A Seminar Report
on

TRUST AND CREDIBILITY (TWITTER)

submitted by

CH.Sai Karthik Dattu
151FA04011

VIGNAN'S
Foundation for Science, Technology & Research
UNIVERSITY
(Estd u/s 3 of UGC Act of 1956)

(ACCREDITED BY **NAAC** WITH **"A"** GRADE)
MHRD **NIRF 88** RANK

DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
VFSTR UNIVERSITY, VADLAMUDI
GUNTUR-522213, ANDHRA PRADESH, INDIA.
2017-2018

CERTIFICATE

This is to certify that the seminar report entitled **“Trust and Credibility(twitter)"** that is being submitted by **Ch.Sai Karthik Dattu** bearing **Regd. No. 151FA04011** in partial fulfilment for the award of III year I semester B.Tech degree in Computer Science and Engineering to Vignan’s Foundation for Science Technology and Research University, is a record of bonafide work carried out by me under the internal guidance of Ms.A.Saranya of CSE Department.

Signature of the faculty guide	Signature of Head of the Department
Ms.A.Saranya	Dr.Venkateswarlu
Asst. Professor	Professor & HoD

TABLE OF CONTENTS

CHAPTER - 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 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.

CHAPTER - 2

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:

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.

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.

CHAPTER - 3

FEATURE BASED STUDY

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. provided a SVM-rank based system TweetCred to assign a credibility score to tweets in a user's timeline. studied features that affected user perception.identified eight features that cannot be automatically identified from tweets, but are perceived by users as important when judging information credibility. 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.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.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. 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. Given a set of human participants of unknown trustworthiness together with their sensory measurements, 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.

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.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.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, 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.

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, 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. 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. ranked tweets through relevance to the query, aiming to identify latent spatial events based on the tweet graph built. Our work is different from in that we intend to evaluate tweet credibility. studied rumor propagation in social networks.identified rumors relevant to Ebola outbreaks using dynamic query expansion. 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. 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.

CHAPTER - 4

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.

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. 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.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.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 proposed to build a social graph modeling web documents, tweets, and users.

By connecting users who share similar contents, the social graphis capable of linking tweets and web documents, filtering informal writing and noise, and inferring unseen relationships between users and tweets from explicit ones. 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.ranked tweets through relevance to the query, aiming to identify latent spatial events based on the tweet graph built. Our work is different from in that we intend to evaluate tweet credibility. studied rumor propagation in social networks. identified rumors relevant to Ebola outbreaks using dynamic query expansion. 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. 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.

CHAPTER - 5

TRUSTWORTHYNESS 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.

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

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 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.

CHAPTER - 6

PERFORMANCE ANALYSIS

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 . 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 using keywords most relevant to “civil unrest,” such as “protest” and “march”.
By inspecting Table 3, we make two observations for tweets obtained by through keyword matching:

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.
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:
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).

CHAPTER - 7

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 based on keyword matching are used as input to the classifier developed in 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.

CHAPTER-8

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.

Enabling context-based trustworthiness estimation to focus on credibility in a specific Quantitative performance comparison of our proposed method with the baseline scheme in the Trustworthiness and Relevance Scores of extracted Tweets (Trust.=Trustworthiness Score, Relev.=Relevance Score).) utilizing credible news reports to infer trustworthiness of tweets exhibiting contextual similarity in textual, spatial and temporal features; and 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.

Chapter 9

REFERENCES

1.F. Benevento, G. Mango, T. Rodrigues, and V. Almeida.Detecting Spammers on Twitter. In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010.

2.R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system.

3.Proceedings of the National Academy of Sciences,105(41):15649–15653, October 2008.

4. B. De Longueville, R. S. Smith, and G. Luraschi. ”OMG,from here, I can see the flames!”: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In LBSN ’09: Proceedings of the 2009 International Workshop on Location Based Social Networks, pages 73–80, New York, NY, USA, 2009. ACM.

5.www.eecs.berkerly.edu
www.emeraldinsight.com
https://books.google.co.in/books
www.sciencedirect.com