Fake News Detection in Twitter

Literature Review

Kristin Kinmont
University of Cape Town
knmkri002@myuct.ac.za

ABSTRACT

Twitter is a micro blogging service which connects millions of users around the world, allowing for the spread of information and news in real-time. While it has become a popular news source, research has show that it is also being used to spread fake news. This has resulted in the need for credibility assessing systems which can identify such news. Many different methods have been proposed for the credability assessment of tweets (posts on Twitter) as well as news on other social media platforms. Many of the methods suggested fall into one of two approaches: linguistic approaches and network approaches. These approaches have, however, been found to be most effective when combined, creating a hybrid approach. Some of the methods used within these approaches include text classifiers, decision trees and directed graphs. Lots of the methods focus on different areas of credibility assessment, and are often optimised for (and potentially limited to) a specific context. A hybrid approach which leverages the strengths of the different methods and approaches may therefore be effective.

CCS CONCEPTS

• **Information systems** → **Web mining**; *Information retrieval*; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Twitter analysis, fake news, credibility assesment

1 INTRODUCTION

The role of social media in our day to day lives has increased rapidly in recent years. It is now used not only for social interaction, but also as an important platform for exchanging information and news [6]. Twitter, a micro blogging service, connects millions of users around the world and allows for the real-time propagation of information and news [10]. These factors have resulted in Twitter playing a critical role in world events, especially crisis events [8] where it has been useful in emergency response and recovery [10] as well as assisting in disaster management [8]. Twitter has, however, not only been used for the spread of valid news, but also deceptive and fake news [3]. This fake news can come in the form of spam [2], astroturf (a technique used in political campaigns to fake support numbers, by making a message appear to have 'grassroots' origins when in reality it originated from one person or organisation) [14], clickbait (content which aims to attract attention and get users to click on a link to increase website traffic) [4] and more. The increase in the volume of fake news has even lead to our current times being labelled 'the age of misinformation' [17] and therefore stresses the importance of assessing the credibility of Tweets [11].

The existence of fake news is not new. Before the use of social media, news was restricted to sources such as the radio, newspapers and TV, where the task of filtering out fake news was assigned to journalists and other news publishers [5]. The rapid increase in user generated content has, however, meant reliance on these traditional filtering techniques is no longer applicable. Research has found that humans are not good at detecting lies in text based on content alone [11] [16] and so there has been a drive to automate news credibility evaluation. Originally development of such technology was isolated and often unknown, slowing progress. Recently, however, collaboration and hybrid approaches have been quite successful [5].

The remainder of this review describes three types of fake news in section 2 and covers different methods and approaches to fake news detection in section 3. An overview of systems developed to assess the credibility of news and users in Twitter is given in section 4. Section 5 then discusses human perceptions of credibility before concluding the review in section 6.

2 TYPES OF FAKE NEWS

[15] define three different types of fake news and suggests that fake news detection services and technologies should cater for such. The three different types are fabrication, hoaxing and satire.

News fabrication uses attention grabbing headings, exaggeration and sensationalism to increase website traffic and profits as well as spread fake news rapidly. It could potentially be identified through the occurrence of 'verbal leakages' [5]. An example of fabrication is clickbait. A preliminary discussion of clickbait, and some of its identifying characteristics, can be seen in [4].

Hoaxes aim to deliberately deceive readers by disguising fake information as news. [17] represents work already done on the detection of hoaxes in Facebook. The paper uses logistic regression and boolean label crowdsourcing (BLC) to show that the users who interact with (e.g. 'like') a post can be used to determine if it is a hoax.

Satire allows for the separation of humorous news from serious. If readers are aware of the humour they are less likely to believe it [15], minimising the potential deceptive impact [16] . This suggests that technology that identifies humour can be used to alert users to the fact and therefore aid in credibility assessment. Research into satire detection includes [16], where it is found that absurdity, grammar and punctuation are the best predicting features for satire.

3 OVERVIEW OF FAKE NEWS DETECTION METHODS

Many different methods have been proposed and used for the detection of fake news in social media. [5], in an attempt to provide a map of the current credibility assessment methods, found that

two major categories of methods emerged: linguistic approaches and network approaches. Both approaches typically make use of machine learning techniques to train classifiers. The paper highlights the importance of a hybrid approach that utilizes the benefits of network and linguistic approaches[5]. When deciding on which approach and method and subsequently which tweet features to use, research has found that the consideration of the data context is important. This is because different features may be important in different contexts and therefore a feature's usefulness can vary within different contexts [12]. It has, however, been found that the following features are good indicators of credibility: URLs, mentions, re-tweets and tweet length [12].

3.1 Linguistic Approaches

Linguistic approaches are those that base assessment on the content of the message (language usage) and aim to identify language 'leakage' or 'predictive deception'. This includes analysis of the data representation, syntax and semantic analysis and classifiers (e.g. based on sets of words) [5]. Methods that take a linguistic approach are sometimes found to be topic specific, limiting their usage in generalised news credibility assessment [5].

Methods that have taken a linguistic approach to fake news detection include [10, 16]. [10] utilizes the idea that fake news is questioned more [3, 10] and uses text classifiers, to identify and tally tweets that are asking questions. When enough people are questioning the news item/tweet, Twitter can warn users of the potential deception. [16] makes use of Natural Language Processing (NLP) with machine learning, to identify satire. 90% precision and 84% recall are achived by focusing on things such as language patterns, sentiment, rhetorical devices and the number of words (ie. the content of the tweet).

3.2 Network Approaches

Network based approaches make use of network properties and behaviour to classify news and would complement content based (linguistic) approaches [5]. These approaches include making use of existing bodies of human knowledge (linked data), metadata analysis [5] and message diffusion patterns [14].

A promising method for detecting fake news is described by [14]. The method analyzes the diffusion of information in social media, and specifically Twitter. Through this analysis, [14] attempt to identify efforts to fake the organic spreading of information in Twitter. This method is specifically used to identify political astroturf. The paper highlights the importance of identifying such stories/messages before they have gained the public's attention. At this point, it becomes nearly impossible to distinguish their spreading patterns from organic ones. The diffusion pattern of a tweet is analyzed through the creation of directed graphs where the nodes are users and edges represent a re-tweet relationship. This method therefore focuses on how the message is delivered rather than its content. The research conducted in this paper led to the creation of 'Truthy' [13] (see section 4).

3.3 Other and Hybrid Approaches

The use of decision trees has also been quite popular and effective. Decision trees are used by [8] to identify fake images with up to 97% accuracy and by [3] to identify tweet credibility with 89% accuracy. Both methods create these trees using user and tweet based features,with [3] also including propagation and tobic based features. Overall, [3] makes use of a much more extensive list of features. Research conducted in [8] also includes the use of Naive Bayes to train the classifier, however, it was outperformed by the J48 Decision Tree approach where the tweet based features were found to be most effective. Due to the variety of features used, from the occurance of positive and negative words to the number of followers of the user, both methods can be seen to take a hybrid approach. Interesting observations to come out of [8] research were that the fake images were mostly re-tweets and that follower relationship contributed as little as 11% to the overall spread of the images, showing that the social network of the users had little impact on spread. [3] observed that credible news is generally propagated by users who are active on Twitter, originates from one or a few users and has many reposts.

Supervised and semi-supervised machine learning algorithms have also been used in credibility assesment of tweets [7] as well as the destection of spammers [2]. Spammers are classified as those who take advantage of trending topics to generate website traffic and revenue. This is done by posting tweets containing trending words and using URLs obscured by URL shorteners, forcing users to load the website in order to identify its content [2]. The machine learning algorithms are trained using labels acquired through crowdsourcing and make use of a variety of tweet attributes/features. [7] uses 45 different features that were selected based off of previous works and that could be derived from a single tweet. These features can be divided into 4 categories: tweet content, tweet metadata, tweet author, tweet network and tweet links. [2] makes use of 62 different attributes, but gives the following as the top three attributes in spammer classification: fraction of tweets with URLs, the age of the account and the average number of URLs per tweet. The fraction of tweets using URLs and the average number of tweets with URLs is significantly higher for spammers compared to non-spammers. Spammers also generally have very new accounts. These two methods differ in that [7] focuses on the credability of the tweet itself while [2] focuses on the user (spammer), stating that identifying spammers rather than spam is more robust to spammers changing their strategies as well as allows for a larger set of attributes.

Trend detection can be used to detect abnormal activity online and can therefore help prevent the abuse of social platforms such as Twitter [1]. This abuse can include the spread of fake news. [1] provides a comparative study of different detection methods but conclude that methods with the best performance use n-gram cooccurrence (words appearing in the same scope) and $df - idf_t$ topic ranking (a time dependent ranking which increases the importance of 'bursty' events). n-grams make use of a linguistic approach while $df - idf_t$ topic ranking makes use of a network approach.

4 AVAILABLE CREDIBILITY CHECKING SYSTEMS

The following section covers three credibility checking systems: Truthy [13], TweetCred [7] and Cognos [6].

Truthy and TweetCred are real-time systems developed to evaluate the credibility of a Tweet. Truthy was specifically developed to track the diffusion of political ('truthy') memes in Twitter in the context of US political elections. From the Truthy website (truthy.indiana.edu), users can inspect memes through multiple different views including a meme's temporal data and diffusion netwrok. 'Truthy' memes are annoted. TweetCred, on the other hand, has a more general application. It comes in the form of a browser extension and assigns a credibility rating between 1 (low credibility) and 7 (high credibility) to tweets in a user's timeline. Users are able to agree or disagree with this rating and suggest their own rating to the system if necessary. It has been evaluated by 1000s of users and was one of the first to be evaluated on such a large scale. Performance was evaluated in terms of response time, effectiveness and usability.

One of the main differences between Truthy and TweetCred is the features used to evaluate tweets. To help ensure TweetCred provides ratings in real-time, only features that can be derived from a single tweet are used in credibility evaluation. It therefore doesn't assume the existence of history data. This is in contrast to Truthy which uses real-time Twitter data as well as 3 months of history [8]. It therefore uses a more complex set of features for its evaluation. For more information on the methods used in the systems, see section 3.

Cognos focuses on a different area of credibility assesment, identifying topic experts in Twitter. Users use the system to query topic experts who are then returned as a ranked list. Fore more information on how this is done, see section 5.

5 HUMAN PERCEPTIONS OF CREDIBILITY

Research has found that there is a difference between the features considered important to humans when assessing credibility and those currently made available in search engines [11]. Humans tend to limit their credibility assessment to features available at a glance. Although this alone may lead to an inaccurate credibility assessment, it would be beneficial to incorporate these features (ie. to consider the end user's perception of credibility [12]) in current techniques and methods as they are features deemed important to users [11]. [7] raises the issue of personalisation in credibility systems. Users should be able to indicate if they trust certain contacts more than others and therefore have this included in the credibility scores appearing on thier timeline.

[11] suggest the following features decrease a user's perception of credibility: abnormal grammar and punctuation, the use of the default picture or avatars for a profile picture and an uneven relationship between followers and followees. Features that improve credibility perceptions include: the author's influence (such as the number of followers, re-tweets and mentions), the author's topic expertise (such as previous tweets relevant to the topic, their Twitter home page bio, if their location is relevant to the topic and their Cognos ranking [6]), the author's reputation (such as whether or not the account has Twitter's verification seal), URLs in the Tweet lead to quality sites and, finally, if other tweets exist with similar content. The experiment used by [11] to gather these features was conducted on a relatively small set of individuals and is therefore not necessarily a true reflection of all Twitter users. The demographics that were considered, however, all produced similar results and therefore suggest the results are applicable to more than just a single demographic.

Reserch that incorperates human perceptions of credibility, and hence utilize the knowledge of the crowds, includes [6, 9]. Cognos [6] utilizes Twitter Lists and could be a very useful system when evaluating a tweet's credibility. Twitter Lists are collections/lists of Twitter accounts, created by Twitter users, generally grouping together accounts on similar topics. Cognos gives promising results that are comparable to and better than those provided by similar systems such as Twitter's search system: Who To Follow (WTF) [6]. [9] goes one step further, combining crowdsourced knowledge with investigative journalism features to automatically identify rumours (defined in the paper as an event with multiple contradicting micro blogs) on social media. These investigative journalism based features were proven to improve credibility prediction. These features include the source credibility, source identity, source diversity, source location and if it is a witness account, the belief in the message and finally, event propagation.

6 CONCLUSION

The detection of fake news in social media is a relatively new research topic; however, its significance and the amount of data available, make it a very popular topic. Many of the papers discussed here provide promising results in the development of a real-time system which can be used to detect fake news and/or rate a news story's credibility. Already systems such as TweetCred and Truthy have been developed for tweet credibility assesment and Cognos for identifying topic experts in Twitter. Multiple approaches and methods have been suggested and used and focus on different aspects of credibility assement. The majority of these methods utlize similar features of the news posts. The selection of suitable features to use in credibility assesment, however, should consider the context of the data as some features are better suited to, and therefore may be more useful in, certain contexts. Each method has its own strengths and weaknesses, suggesting the creation of a hybrid approach which leverages the strengths of the different methods. When developing methods it is also important to think about the credibility perceptions of the end user and incorperate some of these views into the assesment.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Gker, Ioannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing trending topics in Twitter. *IEEE Transactions on Multimedia* 15, 6 (2013), 1268–1282.

[2] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6. 12.

[3] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[4] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[5] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[6] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. Cognos: crowdsourcing search for topic experts in

microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 575–590.

[7] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

[8] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 729–736.

[9] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1867–1870.

[10] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics*. ACM, 71–79.

[11] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 441–450.

[12] John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, and Sibel Adalii. 2012. 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)*. IEEE, 293–301.

[13] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 249–252.

[14] Jacob Ratkiewicz, Michael Conover, Mark R. Meiss, Bruno Gonalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. *ICWSM* 11 (2011), 297–304.

[15] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[16] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News.. In *Proceedings of NAACL-HLT*. 7–17.

[17] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv preprint arXiv:1704.07506* (2017).