ECIS 2024 TREOS

AIS TREO Papers

6-14-2024

Domain-Aware Credibility Assessment For Improved Fake News Detection On Twitter

Anju R
Indian Institute of Technology Madras, anjurvdn@gmail.com

Nargis Pervin
IIT Madras, nargisp@iitm.ac.in

DOMAIN-AWARE CREDIBILITY ASSESSMENT FOR IMPROVED FAKE NEWS DETECTION ON TWITTER

TREO Paper

Anju R, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
anjurvdn@gmail.com

Nargis Pervin, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India,
nargisp@iitm.ac.in

Abstract

In today's digital era, the spread of misinformation on social media platforms has become a critical issue, eroding public trust and altering the dynamics of societal interactions. We present a novel approach for the detection of fake news on Twitter, addressing the growing challenge of false information proliferation in social media. Our methodology emphasizes the importance of credibility, both of users and content, by analyzing interactions and domain information extracted from tweets. We categorize Twitter interactions, leverage domain expertise, and integrate language models like BERT for textual analysis. The approach is multidimensional, combining various elements for enhanced fake news detection. The theoretical foundation is based on the Elaboration Likelihood Model (ELM) and Information Adoption Model (IAM), which elucidate the mechanisms behind news acceptance and propagation on Twitter. We plan to evaluate the proposed model using the CoAID dataset, which contains COVID-19 healthcare misinformation, and IBM Watson Natural Language Understanding (NLU) API for domain information classification. Our goal is to develop a mathematical framework that leverages network interactions, domain information, and language models to improve the detection of fake news on the platform.

Keywords: Fake news, Twitter, Domain-based Credibility, Deep Learning.

1 Extended Abstract

In the digital age, the proliferation of false information has emerged as a significant challenge in social media, undermining public trust and distorting social discourse. False information can manifest in various forms such as false news, fake news, rumor etc. (Bondielli et al 2019). Many people struggle to distinguish between genuine and false information, which exacerbates the problem. The consequences of this are far-reaching, posing a threat to public health, influencing elections, and impacting various other aspects of society. The lack of verification of content shared on social media platforms makes it challenging to use and analyze information effectively. Therefore, the concept of credibility i.e., the perceived level of trustworthiness (Dongo et al 2020), has become crucial in distinguishing genuine information from false information.

Social media platforms like X (formerly, Twitter) and Facebook, with their vast reach and rapid dissemination capabilities, have become fertile ground for the spread of fake news. The pervasive nature of social media amplifies the need for rigorous assessment of information sources, underlining the importance of user and content credibility in mitigating the impact of fake news. As users are often the primary channels for the spread of fake news, understanding their trustworthiness is essential. Similarly, assessing the credibility of the content itself is crucial for identifying fake news. Building on the importance of assessing both user and content credibility, we emphasize that these two factors are interdependent and together contribute to a more accurate detection of fake news. By simultaneously evaluating the trustworthiness of users and the credibility of content, we propose an effective approach to distinguish between genuine and fake news on X(Twitter). We will refer to X as Twitter in the rest of the article.

On Twitter, users engage in a variety of interactions, such as posting tweets, retweeting, replying, and mentioning others. These interactions, along with user and tweet characteristics, are key in assessing credibility. However, many existing methods neglect the importance of network interactions (Verma et al 2023) and fail to account for the interdependence between credibility and tweet content (Cardinale et al 2021, and Li et al 2020). Our approach introduces a novel and comprehensive method to detect fake news on Twitter, overcoming the limitations of existing techniques. By integrating domain information and analyzing the dynamics of user and tweet interactions, our model offers a unique blend of user and tweet credibility assessment combined with textual feature extraction, significantly enhancing the detection of fake news on the platform.

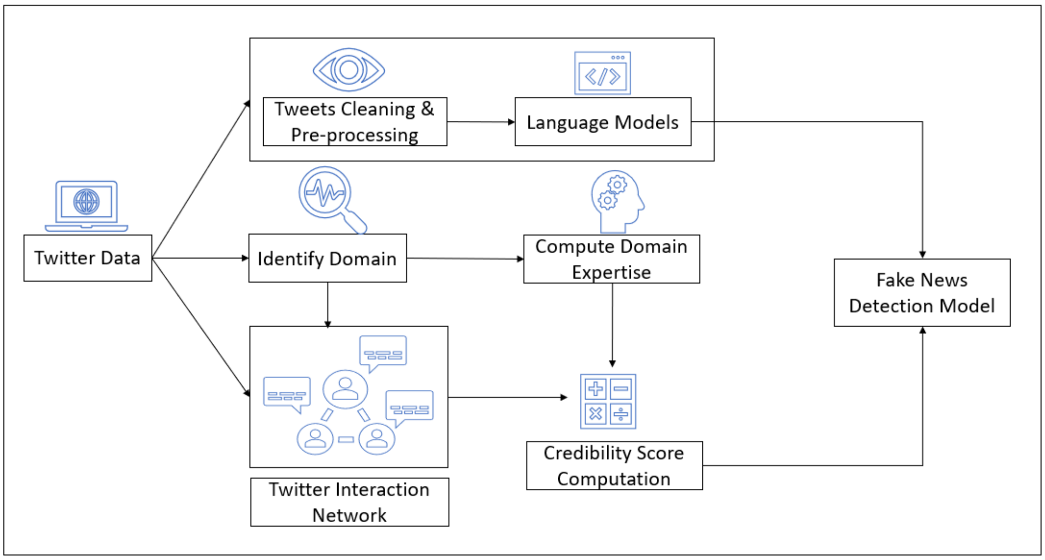
Tweets Cleaning & Pre-processing
Language Models
Twitter Data
Identify Domain
Compute Domain Expertise
Fake News Detection Model
Twitter Interaction Network
Credibility Score Computation

Figure 1. Architecture diagram for fake news detection.

In this study, we aim to develop a comprehensive approach for detecting fake news on Twitter. Our approach leverages interactions and domain information extracted from tweets to assess the trustworthiness of users and their content. The various interactions in Twitter can be categorized into four types: user-user (UU) (user follower network), user-tweet (UT) (user posting a tweet), tweet-user (TU) (a user is mentioned in a tweet), and tweet-tweet (TT) (a tweet is retweeted/replied (tweet) to by another user). Domain information refers to the broad area or subject matter that a tweet focuses on. This information when coupled with the users' activities (post, retweet, reply) plays a major role in understanding expertise of users in specific subject areas. The information about tweet type such as post, retweet or reply, is required to find the relative importance of user in a domain. The domain expertise is determined by analyzing the types of tweets posted by users and their relevance to different domains. This domain expertise is then utilized to quantify the user and tweet interactions in the Twitter Interaction Network. Identifying users' domain expertise can be valuable for identifying whether the user has any prior knowledge of the domain. However, domain knowledge alone may not be sufficient to evaluate the user credibility. Additionally, we plan to use language models such as LSTM, BERT etc. to analyze the textual content of tweets, capturing the depth and nuances of the language. These elements combined together form a multidimensional approach that not only assesses credibility but also enhances the accuracy of fake news detection. Figure 1 presents an overview of our proposed architecture for the fake news detection approach.

To conceptualize the proposed model, we utilize the Elaboration Likelihood Model (ELM) (Petty and Cacioppo, 1983) and Information Adoption Model (IAM) (Sussman & Siegal, 2003) to theoretically underpin the mechanisms behind the acceptance and propagation of news, potentially fake, on Twitter. The ELM posits two distinct routes of persuasion: central and peripheral. The central route involves evaluation of the message content, while the peripheral route relies on contextual cues or heuristics, such as the credibility of the source or the strength of the message endorsement by others. Our model integrates these concepts, with message quality directly impacting perceived usefulness, reflecting the central route. The message quality directly impacts the source expertise in regard to the domain information. Simultaneously, the source trustworthiness, which is informed by a user's historical domain knowledge (source expertise) and past interactions (based on tie strength) leading to the perceived credibility of the message, akin to the peripheral route of ELM. In harmony with ELM, IAM emphasizes the perceived usefulness and perceived credibility as determinants of news adoption, which potentially contribute to the sharing of fake news.

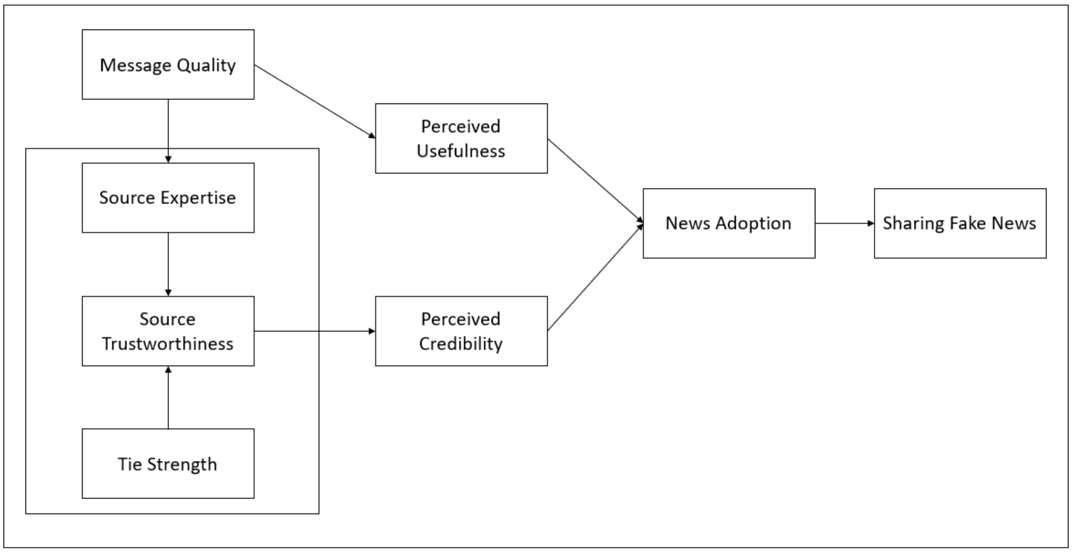

Figure 2. Theoretical framework.

We plan to utilize CoAID (Covid-19 healthcare misinformation Dataset), which contains COVID-19 healthcare misinformation, including fake news on Twitter (Cui et al 2020). The dataset contains data between 1st December 2019 and 30th September 2020, including tweets, retweets, replies, mentions, and followers of the users. The dataset includes annotations indicating whether the content is fake or real, serving as the ground truth for analysis. The data will provide a representative sample of user interactions and content on the platform, which serves as the basis for analysis and evaluation in the study. To identify domain information of tweet, we aim to use tools such as IBM Watson Natural Language Understanding (NLU) API. NLU provides a hierarchical classification along with a score indicating the degree of relevance to the domain (IBM 2021).

Using network interactions along with domain information, we aim to frame a mathematical framework to compute the credibility of user and tweet, which when integrated with the language models enhance the detection of fake news. The proposed model addresses leveraging structure, semantics, and domain information to enhance fake news detection on Twitter.

References

Verma, P.K., Agrawal, P., Madaan, V. and Prodan, R., 2023. "MCred: multi-modal message credibility for fake news detection using BERT and CNN". Journal of Ambient Intelligence and Humanized Computing, 14(8), pp.10617-10629.

Li, P., Zhao, W., Yang, J. and Wu, J., 2020. “CoTrRank: trust ranking on Twitter”. IEEE Intelligent Systems, 36(1), pp.35-45.

Cardinale, Y., Dongo, I., Robayo, G., Cabeza, D., Aguilera, A. and Medina, S., 2021. “T-creo: a twitter credibility analysis framework”. IEEE Access, 9, pp.32498-32516.

Bondielli, A. and Marcelloni, F., 2019. A survey on fake news and rumour detection techniques. Information sciences, 497, pp.38-55.

IBM. (2021). Watson Natural Language Understanding. [Software]. Available at: https://www.ibm.com/watson/services/natural-language-understanding/ [December 2023].

Cui, L. and Lee, D., 2020. “Coaid: Covid-19 healthcare misinformation dataset”. arXiv preprint arXiv:2006.00885.

Sussman, S.W. and Siegal, W.S., 2003. “Informational influence in organizations: An integrated approach to knowledge adoption”. Information systems research, 14(1), pp.47-65.

Petty, R.E., Cacioppo, J.T. and Schumann, D., 1983. “Central and peripheral routes to advertising effectiveness: The moderating role of involvement”. Journal of consumer research, 10(2), pp.135-146.