Towards Customizing Credibility in Different Contexts: Languages, Topics and Locations - A Twitter Case Study

Amal Abdullah AlMansour
King Abdul-Aziz University
Dept. of Computer Science
Jeddah – Saudi Arabia
aalmansour@kau.edu.sa

ABSTRACT

Even though online social networks (ONS) have been increasingly used as a source of news and information, the credibility of that easily-available information may be questionable. In this paper, we review selected literature related to information credibility focusing more on micro-blogging credibility as a part of ongoing research. From this review it can be concluded that credibility is situational and contextual; it varies from one context to another. We propose to examine messages related to confirmed news topics and false rumors topics and then study the effect of three dimensions: language, topic content, and location on the detection features utilized by the existing work. Firstly, we will check if the frequency distribution of selected features remains similar or changes according these dimensions. We then examine different available classifying techniques and study the effect of each context on the credibility classifying results. Another goal is to report on user studies that explore how people from different environments perceive and judge information credibility on Twitter. We believe the analysis of our results will improve existing credibility measurement by helping us in selecting the most suitable features for information credibility classification in each context.

KEYWORDS

Social networks; Micro-blogs; Credibility; Arabic; Machine Learning.

1 INTRODUCTION

With the growing popularity of online social networks over the past few years, sharing and spreading news became faster and easier more than before. Social networks allow their members to easily and freely disseminate real-time news information to a large number of people - in some cases, before traditional media [1]. However, this unfiltered and distributed nature of social networks facilitates spreading baseless rumors besides valid truthful news. As more people search and browse social media platforms (e.g., Twitter, Facebook, etc.) for sensitive topics such as health, politics, business and crises news, the presence of misleading, questionable and inaccurate information may have detrimental effects in people's beliefs and decision-making and may create a public disturbance (e.g., Apple shares1 and Chile earthquake2). As a consequence, there is significant importance to evaluate information comes from social networks sources. However, assessing the information credibility is challenging task since the term “credibility” is a structure based on user’s cognitive state. Users usually invoke cognitive heuristics to assess the credibility. In other words, it is the subjective judgment and assessment from the users [2]. Also, credibility is situational and contextual; it varies from one context to another. A person sometimes accepted certain information as credible primarily by relying on the context in which he/she encountered the information [3].

This study suggests that available automatic credibility assessments need to consider both in their process, people credibility judgments and credibility contexts such as environment, situations, expectations, etc. [3, 4]. As a start, we

1 http://money.cnn.com/2008/10/03/technology/apple/
2 http://irevolution.wordpress.com/2010/06/30/crowdsourcing-detective/
will address assessing information credibility found through one of the popular real-time information resources on the Web - Twitter that not acts only as a social network, but as a news source [5, 6]. Then we will consider other social network domains. In the beginning, we will extract selected features from posts related to both true and false news topics and study how these features are being distributed surrounding different contextual dimensions: cultural, situational, and topical variations. The aim is to develop an understanding of how different communities (both in the U.S.A/U.K and in the Arab countries) act during different situations in order to determine when specific features are useful in determining credibility. Then by using machine learning, mainly supervised learning method, we will study the effect of the context on the credibility classifying results of a given classifying technique. There are different useful features proposed by the previous studies however the effect of the message language, topic, and location on these detection features is not considered. We suggest that assessing information credibility with respect to its context is useful to understand how and why users carry out with their credibility judgments of information arguing that the perceived information credibility will change depending on the context.

A secondary goal is to report on survey data that compare credibility perceptions among English speaking countries and Arab countries audiences. We will try to identify any differences in how users from each country consume micro-blog content, and discuss how to incorporate these findings into classifying Twitter information automatically in different cultural settings. Also in this survey, we are interested to understand rumors diffusion through the social networks and analyze users' attitudes toward such rumors. We want to know: what kinds of content are most likely to be re-tweeted and why, would people re-tweet information even if they are not sure about their truthfulness? Another looked-for goal would be to develop a browser plug-in tool that can recognize message’s language, topic, situation, and location. Then calculate the credibility score depending on the available message’s features and context.

We think that our results will provide new insights to the information credibility research. (1) Introduce credibility model that deals with different contexts and maintains people credibility visions, (2) identify messages’ features that users care about the most within each different context. By identifying these features, we could (3) enrich existing automatic credibility classification techniques, as an example proposing to adjust features’ weights depending on its occurrence and importance to the end users. It should be noted that in this work we do not introduce a new method for assessing credibility automatically, but empirically study how, and to what extent, the existing automatic credibility measurement can be used to identify credible information in different context. This study will address the following research question: What type of available features is most useful for informing credibility judgment’s about Twitter information in different contexts? Do different contexts lead to similar credibility judgments? And to what degree it remains the same?

This paper is organized as follows: section 2 presents related Twitter information credibility research, along with credibility contexts surveys and features. Section 3 and 4 presents a proposed solution along with future work.

2 RELATED WORK

Most previous research related to this work falls into these broad categories: automatic measurement of information credibility, Arabic content credibility, and then we will investigate different credibility contexts surveys along with Twitter features and datasets that have been used for credibility classification in previous studies.

2.1 Automatic measurement of information credibility

In assessing information credibility automatically, different research relies on different methods and factors to identify information credibility. Some used machine learning, mainly supervised learning methods and some used graph analysis while others used only statistical analysis like feature
distributions to predict credibility. Most of these methods are summarized in Table 1.

### Table 1. Assessing information credibility automatically

<table>
<thead>
<tr>
<th>Model Used</th>
<th>Used by - Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph-based / Hybrid (Classification with Graph)</td>
<td>Ratkiewicz et al. 2011 [15]; Gupta, Zhao, &amp; Han 2012 [16]; Ravikumar, Balakrishnan, &amp; Kambhampati 2012 [17]; Ulicny &amp; Kokar [18]; McKelvey &amp; Menczer 2013 [19].</td>
</tr>
<tr>
<td>Weighting-based feature/ Content similarity with credible source: mainly linguistic features</td>
<td>Assigning scores by certain mathematical functions or algorithms to each feature. Al-Eidan, Al-Khalifa, &amp; Al-Salman 2010 [20]; Al-Khalifa &amp; Al-Eidan 2011 [21].</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>Features distributions by ODonovan et al. 2012 [22].</td>
</tr>
</tbody>
</table>

Based on the good results achieved using supervised machine learning classifiers in assessing information credibility, we will focus more on investigating research adopted this method.

### 2.2 Arabic content credibility

A work regarding Twitter credibility for Arabic content has been done by Al-Eidan, Al-Khalifa, & Al-Salman 2010 [20] and Al-Khalifa & Al-Eidan 2011 [21]. They proposed a system to evaluate the credibility of Twitter Arabic news content automatically. Their method is only useful for tweets combined with credible external sources also they didn’t embrace most of prominent features proposed by previous research such as hash-tags, re-tweets, and emoticons. Moreover, there is a need to investigate their credibility formula and if it works after adding more features.

### 2.3 Credibility contexts surveys

To identify influential contexts, several surveys have been carried out: Kang, ODonovan, & Hollerer 2012 [11] provided a user study to analyze the effects of Twitter specific data (users’ information, followers, and re-tweets) on perceived credibility. Other study by Canini, Suh, & Pirolli 2011 [23] evaluated the effect of context variance on perceived credibility through a user experiment. They conducted a user experiment to measure the extent to which different factors (expertise, social status, and visualization) affect both explicit and implicit judgments of credibility.

Pal & Counts 2011 [24, 25] studied how bias due to (author’s name value) impacts the perception of quality of Twitter authors. In their survey study, Morris et al. 2012 [26] manipulated several features of tweets (message topic, user name, and user image) to assess their impact on credibility ratings. In [27], Yang et al. 2013 ran an online study in which participants from U.S. and China rated the credibility of tweets, with each of the factors of interest manipulated (gender, name style, profile image, location, network overlap and message topic). A study designed to examine the effects of the number of followers and the ratio between Twitter followers and follows had on ratings of trustworthiness has been conducted by Westerman, Spence, & Van Der Heide 2012 [28].

Unlike our work, previous surveys mainly manipulate Twitter specific data within their experiment to measure the impact of their manipulation on users’ judgments; here we are more concerned about different external contexts: cultural, languages, and situations on affecting the credibility.

### 2.4 Twitter features and datasets

There are a wide range of features proposed in many different studies to assess credibility of tweets. Most of these studies rely on different Twitter features related to the messages’ author and content [7, 8, 9, 10, 11, 12, and 22] for assessing information credibility. Some mainly focus on the linguistic features of the content [13, 14, 20, and 21] and others focus on one or two Twitter specific factors and check their validity.
such as the work by Kang et al. 2013 [12] and ODonovan et al. 2012 [22], who studied how re-tweet chain length and dyadic exchanges are used as metrics to measure credibility.

In reference to tested datasets, some used different topics for assessing credibility [7, 8, 9, 13, 14, 20, and 21] others only rely on a single topic dataset [10, 11, 12, and 29]. Table 2 shows a sample of Twitter features and datasets that had been used in previous work, focusing on the prominent features in each work, and information credibility level.

Table 2. Twitter features and datasets

<table>
<thead>
<tr>
<th>Used by</th>
<th>Datasets, Prominent Features and Credibility Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gupta &amp; Kumaraguru 2012 [8]</td>
<td>14 different news events Content: characters count, unique characters count, swear words count, inclusion of pronouns, presence of sad / happy emoticons, and presence of URL; User: followers count, and username length. Tweet level; Rank-SVM and relevance feedback.</td>
</tr>
</tbody>
</table>

3 PROPOSED SOLUTION AND SYSTEM ARCHITECTURE

In this study we are planning to answer research questions from two perspectives. First, we will conduct a user study to explore how people from different environments perceive and judge information credibility on Twitter. Second, we will investigate how useful existing works are for classifying information credibility automatically given different contexts. In the process of classifying information credibility automatically, we will need to identify:

1. **Twitter features:** we will rely on the user survey and related work to identify the most prominent features that indicate the influence of the authors and the quality of the tweets. On the user survey, users will rate the importance of each of the different attributes. We will present a list of prominent features from previous studies and for each feature, they will be asked to select to what degree this feature convey tweet credibility. The survey also will take in consideration the different contexts and see if users’ answers remain the same. Provided below are some of the features we are planning to extract and compute:

   a. **Reputation of Authors:** followers count (possible views/popularity/diffusion), friends count, list count (authority/expertise), statuses count, registration age, verified account (popularity and authority), description (expertise), follower-friend ratio.

   b. **Quality of content:** message length, unique characters count, words count, swear words count, presence of URL (reference), hash-tags count, re-tweets count (voting/diffusion), mentions count (responsivity), presence of sad or happy emoticons, presence of question mark?, presence of exclamation mark!, presence of pronouns, uppercase characters count. In terms of the aggregated posts on the topic, important predictors include: fraction of tweets having URL, fraction of negative and positive sentiment words, fraction of tweets with an exclamation mark, presence of URL, re-tweets, and sentiment; User: registration age, followers count, friends count, status count, and listed count.
2. **Contexts:** there are infinite choices for context, so for our analysis we will focus on a carefully chosen subset related to the main elements (content, author, and reader):
   
   a. **Topic Type:** there are multiple interpretations of credibility depending on to the type of information to be evaluated. Example: when a person evaluates health information, maybe expertise source would be more important than any other element.
   
   **H1a:** Messages’ features change depending on the topic content which means different topic content will vary in their credibility measurement.
   
   **H1b:** Individuals’ perceived credibility for message and message’s author will vary according to the topic genre.
   
   b. **Language:** we think that different languages may depend on different linguistic features in measuring credibility. A study done by Alarifi & Alsaleh 2012 [30] in detecting Arabic and English spam web pages showed that the distributions of the content features vary according to the underlying language of the examined page. Moreover, content features that are studied in previous work for English tweets, such as the uppers cases characters count does not apply to the Arabic language.
   
   **H2a:** Messages’ features change depending on the language which means different languages will vary in their credibility measurement.
   
   **H2b:** Individuals’ perceived credibility for message and message’s author will vary according to different languages.
   
   c. **Location and culture:** [US/UK vs. Arabs] users may find information comes from liberal sources to be more credible than when it comes from people living on conservative locations such as the Arab countries [27].
   
   **H3:** Different communities will have different credibility perceptions.

### 3.1 Proposed System Architecture

The proposed system architecture consists of four main phases illustrated in Figure 1.

![Proposed System Architecture](image-url)

1. **Datasets collection:** collect true and rumor messages related to events happened on different countries and categorize them according to language, topic, and location. We will use sets of keywords and hash-tags related to each event.

2. **Labeling:** using the user study, participants from different communities will label the tweets of different topics to measure the believing degree. In general, users will label the date and weight the features. By doing this, we want to know if humans are capable of distinguishing credible tweets or not.

3. **Feature extraction and analysis:** Messages will be represented by a set of computed features.

4. **Classification:** using classifier-based feature, the final datasets will be used by the classifier to train its model and acquire necessary knowledge which will be used to classify the messages credibility. Two widely known classifiers are decision tree and naïve bayes. Using weighting-based feature approach [20, 21] as a second technique, each feature is given an experimental weight based on its
importance (users survey), and frequency of appearance in a tweet. Then we will use a mathematical formula to calculate the credibility score.

4 CONCLUSION AND FUTURE WORK

In this paper, we reviewed the previous work in automatic information credibility techniques related to social media in particular Twitter as a part of on-going research. It is concluded that: (1) even though previous research already proposed different features to assess information credibility in social networks platforms. However, they didn’t investigate the usefulness of these features in informing credibility judgments in different contexts such as culture, language, topic and situation, also incorporating users in weighting the credibility features and let them decide their importance then use them to compute credibility score has not been taken lots of attention in previous work. (2) Mostly, the ground truth for credibility classification was assessed by humans from crowdsourcing websites; and was not rated from trusted media sources or experts in topic domain. (3) It is noted that by using different datasets, some of the used features were prominent regardless of the datasets, but many were not. This is a good insight to our work trying to find the relation between datasets’ topics and the presence/absence of different features. Lastly (4) there are not much research has been conducted regarding credibility of Arabic content, so there is a need to apply previous assessment methods and investigate their usefulness with Arabic content.

4.1 Recommendation and future work

The next step in our research project is to report the results on (1) user credibility survey, (2) experimental study that analyze datasets contain confirmed news topics and false rumors topics from different countries, topics, and languages and statistically examine and compare how the features are being distributed within these dimensions, (3) the effect of the context on the classifying results of a given classifying technique (classifier-based feature and weighting-based feature). (4) Building a system that can recognize message’s features and context, and then calculate the credibility score. And (7) applying the multi-context credibility assessment on other social media platforms.

5 REFERENCES


