Computer Science • 16 (2) 2015 http://dx.doi.org/10.7494/csci.2015.16.2.157

Krzysztof Lorek
Jacek Suehiro-Wiciński
Michał Jankowski-Lorek
Amit Gupta

AUTOMATED CREDIBILITY ASSESSMENT ON TWITTER

Abstract *In this paper, we make a practical approach to automated credibility assessment on Twitter. We describe the process behind the design of an automated classifier for information credibility assessment. As an addition, we propose practical implementation of TwitterBOT, a tool which is able to score submitted tweets while working in the native Twitter interface.*

Keywords Twitter, credibility, Machine Learning Algorithms

Citation Computer Science 16 (2) 2015: 157–168

1. Introduction

The Internet as an information medium has become very influential over the last years, and dependence on the Internet as a source of information in many crucial domains is still bound to grow.

Due to the decentralized nature of the Internet, the credibility of online information has long interested researchers and practitioners. The concern about credibility stems from the fact that Internet and digitization technologies lowered the cost of information dissemination while increasing accessibility to that information. As a result, much more information is available and easily accessible now than ever before. Problems arise because many sites operate without much oversight or editorial review.

Twitter, the most popular microblogging service, lets users broadcast 140-character status messages known as tweets. The system itself, as well as the social network it creates, has been studied extensively as a news distribution mechanism, both for regular news and emergency situations like natural disasters and other high-impact situations [9].

In our work, we decided to select a set of tweet features and check whether it brings additional value in comparison to features related to the content of the tweet. Then, we build a classifier to see if credibility can be assessed automatically based only on tweet features, and compared to a similar classifier using content as well as a combined set of features.

To achieve this goal, we have built a dataset of manually-evaluated tweets which can be a used as a solid base for further analysis (including credibility assessment process automation). During the process, we solved some practical problems, including an algorithm of reconciling different, manually-assigned scores.

Figure 1. Example of TwitterBOT communication interface.

Having a very efficient classifier, we think to take one step further and get practical implementation of our solution. We are in the process of creating a tool called TwitterBOT, which is able to assess the credibility of tweets submitted by any user. The proposed interface message is depicted in Figure 1, and the high-level architecture of the system is shown in Figure 2 below.

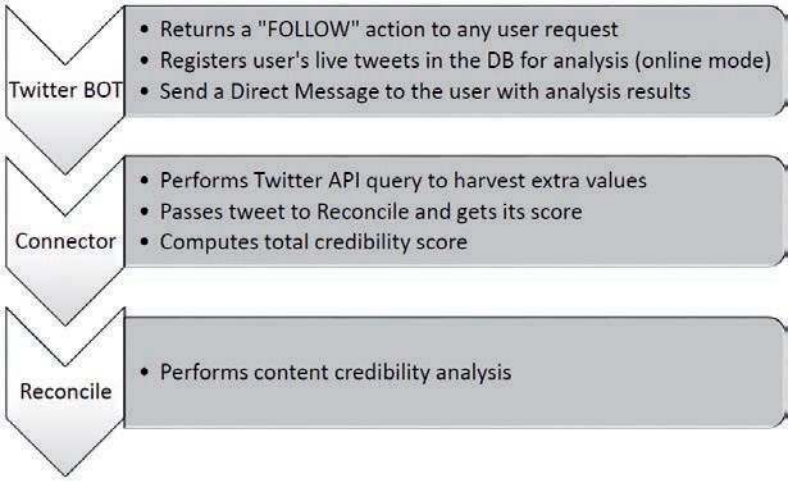

Figure 2. Proposed architecture of TwitterBOT solution (future work).

2. Related work

Social media is increasingly being incorporated into general search engine results. While a potentially valuable source for news and information, this transition removes a critical element of social media: that users are friends or followers of the content author. The result is that users must judge the credibility of content authored by people they do not know [8].

One very interesting approach to automatic credibility assessment was proposed by Gupta et al. [4]. Their goal was to assign a credibility score to each event, so events that are more credible receive a higher score. The authors presented a PageRank-like credibility propagation approach to establish event credibility. They explored the possibility of detecting credible events from Twitter feeds using credibility analysis. A new credibility-analysis model for computing the credibility of linked sets of multi-typed entities has been evaluated.

Further research focused on analyzing microblog postings related to trending topics and classifying them as credible or not credible, based on features extracted from them (as performed by Castillo et al. [2]). The authors used features from tweets and re-tweets, the text of the posts, and citations to external sources. Their results show that there are measurable differences in the way messages propagate, which can be automatically used to classify them as credible or not credible.

Finally, there is some research trying to implement the proposed approaches in practical ways. One interesting system that has recently been studied – TweetCred – was described by Gupta et al. [3]. It works as a browser plugin and computes tweet credibility in near-real time. In comparison, the classifier behind the TwitterBOT that we designed gives better results than TweetCred. This advantage is achieved due to

extra scoring performed by the Reconcile system. The drawback of such a solution is the amount of time necessary to compute credibility behind links included in those tweets being analyzed. Our system does not work in near-real time like TweetCred – it needs a couple of minutes to complete and return credibility score to the user making the request.

The TwitterBot architecture proposed in this article is centralized. However, for increased scalability, the proposed architecture could be distributed using a Peer-to-Peer model [7, 10] in the future.

3. Defining credibility

Researchers have put a lot of effort studying various aspects of the information credibility of Twitter. Kang et al. [6] defines two types of credibility on Twitter:

Definition A Tweet-Level Credibility: A degree of believability that can be assigned to a tweet about a target topic, i.e., an indication that the tweet contains believable information.

Definition B Social Credibility: The expected believability imparted on a user as a result of their standing in the social network, based on any and all available metadata.

3.1. Assessing information credibility

Users currently assess tweet credibility based on trust relationships with authors whose streams they choose to follow. If a social network user is interested in receiving information about a particular topic of interest, a task of primary importance is to decide which other users to behare in a similar way, in order to maximize relevance, credibility, and the quality of information received. Unfortunately, social network users are practically unable to directly observe how well someone is trusted in a particular domain. The 140-character length limit of Twitter posts makes them somewhat unsuitable for analysis with popular topic models. Individual tweets tend to be too short to convey strong information about the precise mixture of latent topics within them. Links between users in a social network serve the function of a vote of support between them, so it should be possible to estimate expertise from observable link data.

Social media is being increasingly incorporated into general search engine results. While a potentially valuable source for news and information, this transition removes a critical element of social media: that users are friends or followers of the content author. The result is that users must judge the credibility of content authored by people whom they do not know [1].

4. Dataset and credibility classes

The process of data collecting was based on gathering real tweets posted on Twitter on one particular subject (in our case, we focused on nature environment preservation)

and dumping them into an external local database for future analysis. We used the twitter river plugin called Elasticsearch to get access to Twitter Stream API. Post subject was verified based on filters that used keywords like "climate change", "carbon dioxide emission", "global warming potential" or "Kyoto protocol". The full list of keywords used consists of more than one hundred subject-related terms. As an output, we got a set of more than 7.000 real tweets, which we randomly narrowed to the final number of 1.206 tweets which went through the manual-tagging stage. Having a well-defined dataset of tweets, our next task was to evaluate each and assign one of the four following credibility levels:

- HC – HIGHLY CREDIBLE – when the tweet focuses on an event or phenomenon and is not a private opinion or comment, its author either follows another post or news or shows the source of information directly in the form of an embedded link, which in turn redirects the reader to a recognizable news service or webpage with a commonly-shared reputation. Also the link doesn't point to a public forum, private blog, or another tweet.
- HNC – HIGHLY NON CREDIBLE – in common understanding, an exact opposite of the highly-credible class, the tweet is a private opinion or comment and doesn't show a source of information. Also, it doesn't contain a clear explanation of how the author came to the conclusion presented in the post.
- N – NEUTRAL – this class covers either general knowledge simply put in another form, or a simple citation of previously-published facts, news, or discovery. By definition, such a post doesn't add anything from the authors' point of view and is just "information" in many cases copied to allow the flow of information.
- C – CONTROVERSIAL – this note covers all tweets that may cause doubts to the reader, and also is given in a situation when the same tweet received two contrary notes from different judges. An important feature of a controversial note is that it is used more often when comparing judgements between people than in relation to one person's individual view.

4.1. Manual tagging

A basic approach to manual credibility tagging assumes the evaluation of each tweet from a human point of view. In the case of our experiment, manual tagging was performed by two individuals who gave notes independently to each tweet from the dataset. After reading the message and checking the embedded links, the critic (a person performing the manual-credibility evaluation) checked and took the following items into consideration:

- visible features of the linked tweet, like profile background, profile photo, profile name, and "account verified" mark;
- tweet content syntax, including sentence correctness and abbreviations used;
- external link features, especially answering the following queries:
  - is the target website reached just after clicking the link or it leads to an interactive ad instead,

– does the website appear to be credible,
– is the text or other content believable (based on similar criteria as the tweet itself),
– does the content's topic match the tweet content.

If for some reason the content of the tweet or the link is not retrievable, it was tagged as ERROR.

4.2. Reconciliation steps

As a result of manual tagging, we received a dataset with independently-assigned notes in the manual tagging process described above. In order to reconcile separate tags, we applied a two-step procedure. In the first step, we checked whether the tweets linked to the same source or the page had the same note. It occurred that differences existed not only between critics, but also the same critic could assign a different credibility score to multiplied tweet (it appeared that it wasn't a rare case – mostly because the same link was communicated by different authors sharing various levels of credibility themselves). When such a case was found, we evaluated the credibility note again and tried to understand the sources of such discrepancies. In many cases, the final note was given based on the majority of individual scores.

4.3. Reconciliation rules

We developed and followed a set of rules:

- In case both notes were the same, the final score remained unchanged.
- If any critic received an ERROR, the tweet was checked again. In case the ERROR persists, the note follows that score. But in the case when the error did not appear, the tweet was scored based only on one, properly-assigned, initial note.
- If a tweet received two different notes, then?
  - In case both notes were absolutely opposite (HC versus HNC), the tweet was analyzed again, and both notes were compared assuming the difference was a result of information interpretation and not the information itself. We assumed this because a tweet had a credible source of information (link or direct citation); on the other hand, it presented a controversial idea, which may be interpreted from various perspectives. If a final agreement wasn't possible, the final score assigned was therefore C.
  - Situation when one note was neutral, in many cases was discussed again, and the final score aimed exactly toward neutral note. It was so because one of the critical scores (either HC or HNC) usually followed personal opinion of the critic, and in direct discussion, it was possible to understand it and a lower level of emotion or other non-meritocratic features.

As the final result, each analysed tweet received one credibility note, which tells in which of the four presented classes that the information resides. Distribution of classes in a given dataset is presented in Table 1.

Table 1

Distribution of classes in the dataset.

ALL	1206	100.0%
C	72	6.0%
HNC	249	20.6%
HC	275	22.8%
N	573	47.5%
ERROR	37	3.1%

5. The classifier and features

For machine learning, we used a random forest implementation in R language similar to the one used in [5]. Because of the stochastic nature of the algorithm, we ran the learning process for each model 50 times with a different seed value. All embedded URLs were extracted from tweet content (if they exist) and processed by the Reconcile platform. This adds the Reconcile features and Reconcile score to the dataset. Furthermore, we have proposed the set of tweet features described in Table 2, which we used as variables in the machine-learning algorithm.

Table 2

Tweet specific features used as credibility markers.

No.	Feature name	Description (a tweet is more credible when...)	Points (Continuous Data)
1	No. of tweets (statuses)	source has higher number of tweets	100 tweets = 1 point
2	No. of followers	source has higher number of followers	100 followers = 1 point
3	No. of followers	source has higher number of followers	100 followers = 1 point
4	Ratio of followers to followers	the ratio is closer to zero	0 = 1 point, 1 = 0 points
7	Is the account verified?	source account is verified	if verified = 1
8	Has "Website" parameter set in the profile?	website is set	if website set = 1
9	Is user's Twiter profile linked to another social service?	when "Website" profile parameter contains "facebook", "linkedin" etc.	if "Website" profile parameter contains "facebook", "linkedin" = 1 point
10	Length of description	source account description is more than 0 characters	if "description" is > 0 chars = 1 point
11	Length of screen name	source screen name differs from username	if "name" is not equal to "screen_name" = 1 point
12	Location is set	source account location is set	if "location" is not null = 1 point

6. Results analysis

Current results of Twitter-only feature-based classification (credibility assessment based on features belonging to the tweet source as described in Table 2) are as follows:

- Distinction of highly credible (HC) tweets equals 51%.
- Distinction of highly not credible (HNC) tweets equals 57%.
- Detection of controversial (C) tweets with balanced collection equals 52%.

In regards to credibility classes (HC, HNC and N), the **Twitter-only** feature-based classifier gives 36% class precision. The best results are given with the neutral class (N) – 66%. The highly not credible class (HNC) is recognized with 25% precision, and highly credible (HC) is detected with only 16% precision.

In the next step, all of the URLs from the dataset tweets were passed to the Reconcile platform for analysis. Results of **Reconcile-only** feature-based classification returned as follows:

- Distinction of highly credible (HC) tweets equals 89%.
- Distinction of highly not credible (HNC) tweets equals 84%.
- Detection of controversial (C) tweets with balanced collection equals 82%.

In regards to credibility classes (HC, HNC, and N), the **Reconcile-only** feature-based classifier gives 84% of class precision. The best results are given with the neutral class (N) – only 1% of class error. Highly not credible class (HNC) is recognized with 72% precision, and highly credible (HC) is detected with 78% precision.

In the final step of analysis, Twitter-only feature-based results were merged with Reconcile-only feature-based results. The findings are as follows:

- Distinction of highly credible (HC) tweets equals 89%.
- Distinction of highly not credible (HNC) tweets equals 87%
- Detection of controversial (C) tweets with balanced collection equals 84%.

In regards to credibility classes (HC, HNC and N), the **combined** classifier gives 89% of recognition precision. The best results are given with the neutral class (N) – class error of 0,3%. Highly not credible class (HNC) is recognized with 82% precision, and highly credible (HC) is detected with 85% precision.

In particular, we found that combining Reconcile-only feature-based results with Twitter-only feature-based results helps to detect HC and HNC classes by raising class precision by 8%–10% (see Figs 1–2) and lowers low Controversial (C) class error by 0,3% (see Figs 3, 4, 5).

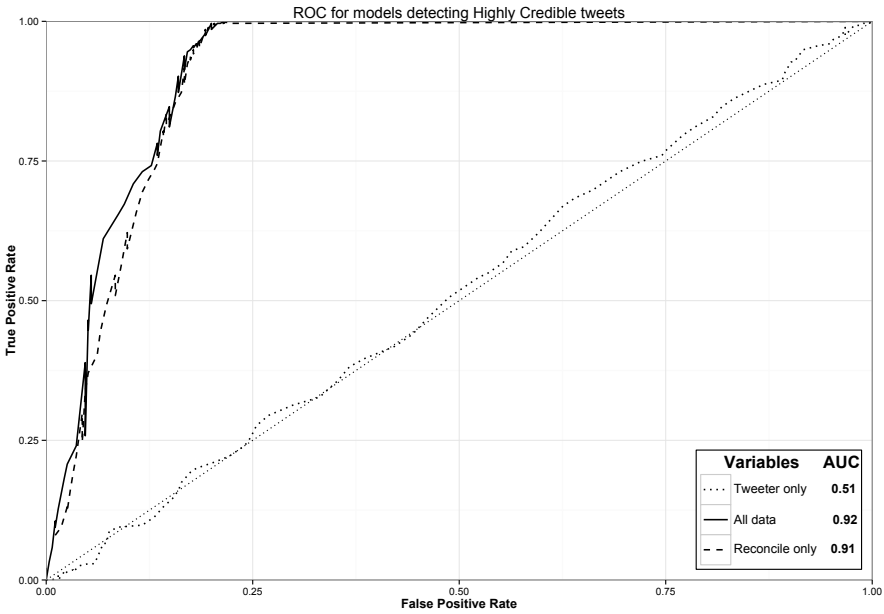

Figure 3. ROC for models detecting highly credible (HC) tweets.

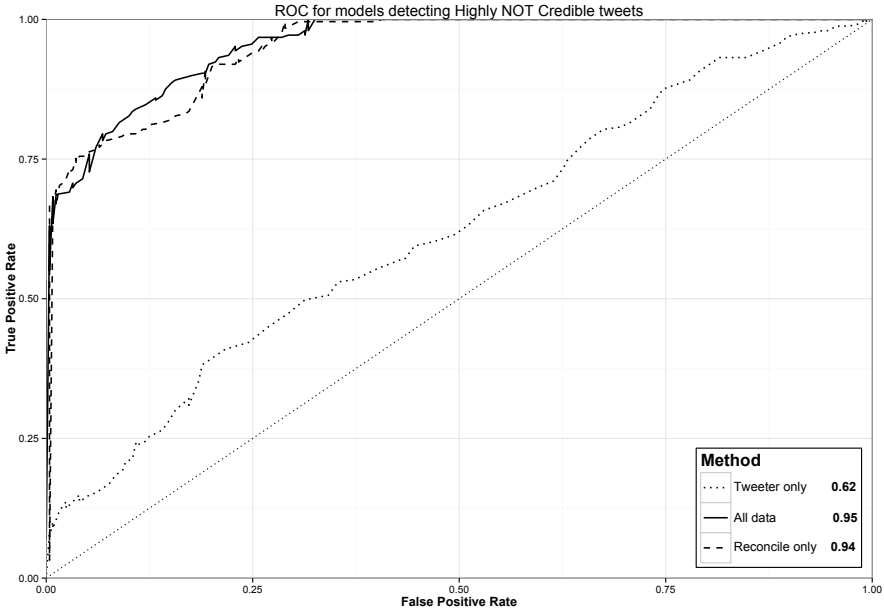

Figure 4. ROC for models detecting highly not credible (HNC) tweets.

7. Conclusion

Twitter has a unique combination of text content and underlying social link structure. In addition, it has a variety of dynamic or ad-hoc structures, making it ideal for the study of information credibility. However, like Web search results, tweets pose an example of a particularly challenging credibility-assessment scenario due to their compact nature.

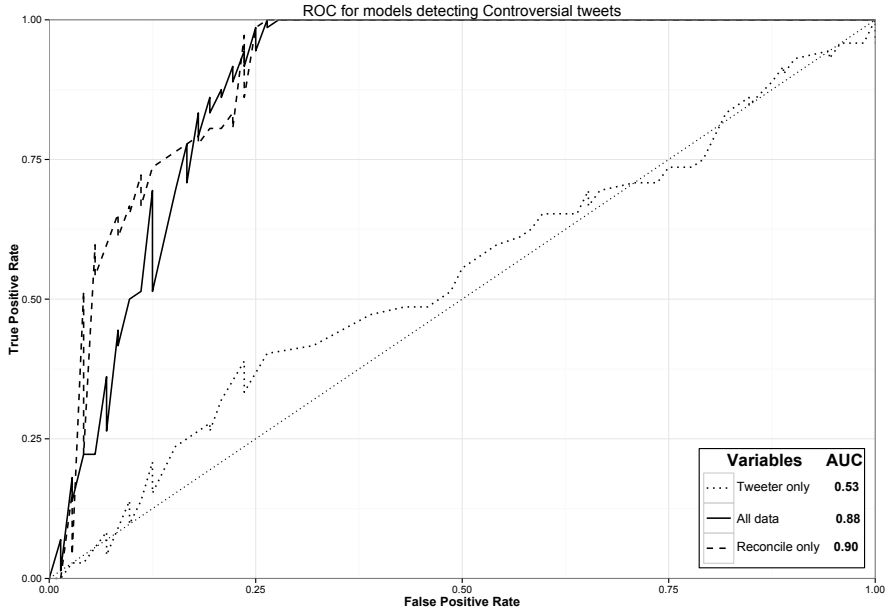

Figure 5. ROC for models detecting controversial (C) tweets.

In our study, we showed that assessing "ground truth" in credibility measurement is possible via manual tagging, even though it requires multiple iterations and comparisons made by independent critics. Somehow, interesting side effects of the manual tagging procedure allowed us to develop a specific set of rules of how to reconcile the notes coming from different sources. Further study of what happens when the number of critics increases may be valuable.

As visualised (see Fig. 6), automated credibility assessment using random forest classifier is possible. However, it is more difficult based only on twitter features, as the results are not statistically more significant than random scoring. However, the combined approach using both content and feature analysis improves credibility assessment considerably. Class detection precision grows between 8%–10%. It may lead us to the very interesting conclusion that, for credibility perception, it is still more important what the user posts and not who they really are.

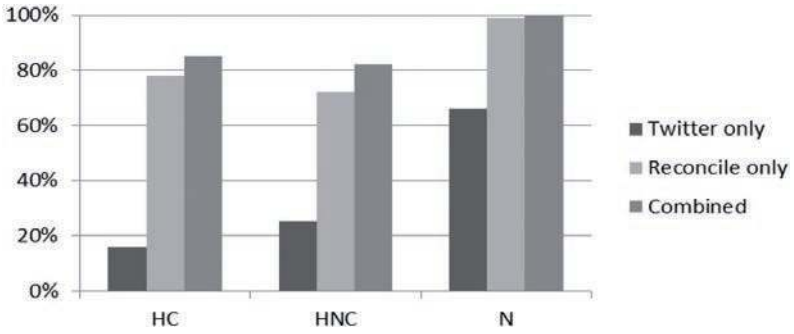

Figure 6. Precision for HC, HNC and N models.

Acknowledgements

This work was supported by the grant Reconcile: Robust Online Credibility Evaluation of Web Content through the Swiss Contribution to the enlarged European Union.

References

[1] Canini K., Suh B., Pirolli P.: Finding Credible Information Sources in Social Networks Based on Content and Social Structure. In: *IEEE International Conference on Privacy, Security, Risk, and Trust*, IEEE, 2011.

[2] Castillo C., Mendoza M., Poblete B.: Information Credibility on Twitter. In: *2011 International World Wide Web Conference*, ACM, 2011.

[3] Gupta A., Kumaraguru P., Castillo C., Meier P.: TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter. *CoRR*, vol. abs/1405.5490, 2014, `http://arxiv.org/abs/1405.5490`.

[4] Gupta M., Zhao P., Han J.: Evaluating Event Credibility on Twitter. In: *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM)*, pp. 153–164, 2012.

[5] Jankowski-Lorek M., Nielek R., Wierzbicki A., Zieliński K.: Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool. In: *Proceedings of the 2014 International Conference on Social Computing*, SocialCom '14, pp. 22:1–22:7, ACM, New York, NY, USA, 2014, `http://doi.acm.org/10.1145/2639968.2640074`.

[6] Kang B., O'Donovan J., Hollerer T.: Modeling Topic Specific Credibility in Twitter. In: *IUI12*, ACM, 2012.

[7] Khan J. I., Wierzbicki A.: Foundations of Peer-to-Peer Computing. In: *Elsevier Journal of Computer Communication*, 2008.

[8] Morris M., Counts S., Roseway A., Hoff A., Shwartz J.: Tweeting is Believing? In: *Understanding Microblog Credibility Perceptions. CSCW12*, ACM, 2012.

[9] O'Donovan J., Kang B., Meyer G., Hollerer T., Adall S.: Credibility in Context: An Analysis of Feature Distributions in Twitter. In: *ASE/IEEE International Conference on Social Computing*, IEEE, 2012.

[10] Wierzbicki A., Szczepaniak R., Buszka M.: Application Layer Mul-ticast For Efficient Peer-to-Peer Applications. In: *IEEE Computer Society*, 2003.

Affiliations

Krzysztof Lorek
Polish-Japanese Institute of Information Technology, Warsaw, Poland, `s8805@pjwstk.edu.pl`

Jacek Suehiro-Wiciński
Polish-Japanese Institute of Information Technology, Warsaw, Poland, `wit@pjwstk.edu.pl`

Michał Jankowski-Lorek
Polish-Japanese Institute of Information Technology, Warsaw, Poland, `fooky@pjwstk.edu.pl`

Amit Gupta
École Polytechnique Fédérale de Lausanne, `amit.gupta@epfl.ch`

Received: 19.01.2015
Revised: 25.03.2015
Accepted: 01.04.2015