IEEE Transactions on Dependable and Secure Computing,Volume:PP,Issue:99,Issue Year:2017

# A Credibility Analysis System for Assessing Information on Twitter

Majed Alrubaian, *Student Member, IEEE*, Muhammad Al-Qurishi, *Student Member, IEEE*
Mohammad Mehedi Hassan, *Member, IEEE* and Atif Alamri, *Member, IEEE*

**Abstract**— Information credibility on Twitter has been a topic of interest among researchers in the fields of both computer and social sciences, primarily because of the recent growth of this platform as a tool for information dissemination. Twitter has made it increasingly possible to offer near-real-time transfer of information in a very cost-effective manner. It is now being used as a source of news among a wide array of users around the globe. The beauty of this platform is that it delivers timely content in a tailored manner that makes it possible for users to obtain news regarding their topics of interest. Consequently, the development of techniques that can verify information obtained from Twitter has become a challenging and necessary task. In this paper, we propose a new credibility analysis system for assessing information credibility on Twitter to prevent the proliferation of fake or malicious information. The proposed system consists of four integrated components: a reputation-based component, a credibility classifier engine, a user experience component, and a feature-ranking algorithm. The components operate together in an algorithmic form to analyze and assess the credibility of Twitter tweets and users. We tested the performance of our system on two different datasets from 489,330 unique Twitter accounts. We applied 10-fold cross-validation over four machine learning algorithms. The results reveal that a significant balance between recall and precision was achieved for the tested dataset.

**Index Terms**—Credibility, reputation, classification, user experience, feature-ranking, Twitter

## 1 Introduction

ONLINE social networks, such as Twitter, have grown highly popular in the 21st century, as the numbers of users who are using them on daily basis attest. Information dissemination through these platforms is their most attractive feature, as it is known to be speedy and cost effective. The fact that users are allowed to express themselves with little to no control is also another very attractive aspect of these platforms [1]. As users are afforded the freedom to publish content with no supervision, the problem of information credibility on social networks has also risen in recent years. Crafty users of these platforms can spread in-formation maliciously for reasons that may not be compatible with the good of society. Users are becoming wary that rumors that are spread through online social networks can have detrimental effects. Research on information credibility is thus the best solution to the problem of how to assess the credibility of information and perhaps mitigate the dissemination of misinformation [2].

Currently, researchers have employed various methodologies in studies on information credibility [2], [3]. Some of them consider the problem to be one of classification that should be solved in an automated fashion using machine learning or graph-based algorithms [3], [5], [6]. Others view it as a cognitive problem requiring human-centric verification [7], [8]. Some authors have looked at how various aspects of social media, such as the effect of the name value and user connectedness, influence users' judgments concerning credibility, [8], [9]. Other researchers have ventured to devise algorithms for assessing credibility, while others have studied the visualization of credibility scores using such means as radar graphs and comparisons between systems such as Fluo and TopicNets [10]. Some researchers have gone so far as to create systems to assess credibility automatically in real time. Such systems include TweetCred [11] and Twitter-Trails [12]. There has also been an enormous amount of research focused on this topic in cases of high-impact events [13], such as earthquakes, floods, and political movements. The main challenge in assessing the credibility of information dissemination on online social networks is the nature of the networks; they are very complex and grow in users and content every day. Among the many challenges related to studying credibility on social networks and the web are the following:

1. The complexity of social networks and the web creates difficulty in identifying resources for use in studying and assessing credibility.
2. OSNs by their very nature evolve dynamically over time and become very large in size, with various structures that make it difficult to obtain the information needed to discern the credibility of users.
3. The credibility of a user is influenced continuously by various factors, such as changes in the social topography, other users' behavior, preferences, and context.
4. Malicious activities can evade existing spam filters through various means. For example, in Twitter, malicious users can purchase followers or use tools to automatically generate fake ac-counts and post tweets with the same meaning but different words.
5. The process of evaluating solutions has also been a problem in terms of resources, given that most researchers are limited in terms of the extent to which they can test their

• *Majed Alrubaian, Muhammad Al-Qurishi, Mohammad Mehedi Hassan, and Atif Alamri are with the Research Chair of Pervasive and Mobile Computing, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia.*
*E-mail: malrubaian.c@ksu.edu.sa, qurishi@ksu.edu.sa, mmhassan@ksu.edu.sa, atif@ksu.edu.sa*

work (Twitter and other OSN limitations).

Thus, it is very difficult to measure the credibility of a user in these networks and to verify his/her posts. As online social networks have become more useful for disseminating information to wider audiences, addressing the above-mentioned challenges to determine the credibility of users in OSNs requires the development of robust techniques for measuring user and content credibility. We propose a hybrid approach to credibility analysis that can be used to identify implausible content on Twitter and prevent the proliferation of fake or malicious information. Our major contributions to this area and the key features of the proposed technique can be summarized as follows:

- We propose a novel credibility assessment system that maintains complete entity-awareness (tweet, user) in reaching a precise information credibility judgment. This model comprises four integrated components, namely, a reputation-based model, a feature ranking algorithm, a credibility assessment classifiers engine, and a user expertise mod-el. All of these components operate in an algorithmic form to analyze and assess the credibility of the tweets on Twitter.
- Using the reputation-based technique, we sought to automatically rank users based on their relevance and expertise on given topics.
- We enhanced our classifier by weighing each feature according to its relative importance. This weighting methodology implements a pairwise comparison that produces a priority vector that ranks the instance's features ac-cording to their relative importance with respect to the user need as well as the topic.
- In our system, an observation is a tweet, and the positive class is credible. In this case, a highly sensitive classifier is more acceptable than precision, because non-credible tweets, if classified as credible, might spread misinformation that goes viral and cause chaos in terms of politics or an emergency. Thus, our priority being to minimize false positives, we might choose to optimize our model with respect to recall or sensitivity.
- We validated our system by applying tenfold cross-validation with four machine-learning algorithms on two different datasets of Twitter content. Our results show that the system that employed a reputation-based filter approach provide a significant and accurate credibility assessment.

The remainder of this paper is organized as follows: Section 2 summarizes related research on credibility assessment on Twitter. Section 3 provides background information on the problem and our sys-tem architecture. Section 4 details our system for measuring trustworthiness on Twitter. Section 5 presents the results of the performance evaluation of the proposed system. Finally, Section 6 concludes the paper.

## 2 Related Works

There have been many extensive studies related to credibility in OSNs. In this section, various approaches have been highlighted in the area of credibility research, such as automated, human-based, and hybrid approaches.

### 2.1 Credibility of the content

In the literature, there is a large body of work on the automated-based approach employing machine learning techniques—specifically, the supervised learning approach [2]. This approach comprises a decision tree, a support vector machine (SVM), and Bayesian algorithms. Castillo et al. [6] was the first such research on Twitter credibility. The paper examined automatic ways of assessing credibility via analysis of microblog postings pertaining to trending topics and classification of the posts as either credible or non-credible, using features extracted from the topics. In essence, the texts of posts, external links cited, and the posting behavior of the user were used in the classification.

Pal and Scott [9] took a different approach to studying credibility on Twitter: they sought to show how name value bias affects the judgments of microblog authors. In this study, the author showed the correlation between name value bias and the number of followers. A similar study by Morris et al. [14] discussed how users perceive tweet credibility. They conducted a survey that showed a disparity in the features used by users to assess credibility and those that are shown by search engines.

Westermann et al. [15] took a different approach to the problem by examining the effect of system-generated reports of connectedness on credibility. The researchers took an experimental approach to designing six mock-up pages on Twitter that varied the ratio between followers and follows and the number of followers. The results revealed that having too many followers or too few led to low assessments of expertise and trustworthiness. Having a narrow gap between follows and followers led to higher assessments of credibility.

Kang et al. [16] discussed ways to model topic-specific credibility on Twitter on an evaluation of three computational models such as a social model, a content-based model, and a hybrid model. The authors used seven-topic specific data sets from Twitter to evaluate these models. The results showed that the social model outperformed the others in terms of predictive accuracy.

Ikegami et al. [17] performed a topic- and opinion-classification-based credibility analysis of Twitter tweets, using the Great Eastern Japan earthquake as a case study. The researchers assessed credibility by computing the ratios of similar opinions to all opinions on a particular topic. The topics were identified using latent Dirichlet allocation (LDA). Sentiment analysis was performed using a semantic orientation dictionary to assess whether a tweet's opinion was negative or positive. An evaluation of this method using kappa statistics showed that it is a good way to assess credibility.

### 2.2 Credibility of the source during an event

Mendoza et al. [14] took a different approach to the problem of assessing information credibility in their study of the behavior of Twitter users in cases of high-impact events. The event considered in this study was an earthquake that occurred in Chile in 2010. The authors studied the activity of Twitter in the hours after the event and combined the results with the results of a study on the dissemination of true information and false rumors on the network at that time. The study established that true information, and false rumors are propagated differently. Tweets that spread false rumors tend to be questioned by other

Fig. 1. Architecture of the proposed system.

network users.

Aditi and Ponnurangam [18] also studied credibility ranking of tweets during high-impact events. Using statistical techniques such as regression analysis, the authors were able to identify important content and source-based features that could be used to predict the credibility of the information in a tweet.

Some other researchers have shown the significance of using both content and social structure in finding credible sources. A good example of this approach is a study by Canini et al. [19] in which an experiment was performed to determine the extent to which these factors influence both explicit and implicit judgments of credibility. Other researchers have analyzed not only ways to measure credibility on Twitter but also ways to communicate scores [20].

### 2.3 Human perception in credibility assessment

O'Donovan et al. [21] sought to achieve synergy be-tween the fields of computer science and the social sciences in a study on competence modeling on Twit-ter. They presented an example of mapping using a Dreyfus model of skill acquisition on four topic specific Twitter datasets.

Kumar and Geethakumar [22] also employed tools from the fields of both computer science and the social sciences in a study on assessment of credibility on Twitter. Their paper discusses how cognitive psychology can be used to detect misinformation, disinformation, and propaganda in online social networks. The cognitive process involved assesses the consistency of a message, the coherency of the message, the credibility of the source, and the general acceptability of message. The paper presents an algorithm that adopts the collaborative filtering feature of social networks to help users detect false content.

### 2.4 Credibility assessment systems

Some researchers have been interested in developing systems that deliver credibility ratings to users in near-real time. TweetCred, a system developed by Gupta et al. [11], is an example of such a solution that is implemented in the form of a browser plugin. The researchers evaluated the performance of the solution among users in a real-world scenario.

CREDBANK is a similar solution presented by Mitra and Gilbert [23]. In principle, CREDBANK is a social media corpus that conglomerates human and machine computation. It has topics, tweets, events, and corresponding human credibility assessments. The corpus is based on real-time tracking of more than one billion tweets over a period of not less than three months, coupled by computational summariza-tion of the tweets and human annotations.

Another system that takes a similar approach is TwitterBot, designed by Lorek et al. [24]. This tool is able to score the credibility of submitted tweets in the native Twitter interface. Fake Tweet Buster is another such tool, designed by Trumper [25], to assess credibility through detection of fake images and users who upload fake information. TwitterTrails, presented by Finn et al. [12] is an-other system for assessing credibility

The different studies discussed earlier have yielded different results because of the different approaches taken by the authors. In general, the previous studies on this subject show that credibility assessment is possible when different dimensions are considered or different approaches are taken in the analysis. The literature also shows that it is possible to build automated systems for measuring and communicating credibility in online social networks.

## 3 Problem Formation

### 3.1 System Architecture

The architecture of our proposed system is illustrated in Figure 1. It consists of five major procedures labeled as follows: 1) tweet collecting and repository, 2) credibility scoring technique, 3) reputation scoring technique, 4) user experience measuring technique, and 5) trustworthiness value, the last of which is an output of the preceding three techniques. In principle, all these mechanisms together represent an iterative process that combines an automated-based methodology for achieving better credibility or trustworthiness results with sophisticated accuracy.

Tweets are collected using two different Twitter application programming interfaces (APIs) [25]: a streaming API and an API for searching for tweets regarding different events. The streaming API is used to collect datasets on given events. The search API is used to collected users' tweets histories simultaneously. On a database server, the data are organized, processed (Step 2), and made available for analysis (Step 3). The prepared data are divided into three groups: tweet con-tent, users who post that content, and the histories of those users (Step 4). These groups of data are passed as inputs to the three techniques to look for signals of truth and credibility (Step 5). The reputation-based technique (Step 6) does not consider aspects such as message content features but does consider factors such as the structure of the network in its model. The credibility technique (Step 7) relies on machine learning methods that are based on training with established ground truth while user expertise method applying both techniques in establishing the reliability of users. Finally, all of the scores obtained using the three techniques are combined to obtain the trust-worthiness value of a given tweet (Step 8).

In this paper, we focus particular attention on how to extract and clean data in Step 2, how to calculation reputation scores in Step 6, the credibility assessment mechanism in Step 7, and finally and most importantly, how the users' experience is calculated.

### 3.2 Definitions of Credibility and Reputation

A crucial part of the system is the assessment of the credibility of tweets and the reputations of the users who posted them. We use the term "credible score" to represent the level of trustwor-

thiness of the posted content. We use the term "reputation score" to represent the level of dependability of the user who posted the content. Another crucial part of the system is user expertise, including a user's credibility assessment and reputation level.

"Credibility" can be defined as "the quality of being trusted and believed in," or "the quality of being convincing or believable." The base word of credibility is "credible," which can be defined as "the quality that someone is able to be believed; convincing," or "capable of persuading people that something will hap-pen or be successful" [34]. Latterly, the most correlated synonyms of credibility have been "trustworthiness" and "believability."

For the purpose of our research on Twitter, we define three classes of credibility with respect to two levels of credibility assessment, as follows.

**Definition 1 Post- (tweet-) level** credibility, denoted by $C(T_i)$, is a numerical score for tweet $T_i$ that reflects how believable and reliable the tweet is, and thus how likely it is that the tweet conveys acceptable information, regarding a certain event.

**Definition 2 User- (account-) level** credibility, denoted by $C(U_i)$, is a numerical score for user $U_i$ that reflects the trustworthiness of a user in an online social network. The lower the trustworthiness of a user is, the more likely it is that the information disseminated by that user is not credible.

**Definition 2 User reputation level**

Users' reputations are based on popularity measures. We describe a popular user as one who is recognized by other users on a similar network. The measures include the Follower-Rank and the Twitter Follower-Followee ratio (TFF). In addition, we consider replies and retweets as measures of a user's popularity.

## 4 Credibility Measuring System

In this section, we describe the main techniques used to achieve our objectives—measuring a Twitter user's reputation and experience, ranking features, assessing a tweet's credibility, and finally obtaining a trustworthiness value for a given piece of content on Twitter. The framework of the system consists of four components: a reputation-based model, a feature-ranking algorithm, a credibility assessment classifiers engine, and a user expertise model. We present each of these components in detail in this section.

### 4.1 Reputation-based model

Measuring user reputation is an important aspect of the problem to be solved because the phenomenon of inspiration is widespread, especially on social net-works. This phenomenon has been verified several times in previous studies. However, there remains a need to investigate the influence measures of social media platforms such as Twitter. Thus, we consider it imperative to discuss the major concepts and characteristics of the Twitter network. The problem ad-dressed here is important because it is often challenging to find measures that can be computed efficiently. Furthermore, some less-than-ideal measures can nonetheless be used to classify users in a reasonable manner.

To measure user expertise and reputation, we use some different measures that are considered to have a huge impact on Twitter. This can be accomplished by measuring reputation through how popular a user is and how sentimental he/she is.

**User Sentiment History**. The sentimentality of a user influences his or her judgments of tweet credibility with respect to an event or topic, especially when the user is inclined favorably or unfavorably toward some sects or groups. Some users have reasons for disseminating information that may be considered misleading and can contribute to chaos, as in the case of the Arab Spring in 2011. Sentiment defines the factors that affect social relationships, psychological states of users, and their orientation. Sentiment also involves an analysis of why a user trusts a trustee or not. In a study on calculation of the number of positive and negative words in a message, based on a predefined "sentiment words" list, researchers found that the least credible messages are associated with negative social events and contain strong negative sentiment words and opinions [17].

For each user $u_i \in U$, we calculate a sentiment score (denoted by $\Delta_u$) based on analysis of his previous tweets, using the following equation:

$$\Delta_{u_i} = \frac{\sum T_{u_i}^{+}}{\sum T_{u_i}^{+} + \sum \left| T_{u_i}^{-} \right|} \qquad (1)$$

where, $T^+$ is a user's positive tweets and $T^-$ is a user's negative tweets, calculated by the Arabic sentiment analysis algorithm SAMAR [27].

**The User Popularity Score** , which is a measure of user popularity, is obtained from a simple arithmetic expression that facilitates the production of basic information concerning social networks, based on a numerical value. Measurements can be combined to define a ranking measure. This measure can be explained in the form of an algorithm that describes criteria suitable for ranking each user on the network regarding his reputation. Suppose that we have the $U$ set of users who have more than one tweet on a given topic p ∈ P. Given a set of tweets (denoted by $T$), we calculate the tweets of each user $t_u$ over $T$. Thus, the initial calculation of user activity is as follows:

$$I^{p}(u_i) = \begin{cases} \sum_{u \in U, p \in P} t^{p}{}_{u_i} / |T|, & if\ t_{u_i} \in |T|; \\ 0, & otherwise \end{cases} \qquad (2)$$

Based on our observations, we consider retweets, favorites, and mentions to be the best indicators from a quantitative perspective. This implies that a tweet that has been retweeted many times is considered to be attractive to the reader. Nevertheless, the most critical indicators are qualitative. One good example is the relationship between the reader and the tweeter of the tweet. Such considerations are a huge hindrance to the definition of a user's expertise.

We first calculate the social popularity $\vartheta$ of $u \in U$ on a given topic $p \in P$ using the following equation:

$$\vartheta^{p \in P}(u_i) = \log(NoFlw(u_i)) / \max(\log(NoFlw(U, p))) \qquad (3)$$

Then, we calculate the event engagement of user u ∈ U on a given topic p ∈ P using the number of favorites φ , the number of retweets ν and the number of mentions ψ concerning topic p as shown in equations 4,5 and 6.

$$\varphi^{p \in P}(u_i) = \log(NoUFav(u_i)) / \max(\log(NoUFav(U, p))) \qquad (4)$$

The number of retweets of his tweets on the same topic $p$ is calculated as shown in equation 5.

$$\upsilon^{p\in P}(u_i) = \log(NoRT(u_i)) / \max(\log(NoRT(U, p)) \quad (5)$$

The number of times the user has been mentioned in others' tweets on the same given topic $p$ is calculated as shown in equation 6.

$$\psi^{p\in P}(u_i) = \log(NoMe(u_i)) / \max(\log(NoMe(U, p)) \quad (6)$$

The event engagement of user $u$ on topic $p$ is determined as a linear combination of the aforementioned components.

$$EE^{p\in P}(u_i) = \varphi^{p}(u_i) + \upsilon^{p}(u_i) + \psi^{p}(u_i) \quad (7)$$

For a given topic p ∈ P, user influence, denoted by ω(u) can be computed as follows:

$$\omega^{p\in P}(u_i) = \frac{\vartheta^{p}(u_i) + EE^{p}(u_i) + I^{p}(u_i)}{Log(\square)} \quad (8)$$

We use □ to denote the number of users considered with respect to topic $p$ as shown in the formula above.
The final step is to rank the users according to their reputations, which can be calculated as follows:

$$\Re^{p}(u_i) = \Delta_{u_i} \times (\omega^{p}(u_i)) \quad (9)$$

Users for whom the value of $\Re^{p}(u_i)$ is lower than 0.1 for users are considered to be unreliable/non-credible sources.

The users with the highest priority $\Re^{p}(u_i)$ values are considered to be the most trusted sources on a given topic, while the users with the lowest priority values are considered to be the least trusted. The ranked list of users is an input to the next technique. Using reputation-based method leads to improved prediction accuracy over all data sets. It can also be seen that for a given data set, it is possible to identify an optimal threshold that minimizes the prediction error. The next credibility assessment technique helps in achieving high recall.

**Algorithm I**

1: **procedure** CALCUSERREPUTATION (*User*, *Tweets*)
2: **If** *Tweets* **is** *empty* **then return** 0
3: **If** *User* **is** *verified* **then return** 1
4: **For each** $u \in$ *User*
5: Calculate UserActivity $I^{p}(u_i) = \sum_{u\in U, p\in P} t^{p}{}_{u_i} / |T|$
6: Calculate UserInfluence $\omega^{p\in P}(u_i) = \frac{\vartheta^{p}(u_i) + EE^{p}(u_i) + I^{p}(u_i)}{\square}$
7: Calculate UserSentmentHitory $\Delta_{u_i} = \frac{\sum T^{+}_{u_i}}{\sum T^{+}_{u_i} + \sum |T^{-}_{u_i}|}$
8: **End For**
9: **User reputation** $\Re^{p}(u_i) = \Delta_{u_i} \times (\omega^{p}(u_i))$
10: **return** $\Re^{p}(u_i)$
**11: end procedure**

## 4.2 Credibility assessment model

Credibility on Twitter has a considerable influence on modern society, given that information has the power to move masses. It is not uncommon for malicious persons to use Twitter as a way to spread misinformation, for purposes such as defamation of brands in business completion or of public figures in political battles. Such information can be received through content that has been edited to suit the target attack strategy. Because information is difficult to verify, naive users may propagate misinformation, and in some cases, even the print media can be drawn into propagating misinformation. These scenarios illustrate the need for credibility assessment algorithms that can provide analysis results on the truth measure of tweets in real time. The challenge in addressing this problem is to develop such a system that yields accurate results.

We believe that when people discuss a topic related to a sensitive event, they are subject to influences that affect what they post. These influences are important in evaluating information credibility. One of these influences is the orientation of people. In relation to some events (such as the Arab Spring in 2011), this factor leads to division of people into two groups, supporters and opponents, and everyone spread news that supports his/her orientation. Users cite external sources using Internet uniform resource locators (URLs). In relation to other events (such as chemical weapons use in Syria or ISIS crimes), people express their emotions using opinion statements that convey positive or negative sentiments. People also question the propagated information and the users who posted it and so on.

## 4.3. Levels of the extracted features

In this section, we divided the extracted features into three levels as follows:

### 4.3.1. Tweet-level

*Text features:* include some characteristics related to the content of the tweet such as the length of a message, the number of replies and/or the number of retweets may reflect the importance of the tweet. In addition, is the tweets contains #tags and "@mentions" as well as URLs and number of static and animated emoticons.

*Sentiment features:* calculating the number of positive and negative words, based on a predefined sentiment words list.

### 4.3.2. User-level

Some of these features are latent and some of them explicitly revealed in user profiles. For example, age, gender, education, political orientation, and even any user preferences are considered as latent attributes. The number of followers, number of friends and the number of retweeted tweets as well as the replies of user's tweets.

### 4.3.3. hybrid-level

Extracting hybrid-level features is the process of aggregating most of the tweet-based features such as the URL fraction of the tweets, the hashtags (#) fraction of the tweets, and the average sentiment score in tweets. The number of duplications, which means that the user may post the same tweets more than once.

To assess the credibility of Twitter content, we per-formed extraction of tweet features at three levels: the post (message) level, the user level, and the hybrid level. For each level, we used aggregated features, such as the number of retweets on a specific topic, as explained in section 4.3. Each user had a personal record of information in his/her profile. We excluded users who had no followers. We take ad-vantage of the post/message-level, topic/event-level, and user-level credibility assessments in forming hybrid credibility measurements.

We considered hybrid-level feature extraction in our credibility assessment process for two reasons. First, at the tweet level, the 140-character length limit of Twitter messages makes them, to a certain degree, inappropriate for analysis with high-impact topic models. From another perspective, individual tweets do not provide sufficient information about the combination of precise latent topics within them. Second, users can easi-

ly obtain thousands of followers in a minute from so-called Twitter follower markets.

Given a set of tweets, the next goal of our system is to classify each of them as either credible or non-credible. Once the content features are extracted and the users' reputations are identified using the reputation-based technique, the next task is to utilize a new mechanism for credibility assessment to score the tweets. The proposed framework employs two key techniques: a feature rank algorithm and a credibility classification engine, to correctly evaluate and score a given piece of content.

### 4.2.1 Feature Ranking Algorithm

We believe that the extracted features should be weighted before calculating the assessment of a given tweet, user, or topic, because of influence of the features on the final judgment of credibility. In our research on credibility of social web content, we concluded that the number of followers is the most important feature, followed by the number of message URLs, retweets, and user mentions. The least influential factors among those considered were concluded to be the time zone, media, and the number of favorites [2]. Therefore, the ranking of the features considered has an important influence on the results of the classification process. Not all of the features are quantitative; some are qualitative and require human intervention to determine their importance with respect to the overall goal. This intervention happens only once in the process. We rely on a human expert to generate a judgment matrix concerning the importance of each feature. Equation 10 illustrates the form of the judgment matrix.

$$\Lambda = \begin{pmatrix} 1 & f_{12} & \cdots & f_{1n} \\ f_{21} & 1 & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & 1 \end{pmatrix} \qquad (10)$$

We use a pairwise comparison method to determine the relative importance of the extracted features. This comparison process generates a measurement scale of relative priorities or weights of those features. The pairwise comparisons are made by ratios of importance (of compared features), which are evaluated on a numeric scale from 1 to 9.

The pairwise comparisons are performed at the feature level, where they are compared, two at a time, with respect to the overall goal of the ranking process. The results of pairwise comparisons are entered into a matrix (the judgment $\Lambda$ matrix, Eq. 10). A feature is equally important when compared to itself, thus, the diagonal of the matrix must consist of "1"values. In addition, the matrix must satisfy the relation $f_{ij}=1/f_{ji}$. The matrix is read as follows: for example, if $f_{ij}$ =5, then feature i is of essential and strong importance over feature j. This also implies that the value $f_{ji}$ will be equal to 1/5. This matrix is the basic unit of the analysis and is also called "the pairwise comparison matrix."

After all the pairwise comparison values are entered into the matrix, a vector of priorities is computed for the features. In mathematical terms, the principal eigenvector (the eigenvector associated with the maximum eigenvalue) is obtained. We normalize the matrix by dividing each element of the matrix by the sum of its column elements, and this becomes the vector of priorities. Algorithm 2 shows the computation of the priority vector. Given a set of attributes, $f = f_1, f_2 \ldots f_n$, where $n$ is the number of features, the analyst repeatedly compares one feature to another until all possible pairwise comparisons are completed.

A consistent and coherent expert should be able to determine his pairwise preferences, i.e. $f_{ij} = v_i/v_j \ \forall i, j,$ precisely. Thus, let us reflect on the consequences of this condition being satisfied on the entries in the pairwise comparison matrix $\Lambda$. If we write $f_{ij}\, f_{jk}$ and use the condition $f_{ij} = v_i/v_j \ \forall i, j$ then we can derive the following:

$$f_{ij} f_{jk} = \frac{v_i}{v_j}\frac{v_j}{v_k} = \frac{v_i}{v_k} = f_{ik} \qquad (11)$$

**Algorithm II**

**1: procedure FEATURERANK ( $\Lambda$ )**
**2:** **For each** *column* $\vec{c} \in \Lambda$
3: $\vec{S} \leftarrow \sum_{i \in \vec{C}} (f_i)$ *w.r.t the row*
**4:** **End For**
**5:** **For each** *feature* $f_i \in \Lambda$
**6:** $\widehat{\Lambda} \leftarrow$ **Normaliz** ( $\Lambda$ ) dividing each entry on the $\vec{S}$
7: **Calculate Geometric Mean** ( $\overline{P_v} = \left(\prod_{j=1}^{n} f_{ij}\right)^{\frac{1}{n}} \Big/ \sum_{i=1}^{n}\left(\prod_{j=1}^{n} f_{ij}\right)^{\frac{1}{n}}$ )
**8:** **End For**
9: $RF \leftarrow$ **Create** a list of ranked features w.r.t $\overrightarrow{P_v}$
**10:** **return** $RF$
**11: end procedure**

A consistent and coherent expert should be able to determine his pairwise preferences, i.e. $f_{ij} = v_i/v_j \ \forall i, j,$ precisely. Thus, let us reflect on the consequences of this condition being satisfied on the entries in the pairwise comparison matrix $\Lambda$. If we write $f_{ij}\, f_{jk}$ and use the condition $f_{ij} = v_i/v_j \ \forall i, j$ then we can derive the following:

$$f_{ij} f_{jk} = \frac{v_i}{v_j}\frac{v_j}{v_k} = \frac{v_i}{v_k} = f_{ik} \qquad (11)$$

Accordingly, if all features weights in the judgment matrix satisfy this condition, we can conclude that

$$f_{ik} = f_{ij} f_{jk}, \forall i, j, k \qquad (12)$$

In the best case, the expert does not contradict himself; however, this condition is not always satisfied and inconsistencies may occur in the matrix for various reasons, including lack of information, clerical error, and lack of concentration. To measure an expert's consistencey we can use the consistency ratio developed in [28][29].

Using Saaty's consistency ratio we have to find the maximum eigenvalue $\lambda_{max}$ and the consistency index of the judgment matrix $\Lambda$. In this respect, we have the following reciprocal matrix:

$$\Lambda v = \begin{pmatrix} v_1/v_1 & v_1/v_2 & \cdots & v_1/v_n \\ v_2/v_1 & v_2/v_2 & \cdots & v_2/v_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n/v_1 & v_n/v_2 & \cdots & v_n/v_n \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} nv_1 \\ nv_2 \\ \vdots \\ nv_n \end{pmatrix} = nv \quad (13)$$

From linear algebra, we conclude that a formulation of the type $\Lambda v = nv$ indicates that $n$ and $v$ are an *eigenvalue* and an *eigenvector* of $A$, respectively. Thus,

$$\Lambda v = \lambda_{max} v \qquad (14)$$

If the matrix is consistent, then $\lambda_{max}=n$ and greater than $n$ otherwise, and that can be measured using the Saaty consistency ratio ***CR***. To obtain the *CR,* we must determine the consistency index *CI,* which can be calculated as follows:

$$CI(\Lambda) = \frac{\lambda_{max} - n}{n-1} \qquad (15)$$

The consistency ratio ***CR*** can be calculated dividing ***CI*** by a real number $RI_n$ (*random consistency index*) as follows:

$$CR(\Lambda) = \frac{CI(\Lambda)}{RI_n} \quad (16)$$

The random consistency index (RI) is randomly generated from the reciprocal matrix using the scale (1/9, 1/8, …, 1, …, 8, 9) (which is similar to the bootstrap concept). In practice, a matrix with values **CR ≤ 0.1** is accepted, and one with values greater than 0.1 is rejected. Algorithm 3 illustrated the calculations of the consistency ratio for a given matrix.

**Algorithm III**

**1: procedure CHECKCONSISTENCY ( $\Lambda$ )**
**2:** $RI \leftarrow \{0,0,0.58,0.9,1.12,1.24,132,1.41,1.45,1.49\}$
**3:** $V \leftarrow \overrightarrow{P_v} \times \Lambda$
**4:** $m \leftarrow SizeOf(V)$
**5:** **For each** *feature* $f_i \in V$
**6:** **For each** *feature* $f_j \in \overrightarrow{P_v}$
**7:** **If i==j then**
**8:** $\overline{X} = \sum_{i \in V, j \in P_v} (f_i / f_j)$
**9:** **EndIf**
**10:** **End For**
**11:** **End For**
**12:** $\lambda_{max} = \mu_{\overline{X}}$ **where** $\lambda_{max}$ **should be close to** $m$
**13:** $CI(V) = \frac{\lambda_{max} - m}{m-1}$
**14:** $CR(V) = \frac{CI(V)}{RI_n}$
**15:** **If** $CR(V) > 0.1$ **then**
**16:** *consistency is not acceptable*
**17:** **Else**
*consistency is acceptable*
**18: end procedure**

### 4.2.2 Credibility classification engine

The credibility assessment task is based on two different supervised classifiers applied to the collection to guarantee high recall. For a given tweet associated with a certain topic, we can automatically determine whether the tweet is credible or not. The main hypotheses are the following:

- We can automatically estimate the credibility score of information disseminated through Twitter.
- The priority vector of the extracted features has an important impact on the process of classifying tweets.

The features that we used to classify each tweet in our collections are described in section 4.3. Most of them are restricted to the Twitter platform.

We trained a supervised classifier to determine the credibility of each tweet. Labels given by human expert evaluators were used to conduct the super-vised training phase. We trained a classifier to con-sider the two classes, credible and non-credible, along with the feature rank process, to increase the relevance of the prediction.

We used the well-known naïve Bayes classifier, along with the feature rank process, to achieve better recall. To determine in which class (credible or non-credible) a tweet T belongs, we calculate the probability that tweet T is in class $C_x$. This is written as P ($C_x$ | T), where x is credible or non-credible. Using the naïve Bayes classifier, we calculate P ($C_x$ | T) as follows:

$$P(C_x \mid T) = (P(T \mid C_x) \times P(C_x)) / (P(T)) \qquad (17)$$

because we are interested in relative, not absolute, values P(T) can safely be ignored:

$$P(C_x \mid T) = P(T \mid C_x) \times P(C_x) \qquad (18)$$

Tweet T is split into a set of features F= {$f_1$, $f_2$,…, $f_n$ }. The probability of each attribute is multiplied by its priority vector (PV) (see the previous section B) as follows:

$$P(f_i) = P(f_i \mid C_x) \times \overrightarrow{P_v}(f_i) \quad (19)$$

Thus, assuming that the attributes appear independently, P (T | $C_x$) is calculated as follows:

$$P(f_1, f_2, \ldots, f_n \mid C_x) = \prod_{i=1}^{n} P(f_i) \qquad (20)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2016.2602338, IEEE Transactions on Dependable and Secure Computing



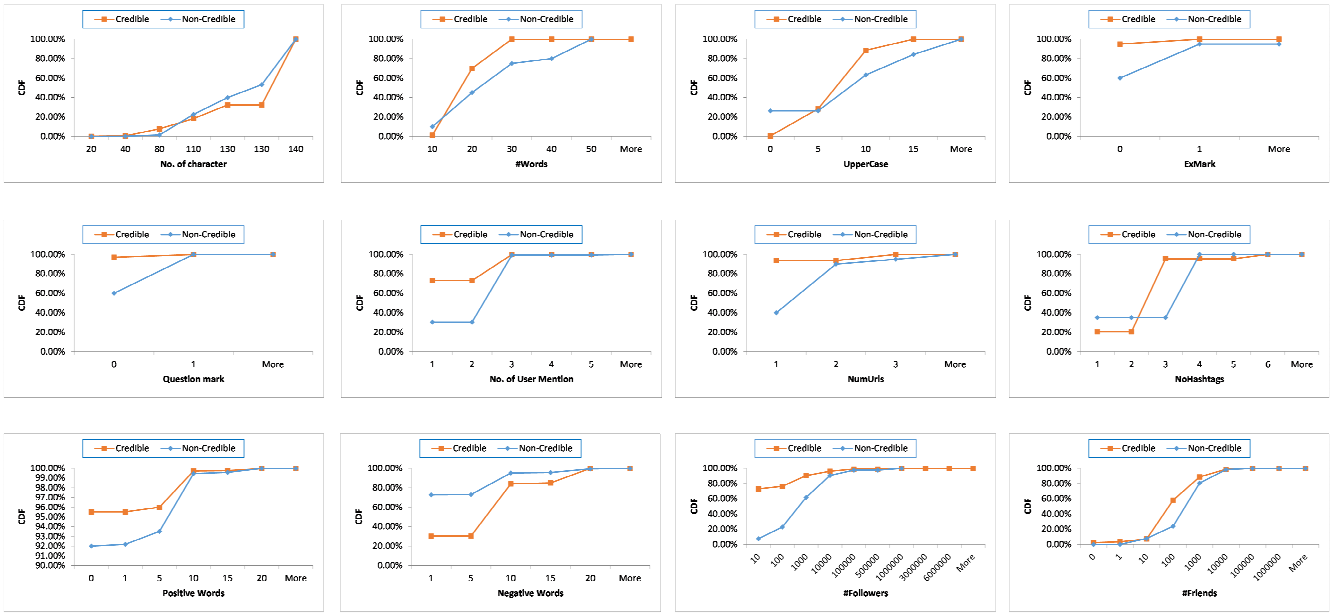


Fig. 2. Sample of Users Profiles Extracted Features Cumulative Distribution Function

For each class, selecting the largest value determines the class for tweet T, as follows:

$$C_{nb} = \arg\max_{x \in C} P(x) \times \prod_{i=1}^{n} P(f_i) \qquad (21)$$

The following algorithm represents the classification process.

**Algorithm IV**

**1: procedure** CLASSIFICATIONENGINE (*tweets, default*)
**2:** **if** *tweets* **are empty then return** *default*
**3:** **For each** $t \in tweets$
**4:** F ← EXTRACTFEATURES(T)
**5:** A ← BUILDJUDGMENTMTRIX(F)
**6:** IF CHECKCONSISTENCY(A)THEN
7: $P_v$ ← CALCULATEPRIORITYVECTOR(F)
**8:** ELSE
**9:** *Refine* **A** *until* CHECKCONSISTENCY(A)=TRUE
**10:** END IF
**11:** End For
**12:** If *CR(A)* < 0.1 then
For each $t \in tweets$
For each f ∈ F

$$P(f_i) = P(f_i \mid C_x) \times \overrightarrow{P_v}(f_i)$$

$$P(f_1, f_2, \ldots, f_n \mid C_x) = \prod_{i=1}^{n} P(f_i)$$

End For

$$C_{nb} = \arg\max_{x \in C} P(x) \times \prod_{i=1}^{n} P(f_i)$$

End For
EndIf
**21:** **return** $C_{nb}$
**22: end procedure**

### 4.2.3 Assessing User history

In this section, we describe how we assess a user's profile to measure his experience on Twitter. Users with high expertise are more credible and vice versa. User expertise can be evaluated using Algorithm V:

**Algorithm V**

**1: procedure** CALCUSERHISCREDIBIITY (*tweets, User, default*)
**2:** **if** *tweet* **is empty then return** *default*
**3:** **For each** $u \in User$
**4:** $RS_u \leftarrow$ CALCUSERREPUTATION($u$, *tweets*)
**5:** End For
**6:** **For each** $u \in User$
**7:** **For each** $t \in tweets$
**8:** $CS_t^u \leftarrow$ CLASSIFICATIONENGINE (T)
**9:** End For
**10:** $I(u) \leftarrow RS_u + CS_t^u$ **user interaction history**
**11:** $Expr(u) \leftarrow 1 - e^{-\theta \times I(u)}$ **, where** • **= 0.5**
**12:** End For
**13: end procedure**

(b) Number of Char (b) Number of Words
(i) Number of Positive Words
(e) Question Mark (f) User Mention
(j) Number of Negative Words

## 5 PERFORMANCE EVALUATION

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2016.2602338, IEEE Transactions on Dependable and Secure Computing



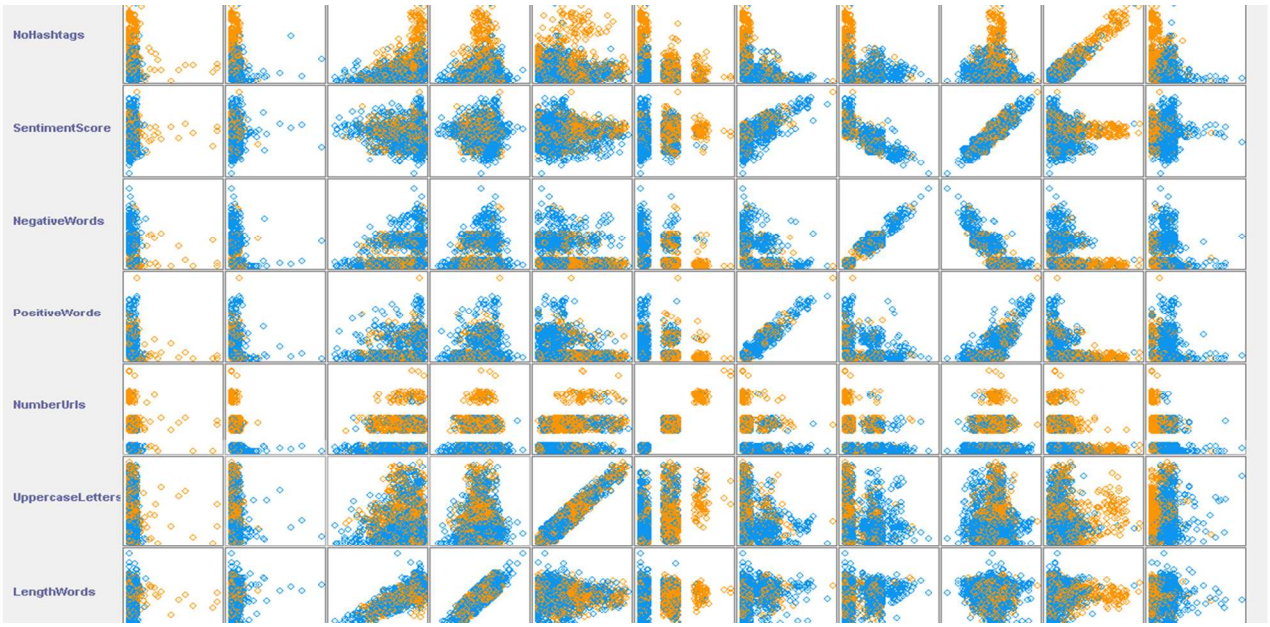


Fig. 3. Comparisons of each feature used in the credibility assessment system. In this plot, lighter areas indicate high credibility and darker areas indicate low credibility assessments.

We evaluate the performance of the trustworthiness assessment system by conducting an assessment of information from the Twitter online social network.

### 5.1 Data Set

We applied our proposed system to a set of real-world data from Twitter regarding the Saudi-led campaign against Houthi rebels in Yemen. The full dataset was divided into three separate datasets. The crawled data consisted of 1,416,443 tweets by 489,330 unique users. Datasets A and T were generated from tweets related to the Yemen Civil War. Dataset T was generated from tweets in mid-December 2015 using keywords "Taiz ( مدينة تعز)." Dataset A, which contains tweets related to the condition of the city of Aden after pro-government fighters recaptured the city, was compiled using "Aden (مدينة عدن)" as the keyword. Those data sets are related to recent events concerning Houthis, and they were generated from tweets related to situations that occurred leading up to and in the immediate aftermath of the Houthi takeover in Yemen. Furthermore, these data sets capture a range of contexts, from humanitarian concerns such as poverty, healthcare shortages, etc., to political events. For the purposes of this research, we divided the collected tweets into two groups: experimental data and prediction data. These datasets have no tweets in common. In another words, $T \cap A = \emptyset$.

After removing the intersection, we randomly sampled 23,000 tweets concerning topic T from the experimental data set and 25,000 tweets concerning topic A. Then, we randomly sampled 4,000 tweets concerning topic T from the prediction dataset and 7,000 tweets concerning topic A. In addition, from each data set, a distinct list of users was examined. Twitter allowed us to pull 3,200 tweets for each user. The following Table 1 summarized the sampled users' profiles from both datasets.

**Table 1: Sampled users' profiles**

| Datasets | #Tested Users Profiles | #of Tweets |
|---|---|---|
| Taiz (T) | 1363 | 3,847,623 |
| Aden (A) | 1480 | 4,022,926 |

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2016.2602338, IEEE Transactions on Dependable and Secure Computing



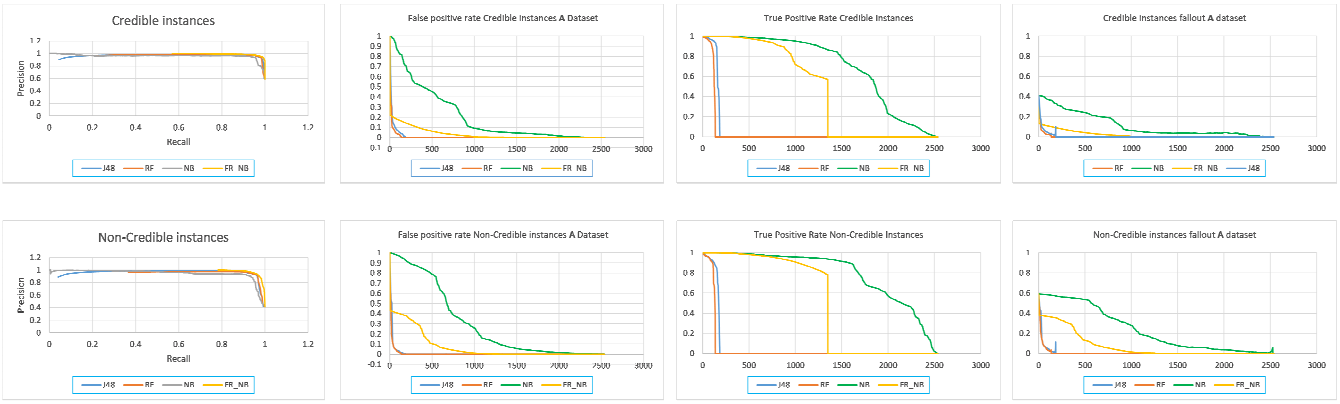


Fig. 4. Dataset ***A*** false positive, true positive, fallout, precision, and recall

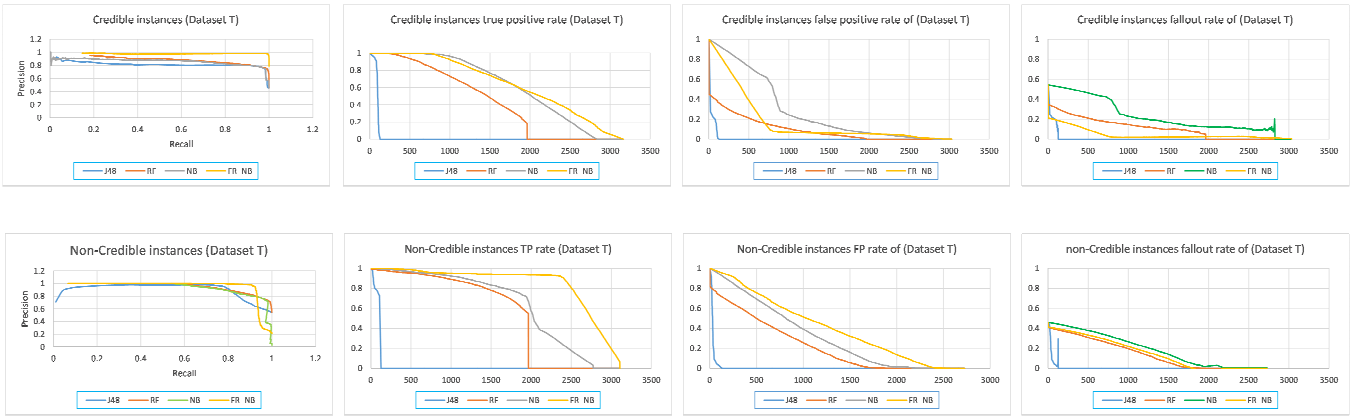


Fig.5. dataset ***T*** false positive, true positive, fallout, precision, and recall

The following sections desc ribe a sample of labeled data distribution and then we discuss results analysis.

## 5.2 Data Analysis

After labeling the credible and non-credible tweets, we further extracted features from them. We were not able to build a social graph from the public data because Twitter's public streaming and searching APIs provide the only access to public tweets, and they are not socially connected. Consequently, it is not possible for us to extract social graph-based features such as local centrality and distance properties. Such expensive features are not suitable for use in the classification process, despite their having more power in discriminating between credible and non-credible tweets. In addition, we preferred to use features that could be computed from the content in a straightforward manner. To ensure the relevance of the content to our topic, we remove greeting tweets and any other redundant tweets. This step is crucial at the end of all systematic methods to guarantee the results and outcomes. The features we extracted from our datasets are listed in section 4.3. Depending on the object from which the features were extracted, the features can be classified into three categories: tweet-level features, user-level features, and hybrid-level features. We analyzed the collected datasets from two per-spectives: the tweets' topics and the users' history profiles.

### 5.2.1 Dataset of users' history

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2016.2602338, IEEE Transactions on Dependable and Secure Computing



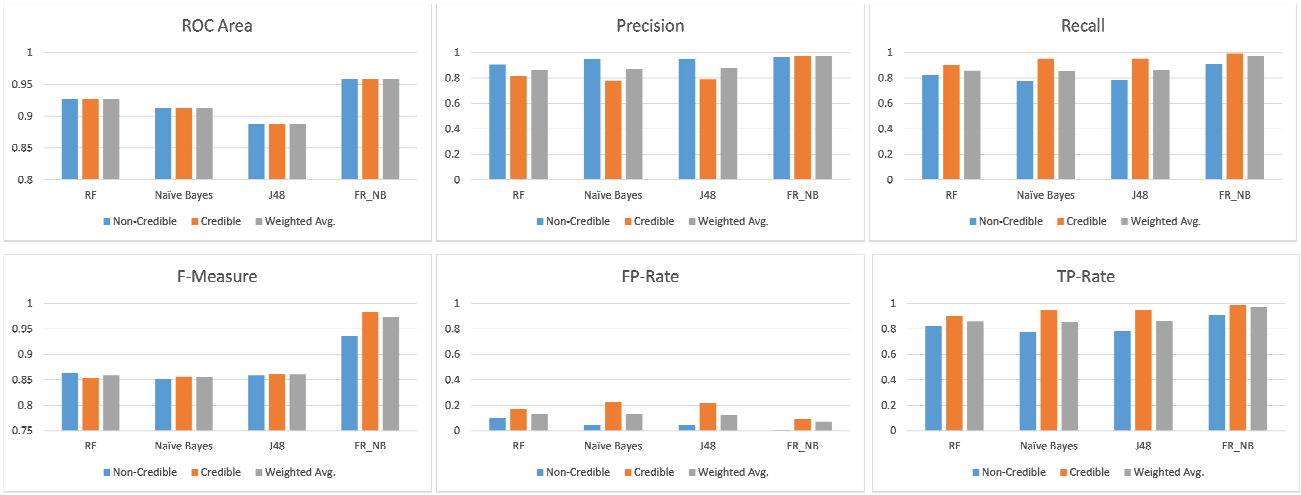


Fig. 6. Overall performance comparison with different approach over dataset ***A***

**Table 2: Statistical results from for the best performing of the credibility assessment system over dataset *A***

| Dataset *A* | Random Forest | Naive Bayes | J48 | FeatureRank_NB |
|---|---|---|---|---|
| **Attributes** | 21 | 21 | 21 | 21 |
| **Correctly Classified Instances** | 95.2133% | 93.0311% | 95.157% | 96.0439% |
| **Incorrectly Classified Instances** | 4.7867% | 6.9689% | 4.843 % | 3.9561% |
| **Kappa statistic** | 0.9009 | 0.8562 | 0.8997 | 0.918 |
| **Mean absolute error** | 0.0608 | 0.0827 | 0.0657 | 0.0558 |
| **Root mean squared error** | 0.1964 | 0.2441 | 0.1979 | 0.1714 |
| **Relative absolute error** | 12.5745% | 17.0944 % | 13.5739 % | 11.5229 % |
| **Root relative squared error** | 39.9328% | 49.6268 % | 40.2417 % | 34.856 % |

We employed a randomly sampled selection of users' profile datasets. To examine the characteristics of these profile features, we plotted the cumulative distribution function (CDF) of one sample of each level, as shown in Figure 2. We can see from Figure 2 (a) that non-credible tweets tend to have fewer characters per tweet than credible tweets. With respect to the feature "Number of words per tweet," there is not much difference between credible tweets and non-credible tweets. The reason for this could be that users who propagate non-credible tweets try to mimic the posting behavior of credible sources, as shown in Figure 2 (b). Almost 99% of the credible tweets have no exclamation marks, whereas 80% of the non-credible tweets have at least one exclamation mark, as shown in Figure 2 (d). Likewise, as shown in Figure 2 (e), non-credible tweets tend to have at least one question mark. Non-credible tweets tend to have more hashtags than credible ones, as shown in Figure 2 (h). However, credible tweets have more user mentions than non-credible tweets, and they are more stable in their distribution with respect to the topic or event than non-credible tweets, because of the many outliers in the distribution of user mentions in non-credible tweets. More than 30% of the non-credible tweets do not have mentions embedded in their sent tweets, while the ratio in credible tweets is 73%. In Figure 2 (j), we can see that non-credible tweets are more likely to be negative than credible tweets, whereas, as we can see in Figure 2 (i), credible tweets tend to be more positive than non-credible ones. Finally, users who propagate non-credible tweets seem to have fewer followers than credible ones, as shown in Figure 2 (k). In this paper, we only describe a sample of the feature characteristics that we tested. The results of the analysis of these features revealed their relative power to discriminate between credible and non-credible news on Twitter.

### 5.2.2 Features of Datasets

In this section, we describe the relationships be-tween the different features of the T Dataset. Figure 3 illustrates the results of a comparative analysis of some selected features of a hybrid- based model. In this figure, the brighter orange nodes correspond to credible tweets, and the blue nodes correspond to non-credible tweets. Clusters appear in some of the scatter plots, indicating that the feature does have some role in assessed credibility. For example, from the plots for the features "NoHashtags" and "No-Mentions," it is clear that tweets with fewer hashtags and fewer mentions tend to be judged to be more credible than others. Tweets with fewer negative words and more followers also align well with report-ed credibility. We observe a positive linear correlation between the number of words and number of char-acters. Meanwhile, the plot of number of uppercase letters versus number of characters exhibits no cor-relation: the cluster of points is almost round, and a line does not fit the points in the plot well. As the correlation coefficient increases, the observations group closer together in a linear pattern. The line is difficult to detect when the relationship is weak (e.g., upper-case letters and positive words), but becomes clearer as relationships become stronger (e.g., sentiment score and number of negative words).

## 5.3 Classification Engine Training

The credibility assessment component of the classification engine requires training before it can be used. Its model was trained on credible and non-credible data sets. The non-credible

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2016.2602338, IEEE Transactions on Dependable and Secure Computing



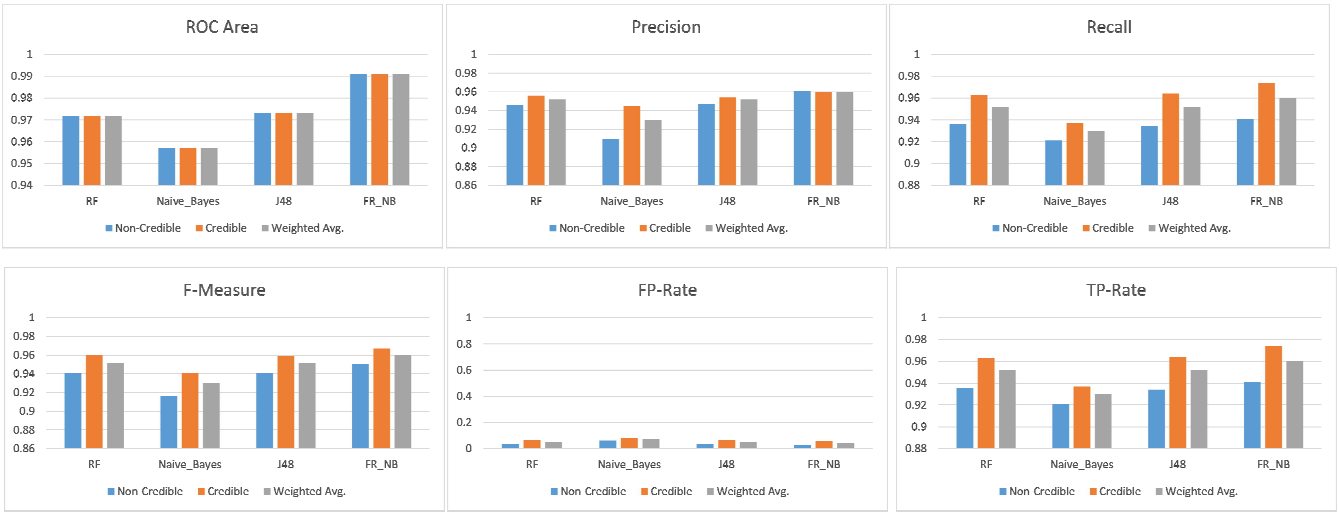


Fig. 7. Overall performance comparison with different approaches over dataset ***T***

Table 3: Statistical results from for the best performing of the credibility assessment system over dataset *T*

| **Dataset *T*** | **Random Forest** | **Naive Bayes** | **J48** | **FeatureRank_NB** |
|---|---|---|---|---|
| **Attributes** | 21 | 21 | 21 | 21 |
| **Correctly Classified Instances** | 85.9163% | 85.4812% | 86.0794% | 91.4187% |
| **Incorrectly Classified Instances** | 14.0837% | 14.5188% | 13.9206% | 8.5813% |
| **Kappa statistic** | 0.7187 | 0.7124 | 0.724 | 0.8911 |
| **Mean absolute error** | 0.1805 | 0.1594 | 0.2105 | 0.0279 |
| **Root mean squared error** | 0.3218 | 0.3489 | 0.3319 | 0.1536 |
| **Relative absolute error** | 36.3504% | 32.0941% | 42.4027 % | 8.3599 % |
| **Root relative squared error** | 64.5922% | 70.0168% | 66.6066% | 37.5793 % |

data set consisted of non-credible tweets and non-credible users, which were identified throughout the creation of the ground truth data set. Some credible/verified users sometimes post non-credible tweets (usually unintentionally), and these tweets go viral due to their popularity. Thus, any tweet should be examined us-ing the classification engine, even if it obtained a 100% score from the reputation-based scoring technique. We did not consider those users who have no social popularity. To obtain the benefit of the accuracy of human-based evaluation to build the ground truth and label data, we selected purview users (e.g., citizens, wit-nesses, experts, media reporters, etc.). In addition, the selected annotators have technical expertise with Twitter. They examined and evaluated tweets using URLs by checking linked webpages, searching the attached media, and comparison with credible sources. Training the models with up-to-date text and metadata feature patterns on Twitter helped to improve their accu racy. In addition, we created two predefined lists of sentiment words and users' orientations words with high frequency on Twitter to help in the assessment of users' history with respect to the context.

## 5.4 Measuring Accurcy

In addition to the reputation-based model, we implemented our credibility assessment system by training four machine learning algorithms. Only approximately 0.006 and 0.001 of the users in both the T and A datasets, respectively, passed the reputation-based filter with a 100% score. Those users were considered verified users. We applied tenfold cross-validation to train and test the classifier over the ground truth set. Given a dataset ***D*∈ {T, A}** and a class model ***m*∈ {credible, non-credible},** we want to choose a model ***m*** that will give a suitable classification for ***D***. To achieve that aim, we need to minimize our expected loss ***EL*** as follows:

$$\varepsilon_m = \min EL(d, f_m(d))$$

where ***d*** is an element of our dataset (tweet or user) ***d*∈ *D*** and $f_m$ ***(d)*** is the classifier function that we can obtain for dataset ***D***. the cross-validation is conducted as follows: First, we randomly permute the dataset. Second, we split the data set into K equally sized and complementary folds, where, in our case, K = 10 and each fold is a subset of the dataset D. Third, in the validation stage, we set the number of rounds equal to the number of folds. In each round, one of ten folds is considered to be the validation set to test the classifier, while the remaining nine folds are used as the training set to train the classifier. Fourth, the results from the ten rounds are averaged to generate the final estimation. For each model m in our classifiers list, the empirical error for each constructed round is averaged follows:

$$\hat{\varepsilon}_m = \frac{1}{K}\sum_{i=1}^{K}\varepsilon_m(i)$$

Next, we choose the class ***m*** that minimize $\hat{\varepsilon}_m$. We then retrain all datasets using the selected ***m***. We chose the cross validation technique for several reasons. First, the samples from the datasets were validated one at a time, unlike random subsampling, which might result in choosing the same sample more than once, which would lead to overestima-tion of the error. Second, all samples in the data set were used for both training and validation. Conse-quently, each round was

processed independently without affecting the following ones.

The results obtained using the classification engine, which contains four learning algorithms—random forest (RF), naïve Bayes (NB), decision tree (J48) and feature-rank naïve Bayes (FR_NB)—are shown in Tables 4 and 5 for datasets A and T, respectively. For dataset A, the supervised classifier achieved accura-cies of 95%, 93%, 95%, and 96%, as shown in Table 2. The kappa statistics show that our classifier engine achieved significant predictability values of 0.90, 0.85, and 0.89 for the first three algorithms and a very significant value of 0.918 for the F_NB algorithm. whereas, for dataset T, the supervised classifier achieved accuracies of 85.9%, 85.4%, 86%, and 91.4%, as shown in Table 3. The kappa statistics show that the predictability values achieved by our classifier engine for the first three algorithms were 0.72, 0.71, and 0.72 and that a significant value 0.89 was achieved for the F_NB algorithm.

The percentage of misclassifications and the error rate between credible and non-credible tweets were very small for dataset A, as shown in Table 2. However, these may be higher for dataset T because T has more noise in its features than dataset A, as we observed during our experiments.

Other accuracy measurem ents of our credibility assessment system include the false positive (FP) rate, true positive (TP) rate, fallout rate, precision, and recall, F-measures, and Receiver Operating Characteristic (ROC) curve. We carried out these series of tests on both the A and T datasets. The TP rate reflects how truly credible a tweet is, i.e., how often the classifier is correct. TP is also referred to as "sensitivity" and reflects how sensitive the classifier is to detection of positive instances. The sensitivity or TP rate is equivalent to recall. The FP rate reflects how truly non-credible a tweet is, i.e., how often the classifier is incorrect. The precision is the rate of how precise the classifier is when predicting credible tweets. In our case, the priority was to balance the sensitiv ity and specificity of our classifier engine, be-cause credible and non-credible observations are critical in terms of emergencies and crises. In binary classification problem such as ours, the default threshold is equal to 0.5. Both sensitivity and specificity are affected by tuning the default threshold for better balancing: lowering the threshold will increase the sensitivity, and raising it will increase the specificity. We can examine the tradeoff between the two different parameters using the ROC metric. We use ROC to choose a threshold that balances these two parameters. The details of the assessment per class are shown in Figs. 4, 5, 6, and 7 for both datasets A and T. As we can see, the classifier yields excellent results for the prediction of both datasets, achieving the best TP rate and FP rate across the classes. F-measures equivalent to 0.864 for RF, 0.852 for NB, 0.859 for J48, and 0.937 for FR_NB illustrate that, es-pecially for the non-credible class, the classifier en-gine achieves a good balance in the precision–recall tradeoff. Likewise, it yields very good results for the credible class, with F-measures of 0.864 for RF, 0.852 for NB, 0.859 for J48, and 0.937 for FR_NB. The classifier also achieves significantly better performance than would be achieved by random chance, as indi-cated by the high kappa statistics in Tables 5 and 6, as well as the high ROC area values, as shown in Figure 7. Moreover, for dataset A, the results obtained are better than for dataset T, as shown in Figure 6. The reason for this is that dataset T is a noisier than A, which makes the classification process more difficult.

## 6 Conclusion

This paper presents the results of a study of the problem of assessing information credibility on Twit-ter. The issue of information credibility has come un-der scrutiny, especially in social networks that are now being used actively as first sources of information. Twitter and other social networks have be-come widely used in disaster mitigation in cases of high-impact events because they make it possible for relevant parties to obtain important information sufficiently quickly to coordinate countermeasures to such events.

To obtain a better understanding of how to assess information credibility on Twitter, we measured and characterized the content and sources of Twitter tweets. By crawling Twitter, we collected data from more than 1,416,443 tweets by 489,330 unique users. In addition, we examined data for 2,843 Twitter users with more than 7,870,549 tweets. Based on the data, we extracted the features that can be of most help in the assessment process. Based on our feature ex-traction process, we designed an automated classification system that consists of four main components: a reputation-based component, a credibility classifier engine, a user experience component, and a feature rank algorithm. The reputation-based technique helps to filter neglected information before starting the assessment process. The classifier engine component distinguishes between credible and non-credible content. The user expertise component yields ratings of Twitter-user expertise on a specific topic. Finally, the feature rank algorithm helps in selecting the best features, based on the relative importance of each feature. The effectiveness of the system was evaluated using testing two datasets. We also applied the system to classification of users' profiles using more than 7,870,549 collected tweets. In the near future we will try to analyze the credibility using time-sensitive and location-based approaches that give more reliable and trusted results.

## Acknowledgment

This work was full financially supported by the King Saud University, through Vice Deanship of Research Chairs.

## References

[1] M. Al-Qurishi, R. Aldrees, M. AlRubaian, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, "A new model for classifying social media users according to their behaviors," in Web Applications and Networking (WSWAN), 2015 2nd World Symposium on, 2015, pp. 1-5

[2] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, "A Multi-stage Credibility Analysis Model for Microblogs," presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 2015

[3] A. A. AlMansour, L. Brankovic, and C. S. Iliopoulos, "Evaluation of credibility assessment for microblogging: models and future directions," in Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business, 2014, p. 32.

[4] Majed AlRubaian, Muhammad Al-Qurishi, Sk Md Mizanur Rahman, and A. Alamri, "A Novel Prevention Mechanism for Sybil Attack in Online Social Network," presented at the WSWAN'2015, 2015

[5] P. T. M. Gayo-Avello, Eni Mustafaraj, Markus Strohmaier, Harald Schoen, D. Peter Gloor, C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media,"

Internet Research, vol. 23, pp. 560-588, 2013.

[6] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011.

[7] S. Y. Rieh, M. R. Morris, M. J. Metzger, H. Francke, and G. Y. Jeon, "Credibility Percep-tions of Content Contributors and Consumers in Social Media," 2014.

[8] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in Proceedings of the ACM 2012 confer-ence on Computer Supported Cooperative Work, 2012, pp. 441-450.

[9] Pal, A. and Counts, S. What's in a @name? How Name Value Biases Judgment of Microblog Authors. in Proc. ICWSM, AAAI (2011)

[10] J. Schaffer, B. Kang, T. Hollerer, H. Liu, C. Pan, S. Giyu, and J. O'Donovan, "Interactive interfaces for complex network analysis: An information credibility perspective," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on, 2013, pp. 464-469

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

[12] Metaxas, Panagiotis Takas, Samantha Finn, and Eni Mustafaraj. "Using TwitterTrails. com to Investigate Rumor Propagation." Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing. ACM, 2015

[13] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, 2012, p. 2

[14] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?," in Proceedings of the first workshop on social media analytics, 2010, pp. 71-79

[15] Westerman, D., Spence, P.R., and Van Der Heide, B.: 'A social network as information: The effect of system generated reports of connectedness on credibility on Twitter', Computers in Human Behavior, 2012, 28, (1), pp. 199-206

[16] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on twitter," in Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 179-188

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

[18] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, 2012, p. 2

[19] K. R. Canini, B. Suh, and P. L. Pirolli, "Finding Credible Information Sources in Social Networks Based on Content and Social Structure," in Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, 2011, pp. 1-8

[20] J. R. Nurse, I. Agrafiotis, M. Goldsmith, S. Creese, and K. Lamberts, "Two sides of the coin: measuring and communicating the trustworthiness of online information," Journal of Trust Management, vol. 1, pp. 1-20, 2014

[21] J. O'Donovan, B. Kang, G. Meyer, T. Hollerer, and S. Adalii, "Credibility in context: An analysis of feature distributions in twitter," in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 2012, pp. 293-301

[22] K. K. Kumar and G. Geethakumari, "Detecting misinformation in online social networks using cognitive psychology," Human-centric Computing and Information Sciences, vol. 4, pp. 1-22, 2014

[23] T. Mitra and E. Gilbert, "CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations," in Ninth International AAAI Conference on Web and Social Media, 2015

[24] K. Lorek, J. Suehiro-Wicinski, M. l Jankowski-Lorek, and A. Gupta, "AUTOMATED CREDIBILITY ASSESSMENT ON TWITTER," Computer Science, vol. 16, pp. 157-168, 2015

[25] D. Saez-Trumper, "Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News on Twitter," 2014

[26] M. Al-Qurishi, M. Al-Rakhami, M. AlRubaian, A. Alarifi, S. M. M. Rahman, and A. Alamri, "Selecting the best open source tools for collecting and visualzing social media con-tent," in Web Applications and Networking (WSWAN), 2015 2nd World Symposium on, 2015, pp. 1-6

[27] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," Computer Speech & Language, vol. 28, pp. 20-37, 2014

[28] T. L. Saaty, "Relative Measurement and Its Generalization in Decision Making Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors The Analytic Hierarchy/Network Process (To the Memory of my Beloved Friend Professor Sixto Rios Garcia)," Revista De La Real Academia De Ciencias Exactas Fisicas Y Naturales Serie a-Matematicas, vol. 102, pp. 251-318, 2008

[29] Software-Defined Mobile Networks Security", ACM/Springer Mobile Networks and Applications, DOI: 10.1007/s11036-015-0665-5, 2016.

**Majed Alrubaian (M'16)** is a Ph.D. candidate in Information Systems Department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia. He received his Master degree in Information Systems from King Saud University, Kingdom of Saudi Arabia. He has some paper publications in refereed IEEE/ ACM/ Springer conference and journals. He is a student member of ACM and IEEE. His research interest includes Social Media Analysis, Data Analytics and Mining, Social Computing, information credibility, and Cyber Security. He is a student member of IEEE.

**Muhammad Al-Qurishi (M'16)** is a Ph.D. candidate in Information Systems Department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia. He received his Master degree in Information Systems from King Saud University, Kingdom of Saudi Arabia. He has some paper publications in refereed IEEE/ACM/Springer conference and journals. His research interest includes Online Social Network, Social Media Analysis and Mining, Human-computer interaction, and Health Technology. He is a student member of IEEE.

**Mohammad Mehedi Hassan (M'12)** is currently an Assistant Professor of Information Systems Department in the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Kingdom of Saudi Arabia. He received his Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea in February 2011. He received Best Paper Award from CloudComp 2014 conference at China. He also received *Excellence in Research Award* from CCIS, KSU in 2015 & 2016. He has published over 100+ research papers in the journals and conferences of international repute. He has served as, chair, and Technical Program Committee member in numerous international conferences/workshops. He has also played role of the Guest Editor of several international ISI-indexed journals. His research areas of interest are cloud federation, multimedia cloud, sensor-cloud, Internet of things, Big data, mobile cloud, cloud security, IPTV, sensor network, 5G network, social network, publish/subscribe system and recommender system. He is a member of IEEE.

**Atif Alamri (M'12)** is an Associate Professor of Information Systems Department, at the College of Computer and Information Sciences, King Saud University. Riyadh, Saudi Arabia. His research interest includes multimedia assisted health systems, ambient intelligence, and service-oriented architecture. Mr. Alamri was a Guest Associate Editor of the IEEE RANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, a Co-chair of the first IEEE International Workshop on Multimedia Services and Technologies for E-health, a Technical Program Co-chair of the 10th IEEE International Symposium on Haptic Audio Visual Environments and Games, and serves as a Program Committee Member of many conferences in multimedia, virtual environments, and medical applications.