9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

BREAKING NEWS: Psycholinguistic and Behavioural Differences in (Un)-Trustworthy Online News Source Interaction

psycholinguistics, online news, behavioural differences, machine learning, online communities

Extended Abstract

The psychological need to share unverified information and the impact thereof has been greatly facilitated with the advent of online activities and online social networks. The pervasiveness of unfiltered and non-curated information in online social networks induces exposure to untrustworthy information (i.e., information that comes sources with a low factual scores). This exposure makes it increasingly challenging for users to discern trustworthy from non-trustworthy online news sources. Moreover, the social character of these platforms also raises questions around the diffusion of such untrustworthy news sources. A better understanding of the characteristics of users and online communities sharing untrustworthy online news sources could help to identify and to prevent the diffusion of untrustworthy content. Existing literature has mainly focused on the experimental investigation of individual differences in the interaction with untrustworthy online news sources. Moreover, these studies commonly focused on specific, individual differences (e.g., political orientation, emotions, reasoning, social cues). To comprehensively analyse the factors that are related to the online dissemination of untrustworthy news, a better understanding of user's characteristics in the complexity of a real-world setting is called for. Here, we aim to answer the following research question:

What are systematic differences in users' demographics, psycholinguistic characteristics, and online posting behaviour regarding interaction with (un)trustworthy news sources?

To address this question, we use open-access social media data from the Reddit platform (Pushshift) and independent data for political bias and trustworthiness of online media (e.g., media bias fact check) to highlight groups of users who tend to share less factual content to those platforms. We extract two groups of users sharing only trustworthy news and users sharing only untrustworthy content. Using a method introduced by Waller and Anderson [4] we adopt communities embeddings to ensures that users in the two groups belong to similar communities. In line with Ashokkumar and Pennebaker's [1] work on group identities in Reddit groups, we use the dictionary-based Linguistic Inquiry and Word Count (LIWC) to quantify psycholinguistic characteristics, related to an individual's cognitive state, feelings, and emotions based on the text in the comments that users produce. Furthermore, to remove any confounding factors, when analyzing the group associated to adherence to untrustworthy news source, we remove the text commented below posts that reference such source. We complement the LIWC scores with general text-based characteristics such as lexical diversity and readability scores. The second set of features focuses on the online behaviour of the Redditors. For this, we extract posting behaviour features based on the popularity of the user, posting frequency, and network-size. Lastly, we use the Bidirectional Encoder Representations from Transformers (BERT) model trained on the RedDust dataset to infer users' demographic characteristics such as age and gender. Previous work attempted to identify drivers of fake believes, Ecker and Lewandowsky [2] classified these drivers into two categories: cognitive drivers and socio-affective drivers. An important socio-affective driver are emotions, in fact people who rely on

emotions are more likely to increase their beliefs in false news. We use the *Positive Emotion* and *Negative Emotion* LIWC scores to measure for the degree of emotionality. Furthermore, we specifically measure *stress* in the comment of a given submission by using a custom made LIWC dictionary which has been used to assess stress in the social media context. Additionally, we use the LIWC variable *Prosocial Behaviours* to measure the how much a user is signalling intentions to help and/or caring about others, particularly at the interpersonal level. To identify which variables from the feature set have the highest ability at separating users in our dataset, we adopt a machine learning classification approach using a Random Forest classifier. In a second iteration of the project, we plan to implement a regression analysis by changing our target variable into a continuous variable. With this change we will be able to analyse the whole spectrum of interaction with untrustworthy news sources rather than the extreme groups. In this version of the paper, we want to quantify the impact of these features on the classification output. To do this we adopt a method from coalition game theory: Shapely values. This method is used in interpretable machine learning and it represents the marginal contribution of each feature to the outcome [3]. The results of this study will contribute to a more nuanced understanding of how users' psycholinguistic characteristics that can be associated to different consumption patterns of trustworthy or untrustworthy news sources, and potentially provide valuable insights for predicting which communities are vulnerable to the spread of misinformation. Figure 1 reports both variable importance and their impact for the prediction. They are sorted in descending order of importance and the values on the X axis represents the impact. The impact is measured by the Shapely Value as the marginal contribution to the final prediction in percentage terms. In our The sign of the value will indicate the direction of the contribution: a positive sign indicates that the feature has impact in favor of the Trustworthy Sharers group; a negative sign indicates that the feature has impact in favor of the Untrustworthy Sharers group. For example, *politic* is the most important variable. Furthermore, users' high values of politics have a negative impact on the classification out, signalling that users with this characteristic can potentially belong to the untrustworthy shares group. The findings can be used to identify groups who are susceptible consuming and spreading disinformation, this can help individuals make informed decisions and avoid exposure to false information. Furthermore, the results can help in developing interventions and policies debunk disinformation and raise awareness in society about media literacy. Lastly, this can also have implications in the field of differential psychology. It can bring insights on the human factors that contribute to individual's vulnerability to adherence to untrustworthy news sources and develop interventions to help an individual's truth discernment.

References

[1] Ashwini Ashokkumar and James W Pennebaker. "Tracking Group identity through natural language within groups". In: *PNAS Nexus* 1.2 (2022).

[2] Ullrich K. Ecker et al. "The psychological drivers of misinformation belief and its resistance to correction". In: *Nature Reviews Psychology* 1.1 (2022), pp. 13–29.

[3] Lloyd S. Shapley. "Notes on the N-Person Game — II: The Value of an N-Person Game". In: (1951).

[4] Isaac Waller and Ashton Anderson. "Quantifying social organization and political polarization in online platforms". en. In: *Nature* 600.7888 (Dec. 2021), pp. 264–268.

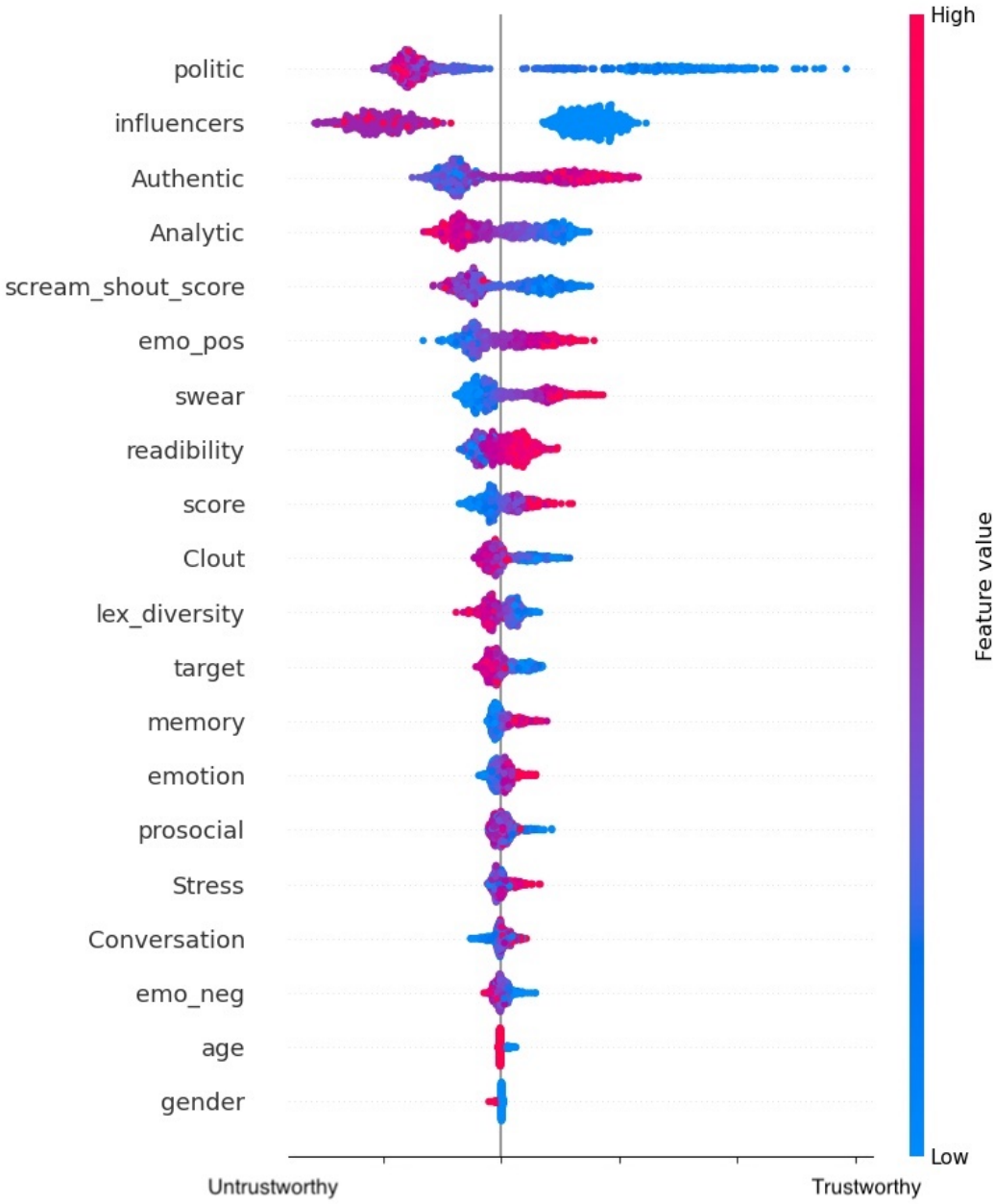

Figure 1: Preliminary results: shapely values. The features are sorted in descending order of importance.The location on the X-axis shows which target class the impact of that feature is associated. They represents percentage marginal contribution against the 50% threshold. The colors of each dot show whether that variable is high (in red) or low (in blue) for a given observation.