) + \lambda_4 \cdot R_4(D) \tag{15}$$

where λ_1, λ_2, λ_3, and λ_4 with $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ are weights to allow tradeoffs among these four rules of trust propagation. As a default setting, we set them to 1/4, which means that they are equally weighted.

5.3. Algorithm Description

In this section, we give an algorithmic description of our iterative trust propagation protocol for Twitter (ITPTwitter). ITPTwitter considers context semantics, social network structuring, and user credibility. It is an iterative process by which trust ranking is propagated though the Twitter heterogeneous information network until convergence.

As illustrated by Algorithm 1, the tweet semantic relationship matrix B (related by words), authorship matrix E, friendship matrix F, replying relationship adjacency

Algorithm 1: ITPTwitter

Input: B, E, F, G, H, and $R_0(D)$
Output: Inferred Trustworthiness $R^*(D)$

Set λ_1, λ_2, λ_3 and λ_4
Set $k = 1$
Initialize $R^{(0)}(D)$ by $R_0(D)$
//Propagate the trustworthiness score iteratively
repeat
  $R_1^{(k)}(D) = B \cdot H \cdot B^T \cdot R^{(k-1)}(D)$
  $R_2^{(k)}(D) = G \cdot R^{(k-1)}(D)$
  $R_3^{(k)}(D) = E \cdot E^T \cdot R^{(k-1)})(D)$
  $R_4^{(k)}(D) = E \cdot F \cdot E^T \cdot R^{(k-1)}(D)$
  $R^{(k)}(D) = \lambda_1 R_1^{(k)}(D) + \lambda_2 R_2^{(k)}(D) + \lambda_3 R_3^{(k)}(D) + \lambda_4 R_4^{(k)}(D)$
  $\delta = |R^{(k)}(D) - R^{(k-1)}(D)|$
  $k = k + 1$
until $\delta \leq \varepsilon$;
Return $R^*(D) = R^{(k)}(D)$

matrix G and diagonal word weight matrix H are initialized and given as input. The trustworthiness of entities $R_0^{(1)}(D)$ is initialized to $R_0(D)$. After setting the initial values of λ_1, λ_2, λ_3, and λ_4 to 1/4 (for equal contribution) and the iteration number $k = 1$, we propagate trust through the social graph iteratively. When the difference between the trustworthiness calculation results in two consecutive iterations falls below a threshold, we stop the iterative process.

Theorem 1. *The iteration process of ITPTwitter will always converge, that is, the trustworthiness score of any user or tweet in the Twitter heterogeneous information network will converge to a stable value.*

Proof:
We rearrange Equation 15 and thus have the following equation:

$$R^{(k)}(D) = \Gamma \cdot R^{(k-1)}(D) \tag{16}$$

where Γ is the transition matrix between the current trustworthiness score and the next iteration's trustworthiness score, i.e.,

$$\Gamma = \lambda_1 \cdot BHB^T + \lambda_2 \cdot G + \lambda_3 \cdot EE^T + (1 - \lambda_1 - \lambda_2 - \lambda_3) \cdot EFE^T \tag{17}$$

According to the "six degrees of separation" theory [27], any user (in a Twitter group) has a path to any other one in a finite number of steps. Hence, EFE is irreducible because its corresponding graph is strongly connected, which ensures that the Markov chain associated with the matrix σ is irreducible and aperiodic [28], [29]. Finally,

the Perron-Frobenius Theorem [28] guarantees the existence of a unique stationary distribution vector for the Markov Chain, meaning that $R^{(k)}(D)$ will always converge to a stable value.

6. Performance Evaluation

In this section, we first introduce our performance evaluation methodology and metrics. Then, we describe the experimental settings. Finally we perform a comparative performance analysis of our similarity-based trust evaluation method against two baseline schemes.

6.1. Methodology and Metrics

Because of the sheer volume of Twitter data, trust ranking of individual tweets and users is impractical. Instead, we resort to identifying trustworthy tweets while excluding rumors and noise for the *Twitter event detection* application. Specifically, for Twitter event detection we apply our similarity-based trust evaluation method described in Sections 4 and 5 to collect top tweets with the highest trustworthiness scores in a topic domain (i.e., civil unrest). Then, we use these high-ranked tweets identified as a training set to a SVM classifier. Next, the trained SVM classifier is applied to new Twitter data to identify emerging events.

We evaluate the effectiveness of our similarity-based trust evaluation method against two baseline schemes:

- Manually ranked tweets: a manually labeled training set is created as input to the same SVM classifier to identify emerging events.
- Tweets generated by [5] based on keyword matching are used as input to the classifier developed in [5] to identify emerging events.

Performance metrics in the experiment include *precision*, *recall*, and *F-score*. Precision quantifies the fraction of detected events (through high-ranked tweets) that match with *ground truth* events. Recall quantifies the percentage of events that are correctly detected. F-score score is the harmonic mean of precision and recall.

6.2. Experimental Settings

6.2.1. Datasets

We use two data sources in the performance evaluation: Twitter and GSR.

- **GSR Dataset:** GSR stands for Gold Standard Report (generated by MITRE [2]), a news dataset specializing in the targeted domain (namely "civil unrest"), in which each GSR event consists of a date, location, and corresponding news reports. A real world event is selected as a GSR event if it is reported by the top 3 news outlets in that country or by influential international media.

- **Twitter Dataset:** We randomly selected 10% of raw Twitter data for inclusion in our database. In total, we collected 305 million tweets for this evaluation. To obtain tweet locations, we extracted GPS geo-tags, location mentions, and user profile locations from original Twitter data. Then the extracted entities are mapped to ground-truth locations through a decision tree. There are two matching schemes. One is exact matching, which maps geo-entities into ground-truth locations with exact string match. Another is approximate matching, where a geo-entity is considered a match to a ground truth location if the distance between them is less than a distance threshold. In summary, about 50% of the tweets in the dataset were labelled with country-level locations and 20% with city-level locations.

For both datasets, we collected data across 10 countries in Latin America from July 2012 to May 2013, including: Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Table 1 lists more detailed information about Twitter data, news reports sources, and events that happened in each country.

Data from July 2012 to December 2012 are utilized as the training set, and that for January 2013 to May 2013 as the test set. We estimate the trustworthiness of tweets in the training set using our similarity-based trust evaluation method described in Sections 4 and 5 and use the most trustworthy tweets as labels to train the SVM classifier. The trained SVM classifier is then applied to the test set to detect events.

For performance comparison, we also created a manually labeled training set as input to train the SVM classifier. We manually picked tweets related to civil unrest as positive (highly trustworthy) such as "With protests in the Zocalo, # YoSoyCan26 requires Iztapalapa dogs to be free" and left those containing some keywords but irrelevant to civil unrest as negative (lowly trustworthy), such as "Measures should be taken to protest trees against winter damage." To strengthen the quality of training data set, each tweet was assigned to 3 different annotators. In total, we collected 11,533 tweets in the training set, of which about 46% are "civil unrest related" (positive examples), and 54% are non-related (negative examples).

There are several parameters that could affect the performance of our method. In the *feature extraction* module, threshold α_c in Equation (3) defines the score boundary η_c between important domain words and trivial ones. Figure 3 plots weight distribution of top 500 domain words. Tuning points are usually chosen to detect those important "outliers". Notice that there are three obvious tuning points in the curve, near the intersection points with threshold line $\alpha_c = 0.1, 5, 10$. A small value of α_c fails to filter away trivial words. When α_c is set to be 0.1, the size of domain word set reaches 239. In this case, trivial words such as "yesterday," "adult," and "down" are selected as domain words. On the other hand, a large value of α_c will remove important words. When α_c is set up to 10, only 15 words are left as domain words. Therefore, we set

[2] http://www.mitre.org/

[3] In addition to domestic Top 3 news outlets, the following global news outlets are also included: The New York Times; The Guardian, The Wall Street Journal, The Washington Post, The International Herald Tribune, The Times of London, Infolatam.

Country	#Tweets (million)	News source [3]	#Events
Argentina	52	Clarín; La Nación; Infobae	365
Brazil	57	O Globo; O Estado de São Paulo; Jornal do Brasil	451
Chile	28	La Tercera; Las Últimas Notícias; El Mercurio	252
Colombia	41	El Espectador; El Tiempo; El Colombiano	298
Ecuador	13	El Universo; El Comercio; Hoy	275
El Salvador	7	El Diáro de Hoy; La Prensa Gráfica; El Mundo	180
Mexico	51	La Jornada; Reforma; Milenio	1217
Paraguay	8	ABC Color; Ultima Hora; La Nacíon	563
Uruguay	3	El Paí; El Observador	124
Venezuela	45	El Universal; El Nacional; Ultimas Notícias	678

Table 1: Distribution of tweets and GSR events across 10 Latin countries. "News source" shows the news agencies utilized as sources for the GSR dataset.

α_c to be 5 ($\eta_c = 0.09$), which returns a medium-size domain word set that contains 52 words.

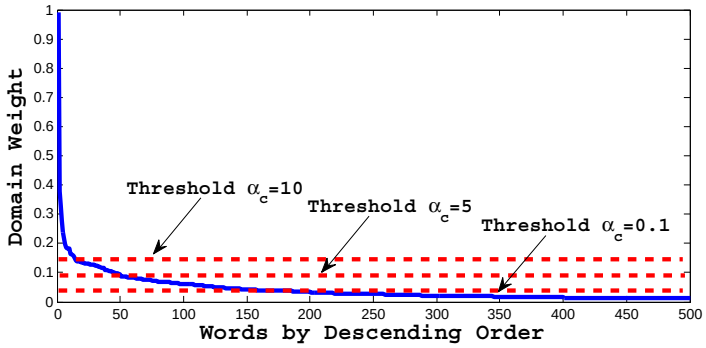


Figure 3: Domain words weight distribution.

In the *trust ranking* module, coefficient λ in Equation (6) is another tunable parameter. To estimate the value of λ, we sample 500 events and fit them to exponential distribution. As a result, we obtain $\lambda = 0.48$ with $R^2 = 0.91$ on average.

6.3. Trustworthiness of Twitter Users

In this section, we validate the assumption that if a Twitter user posts a high percentage of trustworthy tweets, then the user should be more likely to be trustworthy. Although it is almost prohibitive to directly identify whether or not a Twitter user is trustworthy or not, some important Twitter indices are commonly leveraged as surrogates to indicate the Twitter users' trustworthiness. Specifically, the well-recognized Twitter-author indices are: Account Time Length (the time since the

profile was created), Favorite Count, Follower Count, Friends Count, Listed Count (the number of categories interesting the user), and Verified or Not [10].

Therefore, in our experiments, we evaluate whether there is a positive correlation between the tweets' trustworthiness weights and their corresponding Twitter-author trustworthiness indices, and whether this positive correlation is statistically significant. In statistics, rank correlation is commonly utilized to measure the relationship between rankings of different ordinal variables, where a "ranking" is the assignment of the labels "first," "second," "third," etc. Specifically, in our experiments, the Spearman correlation [30] is utilized to evaluate the rank correlations between the tweets' trustworthiness weights and the Twitter-author trustworthiness indices. Moreover, we use p-value to evaluate the statistical significance of the Spearman correlation with the null hypothesis meaning that two sets of data values are Spearman-uncorrelated. A p-value that is equal to or smaller than the significance level (0.03 is used in the paper) means that the null hypothesis is to be rejected, thus supporting the hypothesis that two sets of data values are Spearman-correlated. As can be seen in Table 1, Spearman correlation values between the tweets' and their authors' trustworthiness are mostly larger than 0, demonstrating their positive correlation. Moreover, the p-values are mostly less than 0.03, demonstrating strong statistical significance of this positive correlation.

Among all the indices, "Verified or Not" shows a high Spearman correlation, indicating that a verified user seems more likely to post trustworthy tweets, which is reasonable in real-world situations. "Follower Count" also shows that a user with a large number of followers tends to be more trustworthy. Other indices such as "Account Time Length," "Listed Count," and "Status Count" endorse strong positive correlation between tweets' and their authors' positive correlations, too. Lastly, the index "Favorite count" seems subtle, indicating there is not an apparent correlation between the Twitter users' trustworthiness and their favorite counts. This also makes sense because an untrustworthy user does not necessarily have a small number of favorites in Twittersphere.

6.4. Comparative Performance Analysis

6.4.1. Comparison with Supervise Learning with Manually Labeled Tweets

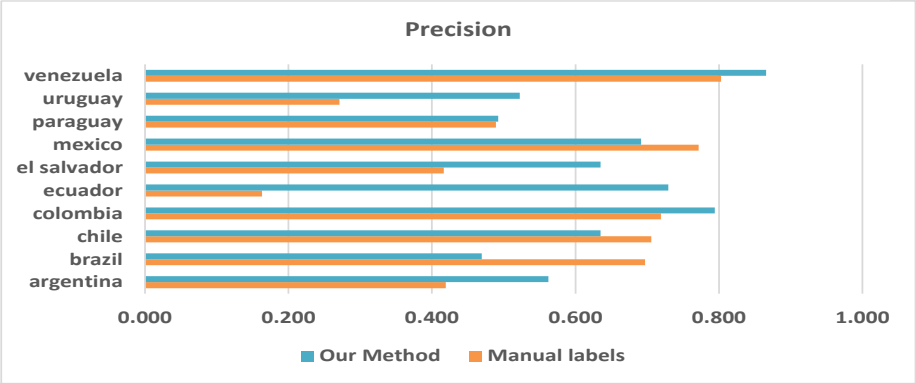
With the topic domain "civil unrest," we compare Twitter event detection performance using tweets ranked by our method with that using manually labeled tweets. The performance comparison is shown in Figure 4. As shown in the figure, our method achieves a higher F-score in 7 out of 10 countries. We make the following two observations:

1. Our method, using automatically ranked tweets, achieves a comparable precision to that of supervised learning using manual labels, and outperforms it in recall and F-score.

2. Our method performs stably across all countries, while the supervised learning using manual labels produces vastly different results across countries. Although the baseline scheme functions better than our method in small countries such as "Paraguay" and "Uruguay," it falls short in large countries like "Mexico" and

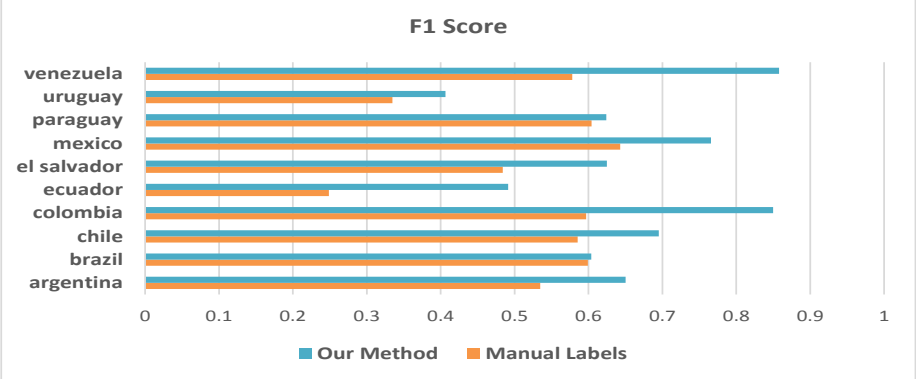
Table 2: Evaluation of the Trustworthiness of Twitter users. Spearman correlation and p-value with the trustworthiness labels show that our top-ranked trustworthy tweets' authors are more trustworthy. Here AR=Argentina, BR=Brazil, CH=Chile, CO=Colombia, EC=Ecuador, EL=El Salvador, ME=Mexico, PA=Paraguay, UR=Uruguay, VE=Venezuela, Acco.=Account time length, Favo.=Favorite Count, Foll.=Follower Count, Frie.=Friend Count, List.=Listed Count, Stat.=Status Count, and Veri.=Account Verified (=1) or Not (=0).

	Index	Acco.	Favo.	Foll.	Frie.	List.	Stat.	Veri.
AR	Spearman	0.189	-0.057	0.224	0.146	0.257	0.11	0.087
	p-value	0.017	0.02	0.008	0.014	0.012	0.01	0.032
BR	Spearman	0.138	-0.11	0.166	0.023	0.183	0.148	0.076
	p-value	0.012	0.011	0.012	0.012	0.004	0.003	0.003
CH	Spearman	-0.002	-0.055	0.125	0.046	0.143	0.08	-0.067
	p-value	0.018	0.009	0.01	0.024	0.005	0.017	0.016
CO	Spearman	-0.022	0.109	0.22	-0.038	0.125	0.034	0.047
	p-value	0.015	0.022	0.027	0.015	0.004	0.035	0.023
EC	Spearman	0.131	0.141	0.474	0.249	0.479	0.309	0.531
	p-value	0.052	0.009	0.011	0.009	0.007	0.032	0.004
EL	Spearman	-0.11	-0.163	0.031	0.002	0.395	0.254	N/A
	p-value	0.025	0.007	0.008	0.012	0.008	0.01	N/A
ME	Spearman	0.122	-0.092	0.188	-0.018	0.193	0.117	0.189
	p-value	0.003	0.004	0.004	0.013	0.002	0.003	0.002
PA	Spearman	0.194	0.202	0.39	0.173	0.346	0.345	0.385
	p-value	0.027	0.014	0.019	0.043	0.019	0.013	0.023
UR	Spearman	0.245	0.053	0.333	-0.003	0.387	0.282	0.38
	p-value	0.009	0.021	0.005	0.036	0.002	0.01	0.011
VE	Spearman	0.171	-0.111	0.259	0.079	0.244	-0.051	0.18
	p-value	0.003	0.005	0.001	0.011	0.001	0.002	0.002



(a) Precision.

(b) Recall.



(c) F-score.

Figure 4: Detection performance comparison.

"Venezuela," which occupy more than 32% of total Twitter data and have 46% of total civil unrest events in Latin America.

In summary, our method outperforms the baseline method in both effectiveness and robustness. Namely, our method can yield better results and work more stably across countries. We attribute this to the fact that our method can generate a large amount of high-quality labels for countries with different languages, while it is hard to manually create enough labels with equivalent diversity.

6.4.2. Comparison with Supervised Learning with Tweets Generated through Keyword Matching

With the topic domain "civil unrest," we compare Twitter event detection performance using tweets ranked by our method with that based on supervised learning with tweets generated through keyword matching [5]. We show that trustworthy tweets identified by our method are of high quality through both quantitative and qualitative analyses.

Take the small "dog protest" event in Mexico as an example, Table 3 lists the top 3 ranked trustworthy tweets generated by our method using the design concept of similarity-based trustworthiness evaluation and trust propagation against the top 3 ranked trustworthy tweets generated by [5] using keywords most relevant to "civil unrest," such as "protest" and "march".

By inspecting Table 3, we make two observations for tweets obtained by [5] through keyword matching:

1. Some tweets are irrelevant to "civil unrest" at all. Take Tweet #3 for example. Its original Spanish text is: "La gente cambia. El amor duele. Los Amigos se **marchan**. Las cosas aveces van mal. Pero recuerda que la vida sigue." Although with one civil unrest keyword "marchan" (becomes "march" after stemming), this tweet is in fact about people's daily feeling.

2. For those tweets indeed related to "civil unrest," most of them reflect influential protests that occurred in countries outside Mexico. For example, Tweet #1 is about a protest in Northern Ireland, and Tweet #2 mentions a protest which happened in Venezuelan. Small events such as the "dog protests" are submerged in these big events.

In contrast, trustworthy tweets retrieved by our method are highly related to the "dog protest" event. These tweets can be summarized into two types:

1. Tweets that talk about the protest itself, such as Tweet #1 and Tweet #2. These tweets contain highly ranked "civil unrest" *domain words* "protesta" (protest) and "marcha" (march), as well as important *event words* "perrors" (dogs) and "Iztapalapa" (location name).

2. Tweets that are related to the reason of triggering the dog protest event, such as Tweet #3. Here we note that the reason for triggering the dog protest event is not mentioned in the news report. According to Tweet #3, we find that, citizens protest for the freedom of innocent dogs, which are captured by the Mexico

government as suspects for killing 4 people. Besides *event words*, these tweets also contain middle-ranked *domain words* such as "Gobierno" (government) and "Mexico," which are weak indications for "civil unrest" when appearing alone, but get stronger when they co-occur in one tweet.

Tweets by [5]	1. Northern Ireland live another **march** day: **Demonstrators protest** since December by a decree ... http://t.co/O2K9hMIq 2. #EnImágenes Students **protest** in several states against the judgment of the Supreme Court http://t.co/clj5XraS 3. RT @FilosofiaTipica: People change. Love hurts. Friends **leave**. Things sometimes go wrong. But remember that life goes on.
Tweets by our method	1. With **protests** in the **Zocalo**, **#YoSoyCan26** requires **government** to free **dogs** of **Iztapalapa**. http://t.co/XPsQ90po #AMLO 2. **#YoSoyCan26 march** in solidarity with Socket for victims' families in **Cerro de la Estrella** and **demand** liberty for **dogs**. 3. RT @politicosmex: To people of **Mexico**, **dogs** are murderers is incredulous : **Government** of the capital is asked to clarify the truth ...http://t.co/m5UbmJXT

Table 3: Comparing tweets collected for the dog protect event in Mexico. *Domain words* are denoted by bold style and *event words* are marked with underline. The tweets, originally in Spanish, have been translated into English using Google Translate.

Table 4 quantitatively compares the trustworthiness and relevance scores of the tweets extracted by the baseline scheme [5] vs. our proposed method. Each tweet was sent to 3 annotators for evaluating whether it is trustworthy or relevant (labeled as True) to the civil unrest topic or not (labeled as False). The tweet is annotated by the labels from the majority of the annotators. The trustworthiness and relevance scores are calculated as the percentage of tweets labeled as True. From Table 4, it is clear that the trustworthiness and relevance of the extracted tweets by the proposed method are much higher than those of the baseline scheme. Our proposed method outperforms the baseline scheme by 15% in trustworthiness, and 6% in relevance. Moreover, the amount of tweets extracted by our method is 14% more than that by the baseline scheme. Our proposed method performs even better in larger countries (e.g., Mexico and Venezuela). We attribute this to our topic-focused similarity-based trust evaluation and trust propagation designs for identifying trustworthy twitters/users, particularly for countries with a large social network.

7. Conclusion

In this paper, we proposed a new design notion of topic-focused similarity-based trust evaluation and trust propagation to rate trustworthiness of tweets and users in Twitter. Compared to existing methods, our approach has three advantages: (1) enabling context-based trustworthiness estimation to focus on credibility in a specific

Table 4: Quantitative performance comparison of our proposed method with the baseline scheme [5] in the Trustworthiness and Relevance Scores of extracted Tweets (Trust.=Trustworthiness Score, Relev.=Relevance Score).

	Baseline [5]			Our proposed method		
Country	Trust.	Relev.	Amount	Trust.	Relev.	Amount
Argentina	67%	71%	1836	69%	82%	1839
Brazil	62%	76%	577	80%	76%	2102
Chile	66%	70%	1803	75%	65%	1714
Colombia	70%	82%	1820	83%	85%	2542
Ecuador	65%	68%	366	69%	65%	242
El Salvador	56%	68%	146	72%	70%	110
Mexico	74%	85%	3109	78%	91%	2349
Paraguay	83%	70%	136	68%	68%	202
Uruguay	71%	68%	139	67%	70%	124
Venezuela	75%	78%	4702	91%	89%	5405
Total	71%	78%	14634	82%	83%	16629

topic domain; (2) utilizing credible news reports to infer trustworthiness of tweets exhibiting contextual similarity in textual, spatial and temporal features; and (3) combining semantic and contextual information with social networking information for trustworthiness propagation. Experiments on Twitter event detection demonstrated that our method can effectively extract trustworthy tweets while excluding rumors and noise. In addition, a comparative performance analysis demonstrated that our method outperforms existing supervised learning schemes using tweets manually labeled or tweets generated based on keyword matching as the training set.

This paper assumes persistent attack behavior, i.e., a malicious user attacks without disguise whenever it has a chance. In the future, we plan to consider more sophisticated attack behaviors such as random, opportunistic, and insidious attack behaviors [31] [32] [33] [34] to further test the robustness of our topic-focused similarity-based trust evaluation scheme.

Acknowledgment

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government. This work is also supported in part by the U. S. Army Research Office under contract numbers W911NF-12-1-0445 and W911NF-11-D-0001/DO 0263.

References

[1] D. Murphy, Twitter: Social Communication in the Twitter Age, 2013.

[2] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, ACM 19th international conference on world wide web, 2010, pp. 591–600.

[3] S. Adali, Measuring behavioral trust in social networks, IEEE International Conference on Intelligence and Security Informatics, 2010.

[4] W. Sherchan, S. Nepal, C. Paris, A survey of trust in social networks, ACM Computing Survey 45 (4) (2013) 47.

[5] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, 20th international conference on world wide web, ACM, 2011, pp. 675–684.

[6] W. Weerkamp, M. de Rijke, Credibility-inspired ranking for blog post retrieval, Information retrieval 15 (3-4) (2012) 243–277.

[7] Y. Duan, L. Jiang, T. Qin, M. Zhou, H.-Y. Shum, An empirical study on learning to rank of tweets, 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 295–303.

[8] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, Tweetcred: Real-time credibility assessment of content on twitter, Social Informatics, Springer, 2014, pp. 228–243.

[9] M. R. Morris, S. Counts, A. Roseway, A. Hoff, J. Schwarz, Tweeting is believing?: understanding microblog credibility perceptions, ACM conference on Computer Supported Cooperative Work, ACM, 2012, pp. 441–450.

[10] S. Ravikumar, K. Talamadupula, R. Balakrishnan, S. Kambhampati, Raprop: ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement, 22nd ACM international conference on Conference on information & knowledge management, 2013, pp. 2345–2350.

[11] S. Aladhadh, X. Zhang, M. Sanderson, Tweet author location impacts on tweet credibility, Proceedings of the 2014 Australasian Document Computing Symposium, ACM, 2014, p. 73.

[12] J. Yang, S. Counts, M. R. Morris, A. Hoff, Microblog credibility perceptions: comparing the usa and china, ACM conference on Computer Supported Cooperative Work, 2013, pp. 575–586.

[13] M.-A. Abbasi, H. Liu, Measuring user credibility in social media, Social Computing, Behavioral-Cultural Modeling and Prediction, Springer, 2013, pp. 441–448.

[14] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. K. Le, T. F. Abdelzaher, J. Han, A. Leung, J. Hancock, et al., Tweet ranking based on heterogeneous networks, International Conference on Computational Linguistics, 2012, pp. 1239–1256.

[15] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, F. Menczer, Detecting and tracking political abuse in social media, 5th International AAAI Conference on Weblogs and Social Media, 2011.

[16] S. M. Shariff, X. Zhang, M. Sanderson, User perception of information credibility of news on twitter, Advances in Information Retrieval, Springer, 2014, pp. 513–518.

[17] D. Wang, T. Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, H. Le, C. C. Aggarwal, On bayesian interpretation of fact-finding in information networks, IEEE 14th International Conference on Information Fusion, 2011, pp. 1–8.

[18] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: a maximum likelihood estimation approach, 11th ACM International Conference on Information Processing in Sensor Networks, 2012, pp. 233–244.

[19] B. Kang, J. O'Donovan, T. Höllerer, Modeling topic specific credibility on twitter, ACM International conference on Intelligent User Interfaces, 2012, pp. 179–188.

[20] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, N. Ramakrishnan, Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling, PloS one 9 (10) (2014) e110206.

[21] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, N. Ramakrishnan, Misinformation propagation in the age of twitter, Computer 47 (12) (2014) 90–94.

[22] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent features of rumor propagation in online social media, Data Mining (ICDM), 2013 IEEE 13th International Conference on, IEEE, 2013, pp. 1103–1108.

[23] H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of research and development 1 (4) (1957) 309–317.

[24] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation 28 (1) (1972) 11–21.

[25] H. Walker, Studies in the history of the statistical method, Williams and Wilkins Co, 1931.

[26] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, 19th international conference on world wide web, 2010, pp. 851–860.

[27] A. Cheng, Six degrees of separation, twitter style (2010).
URL http://www.sysomos.com/insidetwitter/sixdegrees/

[28] C. P.-C. Lee, G. H. Golub, S. A. Zenios, A fast two-stage algorithm for computing pagerank and its extensions, Technical Report, Scientific Computation and Computational Mathematics, Stanford University, 2003, pp. 1–9.

[29] T. H. Haveliwala, Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search, IEEE Transactions on Knowledge and Data Engineering 15 (4) (2003) 784–796.

[30] D. Zwillingerand, S. Kokoska, CRC standard probability and statistics tables and formulae, CRC Press, 1999.

[31] R. Mitchell, I. R. Chen, A Survey of Intrusion Detection in Wireless Network Applications, Computer Communications 42 (2014) 1–23.

[32] F. Bao, I. Chen, M. Chang, J. Cho, Trust-based intrusion detection in wireless sensor networks, IEEE International Conference on Communications, 2011, pp. 1–6.

[33] R. Mitchell, I. R. Chen, Modeling and Analysis of Attacks and Counter Defense Mechanisms for Cyber Physical Systems, IEEE Transactions on Reliability (2015) 1–9.

[34] R. Mitchell, I. R. Chen, Behavior Rule Specification-based Intrusion Detection for Safety Critical Medical Cyber Physical Systems, IEEE Transactions on Dependable and Secure Computing 12 (1) (2015), 16–30.

[35] B. Todd, T.Conrad, H. Kenneth, B. Sven G, Increasing the veracity of event detection on social media networks through user trust modeling, International Conference on Big Data, 2014, pp. 636–643.

[36] C. Yean, C. Yee, I. Tan, Relative Trust Management Model for Twitter: An Analytic Hierarchy Process Approach, International Conference on Frontiers of Communications, Networks and Applications, 2014, pp. 1–6.

[37] Z. Li, X. Zhang, H. Shen, W. Liang, Z. He, A Semi-Supervised Framework for Social Spammer Detection, Advances in Knowledge Discovery and Data Mining (2015), 177–188.

[38] E. Magdalini, L. D, V. Iraklis, A trust-aware system for personalized user recommendations in social networks, Systems, Man, and Cybernetics: Systems, IEEE Transactions on 44 (4) (2014), 409–421.

[39] H. Zhang, C. Xu, J. Zhang, Exploiting trust and distrust information to combat sybil attack in online social networks, Trust Management (2014), 77-92.