

# Combating Misinformation on Social Media by Exploiting Post and User-level Information



A Thesis submitted to the University of Sheffield  
for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Yida Mu

*Supervisor:* Professor Nikolaos Aletras

Department of Computer Science

The University of Sheffield

February 2023

---

## DECLARATION

I, Yida Mu, hereby declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgement, the work presented is entirely my own.

*This thesis is dedicated to my parents for their love and endless support.*

---

## ACKNOWLEDGEMENTS

I would first like to thank my supervisor, Professor Nikolaos Aletras, for his support, guidance, patience, and encouragement throughout my PhD journey. I feel so fortunate to have Nikos as my PhD supervisor.

I would also like to thank the GATE team supervisors: Prof Kalina Bontcheva, Dr Xingyi Song and Dr Carolina Scarton for their invaluable advice. My gratitude also goes to the panel committee members: Prof Haiping Lu, Dr Ziqi Zhang and Dr Arkaitz Zubiaga.

Special thanks to all N2LP team members: *Dr George Chrysostomou, Katerina Margatina, Mali Jin, Ahmed Alajrami, Danae Sánchez Villegas, Constantinos Karouzos, Miles Williams, Hardy Huang, Wenzhe Li, Huiyin Xue, Dr Samuel Mensah and Dr Cass Zhao*; colleagues: *Dr Zeerak Talat, Dr Xutan Peng, Dr Haiyang Zhang, Yue Li, Dr Ruizhe Li, Dr Harish Madabushi, Dr Fernando Alva-Manchego, Ben Wu, Shun Wang, Yizhi Li, Bohao Yang, Zehai Tu, Wanli Sun, Dr Jisi Zhang and Mingjie Chen*; and supportive friends: *Yijie Yu, Yifei Cai, Wei Lin, Yuxiang Liu, Yuting Chen, Shaoxin Chen, Yichi Feng, Qianhui Liu, Chencheng Tao, Miaochun Xu, Jing Wang, Zhe Li, Chen Su, Haoning Wang, Xiaowen Li, Linjia Hu, Tianhao Zhang, Wanying Pei, Fanke Huang, Mengdi Zheng, Yiyun Liu, Ninghan Chen, Dr Pu Niu and Dr Shen Li*.

Finally, I would particularly like to thank my parents for their endless love throughout my life.

---

## ABSTRACT

Misinformation on social media has far-reaching negative impact on the public and society. Given the large number of real-time posts on social media, traditional manual-based methods of misinformation detection are not viable. Therefore, computational approaches (i.e., data-driven) have been proposed to combat online misinformation. Previous work on computational misinformation analysis has mainly focused on employing natural language processing (NLP) techniques to develop misinformation detection systems at the post level (e.g., using text and propagation network). However, it is also important to exploit information at the user level in social media, as users play a significant role (e.g., post, diffuse, refute, etc.) in spreading misinformation. The main aim of this thesis is to: (i) develop novel methods for analysing the behaviour of users who are likely to share or refute misinformation in social media; and (ii) predict and characterise unreliable stories with high popularity in social media. To this end, we first highlight the limitations in the evaluation protocol in popular rumour detection benchmarks on the post level and propose to evaluate such systems using chronological splits (i.e., considering temporal concept drift). On the user level, we introduce two novel tasks on (i) early detecting Twitter users that are likely to share misinformation before they actually do it; and (ii) identifying and characterising active citizens who refute misinformation in social media. Finally, we develop a new dataset to enable the study on predicting the future popularity (e.g. number of likes, replies, retweets) of false rumour on Weibo.

---

# TABLE OF CONTENTS

|          |                               |           |
|----------|-------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>           | <b>1</b>  |
| 1.1      | Research Motivation . . . . . | 3         |
| 1.2      | Research Aims . . . . .       | 5         |
| 1.3      | Definitions . . . . .         | 5         |
| 1.4      | Contributions . . . . .       | 6         |
| 1.4.1    | Paper I . . . . .             | 6         |
| 1.4.2    | Paper II . . . . .            | 7         |
| 1.4.3    | Paper III . . . . .           | 7         |
| 1.4.4    | Paper IV . . . . .            | 8         |
| <b>2</b> | <b>Paper I</b>                | <b>10</b> |
| 2.1      | Introduction . . . . .        | 11        |
| 2.2      | Methodology . . . . .         | 13        |

|          |                                                                      |           |
|----------|----------------------------------------------------------------------|-----------|
| 2.2.1    | Data . . . . .                                                       | 13        |
| 2.2.2    | Data Splits . . . . .                                                | 14        |
| 2.2.3    | Models . . . . .                                                     | 15        |
| 2.2.4    | Hyperparameters and Implementation Details . . . . .                 | 15        |
| 2.2.5    | Evaluation Metrics . . . . .                                         | 17        |
| 2.3      | Results . . . . .                                                    | 17        |
| 2.4      | Error Analysis . . . . .                                             | 18        |
| 2.5      | Conclusion . . . . .                                                 | 19        |
| <b>3</b> | <b>Paper II</b>                                                      | <b>28</b> |
| 3.1      | Introduction . . . . .                                               | 29        |
| 3.2      | Background and Related Work . . . . .                                | 30        |
| 3.2.1    | Disinformation in Social Media . . . . .                             | 30        |
| 3.2.2    | Categorization of Unreliable News Sources . . . . .                  | 31        |
| 3.2.3    | Previous Work on Combating Online Disinformation . . . . .           | 32        |
| 3.3      | Task Description . . . . .                                           | 33        |
| 3.4      | Data . . . . .                                                       | 33        |
| 3.4.1    | Collecting Posts from Unreliable and Reliable News Sources . . . . . | 34        |
| 3.4.2    | Collecting Candidate Users . . . . .                                 | 35        |
| 3.4.3    | Labeling Users . . . . .                                             | 36        |
| 3.4.4    | Text Preprocessing . . . . .                                         | 36        |

|          |                                                            |           |
|----------|------------------------------------------------------------|-----------|
| 3.4.5    | Ethics . . . . .                                           | 36        |
| 3.5      | Methods . . . . .                                          | 37        |
| 3.5.1    | SVM . . . . .                                              | 37        |
| 3.5.2    | Avg-EMB . . . . .                                          | 38        |
| 3.5.3    | BiGRU-ATT . . . . .                                        | 38        |
| 3.5.4    | ULMFiT . . . . .                                           | 39        |
| 3.5.5    | T-BERT and H-BERT . . . . .                                | 39        |
| 3.5.6    | T-XLNet and H-XLNet . . . . .                              | 40        |
| 3.6      | Results . . . . .                                          | 40        |
| 3.6.1    | Experimental Setup . . . . .                               | 40        |
| 3.6.2    | Prediction Results . . . . .                               | 41        |
| 3.6.3    | Error Analysis . . . . .                                   | 41        |
| 3.6.4    | Linguistic Analysis . . . . .                              | 42        |
| 3.7      | Conclusions . . . . .                                      | 45        |
| <b>4</b> | <b>Paper III</b>                                           | <b>53</b> |
| 4.1      | Introduction . . . . .                                     | 54        |
| 4.2      | Related Work . . . . .                                     | 56        |
| 4.2.1    | Misinformation: Definition and Types . . . . .             | 56        |
| 4.2.2    | Misinformation Detection . . . . .                         | 56        |
| 4.2.3    | User Behavior Analysis Related to Misinformation . . . . . | 57        |



|       |                                                |    |
|-------|------------------------------------------------|----|
| 4.3   | Task and Data . . . . .                        | 58 |
| 4.3.1 | Task Description . . . . .                     | 58 |
| 4.3.2 | Weibo Data . . . . .                           | 59 |
| 4.3.3 | Twitter Data . . . . .                         | 61 |
| 4.3.4 | Data Statistics and Topical Coverage . . . . . | 62 |
| 4.3.5 | Text Pre-processing . . . . .                  | 63 |
| 4.4   | Predictive Models . . . . .                    | 64 |
| 4.4.1 | Baseline Models . . . . .                      | 64 |
| 4.4.2 | English Transformers . . . . .                 | 64 |
| 4.4.3 | Chinese Transformers . . . . .                 | 65 |
| 4.4.4 | Handling Long Text . . . . .                   | 65 |
| 4.5   | Experimental Setup . . . . .                   | 66 |
| 4.5.1 | Hyper-parameters . . . . .                     | 66 |
| 4.5.2 | Implementation Details . . . . .               | 67 |
| 4.5.3 | Evaluation Metrics . . . . .                   | 67 |
| 4.6   | Results . . . . .                              | 68 |
| 4.6.1 | Predictive Performance . . . . .               | 68 |
| 4.6.2 | Model Explainability . . . . .                 | 70 |
| 4.6.3 | Error Analysis . . . . .                       | 71 |
| 4.7   | Linguistic Analysis . . . . .                  | 72 |

|          |                                                    |           |
|----------|----------------------------------------------------|-----------|
| 4.7.1    | N-grams . . . . .                                  | 72        |
| 4.7.2    | LIWC . . . . .                                     | 74        |
| 4.8      | Conclusion . . . . .                               | 76        |
| <b>5</b> | <b>Paper IV</b>                                    | <b>85</b> |
| 5.1      | Introduction . . . . .                             | 86        |
| 5.2      | Related Work . . . . .                             | 89        |
| 5.2.1    | Rumor Detection . . . . .                          | 89        |
| 5.2.2    | Modeling Popularity in Social Media . . . . .      | 90        |
| 5.2.3    | Our work . . . . .                                 | 91        |
| 5.3      | Task Description . . . . .                         | 91        |
| 5.4      | Data . . . . .                                     | 91        |
| 5.4.1    | Data Collection . . . . .                          | 91        |
| 5.4.2    | Rumor Information . . . . .                        | 92        |
| 5.4.3    | Defining False Rumor Popularity on Weibo . . . . . | 94        |
| 5.4.4    | Data Pre-processing . . . . .                      | 95        |
| 5.4.5    | Dataset Description . . . . .                      | 95        |
| 5.4.6    | Data Splits . . . . .                              | 96        |
| 5.5      | Experimental Setup . . . . .                       | 96        |
| 5.5.1    | Predictive Models . . . . .                        | 96        |
| 5.5.2    | Rumor Content ( $R$ ) Models . . . . .             | 96        |

|          |                                                                                              |            |
|----------|----------------------------------------------------------------------------------------------|------------|
| 5.5.3    | Combining Rumor Text, User Profile Description and User Attributes ( $R + P + U$ ) . . . . . | 98         |
| 5.5.4    | Hyperparameters & Implementation Details . . . . .                                           | 99         |
| 5.5.5    | Weak Baselines . . . . .                                                                     | 100        |
| 5.5.6    | Model Training and Evaluation Metrics . . . . .                                              | 100        |
| 5.5.7    | Results . . . . .                                                                            | 100        |
| 5.6      | Analysis . . . . .                                                                           | 102        |
| 5.6.1    | Ablation Study . . . . .                                                                     | 102        |
| 5.6.2    | Qualitative Analysis of Model Predictions . . . . .                                          | 103        |
| 5.6.3    | Characterizing Highly Popular False Rumors . . . . .                                         | 106        |
| 5.7      | Implications and Ethics Considerations of Our Study . . . . .                                | 108        |
| 5.7.1    | Theoretical Implications . . . . .                                                           | 109        |
| 5.7.2    | Practical implications . . . . .                                                             | 109        |
| 5.7.3    | Ethics Considerations . . . . .                                                              | 110        |
| 5.8      | Conclusion and Future Work . . . . .                                                         | 110        |
| <b>6</b> | <b>Conclusion</b>                                                                            | <b>121</b> |
| 6.1      | Summary of Thesis . . . . .                                                                  | 121        |
| 6.2      | Future Work . . . . .                                                                        | 122        |

# INTRODUCTION

Social media platforms (e.g., Twitter and Facebook) have become a primary source of news and information for the public. For example, Twitter has enabled end users to consume daily news from a variety of reliable news sources (e.g., @BBC and @Reuters), however, it has also led to the spread of unreliable stories from less credible news sources (e.g., @InfoWars and @ActivistPost) (Rashkin et al., 2017; Volkova et al., 2017). Online misinformation such as false rumours and fake news, if left unverified, can spread faster than reliable news stories (Figueira and Oliveira, 2017; Vosoughi et al., 2018) and cause harm to individuals, organisations, and society as a whole. In the era of social media, it has become easier for such unreliable stories to propagate at an unprecedented rate, making the need for early detection of misinformation even more important (Chen et al., 2018; Xia et al., 2020; Zhou et al., 2019).

To combat online misinformation, researches have employed Natural Language Processing (NLP) techniques to automatically detect misinformation using data-driven methods, which is framed as *computational misinformation analysis* (Shu et al., 2017; Zhang and Ghorbani, 2020; Zhou and Zafarani, 2020; Zubiaga et al., 2018). However, previous work mainly focused on detecting misinformation at the post level, e.g., developing supervised classifiers trained on individual or small sets of posts associated with labels to distinguish true or false information given user-generated contents and

| Examples   |                  | Posts                                                                                                                                                                     |
|------------|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>T1:</b> | Twitter Spreader | Breaking: FBI agent and his wife found deceased after he was suspected of leaking hillary emails URL                                                                      |
|            | Twitter Debunker | @USER fake ... URL*                                                                                                                                                       |
|            | Fact-checking    | URL*: <a href="https://tinyurl.com/479kx2p2">https://tinyurl.com/479kx2p2</a>                                                                                             |
| <b>T2:</b> | Weibo Spreader   | 据报道, 十一名索马里海盗今日在美国法庭招人其实他们是<br>为高盛工作的...<br><b>en:</b> Eleven Somali pirates have reportedly confessed in a US court recently that they were working for Goldman Sachs... |
|            | Weibo Debunker   | #FalseRumor URL*                                                                                                                                                          |
|            | Fact-checking    | URL*: <a href="http://weibo.com/6590980486/H9w1QCgMu">http://weibo.com/6590980486/H9w1QCgMu</a>                                                                           |

**Table 1.1:** Two real-world examples (including misinformation spreader, debunker and fact-checking information) from Twitter (T1) and Weibo (T2) respectively.

network information (Ma et al., 2016, 2017; Thorne et al., 2018; Wang, 2017; Zubiaga et al., 2016).

This thesis mainly focuses on modelling the role of social media users in the propagation of misinformation e.g, analysing users’ reactions to misinformation. Social media users can inadvertently diffuse misinformation by *sharing unreliable news items* without verifying their accuracy, or they can actively participate in *debunking misinformation* with the goal of preventing others from being influenced by misleading content. Table 1.1 displays two real-world examples from Twitter (Vo and Lee, 2019) and Weibo (Mu et al., 2022), including source posts, misinformation spreaders, and debunkers.

Identifying and characterising the behaviour (e.g., retweeting, debunking, etc.) of users who are involved in the dissemination of news items on social networks is vital for the propagation of misinformation prevention at the user level. This work is beneficial in several fields: (i) social media platforms (e.g., Twitter<sup>1</sup> and Facebook<sup>2</sup>) can detect and prevent the spread of potentially unreliable news items in the community; (ii) social scientists and psychologists can employ data-driven approaches to complement the work on characterising personality traits of users who actively spread or refute misinformation on a large scale (Bronstein et al., 2019; Lin et al., 2023; Pennycook et al., 2018; Pennycook and Rand, 2019, 2020); and (iii) fact-checking platforms can

<sup>1</sup><https://help.twitter.com/en/resources/addressing-misleading-info>

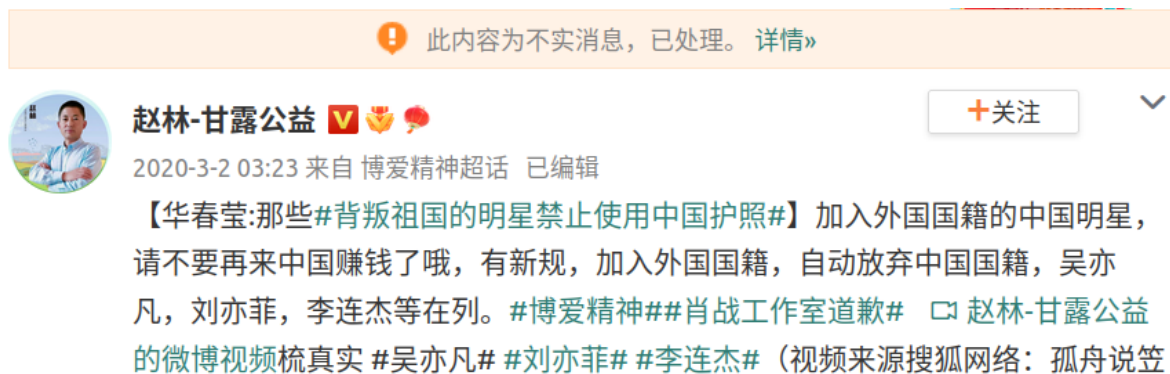
<sup>2</sup><https://www.facebook.com/formedia/blog/working-to-stop-misinformation>

promote personalised recommendation systems to assist self-motivated misinformation debunkers in correcting unreliable news stories (Vo and Lee, 2018, 2019).

## 1.1 Research Motivation

In computational misinformation analysis, there are a number of challenges that we aim to explore:

- To evaluate misinformation detection systems, previous work has typically used a random splits protocol (Lin et al., 2021; Ma et al., 2016, 2017; Rao et al., 2021), which yields topical overlap between the training and test sets. Note that this random data split strategy is not reliable for detecting unseen rumours. For practical considerations, we argue that in order to evaluate the predictive performance of a misinformation detection system, it is necessary to detect not only debunked misinformation, but also new ones. Previous work (Agarwal and Nenkova, 2022; Chalkidis and Søgaard, 2022; Huang and Paul, 2018, 2019; Søgaard et al., 2021) has demonstrated that *temporal concept drift* can significantly affect the classifier performance in various domains e.g., legal, hate speech, medical, etc. However, exploring the impact of the temporal concept drift in the field of *computational misinformation analysis* remains to be explored.
- Notably, modern news recommendation algorithms (such as collaborative filtering approaches and personalised attention networks) developed with user-generated content (e.g., user’s browsing history) (Mao et al., 2021; Qi et al., 2021) can also shape the news stories people see by prioritising personalised posts that are more likely to keep users engaged. This suggests that people who engage more with unreliable news outlets (e.g., propaganda, conspiracy, hoax, etc.) (Aker et al., 2019; Rashkin et al., 2017; Volkova et al., 2017) on social media will be more likely recommended less credible news items. Identifying and characterising individuals that are more likely to repost from untrustworthy news sources is beneficial for the early detection and prevention of misinformation spread in social media, and has yet to be investigated.



**Figure 1.1:** An example of false rumour on the Weibo platform. The text (i.e., ‘! 此内容为不实消息，已处理。’) in orange box denotes ‘this post is unreliable’. Weibo users can click on the orange box, which is linked to the corresponding fact-checking page: <https://service.account.weibo.com/show?rid=K1CaS8wtk7K4k>.

- Who is debunking misinformation in social media? In the era of social media, it has become easier for misinformation to spread at an unprecedented rate, making the need for detecting and debunking misinformation at an early stage even more necessary (Hassan et al., 2017; Thorne and Vlachos, 2018; Zubiaga et al., 2018). Currently, the public generally relies on independent fact-checking platforms (e.g., Snopes<sup>3</sup> and Full Fact<sup>4</sup>) to verify the credibility of such unverified news items, which is highly reliable but costly in terms of time and human resources (e.g., professional journalists). The success of fact-checking platforms has also subsequently motivated individual users to actively debunk misinformation with reliable fact-checking information (see Table 1.1). Identifying and characterising these users who actively refute misinformation in social media is an important task in computational misinformation analysis, which can be combined with existing studies to prevent the propagation of misinformation at the user level.
- In social media, there exist some unreliable stories with higher impact (e.g., reaching more end users) which may have the potential to cause significant harm than those with lower impact (DiFonzo and Bordia, 2011; Einwiller and Kamins, 2008). By identifying and debunking such high-impact unreliable stories early on, it is possible to mitigate their negative effects and prevent them from causing

<sup>3</sup><https://www.snopes.com/>

<sup>4</sup><https://fullfact.org/>

widespread harm in time, especially in social emergencies (Imran et al., 2018; Xie et al., 2017; Xu et al., 2017). Besides, it can be used by social media platforms to flag less credible posts and deliver fact-checking information to alert end users who expose to such posts. Figure 1.1 shows how the Weibo fact-checking platform flags false rumours and provides the public with information to refute them. As a first step, we believe that it is crucial to create a new dataset to support further data-driven studies aimed at detecting unreliable stories with significant future impact in social media.

- In summary, the synergy between user and post level information helps create a more comprehensive and robust approach to misinformation detection. By considering both aspects, researchers and algorithms can gain a deeper understanding of the dynamics of misinformation propagation, enabling more effective and accurate identification and mitigation of false information on social media platforms.

## 1.2 Research Aims

The primary aim of this research is to explore and evaluate the utility of incorporating user and post-level information in the detection of misinformation on social media platforms. The rapid proliferation of misinformation on social media platforms poses significant challenges to the integrity of information dissemination. Traditional approaches to misinformation detection often focus solely on the content of individual posts, which may not provide a comprehensive understanding of the context and the users behind the dissemination. With the potential challenges, we aim to investigate the following research questions:

- *Q1* Does *temporal concept drift* affect the predictive performance of misinformation detection systems?
- *Q2* Can we automatically identify whether a social media (e.g., Twitter and Facebook) user will repost content (e.g., news stories) from *unreliable news*



*sources* by leveraging linguistic information from the user’s posts (e.g., historical timeline)?

- *Q3* Can we automatically identify and characterise individuals who *actively refute misinformation* based on their language usage in social media?
- *Q4* How can we define and predict the future popularity of unreliable posts (e.g., false rumours) given both post and user-level features?

## 1.3 Definitions

For consistency, we provide some definitions for various concepts that we use throughout this thesis:

- **Misinformation** Similar to Wu et al. (2019), we use *misinformation* to represent any false or unreliable information (e.g., fake news, false rumours, conspiracy theories, disinformation, etc.) that is shared and diffused by end users in social media.
- **Rumours and False Rumours** Following Zubiaga et al. (2018), we consider ‘any item of circulated information whose veracity is yet to be verified at the time of posting’ as *rumours* (i.e., unverified posts) in social media. Ultimately, rumours may be true, false, or unverified. Figure 1.1 displays an example of a *false rumour* diffused in Weibo platform.
- **Reliable and Unreliable News Sources** Following Volkova et al. (2017), most reliable news sources are mainstream media and commonly verified by Twitter (e.g., @BBC and @NBC). On the other hand, unreliable news sources (e.g., InfoWars, Russia Today, and Disclose.TV, etc) are categorised (e.g., propaganda, hoax, conspiracy, etc.) by independent organisations (e.g., PropOrNot and FakeNewsWatch).

## 1.4 Contributions

This thesis is organised following the thesis by publications format and consists of a collection of four papers. The main contributions of this thesis are as follows:

### 1.4.1 Paper I

To answer Q1, we provide the first re-evaluation of text classifiers on four widely used rumour detection benchmarks (i.e., Twitter 15&16, Weibo and PHEME (Ma et al., 2016, 2017; Zubiaga et al., 2016)) using chronological splits rather than standard random splits. To this end, we perform a battery of controlled experiments to examine the hypothesis that whether temporal concept drift affects the performance of rumour detection systems. Empirically results uncover that the predictive performance of rumour detection models trained with random data splits is significantly overestimated than chronological splits due to temporal concept drift.

*The paper has been accepted at the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023 Findings). <https://arxiv.org/pdf/2302.03147.pdf> Mu et al. (2023a)*

*My contributions to the work: Conceptualisation, Resources, Methodology, Software, Validation and Writing*

### 1.4.2 Paper II

Individuals who frequently share news stories from untrustworthy news sources (such as *InfoWars* and *Russia Today*) increase the likelihood of spreading misinformation (e.g., disinformation, fake news, etc.). To answer the second question, we made three main contributions:

- We define a novel task aiming to *early detect* Twitter users who propagate tweets

from unreliable news sources using a diverse set of linguistic features.

- Secondly, we create a new dataset comprising 6,266 Twitter users categorised into two groups, (i) users who spread tweets from unreliable news sources and (ii) those who only share content from trustworthy news sources.
- Finally, we conduct a linguistic analysis to highlight the difference in language pattern (i.e., N-grams, Topics, and LIWC) between the two Twitter user categories.

*The paper has been published in PeerJ Computer Science (<https://doi.org/10.7717/peerj-cs.325>) Mu and Aletras (2020)*

*My contributions to the work: Conceptualisation, Data Collection, Methodology, Software, Validation and Writing*

### 1.4.3 Paper III

Identifying and promoting users who are actively refuting misinformation can be used to improve the information environment in a social network and detect less credible posts in time. In the third work, we focus on detecting and analysing social media users who actively refute misinformation in social media. To explore the Q3, main contributions of this paper are as follows:

- We first create a new large dataset from Weibo consisting of approximately 49k users labelled either as active citizens (i.e., people who refute misinformation) or misinformation spreaders;
- Secondly, we re-frame an existing benchmark (Vo and Lee, 2018, 2019) to model the task of identifying active citizens and misinformation spreaders on Twitter;
- Thirdly, we perform experiments with a battery of pre-trained language models for the task. Given that the historical timeline of social media users can be lengthy (e.g., thousands of tweets), we design efficient hierarchical transformer-based systems achieving F1-measure of up to 85 on Weibo and 80 on Twitter.

- Finally, we perform a thorough linguistic analysis to uncover the differences in language usage between two user categories. We also manually conduct an error analysis of the limitations of the hierarchical network in accurately classifying social media users as misinformation debunkers or spreaders.

*This paper has been published in Proceedings of 14th ACM Web Science Conference (<https://doi.org/10.1145/3501247.3531559>) Mu et al. (2022)*

*My contributions to the paper include: Conceptualisation, Data Collection, Methodology, Software, Validation and Writing*

#### 1.4.4 Paper IV

There is a pressing need for the development of trustworthy computational systems for identifying *false rumours with significant impact*, as a supplement to prior research on automated detection of misinformation and verification of unreliable news stories. To address Q4, we made three main contributions in this work:

- We create a new publicly available dataset from Weibo, consisting of 19k false rumours in Chinese that have been debunked by the official Weibo fact-checking platform, along with their popularity score and meta-features.
- Additionally, we perform a battery of supervised models utilising post and user-level information. The best predictive performance is achieved by combining both sources of information with our newly developed pre-trained language model, named as BERT-Weibo-Rumour.
- Finally, we conduct a linguistic analysis to uncover the features of highly popular false rumours compared to those with low popularity. We also observe that some user profile features (e.g., account verified status, number of followers, and number of posts) are positively correlated with the impact of false rumours.

*This paper is under revision by the Expert Systems with Applications*

---

*My contributions to the paper include: Conceptualisation, Data Collection, Methodology, Software, Validation and Writing.*

# EXAMINING TEMPORALITIES ON RUMOR DETECTION

## It's about Time: Rethinking Evaluation on Rumor Detection Benchmarks using Chronological Splits

*Yida Mu, Kalina Bontcheva, Nikolaos Aletras*

Department of Computer Science, The University of Sheffield

{y.mu, k.bontcheva, n.aletras}@sheffield.ac.uk

### Abstract

New events emerge over time influencing the topics of rumors in social media. Current rumor detection benchmarks use random splits as training, development and test sets which typically results in topical overlaps. Consequently, models trained on random splits may not perform well on rumor classification on previously unseen topics due to the temporal concept drift. In this paper, we provide a re-evaluation of classification models on four popular rumor detection benchmarks considering chronological instead of random splits. Our experimental results show that the use of random splits can significantly overestimate predictive performance across all datasets and models. Therefore, we suggest that

rumor detection models should always be evaluated using chronological splits for minimizing topical overlaps.

## 2.1 Introduction

Unverified false rumors can spread faster than news from mainstream media, and often can disrupt the democratic process and increase hate speech (Vosoughi et al., 2018; Zubiaga et al., 2018). Automatic detection of rumors is an important task in computational social science, as it helps prevent the spread of false rumors at an early stage (Ma et al., 2017; Zhou et al., 2019; Karmakharm et al., 2019; Bian et al., 2020).

Current rumor detection approaches typically rely on existing annotated benchmarks consisting of social media data, e.g., Twitter 15 (Ma et al., 2017), Twitter 16 (Ma et al., 2017), Weibo (Ma et al., 2016), and PHEME (Zubiaga et al., 2016) that cover a wide range of time periods. These benchmarks use random splits for train, development and test sets which entail some topical overlap among them (see Table 2.1 for recent previous work). However, the distribution of topics in various NLP benchmarks (e.g., news, reviews, and biomedical) can be significantly affected by time (Huang and Paul, 2018, 2019). This is the phenomenon of temporal concept drift which can be induced by the changes in real-world events. Specifically, this also affects benchmarks on social media with new events such as elections, emergencies, pandemics, constantly creating new topics for discussion.

Existing work has investigated the sensitivity of computational approaches to temporal drift (i.e., the deterioration of their performance due to temporal/topic variation) when evaluated on temporal data splits. (Huang and Paul, 2018; Florio et al., 2020; Chalkidis and Sogaard, 2022). Using random splits also results into posts with almost identical textual content shared during the same period. Table 5.7 displays four pairs of posts with **similar or identical** text content sampled from four different rumor detection benchmarks. This potential information leakage, results in classifying data almost identical to ones already being present in the training set. For practical application reasons, we believe that in order to evaluate a rumor detection system, it is necessary to detect not only long-standing rumors, but also emerging ones.

In this paper, we design a battery of controlled experiments to explore the hypothesis that

whether temporality affects the predictive performance of rumor classifiers. To this end, we re-evaluate models on popular rumor detection benchmarks using chronological data splits i.e., by training the model with earlier posts and evaluating the model performance with the latest posts. Results show that the performance of rumor detection approaches trained with random data splits is significantly overestimated than chronological splits due to temporal concept drift. This suggests that rumor detection approaches should be evaluated with chronological data for real-world applications, i.e., to automatically detect emerging rumors.

## 2.2 Related Work

Gorman and Bedrick (2019) and Sogaard et al. (2021) have showed that using different data split strategies affects model performance in NLP downstream tasks. Previous work has demonstrated that text classifiers performance significantly drops in settings where chronological data splits are used instead of random splits in various domains, e.g., hate speech, legal, politics, sentiment analysis, and biomedical (Huang and Paul, 2018; Lukes and Sogaard, 2018; Huang and Paul, 2019; Florio et al., 2020; Chalkidis and Sogaard, 2022; Agarwal and Nenkova, 2022; Mu et al., 2023). To minimize topical overlaps, a Leave-One-Out (LOO) evaluation protocol has been proposed (Lukasik et al., 2015, 2016). While this topic split strategy could potentially mitigate temporal concept drift, it still yields temporal overlaps between each subset and is practically not applicable to most common rumour detection benchmarks with a large number of topics (e.g., Twitter 15, Twitter 16, Weibo, etc.). We observe that the LOO protocol can be used for a few specific rumor detection benchmarks, such as (PHEME (Zubiaga et al., 2016)), where each post is associated with a corresponding event, e.g., *Ottawa Shooting* and *Charlie Hebdo shooting*. By leveraging the consistent format of datasets collected from the same platform (e.g., Twitter and Weibo), previous work has explored broader temporalities by training a rumor classifier on Twitter 15 and evaluating its performance on Twitter 16. This protocol enables a more comprehensive examination of the generalizability of rumor detection systems, which is crucial for their practical applications in the real world (Moore and Rayson, 2018; Yin and Zubiaga, 2021; Kochkina et al., 2023).



| Paper                       | Twitter 15 | Twitter 16 | PHEME | Weibo |
|-----------------------------|------------|------------|-------|-------|
| Tian et al. (2022)          | ✓          | ✓          | -     | ✓     |
| Zeng and Gao (2022)         | -          | ✓          | ✓     | -     |
| Sheng et al. (2022)         | -          | -          | -     | ✓     |
| Mukherjee et al. (2022)     | -          | -          | ✓     | -     |
| Sun et al. (2022)           | ✓          | ✓          | ✓     | -     |
| de Silva and Dou (2021)     | ✓          | ✓          | -     | -     |
| Ren et al. (2021)           | -          | -          | ✓     | -     |
| Wei et al. (2021)           | ✓          | ✓          | ✓     | -     |
| Li et al. (2021)            | -          | -          | ✓     | -     |
| Rao et al. (2021)           | ✓          | ✓          | -     | ✓     |
| Lin et al. (2021)           | ✓          | ✓          | ✓     | -     |
| Farinneya et al. (2021)     | -          | -          | ✓     | -     |
| Sun et al. (2021)           | -          | -          | ✓     | -     |
| Qian et al. (2021)          | -          | -          | ✓     | -     |
| Song et al. (2021)          | ✓          | ✓          | ✓     | -     |
| Kochkina and Liakata (2020) | ✓          | ✓          | ✓     | -     |
| Yu et al. (2020)            | -          | -          | ✓     | -     |
| Xia et al. (2020)           | -          | ✓          | -     | ✓     |
| Bian et al. (2020)          | ✓          | ✓          | -     | ✓     |
| Lu and Li (2020)            | ✓          | ✓          | -     | -     |

**Table 2.1:** Recent work on rumor detection using random splits.

| Dataset | id   | Post                                                                                                                              | Label     | Leven |
|---------|------|-----------------------------------------------------------------------------------------------------------------------------------|-----------|-------|
| Twi.15  | 407* | r.i.p to the driver <b>who</b> died with paul walker that no one cares about because he wasn't famous.                            | Rumor     | 3     |
|         | 407* | r.i.p to the driver <b>that</b> died with paul walker that no one cares about because he wasn't famous.                           | Rumor     |       |
| Twi.16  | 594* | the kissing islands, greenland. <b>URL</b>                                                                                        | Non-Rumor | 0     |
|         | 604* | the kissing islands, greenland. <b>URL</b>                                                                                        | Non-Rumor |       |
| PHEME   | 498* | <b>happening</b> now in #ferguson URL                                                                                             | Non-Rumor | 9     |
|         | 499* | <b>Right</b> now in #ferguson URL                                                                                                 | Non-Rumor |       |
| Weibo   | 349* | 【喝易拉罐一定要吸管】一妇女喝了罐饮料，被送进医院，离开了世界。研究显示罐上面的毒菌很多 <b>请转给你关心的朋友。</b> <b>Translation: Please forward to your friends you care about.</b> | Rumor     | 10    |
|         | 350* | 【喝易拉罐一定要吸管】一妇女喝了罐饮料，被送进医院，离开了世界。研究显示罐上面的毒菌很多！！ <b>这些你知道吗</b> <b>Translation: Do you know about this?</b>                          | Rumor     |       |

**Table 2.2:** Four pairs of posts from train and test data with similar or identical text content sampled from four rumor detection benchmarks. Post ids with close values indicate that two posts are published in the same period. **Leven** denotes the Levenshtein distance (Levenshtein et al., 1966) on character-level between the two posts with the same label (i.e., lower values indicate higher text similarity and vice versa).

| Splits             | Benchmarks      | Twitter 15 |     |      | Twitter 16 |     |      | PHEME |     |      | Weibo |     |      |
|--------------------|-----------------|------------|-----|------|------------|-----|------|-------|-----|------|-------|-----|------|
|                    | Subsets         | Train      | Dev | Test | Train      | Dev | Test | Train | Dev | Test | Train | Dev | Test |
| Standard Chrono.   | # of Rumors     | 285        | 35  | 52   | -          | -   | -    | 1,420 | 72  | 480  | -     | -   | -    |
|                    | # of Non-Rumors | 234        | 40  | 96   | -          | -   | -    | 2,641 | 508 | 681  | -     | -   | -    |
| Stratified Chrono. | # of Rumors     | 260        | 37  | 75   | 144        | 21  | 40   | 1,380 | 197 | 394  | 1,645 | 235 | 470  |
|                    | # of Non-Rumors | 259        | 37  | 74   | 144        | 21  | 40   | 2,681 | 383 | 766  | 1,619 | 231 | 463  |
| Random Splits      | # of Rumors     | 260        | 37  | 75   | 144        | 21  | 40   | 1,380 | 197 | 394  | 1,645 | 235 | 470  |
|                    | # of Non-Rumors | 259        | 37  | 74   | 144        | 21  | 40   | 2,681 | 383 | 766  | 1,619 | 231 | 463  |

**Table 2.3:** Statistics of subsets. Note that using random splitting yields the same percentage of examples in each category as in the stratified chronological splits.

## 2.3 Methodology

### 2.3.1 Data

We use four most popular rumor detection benchmarks, three in English and one in Chinese. Note that most related work is currently evaluating their rumor detection systems on two or three of these four benchmarks. (see Table 2.1).

**Twitter 15 and Twitter 16:** These datasets contain 1,490 and 818 tweets labeled into four categories including Non-rumor (NR), False Rumor (FR), True Rumor (TR), and Unverified Rumor (UR) introduced by Ma et al. (2017).

**PHEME:** This benchmark contains 5,802 verified tweets collected from 9 real-world breaking news events (e.g., Ottawa Shooting, Ferguson Unrest, etc.) associated with two labels, i.e., 1,972 Rumor and 3,830 Non-Rumor (Zubiaga et al., 2016).

**Weibo:** This dataset includes 4,664 verified posts in Chinese including 2,313 rumors debunked by the Weibo Rumor Debunk Platform<sup>1</sup> and 2,351 non-Rumors from Chinese media (Ma et al., 2016).

**Data Pre-processing** We opt for the binary setup (i.e., re-frame all benchmarks as rumor detection) to distinguish true/false information following Lu and Li (2020); Rao et al. (2021). We pre-process the posts by replacing @mention and hyperlinks with @USER and URL respectively. We also lowercase the tweets from three Twitter benchmarks.

---

<sup>1</sup><https://service.account.weibo.com/?type=5&status=4>

| Statistic                     | Twitter 15 | Twitter 16 | Weibo     | PHEME     |
|-------------------------------|------------|------------|-----------|-----------|
| <i># of source posts</i>      | 1,490      | 818        | 4,664     | 5,802     |
| <i># of True rumors</i>       | 374        | 205        | 2,351     | 3,830     |
| <i># of False rumors</i>      | 370        | 205        | 2,313     | 1,972     |
| <i># of Unverified rumors</i> | 374        | 203        | -         | -         |
| <i># of Non-rumors</i>        | 372        | 205        | -         | -         |
| <i>Time Span (Year)</i>       | 2015       | 2016       | 2012-2016 | 2014-2015 |

**Table 2.4:** Dataset statistics.

### 2.3.2 Data Splits

**Standard Chronological Splits** For Twitter 15 and PHEME, we first sort all posts chronologically and then divide them into three subsets including a training set (70% of the earliest data), a development set (10% of data after train and before test), and a test set (20% of the latest data). There is no temporal overlap between the three subsets.

**Stratified Chronological Splits** On the other hand, we observe that there is no temporal overlap between rumors and non-rumors in Twitter 16 and Weibo datasets. This suggests that it is not possible to use standard chronological splits as in Twitter 15 and PHEME.

Therefore, we apply a **stratified chronological split** strategy for all benchmarks. We first split rumors and non-rumors separately in chronological order. We then divide them into three subsets (a total of six subsets), i.g., all rumors are split into a training set (70% of the earliest rumors), a development set (10% of data after train and before test), and a test set (20% of the latest rumors). Finally, we merge the six subsets into the final three train, development and test sets. Note that this approach will result in no temporal overlap for **each label (i.e., rumor or non-rumor)** among the three final sets. We show the number of each split in Table 2.3.

**Random Splits** Following standard practice (e.g., Bian et al. 2020; Lin et al. 2021; Rao et al. 2021), we **randomly** split data using a 5-fold cross-validation. Note that these splits are made by preserving the percentage of posts in each category. Each split contains a training set (70%), development set (10%) and a test set (20%) with the same ratio as in our chronological

splits.

**Leave-One-Out (LOO) Splits** For reference, we also provide the results of using the LOO evaluation protocol on PHEME dataset (see Table 2.6).

### 2.3.3 Models

The main purpose of our experiments is to improve model evaluation by investigating the effects of temporal drifts in rumor detection by providing an extensive empirical study. Therefore, we opted using strong text classifiers that are generic and can be applied to all of our benchmarks:

- **LR** We train a LR classifier using BOW to represent posts weighted by TF-IDF using a vocabulary of 5,000 n-grams.
- **BERT** We directly fine-tune the BERT base model by adding a linear prediction layer on the top of the 12-layer transformer architecture following (Devlin et al., 2019).
- **BERT+ (BERTweet and ERNIE)** We also experiment with two domain specific models: BERTweet (Nguyen et al., 2020) and ERNIE (Sun et al., 2020) pre-trained on social media data using the same fine-tune strategy as the original BERT model.

### 2.3.4 Hyperparameters and Implementation Details

We train the model on the training set, perform model tuning and selecting on the development set, and evaluate performance on the test set. To evaluate the chronological data splits, we run the model five times with different random seeds for consistency. All chronological splits are available for reproducibility.<sup>2</sup>

For logistic regression, we use word-level and character-level tokenizers for Twitter and Weibo datasets respectively and only consider uni-gram, bi-grams, and tri-grams that appear in more than two posts for each dataset. For BERT, we set learning rate  $lr = 2e - 5$ , batch

---

<sup>2</sup>[https://github.com/YIDAMU/Rumor\\_Benchmarks\\_Temporality](https://github.com/YIDAMU/Rumor_Benchmarks_Temporality)

| Model | Strategy                 | Twitter15   |            |             | PHEME      |            |            |
|-------|--------------------------|-------------|------------|-------------|------------|------------|------------|
|       |                          | P           | R          | F1          | P          | R          | F1         |
| LR    | Random                   | 86.7 ± 2.1  | 85.2 ± 1.8 | 85.0 ± 1.8  | 84.1 ± 1.2 | 79.3 ± 1.0 | 80.9 ± 1.0 |
|       | Standard Chronological   | 56.6 ± 0.8  | 56.3 ± 0.7 | 56.4 ± 0.7  | 67.3 ± 0.1 | 64.0 ± 0.1 | 63.9 ± 0.1 |
|       | Stratified Chronological | 56.3 ± 2.5  | 51.9 ± 0.7 | 41.4 ± 0.4  | 64.5 ± 0.2 | 63.0 ± 0.3 | 63.5 ± 0.3 |
| BERT  | Random                   | 88.2 ± 2.4  | 87.9 ± 2.2 | 87.9 ± 2.2  | 84.8 ± 0.5 | 84.8 ± 1.2 | 84.8 ± 0.8 |
|       | Standard Chronological   | 54.8 ± 4.0  | 55.1 ± 4.3 | 52.9 ± 3.6  | 74.8 ± 1.1 | 75.1 ± 0.8 | 73.7 ± 0.4 |
|       | Stratified Chronological | 58.2 ± 7.3  | 56.1 ± 4.5 | 52.8 ± 5.6  | 75.5 ± 0.6 | 77.7 ± 0.5 | 75.7 ± 1.1 |
| BERT+ | Random                   | 90.8 ± 1.2  | 90.4 ± 1.2 | 90.4 ± 1.2  | 84.6 ± 1.0 | 85.5 ± 0.9 | 85.0 ± 0.8 |
|       | Standard Chronological   | 58.6 ± 1.9  | 58.8 ± 2.1 | 57.4 ± 2.5  | 76.1 ± 1.1 | 74.8 ± 1.5 | 71.6 ± 2.2 |
|       | Stratified Chronological | 61.8 ± 6.5  | 57.9 ± 2.4 | 55.2 ± 1.5  | 75.3 ± 0.9 | 76.9 ± 2.1 | 71.0 ± 3.5 |
| Model | Strategy                 | Twitter16   |            |             | Weibo      |            |            |
|       |                          | P           | R          | F1          | P          | R          | F1         |
| LR    | Random                   | 89.9 ± 1.2  | 89.3 ± 1.5 | 89.3 ± 1.5  | 90.1 ± 0.9 | 90.1 ± 0.9 | 90.1 ± 0.9 |
|       | Stratified Chronological | 62.1 ± 6.9  | 55.8 ± 4.7 | 48.7 ± 11.4 | 79.1 ± 0.1 | 78.1 ± 0.1 | 77.9 ± 0.1 |
| BERT  | Random                   | 91.9 ± 1.0  | 91.5 ± 0.8 | 91.5 ± 0.8  | 92.3 ± 1.2 | 92.2 ± 1.2 | 91.2 ± 1.2 |
|       | Stratified Chronological | 61.0 ± 11.2 | 54.3 ± 4.3 | 47.2 ± 3.5  | 89.0 ± 2.5 | 87.6 ± 2.6 | 87.5 ± 2.6 |
| BERT+ | Random                   | 89.8 ± 2.8  | 89.3 ± 3.2 | 89.3 ± 3.3  | 92.5 ± .4  | 92.5 ± .4  | 92.5 ± .4  |
|       | Stratified Chronological | 49.8 ± 1.7  | 49.9 ± 0.9 | 45.1 ± 2.9  | 88.1 ± 2.5 | 87.6 ± 1.4 | 88.5 ± 1.5 |

**Table 2.5:** Rumor detection prediction results across different data split methods. Green cells indicate that the model trained on random splits performs significantly better than both standard chronological splits and stratified chronological splits ( $p < 0.05$ , t-test).

| Model | PHEME      |            |            |
|-------|------------|------------|------------|
|       | P          | R          | F1         |
| LR    | 68.3 ± 3.8 | 65.1 ± 6.3 | 63.2 ± 6.3 |
| BERT  | 73.4 ± 3.1 | 71.9 ± 6.1 | 70.7 ± 4.9 |
| BERT+ | 75.3 ± 2.2 | 72.6 ± 8.1 | 71.4 ± 7.0 |

**Table 2.6:** Leave-One-Out evaluation protocol on PHEME dataset.

size  $bs = 32$ , and maximum input length as 256 covering the max tokens of all posts. All BERT-style models are trained for 10 epochs using the early stopping method based on the loss on the development set. The best checkpoint model is saved for evaluation on the test set. The average run time of 10 epochs for the BERT model is less than 2 minutes. We employ Bert-Base-Uncased, Bertweet-Base and Chinese-Bert-WWM, Ernie-1.0 models from the HuggingFace library (Wolf et al., 2020). All experiments are conducted on a single NVIDIA V100 GPU with 32GB memory.

### 2.3.5 Evaluation Metrics

For all tasks, we report the averaged macro Precision, Recall and F1 values across five runs using different random seeds.

## 2.4 Results

**Random Splits vs. Chronological Splits** Table 2.5 shows the experimental results across all models and rumor detection benchmarks using **chronological splits** and random **5-fold cross-validation**. Overall, we observe that the use of random splits always leads to a significant overestimation of performance compared to chronological splits (t-test,  $p < 0.05$ ) across all models. Our results corroborate findings from previous work on studying temporal concept drift (Huang and Paul, 2018; Chalkidis and Søgaard, 2022). This suggests that chronological splits are necessary to more realistically evaluate rumor detection models.

We also note that the effect of temporality varies in datasets of different size. For both data splitting strategies, we observe that the difference in performance is 50% higher for the two datasets with hundreds of posts (e.g., Twitter 15 and Twitter 16) and around 10% in ones with thousands of posts (e.g., PHEME and Weibo). For rumor detection tasks, temporality may have a greater impact on small-scale benchmarks than on large-scale benchmarks. For Twitter 16 and Weibo, the use of stratified chronological splits demonstrates significant performance drops compared to random splits due to the temporal concept drift.

For chronological splits, we observe that pre-trained language models (i.e., BERT and BERT+) significantly outperform (t-test,  $p < 0.05$ ) logistic regression in all benchmarks. This is due to the fact that BERT-style models (i) outperform simpler linear models by a large margin in various NLP tasks Devlin et al. (2019); and (ii) have been trained after the development of these four benchmarks implying some information leakage.

**Standard vs. Stratified Chronological Splits** Note that dividing the datasets into standard chronological splits results in subsets that do not preserve the sample percentages for each category (see Table 2.3). The upper part of Table 2.5 displays the difference in model performance between two types of chronological splits on Twitter 15 and PHEME. We observe

| Benchmark |                          | Twitter 15 |    |     | Twitter 16 |    |     | PHEME |     |     | Weibo |     |     |
|-----------|--------------------------|------------|----|-----|------------|----|-----|-------|-----|-----|-------|-----|-----|
| Splits    | Test set                 | total      | #  | %   | total      | #  | %   | total | #   | %   | total | #   | %   |
| Chrono.   | all posts                | 148        | 3  | 2%  | 82         | 6  | 7%  | 1161  | 39  | 3%  | 933   | 41  | 4%  |
|           | # of wrong predictions   | 63         | 2  | 1%  | 34         | 2  | 2%  | 301   | 5   | <1% | 99    | 7   | <1% |
|           | # of correct predictions | 85         | 1  | 1%  | 48         | 4  | 5%  | 860   | 34  | 3%  | 834   | 34  | 4%  |
| Random    | all posts                | 149        | 35 | 23% | 83         | 26 | 30% | 1161  | 181 | 16% | 933   | 129 | 14% |
|           | # of wrong predictions   | 12         | 0  | 0%  | 5          | 1  | 1%  | 150   | 14  | 1%  | 65    | 4   | <1% |
|           | # of correct predictions | 137        | 35 | 23% | 78         | 25 | 30% | 1011  | 167 | 14% | 868   | 125 | 14% |

**Table 2.7:** Error Analysis for all benchmarks. # denotes the number of posts that are similar to posts from training set, i.e., known data. % denote the percentage of similar posts in the test set. We set the threshold value to 20, which indicates that there are two or three different words between the two tweets.

|               | Example                                                                                                                                                                                                        | Test | Train | Correct | Wrong |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|-------|---------|-------|
| <b>Tw1.15</b> | #rip to the driver who died with #paulwalker that no one cares about ...                                                                                                                                       | 4    | 6     | 4       | 0     |
| <b>Tw1.16</b> | steve jobs was adopted. his biological father was abdufattah jandali, a syrian muslim                                                                                                                          | 2    | 13    | 2       | 0     |
| <b>PHEME</b>  | Police are leaving now . #ferguson HTTPURL                                                                                                                                                                     | 4    | 11    | 4       | 0     |
| <b>Weibo</b>  | 【交通新规】2013年1月1日施行:1... 扩散给大家! [广州日报]<br>Translation: [New driving laws] From 1 Jan 2013: Running a red light will result in a fine of 100 RMB and 6 points. ... Spread the news to everyone! [Guangzhou Daily] | 2    | 6     | 2       | 0     |

**Table 2.8:** Four examples of correct predictions using random splits, which artificially removes temporal concept drift. For example, in Twitter 15, there are 4 and 6 similar posts about rumors related to Paul Walker in the test set and the training set respectively.

that using both standard and stratified chronological splits results in similar model predictive performance (t-test,  $p > 0.05$ ). Even though stratified chronological splits contain temporal overlap, it is still not sufficient to improve model performance compared to random splits. This suggests that the temporal drift affects particular classes rather than the entire data set.

## 2.5 Error Analysis

Finally, we perform an error analysis to further investigate the type of errors made by BERT using both random and chronological splits. Table 2.7 shows the number of correct and wrong predictions for each of the two data splitting strategies. We also use the Levenshtein distance<sup>3</sup> to calculate the quantity of posts in the test set that are similar to posts in the corresponding train set.

- We first observe that the temporal concept drift is evident in all rumor detection

<sup>3</sup>We set the threshold value to 20.

benchmarks. Most of the rumors on the same topic are posted in a very short time span.

- In addition, long-standing rumors<sup>4</sup> are only a small part of the data (less than 5%). Second, we note that using random splits leads to topical overlap between the training and test sets (see Table 2.8) resulting in higher model performance.
- Finally, for both random and chronological splits, most of the posts in the test set with overlapping topics in the training set are predicted correctly. In contrast, wrong predictions are often posts with emerging or different topics compared to the posts in the train set.

## 2.6 Conclusion

We have shed light on the impact of temporal drift on computational rumor detection. Results from our controlled experiments show that the use of chronological splits causes substantially drops in predictive performance across widely-used rumor detection benchmarks. This suggests that random splits rather overestimate the model predictive performance. We argue that the temporal concept drift needs to be considered when developing real-world rumor detection approaches. In the future, we plan to study the impact of temporal concept drift on other NLP tasks, such as detecting user reactions to untrustworthy posts on social media (Glenski et al., 2018; Mu and Aletras, 2020; Mu et al., 2022).

## Limitations and Future Work

We provide the first re-evaluation of four standard rumor detection benchmarks in two languages (English and Chinese) from two platforms (Twitter and Weibo). We acknowledge that further investigation is needed in rumor detection datasets in other languages. Besides, rumors on social media also contain a rich amount of contextual information, including comments, user profile information, retweets, and images, which complements the text of

---

<sup>4</sup>Here, long-standing rumors refer to rumors that appear in both the training and test sets when using temporal splits.



the source posts. In the future, we plan to conduct a systematic evaluation of context-based rumor detection approaches, such as utilizing both the source post and contextual information as input, which is a hitherto unstudied research question. Additionally, we aim to address the challenge of dealing with temporal concept drift in such NLP downstream tasks.

## Acknowledgments

We would like to thank Ahmed Alajrami, Danae Sánchez Villegas, Mali Jin, Xutan Peng and all the anonymous reviewers for their valuable feedback.

---

## BIBLIOGRAPHY

- Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.
- Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 549–556.
- Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Parsa Farinneya, Mohammad Mahdi Abdollah Pour, Sardar Hamidian, and Mona Diab. 2021. Active learning for rumor identification on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4556–4565.
- Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

- Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. Identifying and understanding user reactions to deceptive and trusted social news sources. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–181, Melbourne, Australia. Association for Computational Linguistics.
- Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.
- Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.
- Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.
- Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.
- Elena Kochkina, Tamanna Hossain, Robert L Logan IV, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116.
- Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6964–6981, Online. Association for Computational Linguistics.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union.
- Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection. In *Findings of the Association for*

- Computational Linguistics: ACL-IJCNLP 2021*, pages 705–715, Online. Association for Computational Linguistics.
- Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on Twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.
- Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgments of rumours in social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2590–2595, Lisbon, Portugal. Association for Computational Linguistics.
- Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany. Association for Computational Linguistics.
- Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

- Andrew Moore and Paul Rayson. 2018. Bringing replication and reproduction together with generalisability in nlp: Three reproduction studies for target dependent sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1132–1144.
- Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6:e325.
- Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2023. Examining temporalities on stance detection towards covid-19 vaccination. *arXiv preprint arXiv:2304.04806*.
- Yida Mu, Pu Niu, and Nikolaos Aletras. 2022. Identifying and characterizing active citizens who refute misinformation in social media. In *14th ACM Web Science Conference 2022, WebSci '22*, page 401–410, New York, NY, USA. Association for Computing Machinery.
- Rajdeep Mukherjee, Uppada Vishnu, Hari Chandana Peruri, Sourangshu Bhattacharya, Koustav Rudra, Pawan Goyal, and Niloy Ganguly. 2022. Mtlts: A multi-task framework to obtain trustworthy summaries from crisis-related microblogs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 755–763.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.
- Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162.
- Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking network based on level-grained attention-masked BERT for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xiaoying Ren, Jing Jiang, Ling Min Serena Khoo, and Hai Leong Chieu. 2021. Cross-topic rumor detection using topic-mixtures. In *Proceedings of the 16th Conference of the European*

- Chapter of the Association for Computational Linguistics: Main Volume*, pages 1534–1538, Online. Association for Computational Linguistics.
- Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556.
- Nisansa de Silva and Dejing Dou. 2021. Semantic oppositeness assisted deep contextual modeling for automatic rumor detection in social networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 405–415, Online. Association for Computational Linguistics.
- Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.
- Yun-Zhu Song, Yi-Syuan Chen, Yi-Ting Chang, Shao-Yu Weng, and Hong-Han Shuai. 2021. Adversary-aware rumor detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1371–1382, Online. Association for Computational Linguistics.
- Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency matters: A knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1412–1423, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022, WWW '22*, page 2789–2797, New York, NY, USA. Association for Computing Machinery.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.
- Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2022. DUCK: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022*

- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949, Seattle, United States. Association for Computational Linguistics.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3845–3854, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9042–9051, Online. Association for Computational Linguistics.
- Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.
- Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401, Online. Association for Computational Linguistics.
- Fengzhu Zeng and Wei Gao. 2022. Early rumor detection using neural hawkes process with a new benchmark dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4105–4117.

- Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623, Minneapolis, Minnesota. Association for Computational Linguistics.
- Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.



# IDENTIFYING MISINFORMATION SPREADERS

## Identifying Twitter users who repost unreliable news sources with linguistic information

*Yida Mu and Nikolaos Aletras*

Department of Computer Science, The University of Sheffield

### Abstract

Social media has become a popular source for online news consumption with millions of users worldwide. However, it has become a primary platform for spreading disinformation with severe societal implications. Automatically identifying social media users that are likely to propagate posts from handles of unreliable news sources sometime in the future is of utmost importance for early detection and prevention of disinformation diffusion in a network, and has yet to be explored. To that end, we present a novel task for predicting whether a user will repost content from Twitter handles of *unreliable news sources* by leveraging linguistic information from the user's own posts. We develop a new dataset of

approximately 6.2K Twitter users mapped into two categories: (1) those that have reposted content from unreliable news sources; and (2) those that repost content only from reliable sources. For our task, we evaluate a battery of supervised machine learning models as well as state-of-the-art neural models, achieving up to 79.7 macro F1. In addition, our linguistic feature analysis uncovers differences in language use and style between the two user categories.

## 3.1 Introduction

Social media has become an important source for online news consumption, widely adopted by news outlets, individual journalists and end users (Hermida et al., 2012; Kalsnes and Larsson, 2018). The use of social media enhances civic engagement and political participation offering a direct way of communication with millions of users worldwide (Bennett, 2008; Gil de Zúñiga et al., 2012).

A widespread phenomenon in social media platforms is the generation and dissemination of unreliable content (e.g., fabricated or deceptive information, exaggerated headlines, pseudo-science, propaganda) by particular news outlets that typically act as disinformation diffusion sources. Diffusion of disinformation in social media typically begins when a news source publishes a story that subsequently is propagated by users via reposting (e.g., retweeting, sharing) it to their personal networks of friends. It has been observed that disinformation propagates faster compared to credible information amongst users in social media (Vosoughi et al., 2018; Lazer et al., 2018). Furthermore, when a user comes across an unreliable story once, it is enough to increase their later perception of its accuracy (Pennycook et al., 2018). Media that disseminate unreliable content often aim to manipulate people’s opinion and influence election results which has implications to political stability worldwide (Allcott and Gentzkow, 2017; Humprecht, 2018).

Previous studies suggest that factors positively associated with the sharing unreliable news posts on social network include psychological factors (e.g., online trust, self-disclosure, fear of missing out, and ideological extremity) and political orientation (e.g., right-leaning) (Shu et al., 2019; Talwar et al., 2019; Hopp et al., 2020). In this study, we investigate whether user language information can help identify who will repost items from Twitter handles of unreliable news sources. To test this hypothesis, we define a new classification task seeking to predict

whether a user is likely to repost content from *unreliable news sources* given all the history of the user’s posts up to the first repost of a news item, i.e., *before they actually do it*. *Early detection* of users that are likely to repost content from unreliable sources can help: (i) political scientists and journalists to analyse which topics of discussion are related to disinformation on a large scale (Bode and Vraga, 2015); (ii) social media platforms such as Twitter or Facebook to prevent the diffusion of potentially unreliable stories in the network (Castillo et al., 2011; Conroy et al., 2015; Shu et al., 2017); and (iii) psychologists to complement studies on personality analysis (Pennycook and Rand, 2019). The main contributions of our paper are as follows:

- We frame a novel binary classification task for early detection of users sharing content from unreliable news sources using diverse language features extracted from the aggregate of users’ original tweets;
- We evaluate a battery of traditional feature-based and neural predictive models (including hierarchical transformers) that achieve up to 79.7 F1 score;
- We performed a qualitative analysis and found that users who diffuse unreliable news sources are more prevalent in expressing negative emotions and tweeting about politics and religion, while the rest of the users are more likely to express positive emotions and share information about their personal lives.

## 3.2 Background and Related Work

### 3.2.1 Disinformation in Social Media

Social media has become a primary platform for live-reporting (Engesser and Humprrecht, 2015) with the majority of mainstream news media operating official accounts (e.g., @BBC and @Reuters on Twitter). However, social media platforms are also regarded as a fertile breeding ground for the diffusion of unverified, fabricated and misleading information due to its openness and popularity (Zubiaga et al., 2018a). This type of information is often referred to as misinformation.

Misinformation has been defined as an umbrella term to include any incorrect information

that is diffused in social networks (Wu et al., 2019). On the other hand, disinformation is defined as the dissemination of fabricated and factually incorrect information with main aim to *deliberately deceive* its audience (Glenski et al., 2018a).

### 3.2.2 Categorization of Unreliable News Sources

Unreliable news sources are categorized by their intention and the degree of authenticity of their content (Rubin et al., 2015; Rashkin et al., 2017). Rubin et al. (2015) define three categories of deceptive news: (1) serious fabrications including unverified claims coupled with exaggerations and sensationalism; (2) large-scale hoaxes that are masqueraded as credible news which could be picked up and mistakenly disseminated; and (3) humorous fakes that present fabricated purposes with no intention to deceive. Rashkin et al. (2017) extended these three groups of misinformation into a more fine-grained classification:

- Propaganda news uses misleading information and writing techniques (Da San Martino et al., 2019) to promote a particular agenda (Glenski et al., 2018b). Propaganda news sources that mostly share unreliable stories often aim to manipulate people’s opinions and influence election results posing a threat to political stability worldwide (Allcott and Gentzkow, 2017; Humprecht, 2018).
- Clickbait is defined as using exaggerated headlines for grabbing user attention and misleading public opinion (Glenski et al., 2018b).
- Conspiracy theories can be understood as a kind of distorted interpretation of real events from people with ulterior motives such as political and religious groups (Goertzel, 1994; Byford, 2011).
- Satire news commonly mimics professional news press, incorporating irony and illogical contents for humour purposes (Tandoc Jr. et al., 2018; Burfoot and Baldwin, 2009).

Recent efforts on detecting and index unreliable news sources rely on crowdsourcing and experts<sup>1</sup> to annotate the reliability of the news media (Volkova et al., 2017; Baly et al., 2018; Glenski et al., 2018b).

---

<sup>1</sup>For example <http://www.fakenewswatch.com/>, <http://www.propornot.com>, <https://mediabiasfactcheck.com>, etc.

### 3.2.3 Previous Work on Combating Online Disinformation

Previous work on combating diffusion of disinformation in social media (Castillo et al., 2011; Conroy et al., 2015; Shu et al., 2017) has focused on characterizing the trustworthiness of (1) news sources (Dong et al., 2015; Baly et al., 2018); (2) news articles (Rashkin et al., 2017; Horne et al., 2018; Potthast et al., 2018; Pérez-Rosas et al., 2018); and (3) individual claims including news article headlines and rumors (Popat et al., 2016; Derczynski et al., 2017; Volkova et al., 2017; Zubiaga et al., 2018b; Thorne and Vlachos, 2018). Zhou et al. (2019) present a novel task for detecting the check-point which can early-detect a rumor propagated in a social network. Da San Martino et al. (2019) develop models for detecting up to 18 writing techniques (e.g., loaded language, slogans, flag-waving, exaggeration, etc.) used in propaganda news. Similarly, Pathak and Srihari (2019) introduced a corpus of news articles related to US politics containing false assertions which are written in a compelling way.

At the user level, social scientists and psychologists have utilised traditional methods, such as recruiting participants for online surveys and interviews, to explore cognitive factors which may influence people’s ability to distinguish fake news (Pennycook et al., 2018). For instance, the lack of analytic thinking plays a vital role in recognition of misinformation (Pennycook and Rand, 2019). Previous data-driven studies include (1) analysing bots participation in social media discussion (Howard and Kollanyi, 2016) and distinguishing between automated and human accounts (Mihaylov and Nakov, 2016); (2) identifying user reactions (e.g., agreement, answer, appreciation, humor, etc) to reliable/unreliable news posts (Glenski et al., 2018b); and (3) analyzing the demographic characteristics of users propagating unreliable news sources (Glenski et al., 2018a), e.g., low-income and low-educated people are more likely to propagate unreliable news sources on social networks.

In our paper, we tackle the problem of early detecting users who are likely to share post from unreliable news sources which is rather different to the focus of previous work on disinformation detection and analysis.

### 3.3 Task Description

Our aim is the early detection of social media users that are likely to repost content from unreliable news sources before they actually share any other news items at all. To that end, we define a novel binary classification task for predicting whether a social media user will propagate news items from unreliable or reliable news sources using solely **language information**.<sup>2</sup>

We assume a training set of  $n$  users  $U = \{(x_1, y_1), \dots, (x_n, y_n)\}$  where  $x_i$  is a vector representation of language information extracted from user's  $i$  timeline consisting of posts up to the first repost of any news item, and  $y_i \in \{\mathbf{reliable}, \mathbf{unreliable}\}$  is an associated user label. Given  $U$ , we learn a function  $f$  that maps a new user  $j$  into one of the two categories  $\hat{y} = f(x_j)$  using any suitable supervised machine learning algorithm.

We consider the posts up to the first share of any news item, ensuring that we only use prior information that is not linked to any news source. One could also introduce a cut-off in time or keep the top  $k$  posts but we choose to use all the available information possible. We opted to define a binary task (i.e., reliable vs. unreliable) rather than a fine-grained classification task (i.e., propaganda, hoax, clickbait, and reliable) because propagating any type of disinformation might be equally harmful. For similar reasons, we are not focusing on modeling the proportion of posts from reliable/unreliable sources in users' Twitter Timeline.

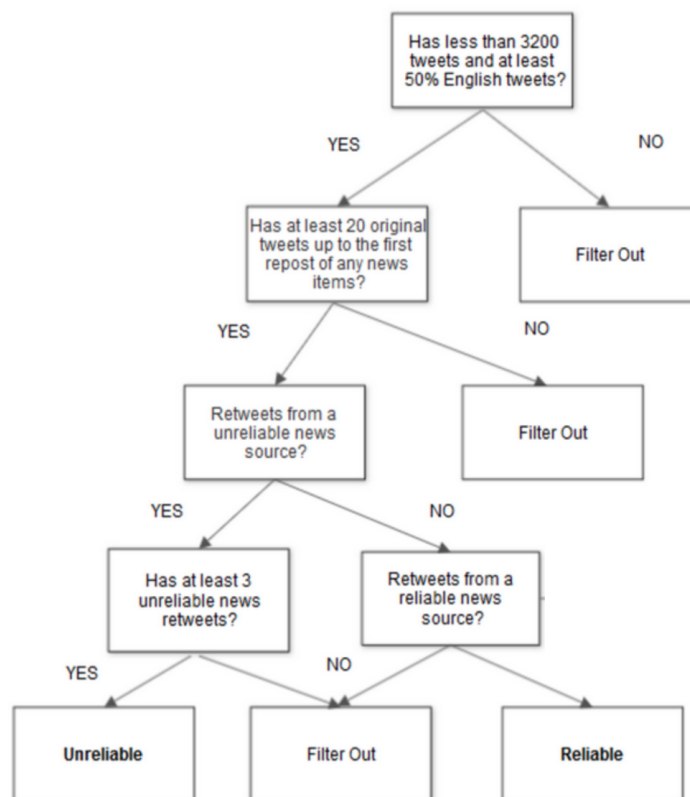
### 3.4 Data

At present, there is no existing dataset to model our predictive task. For the purposes of our experiments, we develop a new dataset of Twitter users who have retweeted posts from unreliable or reliable news sources. We opted for Twitter because the majority of accounts and posts are publicly available and it has been extensively used in related work (Volkova et al., 2017; Rashkin et al., 2017; Glenski et al., 2018b).

Our data collection process consists of three main steps (summarized in Figure 3.1): (1)

---

<sup>2</sup>Note that one could use a user's social network information but this is out of the paper's scope because we are interested in analysing differences in language use between the two groups of users.



**Figure 3.1:** User filtering and labeling flowchart.

collection of posts from reliable and unreliable news sources; (2) collection of candidate users that have shared at least one of the posts collected during the first step; (3) assignment of users to the reliable and unreliable categories.

### 3.4.1 Collecting Posts from Unreliable and Reliable News Sources

To identify users that have shared content from a particular news source, we first need to collect posts from reliable and unreliable news sources. For that purpose, we use a widely-used and publicly available list of *English news media* Twitter accounts provided by Volkova et al. (2017); Glenski et al. (2018b), which contains 424 English news media sources categorized in *unreliable* (satire, propaganda, hoax, clickbait) and *reliable*, following Rubin et al. (2015). For each news source, we retrieve the timeline (most recent 3,200 tweets) using the Twitter public

API. We then filter out any retweets to ensure that we can collect only original posts from each Twitter account.

In this list, unreliable news sources (e.g., Infowars, Disclose.tv) have been annotated by digital journalism organisations (e.g. PropOrNot, fakenewswatch.com, etc.), while the reliable news media accounts (e.g., BBC, Reuters) have all been verified on Twitter and used in Glenski et al. (2018b). Since satire news sources (e.g., The Onion, Clickhole) have humorous purposes (no desire to deliberately deceive (Rashkin et al., 2017)), we exclude them as in Glenski et al. (2018a) resulting into 251 trusted and 159 unreliable sources. Note that the list does not exhaustively cover all the available sources but it is a representative sample for the purpose of our experiments. We also use the characterization of an entire news source as reliable/unreliable following Rashkin et al. (2017); Volkova et al. (2017) and not individual posts.

### 3.4.2 Collecting Candidate Users

We retrieve an initial set of approximately 15,000 candidate users by looking into the most recent user accounts who have retweeted at least an original tweet from each news source. Due to the public Twitter API limits, we do not have access to user likes of news items. Based on the user profile information, we filter out users with more than 3,200 tweets due to the Twitter public API limits, since we need access to the entire timeline to decide the category the user belongs to (see section Labeling Users). For the remaining users, we collect their entire timeline (up to 3,200 tweets) and we filter-out any user with majority of non-English tweets (i.e., tweets labelled as ‘en’ or ‘en-gb’ by Twitter). Then for each user, we count the number of retweets from reliable and unreliable news sources respectively. Subsequently, we remove all user’s retweets (including tweets containing RT) and we keep only the tweets up to the first retweet of a news source for each user. Moreover, we only keep users with more than 10 original tweets.

### 3.4.3 Labeling Users

Our classification task is defined as the early detection of users posting unreliable news sources before they actually do it for the first time. Therefore, candidate users are assigned into two



|               | <b>Unreliable</b> | <b>Reliable</b> |
|---------------|-------------------|-----------------|
| <b>Users</b>  |                   |                 |
| Total         | 2,798             | 3,468           |
| <b>Tweets</b> |                   |                 |
| Min           | 10                | 10              |
| Max           | 2,600             | 2,613           |
| Mean          | 172               | 252             |
| Median        | 71                | 116             |
| Total         | 481,199           | 875,281         |
| <b>Tokens</b> |                   |                 |
| Min           | 17                | 10              |
| Max           | 37,576            | 251,030         |
| Mean          | 1,796             | 2,779           |
| Median        | 657               | 1,150           |
| Total         | 5,024,341         | 9,713,595       |

**Table 3.1:** Dataset statistics.

categories (*Unreliable*, *Reliable*):

- **Unreliable.** Users that have *retweeted unreliable sources at least three times* (to ensure that this is a consistent behaviour) including the case *when a user has shared both reliable and unreliable sources* (modeling the ratio of unreliable/reliable it is out of the scope of early detection) are assigned to the unreliable class.
- **Reliable.** Users that have retweeted *only* reliable news sources are assigned to the reliable category.

Given that Twitter users can also share shortened URLs from unreliable news websites (e.g., [www.infowars.com](http://www.infowars.com)), we collect and expand all shortened URLs (e.g., 'https://t.co/example') extracted from the posts of users labeled as reliable. We then remove all users who have shared any URLs from unreliable news websites. Our data collection process yielded a set of 6,266 users (3,468 and 2,798 for reliable and unreliable respectively) with a total of 1,356,480 tweets (see Table 3.1).

### 3.4.4 Text Preprocessing

We pre-process all tweets from all users by first lowercasing text and then tokenizing. Furthermore, we remove any stop words<sup>3</sup> and replace all URLs and @-mentions with URL and USR markers respectively. See Table 3.1 for token statistics per user.

### 3.4.5 Ethics

Previous work on the study of who spreads misinformation in social networks has used data collected through survey questionnaires (i.e., self-report data) and trace data (i.e., user-generated content) (Talwar et al., 2019; Chen et al., 2015; Shu et al., 2019; Hopp et al., 2020). We employ similar standard practices on studying social media user behavior. Our work has received approval from the University of Sheffield Research Ethics Committee (Ref. No 025470) and complies with Twitter data policy for research.<sup>4</sup> To ensure anonymization, we encrypt and separate labels from user data. The two files are linked by an anonymous ID as suggested by Benton et al. (2017). Note that we will not share the data for non-research purposes.

## 3.5 Methods

### 3.5.1 SVM

We use Support Vector Machines (SVM) with an Radial Basis Function (RBF) kernel (Joachims, 2002) for all of our feature-based models which can be considered as baselines. We extract three types of language features: (1) Bag-Of-Words (BOW); (2) topics; and (3) Linguistic Inquiry and Word Count (LIWC), following a similar approach to recent work in computational social science (Rashkin et al., 2017; Pérez-Rosas et al., 2018; Zhang et al., 2018; Holgate et al., 2018):

---

<sup>3</sup>We use the NLTK English stopwords list.

<sup>4</sup><https://developer.twitter.com/en/developer-terms/agreement-and-policy>

- We use **BOW** to represent each user as a TF-IDF weighted distribution over a 20,000 sized vocabulary with the most frequent unigrams, bigrams and trigrams. We only consider n-grams appearing in more than five and no more than 40% of the total users.
- We also represent each user over a distribution of 200 generic word clusters (**Topics**<sup>5</sup>) computed on a Twitter corpus and provided by Preoțiuc-Pietro et al. (2015) for unveiling the thematic subjects that the users discuss.
- We finally represent each user over a distribution of 93 psycho-linguistic categories represented by lists of words provided by the Linguistic Inquiry and Word Count (**LIWC**) 2015 dictionary (Pennebaker et al., 2001).

We then train SVMs using the three types of features: **SVM-BOW**, **SVM-Topics** and **SVM-LIWC** individually and in combination (**SVM-All**).

### 3.5.2 Avg-EMB

As our first neural model, we use a simple feed forward network (Avg-EMB) which takes as input the concatenation of all the tokenized tweets of a user. Words from users' tweets are first mapped into embeddings to compute an average embedding which represents the textual content posted by a user. Subsequently, the average embedding is passed to the output layer with a sigmoid activation function for binary classification.

### 3.5.3 BiGRU-ATT

Furthermore, we train a bidirectional Gated Recurrent Unit (Cho et al., 2014) with self-attention (Xu et al., 2015) (BiGRU-ATT).<sup>6</sup> The input is first mapped into word embeddings which are then passed through a BiGRU layer. A user content embedding is computed as the sum of the resulting context-aware embeddings weighted by the self-attention scores. The user content embedding is then passed to the output sigmoid layer.

---

<sup>5</sup>Early experimentation with topic models did not yield highly coherent topics.

<sup>6</sup>We also tested a Hierarchical Attention Network (Yang et al., 2016) achieving similar performance to BiGRU-ATT.

### 3.5.4 ULMFiT

The Universal Language Model Fine-tuning (ULMFiT) (Howard and Ruder, 2018) is a transfer learning approach that uses a Average-Stochastic Gradient Descent Weight-Dropped Long Short-Term Memory (AWD-LSTM) (Merity et al., 2017) encoder pre-trained on a large corpus using a language modelling objective. Following the standard adaptation process of ULMFiT, we first fine-tune the AWD-LSTM on language modelling using our dataset, and then we adapt the classifier into our binary task by replacing the output layer. We finally fine-tune ULMFiT using the gradual unfreezing method proposed in Howard and Ruder (2018).

### 3.5.5 T-BERT and H-BERT

Deep Bidirectional Transformers (BERT) (Devlin et al., 2018) is a state-of-the-art masked language model based on Transformer networks (Vaswani et al., 2017) pre-trained on large corpora, i.e., Books Corpus and English Wikipedia. Given the maximum input sequence length of BERT is 512, we first use a truncated version of BERT (T-BERT), which only takes the first 512 word pieces of each user as input. For our specific binary classification task, we add a fully-connected layer with a sigmoid activation on top of the user contextualized embedding obtained by passing input through BERT.

In order to take into account all the available textual information, we also employ a hierarchical version of BERT (H-BERT) since the majority of users' concatenated tweets are longer than 512 tokens. We split the input sequence (i.e., the collection of users' historical tweets) into  $N = L / 510$  chunks<sup>7</sup> of a fixed length e.g., 512 including task special tokens (e.g., [CLS] tokens for BERT). For each of these word chunks, we obtain the representation of the [CLS] token from the fine-tuned BERT on our dataset. We then stack these segment-level representations into a sequence, which serves as input to a mean pooling layer. We finally add a standard linear layer with sigmoid activation.

---

<sup>7</sup>where  $N$  denotes the number of chunks and  $L$  the number of tokens.

### 3.5.6 T-XLNet and H-XLNet

XLNet is a generalized autoregressive language model (Yang et al., 2019) similar to BERT which has achieved state-of-the-art performance in multiple NLP tasks. XLNet uses a perturbed language model objective instead of masked language model used in BERT. Similar to BERT-based models, we employ both truncated and hierarchical versions of XLNet (i.e., T-XLNet and H-XLNet respectively) adapting them to our task using sigmoid output layers.

## 3.6 Results

### 3.6.1 Experimental Setup

We split our data into train (70%), development (10%), and test (20%) sets. The development set is used for tuning the hyper-parameters of the models.

Following a similar hyper-parameter tuning method to recent work in computational social science (Vempala and Preotiuc-Pietro, 2019; Maronikolakis et al., 2020), we tune the penalty parameter  $C \in \{10, 1e2, 1e3, 1e4, 1e5\}$  and  $n$ -gram range  $\in \{(1,1), (1,2), (1,3), (1,4)\}$  of the SVMs, setting  $C = 1e4$  and  $n$ -gram range = (1, 2). For BiGRU-ATT, we tune the GRU *hidden unit size*  $\in \{50, 75, 100\}$  and *dropout rate*  $\in \{0.2, 0.5\}$  observing that 50 and 0.5 perform best respectively. For Ave-EMB and BiGRU-ATT, we use Glove embeddings (Pennington et al., 2014) pre-trained on Twitter ( $d = 200$ ). For all neural models, we use binary cross-entropy loss and the Adam optimizer (Kingma and Ba, 2015) with default learning rate 0.001 (except of the fine-tuning of ULMFiT, BERT and XLNet models where we use the original learning rates). We use a batch size of 8 for BERT, XLNet models and 64 for the rest of the neural models respectively.

We repeat the training and testing of each model three times by setting different random seeds and finally report the averaged macro precision, recall and F1-score. All dataset splits and random seeds will be provided for reproducibility.

### 3.6.2 Prediction Results

Table 3.2 presents the results of the SVM with all the feature combinations (BOW, Topics, LIWC and All) and the neural models.

In general, neural models achieve higher performance compared to feature-based models (SVM). Specifically, the T-BERT model achieves the highest F1 score overall (79.7) surpassing all the feature-based models as well as other neural network-based methods. This demonstrates that neural models can automatically unveil (non-linear) relationships between a user’s generated textual content (i.e., language use) in the data and the prevalence of that user retweeting from reliable or unreliable news sources in the future.

The simpler neural network model, Avg-EMB achieves a lower F1 score (75.5) compared to the other neural models, i.e., BiGRU-ATT, BERT, XLNet and ULMFiT. This happens because the latter have more complex architectures and can effectively capture the relations between inputs and labels while the former ignores word order. Furthermore, ULMFiT, BERT and XLNet models have been pre-trained on large external corpora so they can leverage this extra information to generalize better. Finally, we do not notice considerable differences in performance between the truncated and hierarchical versions of the transformer-based models (BERT and XLNet) suggesting that a small amount of user generated content is enough for accurately predicting the correct user class.

Best single performing feature-based model is SVM-ALL (75.9). Moreover, SVM with BOW, Topics and LIWC achieve lower performance (75.8, 71.2 and 69.6 respectively).

### 3.6.3 Error Analysis

We performed an error analysis on the predictions of our best model, T-BERT. We notice that users in the unreliable class who are classified as reliable are those who repost from both reliable and unreliable sources. These users have an average of 40 future retweets from reliable news sources which is higher than the average number (31 retweets) in the entire dataset. Therefore, it is likely that such users use similar topics of discussion with reliable users. On the other hand, there is a of total 454 unreliable users who have no retweets from reliable sources in our dataset, interestingly, only four of them are classified wrongly. We also observe

| Model                | P               | R               | F1              |
|----------------------|-----------------|-----------------|-----------------|
| <b>Baselines</b>     |                 |                 |                 |
| <b>SVM</b>           |                 |                 |                 |
| BOW                  | 75.8 ± 0.0      | 75.9 ± 0.0      | 75.8 ± 0.0      |
| Topics               | 71.8 ± 0.0      | 71.1 ± 0.0      | 71.2 ± 0.0      |
| LIWC                 | 69.8 ± 0.0      | 69.6 ± 0.0      | 69.6 ± 0.0      |
| All                  | 75.9 ± 0.0      | 75.8 ± 0.0      | 75.9 ± 0.0      |
| <b>Neural Models</b> |                 |                 |                 |
| Avg-EMB              | 76.3 ± 1.2      | 75.3 ± 1.1      | 75.5 ± 1.2      |
| BiGRU-ATT            | 78.0 ± 0.7      | 77.8 ± 0.3      | 77.8 ± 0.5      |
| ULMFiT               | 77.9 ± 0.2      | 77.2 ± 0.6      | 77.4 ± 0.5      |
| T-BERT               | <b>79.7±0.2</b> | <b>79.8±0.1</b> | <b>79.7±0.1</b> |
| H-BERT               | 79.5 ± 0.4      | 78.7 ± 0.6      | 78.9 ± 0.5      |
| T-XLNet              | 79.6 ± 0.3      | 79.8±0.2        | 79.7 ± 0.3      |
| H-XLNet              | 79.3 ± 0.3      | 78.9 ± 0.4      | 79.0 ± 0.3      |

**Table 3.2:** Macro precision - **P**, recall - **R** and F1-score - **F1** (mean ± standard deviation over three runs) for predicting whether a Twitter user belongs to the reliable or unreliable class.

that it is harder for our model to classify correctly reliable users when they have only posted a small number of original tweets (e.g., 10-60).

### 3.6.4 Linguistic Analysis

Finally, we perform a linguistic feature analysis to uncover the differences in language use between users in the two classes, i.e. reliable and unreliable. For that purpose, we apply univariate Pearson’s correlation test to identify which text features (i.e., BOW, Topics and LIWC) are high correlated with each class following Schwartz et al. (2013). Tables 3, 4 & 5 display the top-10 n-grams, LIWC categories and Topics (represented by the most central words as in Preoțiuc-Pietro et al. (2015)) respectively. All Pearson correlations ( $r$ ) presented in tables are statistically significant ( $p < 0.001$ ).

| n-grams    |       |          |       |
|------------|-------|----------|-------|
| Unreliable | r     | Reliable | r     |
| war        | 0.140 | school   | 0.150 |
| media      | 0.137 | gonna    | 0.133 |
| government | 0.135 | myself   | 0.133 |
| truth      | 0.133 | wanna    | 0.131 |
| israel     | 0.123 | feel     | 0.131 |
| liberal    | 0.122 | excited  | 0.131 |
| msm        | 0.121 | mom      | 0.127 |
| liberals   | 0.113 | mood     | 0.122 |
| muslim     | 0.113 | okay     | 0.121 |
| islam      | 0.112 | rn       | 0.121 |

**Table 3.3:** N-grams associated with unreliable and reliable categories sorted by Pearson’s correlation ( $r$ ) between their normalized frequency and the labels ( $p < .001$ ).

## BOW

Table 3.3 shows the ten most correlated **BOW** features with each class. We observe that users reposting unreliable news sources in the future are more prevalent in tweeting about politics (note that we exclude user retweets in our study). For example, they use words related to the established political elite (e.g., *liberal*, *government*, *media*, *msm*<sup>8</sup>) and Middle East politics (e.g., *islam*, *israel*). This may be partially explained by studies which find that people who are more ideologically polarized might be more receptive to disinformation (Marwick, 2018) and engage more with politics on social media (Preotiuc-Pietro et al., 2017). Users using language similar to the language used by unreliable and hyperpartisan sources can be explained by the fact that these users might already consume news from unreliable sources but they have not reposted any of them yet (Potthast et al., 2018; Pennycook et al., 2018).

Users belonging in the reliable news sources category use words related to self-disclosure and extraversion such as personal feelings and emotions (e.g., *mood*, *wanna*, *gonna*, *i’ll*, *excited*). Moreover, words such as *birthday*, *okay* denote more frequent interaction with other users, perhaps friends.

<sup>8</sup>MSM is an Internet acronym for “mainstream media”.



| Topics |                                                                                                                 |       |
|--------|-----------------------------------------------------------------------------------------------------------------|-------|
| #      | Unreliable                                                                                                      | r     |
| 175    | religious, colonialism, christianity, judaism, persecution, fascism, marxism, nationalism, communism, apartheid | 0.244 |
| 118    | #libya, libyan, libya's, loyalists, palestinians, iran's, gaddafi's, al-qaeda, libya, repression                | 0.21  |
| 138    | republican, democratic, gop, congressional, judiciary, hearings, abolishing, oppose, legislation, governors     | 0.196 |
| 106    | allegations, prosecution, indictment, alleged, convicted, allegation, alleges, accused, charges, extortion      | 0.184 |
| 18     | harper, congressman, abbot, mccain, cain, turnbull, spokesman, corbett, president, chairman                     | 0.183 |
| 179    | gov't, govt, government, government's, govt's, privatisation, bureaucrats, draconian, safeguards, bureaucracy   | 0.173 |
| 160    | latvian, bulgarian, croatian, turkish, malaysian, estonia, hungarian, basque, cypriot, romanian                 | 0.166 |
| 196    | govern, compromises, ultimately, unwilling, distrust, thereby, establish, assert, willingness, inaction         | 0.165 |
| 78     | self-serving, hypocritical, moronic, idiocy, bigoted, blatant, reactionary, dismissive, uninformed, pandering   | 0.149 |
| 176    | armed, gunmen, killings, suspected, bombings, police, detained, authorities, policemen, arresting               | 0.148 |
| #      | Reliable                                                                                                        | r     |
| 120    | physics, sociology, maths, biology, math, chem, calculus, geog, worksheet, worksheet                            | 0.143 |
| 101    | 4hours, #naptime, #sleepy, 4hrs, 6hrs, #exhausted, #tired, 3hours, 3hrs, #sotired                               | 0.14  |
| 2      | tomorrows, tmw, tomorrow, tomor, tomrw, #hopefully, 4day, #tgif, arvo, tmrw                                     | 0.135 |
| 53     | giggling, giggled, hysterically, squealing, sobbing, moaned, gasped, screaming, awkwardly, angrily              | 0.125 |
| 1      | tights, cardigan, slacks, sleeveless, sweater, plaid, skirt, v-neck, leggings, skinnies                         | 0.119 |
| 65     | #foodtweets, #foodtweet, yummm, yummmm, #nomnom, spaghetti, sandwich, #yum, yummmmmm, #yummy                    | 0.119 |
| 9      | horribly, dreadfully, slighty, terribly, hungover, hungover, majorly, majorly, horrid                           | 0.118 |
| 27     | 1:30, 6:15, 3:30, 8:45, 7:45, 4:30, 8:15, 9:45, 5:30, 2:30                                                      | 0.116 |
| 166    | chocolate, strawberry, choc, toffee, cinnamon, almond, parfait, butterscotch, choco, strawberries               | 0.112 |
| 33     | b'day, birthday, birthdaaaay, birthdayyyyy, b-day, birthday, birthdayyyy, birthdaaay, bday, birfday             | 0.102 |

**Table 3.4:** Topics associated with unreliable and reliable categories sorted by Pearson's correlation ( $r$ ) between the topic normalized frequency and the labels. All correlations are significant ( $p < .001$ , t-test, Simes corrected).

## Topics

Table 3.4 shows the ten most correlated **topics** with each class. Topics related to politics such as political ideology (#138, #175), government (#179) and justice (#106) are correlated with users that will propagate unreliable sources, aligned with the n-grams analysis. We also observe a high correlation of such users with the topic related to impolite personal characterizations (#78). This corroborate results of a recent study that showed political incivility on Twitter is correlated to political polarization (Vargo and Hopp, 2017).

Users who will repost reliable sources discuss topics related to their day-to-day life such as education (#120), food (#65 and #166) and fashion (#1). Some topic words (e.g., sleep, exhausted and tired from #101) reveal that users emotional or physical states caused from work or study. In other words, these users tend to share more frequently information about their daily life, time and schedule (#101, #2 and #27).

| LIWC                     |       |                      |       |
|--------------------------|-------|----------------------|-------|
| Unreliable               | r     | Reliable             | r     |
| <i>Analytic</i>          | 0.242 | <i>Informal</i>      | 0.200 |
| <i>Power</i>             | 0.203 | <i>NetSpeak</i>      | 0.192 |
| <i>Words&gt;6letters</i> | 0.184 | Word Count           | 0.129 |
| <i>Space</i>             | 0.153 | <i>Authentic</i>     | 0.093 |
| <i>Drives</i>            | 0.140 | <i>Ingest</i>        | 0.087 |
| <i>Risk</i>              | 0.125 | <i>Bio</i>           | 0.080 |
| <i>Religion</i>          | 0.125 | <i>Feel</i>          | 0.073 |
| <i>Money</i>             | 0.117 | <i>WordsPerSent.</i> | 0.071 |
| <i>Death</i>             | 0.105 | <i>Leisure</i>       | 0.067 |
| <i>Neg.Emotion</i>       | 0.097 | <i>Time</i>          | 0.064 |

**Table 3.5:** LIWC features associated with unreliable and reliable categories sorted by Pearson’s correlation ( $r$ ) between the normalized frequency and the labels ( $p < .001$ ).

## LIWC

Table 3.5 shows the ten most correlated **LIWC** categories with each class. LIWC categories such as *Power* and *Drives* are more prevalent in users that will share unreliable sources. We also observe the difference in using casual language, e.g., *Netspeak* and *Informal* categories which are more often used by users that will share trusted sources.

## 3.7 Conclusions

We have presented a new study on the early detection of users reposting unreliable news sources. We have created a new dataset with users labeled into the two categories, i.e., reliable and unreliable. For this binary classification task, a Transformer-based pretrained model (i.e., BERT) achieves up to 79.7 macro F1. Finally, our linguistic feature analysis unveiled the main characteristics and differences between language features (i.e., BOW, Topics and LIWC) in the two groups of users.

In the future, we plan to extend this work by performing a fine-grained classification into hoax, propaganda and clickbait (Glenski et al., 2018a); and explore whether language and social network information are complementary. We also plan to extend the current work by

incorporating network information and graph features (e.g., followers and retweeters). Since the list of reliable news sources is not complete, we intend to enrich the current list of news outlets in a multi-lingual setting.

---

## BIBLIOGRAPHY

- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *EMNLP*, pages 3528–3539.
- W Lance Bennett. 2008. Changing citizenship in the digital age. *Civic life online: Learning how digital media can engage youth*, 1(1-24).
- Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Workshop on Ethics in Natural Language Processing*, pages 94–102.
- Leticia Bode and Emily K. Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.
- Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164.
- Jovan Byford. 2011. *Conspiracy theories: A critical introduction*. Springer.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *WWW*, pages 675–684.

- Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Why students share misinformation on social media: Motivation, gender, and study-level differences. *The journal of academic librarianship*, 41(5):583–592.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.
- Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours. In *SemEval*, pages 69–76.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949.
- Sven Engesser and Edda Humprecht. 2015. Frequency or skillfulness: How professional news media use Twitter in five Western countries. *Journalism studies*, 16(4):513–529.
- M. Glenski, T. Weninger, and S. Volkova. 2018a. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, pages 1–12.

- Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018b. Identifying and understanding user reactions to deceptive and trusted social news sources. In *ACL*, pages 176–181.
- Ted Goertzel. 1994. Belief in conspiracy theories. *Political Psychology*, pages 731–742.
- Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6):815–824.
- Eric Holgate, Isabel Cachola, Daniel PreoȚiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? Analyzing and inferring the intentions of vulgar expressions. In *EMNLP*, pages 4405–4414.
- Toby Hopp, Patrick Ferrucci, and Chris J Vargo. 2020. Why do people share ideologically extreme, false, and misleading content on social media? a self-report and trace data-based analysis of countermedia content dissemination on facebook and twitter. *Human Communication Research*, 46(4):357–384.
- Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *WWW*, pages 235–238.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.
- Philip N Howard and Bence Kollanyi. 2016. Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*.
- Edda Humprecht. 2018. Where ‘fake news’ flourishes: a comparison across four western democracies. *Information, Communication & Society*, pages 1–16.
- Thorsten Joachims. 2002. *Learning to Classify Text using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers.
- Bente Kalsnes and Anders Olof Larsson. 2018. Understanding news sharing across social media: Detailing distribution on Facebook and Twitter. *Journalism Studies*, 19(11):1669–1688.
- D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild,

- Michael Schudson, Steven A Sloman, Cass R Sunstein, Emily A Thorson, Duncan J Watts, and Jonathan L Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. *arXiv preprint arXiv:2004.13878*.
- Alice E Marwick. 2018. Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technology Review*.
- Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.
- Todor Mihaylov and Preslav Nakov. 2016. Hunting for troll comments in news community forums. In *ACL*, pages 399–405.
- Archita Pathak and Rohini K Srihari. 2019. Breaking! presenting fake news corpus for automated fact checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543.
- Gordon Pennycook, Tyrone Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology*.
- Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *COLING*, pages 3391–3401.
- Kashyap Papat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *CIKM*, pages 2173–2178.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, pages 231–240.

- Daniel Preoțiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *ACL*, pages 1754–1764.
- Daniel Preoțiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of Twitter users. In *ACL*, volume 1, pages 729–740.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937.
- Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439.
- Shalini Talwar, Amandeep Dhir, Puneet Kaur, Nida Zafar, and Melfi Alrasheedy. 2019. Why do people share fake news? associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, 51:72–82.
- Edson C Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining “fake news”: A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.
- James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *COLING*, pages 3346–3359.
- Chris J. Vargo and Toby Hopp. 2017. Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on Twitter: A Congressional district-level analysis. *Social Science Computer Review*, 35(1):10–32.



- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Alakananda Vempala and Daniel Preoțiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *ACL*, volume 2, pages 647–653.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.
- Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *ACL*, pages 1350–1361.
- Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

*for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

Homero Gil de Zúñiga, Nakwon Jung, and Sebastián Valenzuela. 2012. Social media use for news and individuals’ social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, 17(3):319–336.

# IDENTIFYING MISINFORMATION DEBUNKERS

## Identifying and Characterizing Active Citizens who Refute Misinformation in Social Media

*Yida Mu<sup>1</sup>, Pu Niu<sup>2</sup> and Nikolaos Aletras<sup>1</sup>*

<sup>1</sup> Department of Computer Science, The University of Sheffield

<sup>2</sup> Central China Normal University

### Abstract

The phenomenon of misinformation spreading in social media has developed a new form of active citizens who focus on tackling the problem by refuting posts that might contain misinformation. Automatically identifying and characterizing the behavior of such active citizens in social media is an important task in computational social science for complementing studies in misinformation analysis. In this paper, we study this task across different social media platforms (i.e., Twitter and Weibo) and languages (i.e., English and Chinese) for the first time. To this end, (1) we develop and make publicly available a new dataset of Weibo

users mapped into one of the two categories (i.e., misinformation posters or active citizens); (2) we evaluate a battery of supervised models on our new Weibo dataset and an existing Twitter dataset which we repurpose for the task; and (3) we present an extensive analysis of the differences in language use between the two user categories

## 4.1 Introduction

The diffusion of misinformation in social media has far-reaching implications on society (e.g., political polarization, election manipulation). Misinformation propagates faster than credible information among users in social media (Vosoughi et al., 2018), whilst coming across a non-factual story once, it is enough to increase later perception of its accuracy (Pennycook et al., 2018).

To combat misinformation, several fact-checking platforms (e.g., Snopes<sup>1</sup> and the Weibo Rumour Reporting Platform<sup>2</sup>) have been created with the aim to provide evidence on why particular claims are not factually correct (i.e., debunking or fact checking). This has subsequently resulted in a new form of active citizenship with large number of social media users directly reporting suspicious posts or actively sharing posts with evidence to refute claims made by other users which are likely to contain misinformation. Other examples of active citizenship include civic engagement, political activism, community help, volunteering and neighborhood associations Johansson and Hvinden (2007). In scope of social media, we consider users who actively debunk misinformation as active citizens since they work to make a difference (i.e., debunking misinformation) in online communities (e.g., social media platforms).

Automatically identifying and analyzing the behavior of active citizens in social networks is important for diffusion of misinformation prevention at the user level (Singh et al., 2020; Rangel et al., 2020; Giachanou et al., 2020). It can be used by (1) social media platforms (e.g., Facebook<sup>3</sup>) to track suspicious posts at an early stage (e.g., reports of suspicious posts

---

<sup>1</sup><https://www.snopes.com/>

<sup>2</sup><http://service.account.weibo.com>

<sup>3</sup><https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works>

from end users); (2) psychologists to complement studies on analyzing personality traits of those users who spread or debunk unreliable posts (Pennycook and Rand, 2019, 2020); and (3) fact-checking websites to develop personalized recommendation systems to assist active citizens in correcting suspicious posts (Vo and Lee, 2018; You et al., 2019; Karmakharm et al., 2019).

Previous work on automatically identifying active citizens who refute misinformation has focused only on a single social media platform (i.e., Twitter) using a relatively small dataset (e.g., with only 454 users) consisting of users tweeting in English (Giachanou et al., 2020). In addition, most of these previous studies have used supervised machine learning models with features extracted from text (e.g., bag-of-words, topics, psycho-linguistic information) and task-specific neural models trained from scratch without exploring state-of-the-art pretrained large language models (Devlin et al., 2019).

The purpose of this paper is to study the differences in language use between the two user categories: (i) users who share suspicious posts (i.e., misinformation posters) and (ii) users who actively debunk misinformation (i.e., active citizens) To this end, we pose the following two research questions:

- Can we automatically identify active citizens and misinformation posters based on their language use in social media?
- Can we characterize the linguistic differentiation between the two groups of users?

To answer these research questions, we make several main contributions:

- We develop a new large publicly available dataset from Weibo consisting of 48,334 users labeled either as active citizens or misinformation posters;
- We repurpose<sup>4</sup> an existing dataset developed by Vo and Lee (2019) to model the task of predicting active citizens and misinformation posters on Twitter;
- We evaluate several state-of-the-art pretrained neural language models adapted to the task. Due to the fact that the user text can be very long (e.g., thousands of posts), we

---

<sup>4</sup>The dataset has been used in computational misinformation analysis for automatic generation of fact-checking tweets and recommender systems for fact-checking (Vo and Lee, 2018, 2020a).

develop efficient hierarchical transformer-based networks achieving up to 85.1 and 80.2 macro F1 scores on Weibo and Twitter respectively;

- We finally provide an extensive linguistic analysis to highlight the differences in language use between active citizens and misinformation posters. We also provide a qualitative analysis of the limitations of our best models in predicting accurately whether a user is an active citizen or a misinformation poster.

## 4.2 Related Work

### 4.2.1 Misinformation: Definition and Types

Misinformation in social media can be generally defined as any false or incorrect information (e.g., fabricated news, rumors, etc.) that is published and propagated by end users (Wu et al., 2019). Particularly, unverified posts in social media (i.e., rumors) are defined as any item of circulated information whose veracity is yet to be verified at the time of posting (Zubiaga et al., 2018).

### 4.2.2 Misinformation Detection

Previous work on computational misinformation detection has focused on predicting the credibility or bias of news articles (Rashkin et al., 2017; Pérez-Rosas et al., 2018; Baly et al., 2020) and news sources (Baly et al., 2018; Aker et al., 2019). To prevent wide spread of misinformation, propagation-based detection methods are employed to enable early misinformation detection in social media (Zhou et al., 2019; Tian et al., 2020; Xia et al., 2020). In addition to using textual information, previous work on automated fact-checking also jointly use images and user profile information extracted from metadata associated with unreliable posts (Lee et al., 2011; Vo and Lee, 2020b).

Common automated fact-checking frameworks rely on external knowledge to determine the credibility of an unverified post, and they usually include one or more information retrieval models (Hanselowski et al., 2018; Nie et al., 2019). Pre-trained language models (e.g., BERT

(Devlin et al., 2019)) have recently been applied for fact-checking without using any external knowledge (Jobanputra, 2019; Lee et al., 2020; Williams et al., 2020) since they encapsulate factual knowledge from the massive amount of data used for pre-training, e.g., the English Wikipedia and Books Corpus (Devlin et al., 2019).

These misinformation detection tasks mentioned above are usually performed on existing datasets, e.g., Liar (Wang, 2017) and FEVER (Thorne et al., 2018) that typically contain claims associated with a label denoting if it is factual or not. These datasets do not usually include information on the user made the claim. Based on the publicly accessible Weibo Rumor Reporting Platform, Liu et al. (2015) developed a Weibo rumor dataset with 7,055 misinformation posters and 4,559 active citizens, however, many users no longer exist as these rumor cases were collected between 2011 and 2013. Similarly, Song et al. (2019) collected 3,387 rumor cases with their corresponding original publishers and 2,572,047 users who repost these fact-checked rumors.

### 4.2.3 User Behavior Analysis Related to Misinformation

Previous work in sociology and psychology have mostly used traditional survey-based methods to explore the personality traits (Pennycook et al., 2018; Talwar et al., 2019) and behavior (Altay et al., 2022; Tandoc Jr et al., 2020) of misinformation posters. A US-based survey shows that consumers of reliable mainstream news media are more likely to use fact-checking websites for checking the factuality of news claims (Robertson et al., 2020). Besides that, social media users are more inclined to trust debunked information that was shared by their network of friends rather than strangers (Margolin et al., 2018).

Existing work on using computational approaches to misinformation analysis has analyzed the difference of users' reactions (e.g., reply or retweet) to unreliable news sources and mainstream media as well as their characteristics (e.g., user demographic information) (Glenski et al., 2018a,b). To detect malicious accounts on social media, Addawood et al. (2019) and Luceri et al. (2020) have focused on identifying political trolls that diffuse misinformation and politically biased information during the US 2016 democratic election. Mu and Aletras (2020) and Rangel et al. (2020) focus on identifying Twitter users who diffuse unreliable news stories either on post level or news media level.

Vo and Lee (2019) uncover the positive impact of misinformation active citizens on preventing the spread of false news. They found that around 7% of the original tweets (among 64k tweets) are irretrievable within five months of being debunked due to the suspension of the Twitter account and the deletion of tweets. This suggests that developing downstream tools (e.g., automatic generation of personalized fact-checking tweets (Vo and Lee, 2020b,a)) can encourage misinformation active citizens to actively prevent the spread of misinformation and help social media platforms to suspend malicious users. Moreover, active citizens who are active in sharing fact-checking information are found to use less informal language including swear words and are more likely to engage in debunking reliable posts about politics and fauxtography, i.e., photo edited images with misleading content (Vo and Lee, 2018). Giachanou et al. (2020) explore the impact of using linguistic features and user personality traits on identifying fake news posters and checkers based on 2,357 Twitter users.

Our work, on the other hand, is the first attempt to model active citizens who refute misinformation across different social media platforms and languages using our newly developed Weibo dataset and a substantially larger Twitter dataset than the one used by Giachanou et al. (2020) which has not been employed yet for this task.

## 4.3 Task and Data

### 4.3.1 Task Description

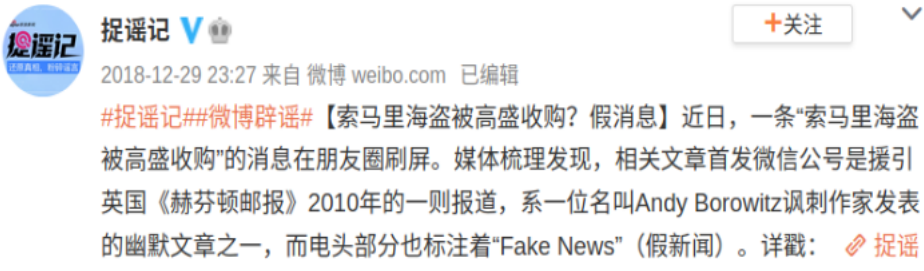
Following Giachanou et al. (2020), we frame a binary classification task aiming to distinguish between users that tend to diffuse misinformation (i.e., misinformation posters) and users who actively tend to refute such unreliable posts (i.e., active citizens) using language information. Note that one could also use a user’s social network information for modeling the predictive task but this is out of the paper’s scope because we are interested in analyzing differences in language use between the misinformation posters and active citizens across different social media platforms. Given a set of social media users, our task is to train a supervised classifier that can learn relations between users’ linguistic patterns (i.e., the collection of users’ original posts) and the corresponding class (i.e., misinformation posters or active citizens).



**Original Post:** 据报道，十一名索马里海盗近日在美国法庭招认其实他们是高盛工作的。海盗们招认，其实他们的工作性质很简单，就是强行攻击那些和高盛已经做空的交易有关的船只，以从中获利 ...

Eleven Somali pirates have reportedly confessed in a US court recently that they were in fact working for Goldman Sachs. The pirates confessed that the nature of their work was in fact very simple, which was to forcibly attack ships that were associated with trades already shorted by Goldman Sachs in order to profit from them...

**Fact-checking Report:** #Hashtag# URL\*  
 URL\*: <https://weibo.com/6590980486/H9wlQCgMu>



The content of the URL suggests that the original post was a rumour that had been already debunked.

**Figure 4.1:** An example of a pair of misinformation poster (in the blue box) and debunker on Weibo with the corresponding fact-checking information (in the red box)

### 4.3.2 Weibo Data

In 2012, Weibo developed an official community management center to receive reports from end users for all kinds of malicious posts including misinformation, hate speech and content plagiarism. To combat misinformation, Weibo provides a fact-checking platform for its users to report any suspicious misinformation posts which are then subsequently fact-checked officially by the platform. These Weibo posts are eventually labeled as *true* or *false*. Alternatively, a post may also remain *unverified* until it has been fact-checked. In case that a post has been deemed to be labeled as *false*, it is also accompanied with debunking information refuting the claim. Figure 4.1 shows an example including the original publisher (e.g., misinformation poster), active citizen and fact-checking information on Weibo.



**Figure 4.2:** An example of a pair of misinformation poster and debunker on Twitter with the corresponding fact-checking information

**Collecting misinformation posters and active citizens** For the purpose of our experiments, we collect 38,712 debunked cases (between 2012 and 2020) from the official Weibo Community Management Center<sup>5</sup> following a similar data collection approach as previous work (Wu et al., 2015; Ma et al., 2016). We keep only those posts that have been judged as *false* and then collect the corresponding poster (i.e., the original user that has published the post) and up to 20 recent active citizens (i.e., users that have officially reported that the post contains misinformation). Note that the official Weibo platform only allows access to the earliest 20 active citizens, even though some suspicious posts have been reported and refuted by more than 20 users. However, we notice that only in 709 cases there are more than 20 active citizens (i.e., less than 2% of all debunked posts). Given that some user accounts may have been suspended or become private, we remove those from the data.

Note that there is a difference in the definition of active citizens in Weibo and Twitter datasets. In Weibo, all active citizens are those who report misinformation to the Weibo Official Fact-checking Platform. In Twitter, active citizens are defined as ones who cite fact-checking URLs to refute misinformation. For consistency, we label all of these users as active citizens since they both **actively** try to refute misinformation. Weibo active citizens are not required to provide evidence or fact-checking URLs but they are free to report a post on a suspicion that it contains misinformation.

<sup>5</sup><https://service.account.weibo.com/?type=5&status=4>

**Table 4.1:** Data Statistics.

|              | Weibo   |         | Twitter |        |
|--------------|---------|---------|---------|--------|
| #Users       | Poster  | AC      | Poster  | AC     |
|              | 22,632  | 25,702  | 15,696  | 17,293 |
| #Posts       |         |         |         |        |
| Min          | 31      | 31      | 30      | 30     |
| Max          | 2,000   | 2,000   | 3,200   | 3,200  |
| Mean         | 596     | 576     | 2,932   | 2,824  |
| Total        | 13.5M   | 14.8M   | 46.0M   | 48.8M  |
| #Tokens/User |         |         |         |        |
| Min          | 127     | 126     | 663     | 674    |
| Max          | 104,947 | 104,801 | 81,052  | 81,028 |
| Mean         | 13,643  | 10,127  | 33,759  | 35,652 |
| Median       | 4705    | 3,730   | 32,726  | 35,312 |

**Collecting User Posts** We use the Weibo API<sup>6</sup> to collect up to 2,000 posts for each user since the median number of user posts are 968 and 855 for the two user categories (i.e., poster and debunker) respectively. We only consider users with more than 30 original posts and filter out all users who have both spread and debunk misinformation posts. After removing duplicate users, the final dataset contains 22,632 distinct posters and 25,702 distinct active citizens respectively.

### 4.3.3 Twitter Data

**Collecting misinformation posters and active citizens** To label Twitter users as misinformation posters or active citizens, we use a publicly available dataset with totally 73,203 users provided by Vo and Lee (2019).

Vo and Lee (2019) first use the Hoaxy System (Shao et al., 2016) to collect fact-checking tweets (FC-tweets) that contain links to relevant fact-checking information from PolitiFact and Snopes. These FC-tweets contain users who post URLs from fact-checking websites as credible evidence to refute misinformation posts in public conversations on Twitter (i.e., active citizens). They also contain the original users whose posts are debunked (i.e., misinformation

<sup>6</sup><https://open.weibo.com/development/businessdata>

posters). Figure 4.2 shows an example that contains the original post, fact-checking tweets and the corresponding debunking information from Snopes. According to Vo and Lee (2019), this dataset only contains misinformation active citizens who post English tweets with corresponding URLs linking to evidence (e.g., news article) that refutes a false claim. During data exploration, we observe that some Twitter users refer the fact-checking URLs to support the personal claims of the original posters, i.e., the original message has been proven correct. Therefore, we only consider users who share fact-checking URLs to refute tweets containing misinformation, i.e. those who are flagged as *False* by the corresponding fact-checking platform. In this way, we ensure that the selected active citizens have a clear intention to refute misinformation.

**Collecting User Posts** For each Twitter user, we use the Twitter Public API<sup>7</sup> to collect up to 3,200 tweets due to limits excluding any retweets. Moreover, we filter out users with less than 30 original tweets, users that may appear in both groups and keep users with a majority of English tweets (e.g., tweets that are labeled as ‘en’ or ‘en-gb’ by Twitter). As in the Weibo dataset, we also remove all users that both spread and debunk misinformation since we currently focus on the binary setting as in Giachanou et al. (2020) since we found that less than 10% of all users fall into this category (i.e., both spread and debunk misinformation) in the two datasets.<sup>8</sup> This process yielded 15,696 posters and 17,293 active citizens respectively. This is approximately 100 and 15 times larger than the datasets used in prior work (Giachanou et al., 2020; Rangel et al., 2020).

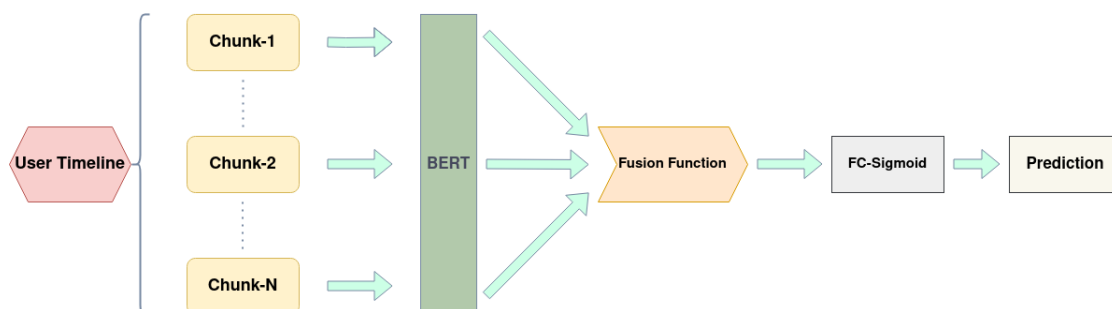
#### 4.3.4 Data Statistics and Topical Coverage

Table 4.1 shows the descriptive statistics of the data including the number of tweets and tokens obtained from Weibo and Twitter. Both datasets cover a broader range of topics rather than a narrow subset of political misinformation. The Twitter dataset provided by Vo and Lee (2019) is developed based on two popular fact-checking platforms (i.e., Snopes and PolitiFact) and covers more topics (e.g., medical and business) other than politics. However, we also notice that more than 50% of cases are related to politics given that one of the fact-checking platform (i.e., PolitiFact) is totally political oriented. As for the Weibo dataset, we collect all

---

<sup>7</sup><https://developer.twitter.com/en/docs>

<sup>8</sup>We leave this multi-label classification task for future work



**Figure 4.3:** Overview of the hierarchical transformer architecture used in our work.  $N = L / 510$ , where  $N$  denotes the number of chunks and  $L$  denotes the number of tokens. The Fusion Function denotes how we fuse the chunk-level information into the global representation, i.e., using Max Pooling, Mean Pooling and LSTM-Attention.

available cases of misinformation from the Weibo fact-checking platform from 2012 to 2020. According to Liu et al. (2015), more than 90% of debunked Weibo misinformation are related to politics, economics, pseudo-science and social life.

### 4.3.5 Text Pre-processing

**Weibo Posts** We pre-process the Weibo raw posts by converting all traditional Chinese words to simplified Chinese and then tokenizing using JIEBA, a Chinese text processing toolkit.<sup>9</sup> We keep all non-Chinese words since we notice some common English words (e.g., *python*, *hello*, and *world*) are including in the training corpus of both Chinese and English transformer models.

**Tweets** For the Twitter data, we follow a similar pre-processing pipeline<sup>10</sup> as in Nguyen et al. (2020). In brief, we pre-process the tweets by first lowercasing and then tokenizing using the *TweetTokenizer* from NLTK toolkit (Bird et al., 2009). Besides, we further normalize each Tweet by replacing each emoji,<sup>11</sup> URL and @-mention with special tokens, i.e., *single word token*, *@USER* and *HTTPURL* respectively.

<sup>9</sup><https://github.com/fxsjy/jieba>

<sup>10</sup><https://github.com/VinAIRResearch/BERTtweet>

<sup>11</sup>We use the emoji Python package <https://pypi.org/project/emoji/>

## 4.4 Predictive Models

### 4.4.1 Baseline Models

**Logistic Regression** We apply logistic regression with L2 regularization penalty using Bag-of-Words (BOW) to represent each user as a TF-IDF weighted vector over a 10,000 sized vocabulary. We only keep n-grams appearing in more than 5 times and no more than 40% of the total users. We also represent each user over a distribution of manually created lexical categories represented by lists of words provided by the Linguistic Inquiry and Word Count (LIWC) 2015 dictionary (Pennebaker et al., 2001). LIWC has been extensively used in psycho-linguistic studies.

**BiLSTM-ATT** Furthermore, we train a Bidirectional Long Short Term Memory network (Hochreiter and Schmidhuber, 1997) with self-attention (BiLSTM-ATT) from scratch. The BiLSTM-ATT takes as input the users' historical posts, maps their words to pre-trained word embeddings and subsequently passes them through a bidirectional LSTM layer. A user embedding is computed as the sum of the resulting context-aware embeddings weighted by the self-attention scores. The user embedding is then passed to the linear prediction layer with sigmoid activation.

### 4.4.2 English Transformers

**BERT** Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) is a masked language model using a Transformer Network (Vaswani et al., 2017) pre-trained on the BooksCorpus and English Wikipedia.

**RoBERTa** The Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) is a BERT-style language model trained with fine-tuned hyper-parameters, larger batch size, and longer sequence compared to the original BERT. RoBERTa is trained on a combination of the original corpus used to train BERT and extra texts including English new articles and web content (Liu et al., 2019).

**Longformer** The Long-Document Transformer (Longformer) (Beltagy et al., 2020) is pretrained using the original RoBERTa checkpoint (Liu et al., 2019) with a sliding window attention pattern (same window size of 512 as RoBERTa) and extra positional embeddings to support a maximum length of 4,096 (from 512). Longformer can handle a longer text sequences and achieves state-of-the-art performance on long document downstream NLP tasks (Beltagy et al., 2020).

### 4.4.3 Chinese Transformers

**CBERT** For our Weibo prediction task, we first employ a Chinese BERT (CBERT) (Cui et al., 2019) model pretrained using a whole word masking strategy. CBERT is trained from the existing checkpoint of the Bert-Base-Chinese<sup>12</sup> model, which has the same structure (e.g., layers and parameters) as the original BERT.

**ERNIE** Enhanced Representation through Knowledge Integration (ERNIE) (Sun et al., 2019b) is designed to learn language representations using knowledge masking strategies, i.e., entity-level masking and phrase-level masking. ERNIE is trained on both formal (e.g., Baidu Baike, a platform similar to Wikipedia and Chinese news articles) and informal (e.g., posts from Tieba, an open discussion forum similar to Reddit) Chinese corpora.

### 4.4.4 Handling Long Text

Transformer-based models cannot handle long sequences in a single standard GPU card due to the large memory requirements. To deal with this issue in our datasets, we experiment with truncated and hierarchical methods for all transformer-based models.

**Truncated Transformers** Following similar work on modeling long texts by Sun et al. (2019a), we first employ a simple truncation method that cuts off the input to the maximum length supported by BERT and Longformer (e.g. 512 and 4,096 tokens). Following the same

---

<sup>12</sup>[https://storage.googleapis.com/bert\\_models/2018\\_11\\_03/chinese\\_L-12\\_H-768\\_A-12.zip](https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip)

strategy as in Devlin et al. (2019), we fine-tune transformer-based models by adding a linear prediction layer on the model special classification token, e.g., [CLS] of BERT and <s> of Longformer respectively.

**Hierarchical Transformers** Given that the majority of users’ concatenated posts contain more than 512 tokens, we also use a hierarchical transformer structure (Pappagari et al., 2019; Mulyar et al., 2019) (see Figure 4.3) for our long document classification task. We first split the input sequence (i.e., the collection of users’ original posts) into  $N = L / 510$  chunks<sup>13</sup> of a fixed length e.g., 512 including task special tokens (e.g., [CLS] tokens for BERT) and 4096 tokens for Longformer. For each of these word chunks, we obtain the representation of the [CLS] token from the fine-tuned BERT on our dataset. We then stack these segment-level representations into a sequence, which serves as input to a LSTM layer with a self-attention mechanism to learn a user-level representation. Finally, we add two fully connected layers with ReLU and sigmoid activations respectively on top of LSTM layer as in Pappagari et al. (2019).

Following Sun et al. (2019a), we also test two simple hierarchical methods by directly using max pooling and mean pooling to stack the [CLS] embeddings of all the chunks of each user into a document-level representation.

## 4.5 Experimental Setup

### 4.5.1 Hyper-parameters

For both Twitter and Weibo datasets, we train the models on the training set (70%) and tune the hyper-parameters on the validation set (10%). We tune the regularization parameter  $\alpha \in \{1e-1, 1e-2, 1e-3, 1e-4, 1e-5\}$  of the Logistic Regression, setting  $\alpha = 1e - 4$ . For BiLSTM-ATT, we use 200-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on 2-billion tweets and 300-dimensional Chinese Word Vectors<sup>14</sup> (Li et al., 2018) pre-trained on Weibo data. We tune the LSTM *hidden unit size*  $\in \{50, 100, 150\}$  and *dropout rate*  $\in \{0.2, 0.5\}$

<sup>13</sup>N denotes the number of chunks and L the number of tokens.

<sup>14</sup><https://github.com/Embedding/Chinese-Word-Vectors>



observing that 150 and 0.5 perform best respectively. For transformer-based models, we use BERT-Base-Uncased, RoBERTa-Base and Longformer-Base-4096 models fine-tuning them with learning rate  $lr \in \{5e-5, 3e-5, \text{ and } 2e-5\}$  as recommended in Devlin et al. (2019), setting  $lr = 2e - 5$ . For the Chinese language models, we use Chinese-BERT-WWM-EXT and ERNIE-1.0 models fine-tuning them with learning rate  $lr \in \{5e-5, 3e-5, \text{ and } 2e-5\}$  as in Cui et al. (2019), setting  $lr = 2e - 5$ . The maximum sequence length is set to 512 (including task special tokens, e.g., [CLS]) except the Longformer-Base-4096 which can handle a 4,096 input sequence length.

We use a batch size of 16 for all transformer-based models except the Longformer where we use batch size of 4. During training of the neural models, we use early stopping based on the validation loss and then use the saved checkpoint to compute the model predictive performance on the test set.

## 4.5.2 Implementation Details

We perform all the experiments on a single NVIDIA V100 graphics card. We use the implementation of transformer-based models available from the HuggingFace library (Wolf et al., 2019).

## 4.5.3 Evaluation Metrics

We run each model with the best hyper-parameter combination three times on the heldout set (20%) using different random seeds, and report the averaged macro precision, recall and F1 score (mean  $\pm$  standard deviation).

**Table 4.2:** Weibo binary classification results (mean  $\pm$  standard deviation). ‡ denotes that the HierERNIE LSTM performs significantly better than truncated ERNIE (t-test;  $p < .05$ )

| Model                            | P (macro)                        | R (macro)                        | F1 (macro)                       |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| <b>Baseline Models</b>           |                                  |                                  |                                  |
| LR-BOW                           | 82.8 $\pm$ 0.1                   | 82.7 $\pm$ 0.1                   | 82.7 $\pm$ 0.1                   |
| LR-LIWC                          | 75.7 $\pm$ 0.1                   | 73.7 $\pm$ 0.2                   | 74.1 $\pm$ 0.3                   |
| BiLSTM-ATT                       | 83.1 $\pm$ 0.2                   | 82.9 $\pm$ 0.1                   | 82.9 $\pm$ 0.1                   |
| <b>Truncated Transformers</b>    |                                  |                                  |                                  |
| CBERT                            | 79.9 $\pm$ 0.1                   | 79.9 $\pm$ 0.2                   | 79.9 $\pm$ 0.1                   |
| ERNIE                            | 80.4 $\pm$ 0.1                   | 80.5 $\pm$ 0.2                   | 80.3 $\pm$ 0.1                   |
| <b>Hierarchical Transformers</b> |                                  |                                  |                                  |
| HierCBERT Max Pool               | 83.3 $\pm$ 0.2                   | 83.2 $\pm$ 0.2                   | 83.2 $\pm$ 0.3                   |
| HierCBERT Mean Pool              | 84.0 $\pm$ 0.1                   | 84.0 $\pm$ 0.1                   | 84.0 $\pm$ 0.1                   |
| HierCBERT LSTM                   | 84.5 $\pm$ 0.1                   | 84.2 $\pm$ 0.2                   | 84.3 $\pm$ 0.1                   |
| HierERNIE Max Pool               | 83.8 $\pm$ 0.1                   | 83.7 $\pm$ 0.2                   | 83.7 $\pm$ 0.2                   |
| HierERNIE Mean Pool              | 84.2 $\pm$ 0.2                   | 84.3 $\pm$ 0.2                   | 84.2 $\pm$ 0.2                   |
| HierERNIE LSTM ‡                 | <b>85.2 <math>\pm</math> 0.1</b> | <b>85.1 <math>\pm</math> 0.1</b> | <b>85.1 <math>\pm</math> 0.1</b> |

## 4.6 Results

### 4.6.1 Predictive Performance

Tables 4.2 and 4.3 show the results obtained by all models in the Weibo and Twitter datasets.

In Twitter, HierLongformer LSTM achieves the highest F1 score overall (80.2) surpassing all the baseline models as well as the simpler hierarchical architectures, e.g., using mean and max pooling. For each of the transformer-based model, we observe that the hierarchical transformer architectures (e.g, LSTM, max pooling and mean pooling) outperform the truncated models across all metrics. Their hierarchical structure allows them to exploit all the available textual information from each user that impacts performance. The Longformer model that supports longer input sequences achieves better predictive results than the other transformer models that support shorter input sequences (e.g., BERT and RoBERTa). This is similar to results obtained by Beltagy et al. (2020); Gutierrez et al. (2020) where the Longformer consistently outperforms other BERT-style models in long document classification tasks.

**Table 4.3:** Twitter binary classification results (mean  $\pm$  standard deviation). ‡ denotes that the HierLongformer LSTM performs significantly better than truncated Longformer (t-test;  $p < .05$ ).

| Model                            | P (macro)                        | R (macro)                        | F1 (macro)                       |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| <b>Baseline Models</b>           |                                  |                                  |                                  |
| LR-BOW                           | 75.8 $\pm$ 0.1                   | 74.9 $\pm$ 0.1                   | 74.9 $\pm$ 0.1                   |
| LR-LIWC                          | 68.4 $\pm$ 0.1                   | 67.4 $\pm$ 0.1                   | 67.4 $\pm$ 0.2                   |
| BiLSTM-ATT                       | 76.0 $\pm$ 0.2                   | 75.0 $\pm$ 0.2                   | 75.1 $\pm$ 0.2                   |
| <b>Truncated Transformers</b>    |                                  |                                  |                                  |
| BERT                             | 73.1 $\pm$ 0.1                   | 72.5 $\pm$ 0.2                   | 72.5 $\pm$ 0.2                   |
| RoBERTa                          | 74.5 $\pm$ 0.3                   | 73.7 $\pm$ 0.2                   | 73.7 $\pm$ 0.2                   |
| LongFormer                       | 77.9 $\pm$ 0.1                   | 77.0 $\pm$ 0.2                   | 77.0 $\pm$ 0.2                   |
| <b>Hierarchical Transformers</b> |                                  |                                  |                                  |
| HierBERT Mean Pool               | 77.7 $\pm$ 0.2                   | 77.4 $\pm$ 0.2                   | 77.5 $\pm$ 0.2                   |
| HierBERT Max Pool                | 73.5 $\pm$ 0.2                   | 72.4 $\pm$ 0.3                   | 73.6 $\pm$ 0.3                   |
| HierBERT LSTM                    | 78.1 $\pm$ 0.2                   | 77.5 $\pm$ 0.1                   | 77.6 $\pm$ 0.2                   |
| HierRoBERTa Mean Pool            | 79.1 $\pm$ 0.3                   | 78.1 $\pm$ 0.2                   | 78.2 $\pm$ 0.2                   |
| HierRoBERTa Max Pool             | 77.5 $\pm$ 0.3                   | 75.7 $\pm$ 0.3                   | 75.9 $\pm$ 0.3                   |
| HierRoBERTa LSTM                 | 78.9 $\pm$ 0.2                   | 78.5 $\pm$ 0.3                   | 78.8 $\pm$ 0.2                   |
| HierLongformer Mean Pool         | 80.3 $\pm$ 0.2                   | 79.8 $\pm$ 0.1                   | 79.9 $\pm$ 0.1                   |
| HierLongformer Max Pool          | 79.3 $\pm$ 0.1                   | 79.0 $\pm$ 0.1                   | 79.0 $\pm$ 0.1                   |
| HierLongformer LSTM ‡            | <b>80.5 <math>\pm</math> 0.1</b> | <b>80.1 <math>\pm</math> 0.1</b> | <b>80.2 <math>\pm</math> 0.1</b> |

In Weibo, HierERNIE LSTM achieves the highest F1 score overall surpassing all other models. In addition, we observe two baseline models (LR-BOW and BiLSTM-ATT) achieve a slightly lower performance than the hierarchical transformers e.g., 82.7 and 82.9 F1-score respectively. This suggests that the relationship between users’ language use and labels can be learned more efficiently by using a simple classifier (e.g., LR) that has access to all users’ posts, compared to a more complex model that does not use all available information. We also observe that, in general, the use of different hierarchical methods (especially the LSTM takes into account the sequence order) improve the performance of truncated transformer models. This suggests that the order of the posts and their dependencies matter.

Lastly, we observe that the models with similar structure and characteristics trained on Weibo data are on average more accurate than the Twitter data (approximately 5%). This highlights that input language (i.e., Chinese vs. English) and its peculiarities play an important role in the performance of text classification models.

## 4.6.2 Model Explainability

For both datasets, we analyze the most important input tokens that contribute to the model prediction (i.e., HierERNIE LSTM in Weibo and HierLongformer LSTM in Twitter) by employing a widely used gradient-based explainability method i.e., the *InputXGrad with L2 Norm Aggregation* (Kindermans et al., 2016) that has been found to provide faithful explanations for transformer-based models in NLP tasks (Chrysostomou and Aletras, 2021, 2022). The *InputXGrad* ( $\mathbf{x}\nabla\mathbf{x}$ ) ranks the input tokens by computing the derivative of the input with respect to the model predicted class and then multiplied by the input itself, where  $\nabla x_i = \frac{\partial \hat{y}}{\partial x_i}$ . We then get the L2 normalized aggregation of the scores across the embedding dimensions similar to Chrysostomou and Aletras (2021).

**Twitter** In Twitter, the InputXGrad scores indicate that some hashtags and emojis (note that we have detokenized wordpieces when calculating importance scores) have a higher impact on model predictions. For example, politics-related terms and hashtags (e.g., 🇺🇸, #POTUS) play an important role when the model predicts Twitter users as misinformation posters. This is similar to the result from Addawood et al. (2019), showing that Twitter users using a higher number of political hashtags are more likely to be identified by the model as political trolls. On the other hand, some tokens related to daily activities (e.g., #yoga, #marchforscience, #vegetarian) and social issues (e.g., #blacklivematters) are more prevalent in misinformation active citizens.

**Weibo** In Weibo, when the model predicts users as misinformation posters, some tokens that express emotions ((e.g., *surprise, unhappy and amazing*)) become the key factors. In contrast, model assigns importance to some popular buzzwords (e.g., *hahahaha, xswl (i.e., LMAO in Chinese abbreviations)*) and celebrities (e.g., *tfboys, uzi, and blackpink*) when it predicts users as misinformation active citizens. These users who tend to debunk misinformation appear to be common users using Weibo for social interactions with friends.<sup>15</sup>

---

<sup>15</sup>The letters U and S, which can be used as part of a regional indicator pair to create emoji flags for various countries.

**Table 4.4:** N-grams associated with Twitter misinformation posters and active citizens sorted by Pearson’s correlation ( $r$ ) between the normalized frequency and the labels ( $p < .001$ ).

| n-grams                       |       |                 |       |
|-------------------------------|-------|-----------------|-------|
| Posters                       | r     | Active Citizens | r     |
| illegals                      | 0.173 | slightly        | 0.154 |
| msm (mainstream media)        | 0.165 | empathy         | 0.154 |
| <b>U</b> (regional indicator) | 0.158 | theories        | 0.144 |
| <b>S</b> (regional indicator) | 0.154 | generally       | 0.141 |
| soros                         | 0.143 | equivalent      | 0.137 |
| brennan                       | 0.142 | necessarily     | 0.135 |
| communist                     | 0.140 | confusing       | 0.135 |
| schumer                       | 0.139 | fewer           | 0.131 |
| leftist                       | 0.130 | quotes          | 0.129 |
| rino                          | 0.128 | actively        | 0.129 |

**Table 4.5:** N-grams associated with Weibo misinformation posters and active citizens sorted by Pearson’s correlation ( $r$ ) between the normalized frequency and the labels ( $p < .001$ ). We translate all the Chinese N-grams into English. ‘Rightmost’ indicates that a Weibo user is asking their followers to check out a post they have shared (Yu et al., 2019).

| n-grams      |       |                 |       |
|--------------|-------|-----------------|-------|
| Posters      | r     | Active Citizens | r     |
| cherish      | 0.177 | WTF             | 0.256 |
| understand   | 0.172 | LMAO            | 0.249 |
| present      | 0.171 | 👉               | 0.239 |
| this morning | 0.154 | 🕒               | 0.232 |
| because      | 0.149 | rightmost       | 0.225 |
| contact      | 0.148 | 😏               | 0.222 |
| rose         | 0.145 | awesome         | 0.218 |
| strong       | 0.143 | Ahhhh           | 0.214 |
| 😏            | 0.142 | 😏               | 0.202 |
| creation     | 0.139 | F*ck            | 0.200 |

### 4.6.3 Error Analysis

We also perform an error analysis by inspecting cases of wrong predictions in both datasets. We first observe that Twitter active citizens who are wrongly classified as posters are more prevalent in posting about politicians (e.g., *Obama*, *Clinton* and *POTUS*) and some hashtags

(e.g., #VOTEBIDEN, #BIDEN2020) related to the democratic party in the U.S.. These users are misclassified by the model possibly due to similar language use with those spreading misinformation. We also notice that Weibo misinformation posters who are misclassified as active citizens use cyber slang while are also more likely to express emotions e.g., *Ahhh* and 🤔. We finally observe that a higher proportion of Weibo misinformation posters who are wrongly classified as active citizens are verified users (15%) (note that higher percentage of posters (21%) are verified users than active citizens (10%)).

## 4.7 Linguistic Analysis

We further perform a linguistic analysis to uncover the differences in language use between users in the two categories, i.e. misinformation posters and active citizens. To that end, we employ univariate Pearson’s correlation test to characterize which linguistic features (i.e., BOW and LIWC<sup>16</sup>) are high correlated with each class following Schwartz et al. (2013). This approach has been widely used in similar NLP studies (Preoțiuc-Pietro et al., 2019; Maronikolakis et al., 2020; Jin et al., 2022).

### 4.7.1 N-grams

Table 4.4 shows that Twitter users who diffuse misinformation are more prevalent in posting about politics (e.g., **US**, *Soros* and *Brennan*). This is similar to findings by Mu and Aletras (2020), which showed that people who often retweeted news items from unreliable news sources (e.g., Infowars, Disclose.tv) are more likely to discuss politics. Moreover, active citizens on Twitter use more frequently adverbs (e.g., *slightly*, *generally*, and *necessarily*) and words that denote uncertainty (e.g., *confusing*). Table 4.5 shows that Weibo active citizens are more likely to use words related to self-disclosure, e.g., *WTF*, *LMAO* and *awesome* and net-speak words e.g., 🍌 and *rightmost*. These buzzwords are more popular among average Weibo users who share interesting posts with their friends or reply to something entertaining. Note that most of Weibo active citizens are not official accounts (i.e., unverified users) which rarely use these words. Similarly, Weibo active citizens also use emojis that express uncertainty, e.g., 🤔,

---

<sup>16</sup>We use the LIWC English (Pennebaker et al., 2001) and Simplified Chinese (Huang et al., 2012) dictionaries.







**Table 4.6:** English LIWC features associated with Twitter misinformation posters and active citizens sorted by Pearson’s correlation ( $r$ ) between the normalized frequency and the labels ( $p < .001$ ).

| LIWC        |       |                 |       |
|-------------|-------|-----------------|-------|
| Posters     | r     | Active Citizens | r     |
| power       | 0.201 | tentat          | 0.250 |
| drives      | 0.186 | differ          | 0.217 |
| colon       | 0.133 | adverb          | 0.201 |
| clout       | 0.132 | cogproc         | 0.200 |
| we          | 0.132 | insight         | 0.194 |
| exclam      | 0.126 | ipron           | 0.182 |
| otherP      | 0.125 | conj            | 0.181 |
| female      | 0.115 | compare         | 0.174 |
| relig       | 0.114 | function        | 0.160 |
| affiliation | 0.113 | comma           | 0.151 |

**Table 4.7:** Simplified Chinese LIWC features associated with Weibo misinformation posters and active citizens sorted by Pearson’s correlation ( $r$ ) between the normalized frequency and the labels ( $p < .001$ ).

| LIWC    |       |                 |       |
|---------|-------|-----------------|-------|
| Posters | r     | Active Citizens | r     |
| social  | 0.251 | model_pa        | 0.338 |
| prep    | 0.208 | informal        | 0.334 |
| health  | 0.180 | assent          | 0.332 |
| space   | 0.170 | progm           | 0.328 |
| you     | 0.168 | nonflu          | 0.307 |
| female  | 0.168 | insight         | 0.298 |
| achieve | 0.166 | practice        | 0.259 |
| sexual  | 0.162 | tensem          | 0.258 |
| friend  | 0.160 | adverb          | 0.220 |
| drives  | 0.160 | swear           | 0.209 |

On the other hand, Weibo active citizens use more frequently words belonging to LIWC categories such as *informal* and *nonflu* (*nonfluent*) that are similar to their correlated N-grams (see Table 4.5).

## 4.8 Conclusion

In this paper, we have presented an extensive study on identifying and characterizing misinformation posters and active citizens across two different social media platforms (i.e., Twitter and Weibo) and languages (i.e. Chinese and English) for the first time. We developed a new Weibo dataset with users labeled into the two categories and repurposed an existing Twitter dataset for the task. Our hierarchical transformer model performs best, achieving up to 80.2 and 85.1 macro F1 score on Twitter and Weibo datasets respectively. Finally, we perform a linguistic feature analysis unveiling the major differences in language use between the two groups of users across platforms.

In the future, we plan to (i) explore cross-lingual settings for the task as well as including information from different modalities such as images (Sánchez Villegas and Aletras, 2021; Sánchez Villegas et al., 2021); (ii) extend the current task into a fine-grained setting (i.e., a multi-class classification task); and (iii) analyse the differences in behaviours between Weibo and Twitter users (e.g., the speed of reactions to online misinformation).

## Ethics Considerations

Our work has received ethical approval from the Ethics Committee of our department (Reference Number 025470) and complies with the Weibo and Twitter data policies for research. To ensure the anonymity of the data, we only share the user’s ID, rather than the username that appears on the platform. We do not share the data for non-research purposes.

## Acknowledgments

We would like to thank Danae Sánchez Villegas, Mali Jin, George Chrysostomou, Xutan Peng and all the anonymous reviewers for their valuable feedback. Pu Niu is the corresponding author and supported by China Postdoctoral Science Foundation (No. 2021M701371).

---

## BIBLIOGRAPHY

- Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *AAAI-ICWSM*, volume 13, pages 15–25.
- Ahmet Aker, Kevin Vincentius, and Kalina Bontcheva. 2019. Credibility and transparency of news sources: Data collection and feature analysis. In *NewsIR@SIGIR*, pages 15–20.
- Sacha Altay, Anne-Sophie Hacquin, and Hugo Mercier. 2022. Why do so few people share fake news? it hurts their reputation. *new media & society*, 24(6):1303–1324.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *EMNLP*, pages 4982–4991.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *EMNLP*, pages 3528–3539.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.
- David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory*, 6(3):203–242.

- George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.
- George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10545–10553.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.
- Anastasia Giachanou, Esteban A Ríssola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. 2020. The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In *NLDB*, pages 181–192. Springer.
- Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018a. Identifying and understanding user reactions to deceptive and trusted social news sources. In *ACL*, pages 176–181.
- Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018b. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, 5(4):1071–1082.
- Bernal Jimenez Gutierrez, Jucheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. Document classification for covid-19 literature. In *EMNLP Findings*, pages 3715–3722.
- Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.
- Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. In *1st FEVER Workshop*, pages 103–108.

- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker. 2012. The development of the chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*.
- Mali Jin, Daniel Preoțiuc-Pietro, Seza Dogruoz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *ACL*. Association for Computational Linguistics.
- Mayank Jobanputra. 2019. Unsupervised question answering for fact-checking. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 52–56.
- Håkan Johansson and Björn Hvinden. 2007. What do we mean by active citizenship. In *Citizenship in Nordic welfare states: Dynamics of choice, duties and participation in a changing Europe*, pages 32–51. Routledge.
- Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.
- Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.
- Kyumin Lee, Brian Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *AAAI-ICWSM*, volume 5.
- Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41.
- Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *ACL*, pages 138–143.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Zhiyuan Liu, Le Zhang, CunChao TU, and MaoSong SUN. 2015. Statistical and semantic analysis of rumors in chinese social media. *Scientia Sinica Informationis*, 45(12):1536–1546.
- Luca Luceri, Silvia Giordano, and Emilio Ferrara. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *AAAI-ICWSM*, volume 14, pages 417–427.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pages 3818–3824.
- Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, 35(2):196–219.
- Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.
- Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6:e325.
- Andriy Mulyar, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. 2019. Phenotyping of clinical notes with improved document classification models using contextualized neural language models. *arXiv preprint arXiv:1910.13664*.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *EMNLP*, pages 9–14.
- Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*, volume 33, pages 6859–6866.
- Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *ASRU*, pages 838–844. IEEE.

- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.
- Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.
- Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.
- Gordon Pennycook and David G Rand. 2020. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *COLING*, pages 3391–3401.
- Daniel Preoțiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.
- Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CLEF*.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937.
- Craig T Robertson, Rachel R Mourão, and Esther Thorson. 2020. Who uses fact-checking sites? the impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. *The International Journal of Press/Politics*, 25(2):217–237.

- Danae Sánchez Villegas and Nikolaos Aletras. 2021. Point-of-interest type prediction using text and images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. Analyzing online political advertisements. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680, Online. Association for Computational Linguistics.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).
- Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.
- Maneet Singh, Rishemjit Kaur, and SRS Iyengar. 2020. Multidimensional analysis of fake news spreaders on twitter. In *CSoNet*, pages 354–365.
- Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019a. How to fine-tune bert for text classification? In *NLPCC*, pages 194–206. Springer.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.
- Shalini Talwar, Amandeep Dhir, Puneet Kaur, Nida Zafar, and Melfi Alrasheedy. 2019. Why do people share fake news? associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, 51:72–82.
- Edson C Tandoc Jr, Darren Lim, and Rich Ling. 2020. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3):381–398.



- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*, pages 809–819.
- Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *ECIR*, pages 575–588. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.
- Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *SIGIR*, pages 275–284.
- Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: Analysis and generation of fact-checking language. In *SIGIR*, pages 335–344.
- Nguyen Vo and Kyumin Lee. 2020a. Standing on the shoulders of guardians: Novel methodologies to combat fake news. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 183–210. Springer.
- Nguyen Vo and Kyumin Lee. 2020b. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. *arXiv preprint arXiv:2010.03159*.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *ACL*, pages 422–426.
- Evan Williams, Paul Rodrigues, and Valerie Novak. 2020. Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. *arXiv preprint arXiv:2009.02431*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on Sina Weibo by propagation structures. In *ICDE*, pages 651–662. IEEE.

- 
- Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.
- Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network with applications to early rumor detection. In *EMNLP*, pages 9042–9051.
- Di You, Nguyen Vo, Kyumin Lee, and Qiang Liu. 2019. Attributed multi-relational attention network for fact-checking url recommendation. In *CIKM*, pages 1471–1480.
- Shuo Yu, Hongyi Zhu, Shan Jiang, Yong Zhang, Chunxiao Xing, and Hsinchun Chen. 2019. Emoticon analysis for chinese social media and e-commerce: The azemo system. *ACM Transactions on Management Information Systems (TMIS)*, 9(4):1–22.
- Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *NAACL*, pages 1614–1623.
- Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.

# PREDICTING THE POPULARITY OF FALSE RUMORS

## Predicting and Analyzing the Popularity of False Rumors in Social Media

*Yida Mu<sup>1</sup>, Pu Niu<sup>2</sup>, Kalina Bontcheva<sup>1</sup> and Nikolaos Aletras<sup>1</sup>*

<sup>1</sup> Department of Computer Science, The University of Sheffield

<sup>2</sup> Central China Normal University

### Abstract

Malicious online rumors with high popularity, if left undetected, can spread very quickly with damaging societal implications. The development of reliable computational methods for early prediction of the popularity of false rumors is very much needed, as a complement to related work on automated rumor detection and fact-checking. To this end, we (1) propose a new regression task to predict the future popularity of false rumors given both post and user-level information; (2) introduce a new publicly available dataset in Chinese that includes 19,256 rumor cases from Weibo, the corresponding profile information of the original spreaders

and a rumor popularity score as a function of the shares, replies and reports it has received; (3) develop a new open-source domain adapted pre-trained language model, i.e., BERT-Weibo-Rumor and evaluate its performance against several supervised classifiers using post and user-level information. Our best performing model achieves the lowest RMSE score (1.55) and highest Pearson’s  $r$  (0.633), outperforming competitive baselines by leveraging textual information from both the post and the user profile. Our analysis unveils that popular rumors consist of more conjunctions and punctuation marks, while less popular rumors contain more words related to the social context and personal pronouns.

## 5.1 Introduction

Social media platforms (e.g., Twitter, Facebook and Weibo) play an important role in information dissemination related to important events, social emergencies and natural disasters (Middleton et al., 2013; Imran et al., 2015; Castillo, 2016; Wang et al., 2016). However, online rumors (i.e., posts with unverified veracity) have been shown to spread faster than reliable information and can thus mislead the public especially when ultimately proven false (Vosoughi et al., 2018).

The timely publication of fact-checks of such false rumors can both raise user awareness and help prevent rumors from spreading further (Vo and Lee, 2020). Vo and Lee (2019) showed that debunked tweets (i.e., Twitter posts containing false rumors) are more likely to be deleted and for their original spreaders to be suspended. To combat false rumors spreading in social media, independent (e.g., PolitiFact<sup>1</sup>) or in-house (e.g. Weibo) fact-checking platforms have been created with the purpose to debunk suspicious posts.

Figure 5.1 shows an example of a debunked false rumor on the Weibo fact-checking platform.<sup>2</sup> The top box shows the rumor: “*Hua Chunying (i.e., the Foreign Ministry Spokesperson of PRC) announces a ban on Chinese stars with multiple nationality from returning to China to do business*”. The exclamation mark sign ‘!’ in the top box denotes that ‘This is a debunked rumor’ and users can click on it<sup>3</sup> to get relevant fact-checking

---

<sup>1</sup><https://www.politifact.com>

<sup>2</sup><http://weibo.com/1074273855/IwFJ6dsuQ>

<sup>3</sup><https://service.account.weibo.com/show?rid=K1CaS8wtk7K4k>

information. The blue box (middle left) denotes the information of the user reported the rumor and the number of all reports from different users. The orange box (middle right) displays the information of the user who posted the false rumor. The green box (bottom) indicates that the original post is debunked as a false rumor and the user (in the orange box) who posted the false rumor will lose 10 ‘credit points’ and can not post or be followed for the next 15 days.

Fact-checking platforms typically verify such rumors manually, which is highly reliable but expensive in terms of time and costs (Pavleska et al., 2018; Vo and Lee, 2018). Therefore, fact checkers are increasingly being assisted by automated rumor detection and veracity (i.e., whether a rumor is true or false) prediction systems for retrieving rumor related information more efficiently (Zubiaga et al., 2018). To further improve their efficiency, professionals also need a way to prioritize for debunking those detected rumors which are likely to become highly popular and reach a large audience (Parikh et al., 2019; Smith and Bastian, 2022).

To address the latter challenge, the focus of this paper is on developing computational methods for predicting the popularity of false rumors as soon as they are detected. Identifying false rumors with a higher impact in the early stage allows social media platforms to timely deliver fact-checking information to the public. Previous work has focused on predicting the popularity of social media posts (e.g., tweets, YouTube videos) (Trzcinski and Rokita, 2017; Gao et al., 2021) and individual social media users (Lampos et al., 2014) with applications in advertising and recommend systems. For the purpose of our task, we only consider immediately available contextual information (e.g., rumor content and user profile information), which is crucial for the early detection of false rumors with high popularity. To the best of our knowledge, the regression task of false rumor popularity prediction (i.e., early detection of popularity) has yet to be explored.

To this end, we pose the following three research questions:

- $RQ_1$  How can we define the popularity of false rumors on Weibo?
- $RQ_2$  How can we predict the future popularity of false rumors based on post and user level information?
- $RQ_3$  What are the most important markers that correlate with high and low-popularity rumors?



**Figure 5.1:** Example false rumor on the Weibo fact-checking platform (English translation included in the main body of the paper).

To answer these research questions:

- We develop a new publicly available dataset<sup>4</sup> from Weibo, which includes 19,256 debunked false rumors in Chinese associated with a popularity score and Meta-features;
- We evaluate several supervised models using post and user-level information and their combination. Combining these two sources of information with our new pre-trained language model (i.e. BERT-Weibo-Rumor) achieves the best overall performance;
- We perform a linguistic analysis to unveil the characteristics of highly popular false rumors compared to those with low popularity. We also unveils that some user profile characteristics (e.g., verified status, number of followers and number of historical posts) are positively correlated with the future popularity of false rumors.

<sup>4</sup>Our dataset and source code will be publicly released.

**Paper outline** The rest of the paper is organized as follows. In Section 5.2, we discuss previous work related to rumor detection. The task description is introduced in Section 5.3. We describe the development of our Weibo dataset in Section 5.4. We discuss the model details, hyperparameters tuning, and results in Section 5.5. In Section 5.6, we conduct extensive analysis (including ablation study, error analysis, and qualitative analysis of model prediction to gain insights for future work. We also discuss the ethics considerations, theoretical and practical implications of this work in Section 5.7. Finally, we conclude and some up with some future directions in Section 5.8.

## 5.2 Related Work

### 5.2.1 Rumor Detection

A rumor is generally defined as any social media post whose credibility is yet to be verified at the time it was published (Zubiaga et al., 2018). Typically, rumor detection systems first predict if a given post is a rumor or not, and second whether it is true or false. Prior work on automatic rumor detection generally falls into one of the following categories:

- feature-based methods that rely on linguistic (e.g., text) and visual (e.g., images) information to detect unreliable posts (Rashkin et al., 2017; Volkova et al., 2017; Qi et al., 2019);
- knowledge-based methods that leverage external knowledge (e.g., Wikipedia) to determine rumor veracity (Sun et al., 2021; Wei et al., 2021; Hu et al., 2021; Jiang et al., 2022);
- user propagation-based methods that consider the diffusion of the rumor (e.g., time-series analysis) (Lin et al., 2021; Chen et al., 2021; Nobre et al., 2022).
- early rumor detection methods (Xia et al., 2020; Silva et al., 2021) that aim to detect rumors as soon as they are posted online. These approaches tend to employ a combination of features (e.g., user features and time-series data) from different periods during the rumor propagation cycle to detect the earliest point in time that a particular post has actually become a rumor (Zhou et al., 2019; Yuan et al., 2020).

Automatic rumor detection methods are usually evaluated on existing annotated datasets, e.g., Weibo (Ma et al., 2016, 2017), FEVER (Thorne et al., 2018), Twitter15&16 (Ma et al., 2016), PHEME (Zubiaga et al., 2016) and LIAR (Wang, 2017). Kochkina et al. (2023) evaluate the generalizability of neural-based rumor classifiers across different benchmarks. Recently, interpretable rumor detection methods (e.g., attention-based and rule-based) have also been explored (Atanasova et al., 2020, 2022; Silva et al., 2021; Ayoub et al., 2021), for generating explanations in aid of fact-checking by highlighting evidence. Some of these methods are often embedded in real-world fact-checking platforms e.g., Propagation2Vec (Silva et al., 2021) and Defend (Shu et al., 2019).

Apart from modeling individual posts, previous work has also explored modeling user behavior, e.g. analyzing user reactions and stance towards unreliable posts (Glenski et al., 2018; Mu et al., 2022; Bazmi et al., 2023) to show that a higher percentage of human users retweet news posts from credible sources (e.g., @BBC and @Reuters) as compared to bots.

### 5.2.2 Modeling Popularity in Social Media

Another strand of related work has focused on predicting the popularity of multimodal online content, e.g., YouTube Videos (Pinto et al., 2013; Kong et al., 2018), tweets (Zhao et al., 2015), Facebook posts (Trzciński and Rokita, 2017) and Weibo posts (Bao et al., 2013; Gao et al., 2014).

Existing work usually relies on post’s user engagement metrics (e.g., shares, replies, views, likes, etc.) to represent its popularity (Yan et al., 2016; Gao et al., 2021). Another metric is engagement rate which is calculated as the sum of the user engagement metrics received divided by the number of views of the post (Alkhodair et al., 2020). To model the popularity score of online posts, post-level features (e.g., textual and visual information) (Pinto et al., 2013; Piotrkowicz et al., 2017) and user features (e.g., profile information) (McParlane et al., 2014; Gelli et al., 2015; Li et al., 2017) are commonly used as they are publicly available.

At the level of individual users, Weng et al. (2010) and Lampos et al. (2014) quantify Twitter user impact as a function of the number of followers and friends. They both predict and analyze user impact through user profile and post-level features.



**Table 5.1:** Specifications of Existing Weibo-based Rumor Datasets. *Note.* ‡ denotes that our dataset contains more user-level features e.g., ‘user Credit Score’, ‘# of Likes Received’ (i.e., user attributes features ( $U$ ) from  $U_6$  to  $U_{12}$  in Table 5.2.)

| Dataset                   | # of False Rumors | Time Span        | User-level Features | # of Engagements |
|---------------------------|-------------------|------------------|---------------------|------------------|
| Ma et al. (2016)          | 2,313             | 2012-2016        | ✓                   | ✓                |
| Jin et al. (2017)         | 4,749             | 2012-2016        | x                   | x                |
| Rao et al. (2021)         | 3,034             | 2016-2021        | ✓                   | ✓                |
| Song et al. (2021)        | 1,538             | N/A              | x                   | x                |
| Lu et al. (2021)          | 1,975             | 2012-2020        | x                   | x                |
| <b>WeiboRumors (Ours)</b> | <b>19,256</b>     | <b>2010-2021</b> | <b>✓ ‡</b>          | <b>✓ ‡</b>       |

**Rumor Popularity** Previous work on predicting the future popularity of false rumors is limited. Alkhodair et al. (2020) present a classification task for predicting the engagement rate of tweets (high, moderate and low) through solely textual information. However, the predicted engagement rate which is calculated by dividing the sum of the engagement by the sum of the views received on the post, cannot be applied for Weibo posts as it requires the number of views on the post which is not available through the official Weibo API. Parikh et al. (2019) define the impact of online false news articles based on three metrics including (i) the topic of the news items (e.g., politics, economics, science, etc.); (ii) the reputation of the news website that posted the news and (iii) the proliferator’s popularity, i.e., the number of followers of users who shared the false news.

### 5.2.3 Our work

We note that while some rumors do spread widely (i.e. gain a lot of attention), many others only reach a very small audience. Therefore, it is equally important to detect the **future popularity of false rumors** on social media, so that they can be prioritized for debunking. Given that most fact-checking platforms (e.g., Weibo rumor debunking platform, PolitiFact, etc.) rely on human resources to manually check the veracity of rumors, social media platforms can first address false rumors with higher popularity. Note that we only use information that is immediately available which is crucial for the early detection of false rumors with high popularity. This task is yet to be explored in computational social science.

## 5.3 Task Description

We define false rumor popularity prediction as a regression task. Given a false rumor  $X = \{(R, P, U)\}$  consisting of textual information  $R$  (i.e., a sequence of tokens representing the actual rumor), user profile description  $P$  (i.e., a sequence of tokens representing the personal description provided by the user) and user attributes  $U$  (e.g., number of followers, posts, etc.), we aim to learn a supervised function  $f$  that can predict the popularity score  $Y$  of a false rumor. The value of the popularity score is calculated using rumor engagement attributes  $E$ , which include the number of shares, number of replies, and number of reports, based on Equation (1).

## 5.4 Data

### 5.4.1 Data Collection

### 5.4.2 Data Collection

For our experiments, we create a new dataset using the fact-checking platform provided by Weibo.<sup>5</sup> We opted using Weibo since it is the largest Chinese-based social media platform and its fact-checking platform has enabled the development of many rumor detection datasets (Ma et al., 2016; Rao et al., 2021).

However, these previously published datasets are relatively small (e.g., there are 2,313 and 3,034 false rumor cases from Ma et al. (2016) and Rao et al. (2021) datasets, respectively) and lack the **metadata information** required for our task. For instance, the Song et al. (2021) and Lu et al. (2021) datasets are not suitable for the regression task of predicting the level of popularity of false rumors, as they do not provide information on the number of engagements (e.g., Shares, replies, etc.) received by rumors. We further elaborate on the details of previously publicly available datasets in Table 5.1.

---

<sup>5</sup><https://service.account.weibo.com/?type=5&status=4>

**Table 5.2:** Information associated with each false rumor in our dataset.

| Description                                         |                                                                            |
|-----------------------------------------------------|----------------------------------------------------------------------------|
| <b>Rumor Engagement Attributes (<math>E</math>)</b> |                                                                            |
| $E_1$                                               | # of Shares                                                                |
| $E_2$                                               | # of Replies                                                               |
| $E_3$                                               | # of Reports                                                               |
| <b>Rumor Content</b>                                |                                                                            |
| $R$                                                 | Text representing the actual rumor                                         |
| <b>User Profile Description</b>                     |                                                                            |
| $P$                                                 | Text describing user’s bio                                                 |
| <b>User Attributes (<math>U</math>)</b>             |                                                                            |
| $U_1$                                               | # of Followers (i.e., other users who follow this account)                 |
| $U_2$                                               | # of Followees (i.e., one can follow others)                               |
| $U_3$                                               | # of Bi_Followers (i.e., users who follow each other)                      |
| $U_4$                                               | # of Statuses (i.e., the number of posts)                                  |
| $U_5$                                               | # of Favorites (i.e. one can like posts from other users)                  |
| $U_6$                                               | Credit Score                                                               |
| $U_7$                                               | Verified Status (i.e., Verified or Unverified)                             |
| $U_8$                                               | # of Shares Received                                                       |
| $U_9$                                               | # of Likes Received                                                        |
| $U_{10}$                                            | # of Replies Received                                                      |
| $U_{11}$                                            | # of Likes Received in Replies                                             |
| $U_{12}$                                            | # of All Reactions Received (i.e., the sum of $U_8, U_9, U_{10}, U_{11}$ ) |

The Weibo fact-checking platform allows end-users to report suspicious posts (i.e., rumors), which are subsequently checked by professional journalists to verify their veracity and provide fact-checking information. In cases where the information of a post is deemed to be *false*, it is also flagged as a false rumor including information that refutes any claims that it contains. Note that rumors are usually defined as online posts whose veracity is yet to be verified at the time of posting (i.e., they can ultimately turn out to be *true*, *false* or *not verifiable*) (Zubiaga et al., 2018). However, in our dataset all rumors are *false*, i.e., the source post contains debunked false information (see Figure 5.1).

We collect a total of 40,936 cases of false rumors using the official Weibo API.<sup>6</sup> All cases have been debunked and cover a period between May 2012 and November 2021.

<sup>6</sup><https://open.weibo.com/>

**Table 5.3:** Descriptive Statistics of the Popularity Score ( $Y$ ) in Train, Dev, and Test splits.  $Y$  denotes the popularity score of the false rumors.

| Popularity Score Distribution |                |                |                |                |                |                |                |                |            |
|-------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|------------|
| <b>All (19,256)</b>           |                |                |                |                |                |                |                |                |            |
| Popularity Score              | $0 \leq Y < 1$ | $1 \leq Y < 2$ | $2 \leq Y < 3$ | $3 \leq Y < 4$ | $4 \leq Y < 5$ | $5 \leq Y < 6$ | $6 \leq Y < 7$ | $7 \leq Y < 8$ | $8 \leq Y$ |
| # of Rumors                   | 5,067          | 6,144          | 2,753          | 1,774          | 1,274          | 934            | 586            | 372            | 351        |
| Proportion (%)                | 26.3           | 31.9           | 14.3           | 9.2            | 6.6            | 4.9            | 3              | 1.9            | 1.8        |
| <b>Train (15,404)</b>         |                |                |                |                |                |                |                |                |            |
| # of Rumors                   | 4,036          | 4,911          | 2,225          | 1,436          | 1,011          | 744            | 462            | 288            | 291        |
| Proportion (%)                | 26.2           | 31.9           | 14.4           | 9.3            | 6.6            | 4.8            | 3.0            | 1.9            | 1.9        |
| <b>Dev (1,926)</b>            |                |                |                |                |                |                |                |                |            |
| # of Rumors                   | 495            | 651            | 262            | 157            | 132            | 106            | 54             | 39             | 30         |
| Proportion (%)                | 25.7           | 33.9           | 13.6           | 8.2            | 6.9            | 5.5            | 2.8            | 2.0            | 1.6        |
| <b>Test (1,926)</b>           |                |                |                |                |                |                |                |                |            |
| # of Rumors                   | 536            | 582            | 266            | 181            | 132            | 84             | 70             | 45             | 30         |
| Proportion (%)                | 27.8           | 30.2           | 13.8           | 9.4            | 6.9            | 4.4            | 3.6            | 2.3            | 1.6        |

**Table 5.4:** Descriptive statistics (i.e., Min, Mean, Median, and Max) of the number of tokens in the rumor content ( $R$ ).

| Descriptive Statistics of Tokens |     |       |        |     |
|----------------------------------|-----|-------|--------|-----|
|                                  | Min | Mean  | Median | Max |
| <b>All</b>                       | 7   | 105.2 | 122    | 259 |
| <b>Train</b>                     | 5   | 105.2 | 122    | 259 |
| <b>Dev</b>                       | 5   | 105.3 | 123    | 183 |
| <b>Test</b>                      | 7   | 105.5 | 124    | 184 |

### 5.4.3 Rumor Information

For each false rumor, we collect rumor engagement attributes ( $E$ ), the rumor content ( $R$ ), the user profile description ( $P$ ) and user attributes ( $U$ ).

Rumor engagement attributes ( $E$ ) include the number of shares ( $E_1$ ), replies ( $E_2$ ) that the post received, and the number of times users have reported ( $E_3$ ) the post on the fact-checking platform. Note that, one rumor can be reported by different users in the Weibo fact-checking platform.

We also collect the text of the false rumor ( $R$ ) and the user profile description ( $P$ ). User attributes ( $U$ ) consist of (1) user social connections (from  $U_1$  to  $U_5$ ) including the number of followers (i.e., other users who follow this Weibo account), followees (i.e., one can follow other users), bi-followers (i.e., users who follow each other); (2) user engagement (from  $U_8$  to  $U_{12}$ )

information including the number of posts, and the number of reactions (e.g., shares, replies, etc from other users) received; (3) Weibo account attributes e.g., Verified Status ( $U_7$ ) (i.e., Verified or Unverified) and Credit Score ( $U_6$ ). Note that the ‘Credit Score’ ( $U_6$ ) is a unique user-level attribute on Weibo. Weibo users lose some of their credit score for posting false rumors. When a user’s credit score falls below a certain threshold, they are not able to post for a period of time.

We only consider the rumor content ( $R$ ), user profile description ( $P$ ) and user attributes ( $U$ ) as they are immediately available when false rumors are published on Weibo. These features can be used to train predictive models for detecting highly popular false rumors in an early stage. Table 5.2 shows a summary of all information collected for each rumor.

#### 5.4.4 Defining False Rumor Popularity on Weibo

In social networks, the user engagement (e.g., shares, replies, etc.) on source posts is visible to all users and is widely employed in characterizing the popularity of a given post (Zaman et al., 2014; Yan et al., 2016; Alkhodair et al., 2020; Gao et al., 2021). For example, Alkhodair et al. (2020) and Gao et al. (2021) define the popularity score through the total count of engagements likes, shares, and comments received by the post on Twitter. These engagement attributes are made publicly available via the Twitter API. Gao et al. (2021) showed that the number of reactions a post receives usually grows in early stages (within 24 hours of posting) and remains almost constant after a specific period of time (within 10 days of posting), i.e., stable stage. Similar to the previous work, we use the number of shares ( $E_1$ ), replies ( $E_2$ ) and reports ( $E_3$ ) at the stable stage (i.e., at the time we collected the data) as indicators of rumor popularity. More formally, popularity  $Y_i$  of a given false rumor  $X_i$  is defined as:

$$Y_i = \ln[(E_1 + E_2 + E_3^2) + \lambda] \quad (5.1)$$

where  $E_1$ ,  $E_2$ , and  $E_3$  denote the number of the shares, replies, and reports of the rumor  $X_i$ ;  $\lambda$  is set to 1 so that the log function always yields a positive value. Note that we give  $E_3$  a higher weight<sup>7</sup> than  $E_1$  and  $E_2$ . For a given rumor, we assume that if the fact-checking platform receives more reports, this indicates that the rumor has already received a lot of

<sup>7</sup>We believe that some users are unsure about the veracity of suspicious rumors. Therefore, they report them and ask the official Weibo fact-checking platform for fact-checking information (see Figure 1 for the pipeline of fact-checking on the Weibo platform).

attention and more users might be unsure about its credibility so they request for it to be fact-checked.

In our initial data exploration, we observed that the number of likes for rumors prior to 2014 was zero, as the ‘Like a Post’ feature on Weibo was introduced in 2014. For consistency, we do not consider the number of likes when measuring the popularity of rumors given that our dataset contains rumors dating back to 2012. Moreover, the number of views on source posts is another metric that defines popularity scores in social media, especially on YouTube (Pinto et al., 2013; Kong et al., 2018). However, we do not consider it in our paper, as there is no access to the number of views of posts from other users through the Weibo API.

### 5.4.5 Data Pre-processing

All textual information (i.e., rumor content ( $R$ ) and user profile description ( $P$ )) are pre-processed by removing URLs and user @mention. All non-simplified Chinese characters are kept (e.g., traditional Chinese, English, Japanese, etc.) since they appear in the vocabulary list of pre-trained language models Sun et al. (2020); Cui et al. (2020). The Chinese text is segmented by using the BERT Tokenizer<sup>8</sup> from the HuggingFace library (Wolf et al., 2020). For user attributes ( $U$ ) (see Table 5.2), we normalize all numerical variables (e.g., number of friends, followers, statuses, etc.) and transform the Boolean values (e.g., Verified Status ( $U_7$ )) into integer values. Note that one can utilize visual information (e.g., images and videos) in the same task. However, we do not consider these features as some rumor cases do not contain these characteristics, or are no longer retrievable.

### 5.4.6 Dataset Description

We remove all false rumors if either the post or the user no longer exist since we need both sources of information for modeling purposes. The final dataset contains 19,256 unique rumors. Each rumor case is linked with the meta-features including (i) rumor engagement attributes ( $E$ ), (ii) rumor content ( $R$ ), (iii) user profile description ( $P$ ) and (iv) user attributes ( $U$ ). Table 5.2 displays the categories of meta-features collected via the Weibo API. All rumors

---

<sup>8</sup>[https://huggingface.co/docs/transformers/model\\_doc/bert#transformers.BertTokenizer](https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer)

and corresponding meta-features will be made publicly available for further investigation by the community.

### 5.4.7 Data Splits

We random split the rumor dataset into three subsets: (1) train (80%), (2) dev (10%) and (3) test (10%). Table 5.3 and Table 5.4 display the distribution (i.e., quantity and proportion) of the popularity score and the descriptive statistics (i.e., Mean, Median, and Max) of the number of tokens from the train, development and test sets respectively. The comparison of the statistical distributions of the three subsets showed no significant imbalance. We also notice that around 25% of the false rumors are with low popularity scores, i.e., no or few shares, which demonstrates the importance of identifying false rumors with high likelihood to become popular.

## 5.5 Experimental Setup

### 5.5.1 Predictive Models

Since this is the first work on false popularity prediction on Weibo, there is no directly comparable method. Therefore, we opt to evaluate a battery of baseline models to encode textual and user metadata that have been used in previous work on computational misinformation analysis (Rashkin et al., 2017; Alkhodair et al., 2020; Rao et al., 2021).

### 5.5.2 Rumor Content ( $R$ ) Models

To represent textual information, we evaluate three standard encoding methods (1) One-Hot Encoding (2) Pre-trained Word Vectors (Word2Vec), (3) Pre-trained Language Models.

**SVR+BOW** We first employ Support Vector Machine for Regression (SVR) (Cortes and Vapnik, 1995) with an RBF kernel using Bag-of-Words (BOW) weighted using TF-IDF. We

use a vocabulary of size 10k most frequent n-grams.

**EMB+BiLSTM** We map the text into pre-trained Chinese word embeddings<sup>9</sup> (Li et al., 2018), and then pass them through a bidirectional Long Short-Term Memory network (EMB+BiLSTM) (Hochreiter and Schmidhuber, 1997) with a self-attention mechanism. The final weighted representation is then passed through a linear layer for rumor popularity prediction.

**Pre-trained Language Models** Following Devlin et al. (2019), we directly fine-tune pre-trained transformer-based models on the popularity prediction task by feed [CLS] token representation of the last transformer layer to a linear prediction layer for regression. We evaluate the following models:

- Chinese BERT (Devlin et al., 2019), pre-trained on the Chinese Wikipedia using character-level tokenization;
- Chinese-BERT-WWM (Cui et al., 2020), an extension of the Chinese BERT model pretrained on larger corpora (e.g., news articles, Baidu Baike, etc.) using the Whole Word Masking (WWM) objective;
- Enhanced Representation Through Knowledge Integration (ERNIE) (Sun et al., 2020), pretrained using both entity-level and phrase-level masking;
- MacBERT (Cui et al., 2021), pre-trained using a text correction task with both WWM and n-gram masking methods.

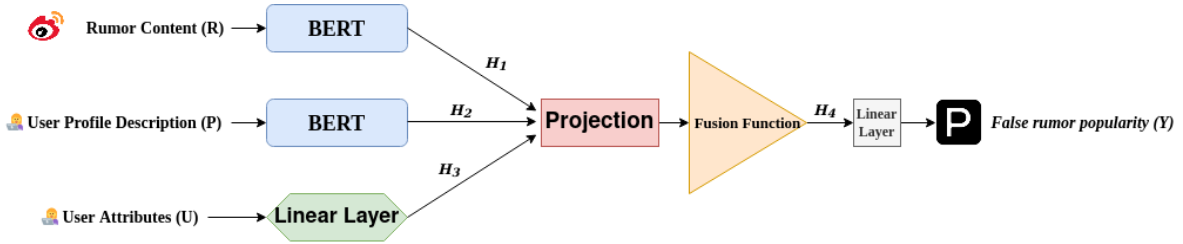
**BERT-Weibo-Rumor (Ours)** Following the task adaptive pre-training (Gururangan et al., 2020), we continually pre-train<sup>10</sup> the MacBERT (the one that achieves the best predictive performance in Table 5.5) on the (1) raw 10GB Weibo corpus and; (2) the training set of our specific rumor popularity prediction task. We first train the MacBERT checkpoint on Weibo raw data for one epoch and then further train the MacBERT model on the task-specific

---

<sup>9</sup>We use 300-dimensional Chinese Word Vectors trained on a Weibo corpus. <https://github.com/Embedding/Chinese-Word-Vectors>

<sup>10</sup>We use the open source code from Huggingface. <https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling>





**Figure 5.2:** Combining rumor content and user profile description and user attributes for rumor popularity prediction. ‘Projection’ denotes that we project  $H_1$ ,  $H_2$  and  $H_3$  into the same dimension.

training set for 40 epochs. For each epoch, we randomly mask 15% words. We then fine-tune our BERT-Weibo-Rumor model<sup>11</sup> using the same strategy as the original BERT model (Devlin et al., 2019).

### 5.5.3 Combining Rumor Text, User Profile Description and User Attributes ( $R + P + U$ )

We also propose a new model that combines rumor content ( $R$ ), user profile description ( $P$ ) and user attributes ( $U$ ). We first obtain two contextualized representations (i.e., the [CLS] token)  $H_1$  and  $H_2$  for the post itself and the user profile description respectively by passing the text through two transformer-based encoders. Here, we use our ‘BERT-Weibo-Rumor’ model that achieves the best performance using only the text from the post (see Table 5.5). The user attributes ( $U$ ) which are represented by a feature vector are projected into a 128-dimensional representation ( $H_3$ ). To obtain the final combined representation  $H_4$  of the input, we first project  $H_1$ ,  $H_2$ , and  $H_3$  into dense vectors of the same dimension and experiment with four different fusion methods:

- We directly concatenate (Concat) the representation of posts ( $H_1$ ), users’ description ( $H_2$ ) and user’ profile information ( $H_3$ ) into a single vector ( $H_4$ );
- We separately employ a mean pooling layer (Mean Pooling) and a max pooling layer (Max Pooling);

<sup>11</sup>This model will be released via the HuggingFace platform, which can be reused by the community.

- Finally, we use a self-attention mechanism (Attention) to learn a weighted combination of  $H_1$ ,  $H_2$ , and  $H_3$ .

The combined representation ( $H_4$ ) is finally passed through a fully-connected layer to obtain rumor popularity predictions using a standard mean square error (MSE) loss function.

$$\mathcal{L}_{MSE} = \sum_{i=1}^D (y_i - \hat{y}_i)^2 \quad (5.2)$$

Figure 5.2 shows the structure of the proposed neural architecture.

#### 5.5.4 Hyperparameters & Implementation Details

All model hyperparameters are tuned on the development set. We tune the regularization parameter  $C \in \{1, 1e1, 1e2, 1e3\}$  and the  $ngram \in \{(1,1), (1,2), (1,3)\}$  of the SVR, setting  $C = 1$  and  $(1,3)$ . We tune the EMB-BiLSTM *Hidden Size*  $\in \{64, 128, 256\}$  and *Dropout*  $\in \{0.2, 0.5\}$  with 256 and 0.2 perform best respectively. For all transformer models, we use the ‘base’ versions with the same architecture and the number of parameters (i.e., 12-layer, 768-dimensional, and 110M model parameters). We fine-tune all transformer based models using the implementations from the HuggingFace library (Wolf et al., 2020). We tune their learning rate range i.e.,  $lr \in \{2e-5, 3e-5, \text{and } 5e-5\}$  as in Devlin et al. (2019), setting  $lr = 5e-5$  for ERNIE,  $lr = 3e-5$  for MacBERT and  $lr = 2e-5$  for the rest of the models. The input sequence length of the post and user description are set to 256 and 64 covering the maximum length of 99% of all false rumors cases in our dataset (see Table 5.4). For the fusion network (see Figure 5.2), we finetune all the model parameters including the two different BERT encoders for the rumor content ( $R$ ) and user profile description ( $P$ ). We use a batch size of 32 for transformer-based models and 128 for the EMB-BiLSTM. All neural networks models are trained by minimizing the Mean Squared Error (MSE) loss using the Adam optimizer (Kingma and Ba, 2015) on a single Nvidia V100 GPU.

### 5.5.5 Weak Baselines

For reference, we also use the mean and median of the popularity scores in the training set as the predicted values of all instances in the test set (i.e., weak baselines).

### 5.5.6 Model Training and Evaluation Metrics

We train all of our models three times by performing hyperparameter tuning on the development set using different random seeds. We evaluate model performance using three standard metrics to measure the difference between the actual ( $x$ ) and predicted ( $y$ ) popularity values on the test set. We report the average (mean  $\pm$  standard deviation across the three runs).

- (i) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5.3)$$

- (ii) Mean Absolutely Error (MAE):

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5.4)$$

- (iii) Pearson correlation coefficient (Pearson's  $r$ ):

$$Pearson's\ r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (5.5)$$

### 5.5.7 Results

Table 5.5 shows the results obtained by all models on the false rumor popularity prediction task. We first observe that all models perform substantially better than the two weak baselines (i.e. assigning to all test instances the Mean and Median popularity computed in the training data).

**Table 5.5:** Average performance (RMSE, MAE, and Pearson’s  $r$ ) for the task of rumor popularity prediction. ‡ denotes that the Concat model performs significantly better than MacBERT (t-test;  $p < .05$ ).

| Model                                 | RMSE                              | Pearson’s $r$                      | MAE                               |
|---------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|
| <b>Weak Baselines</b>                 |                                   |                                    |                                   |
| Mean                                  | $1.98 \pm 0.0$                    | $0.0 \pm 0.0$                      | $1.56 \pm 0.0$                    |
| Median                                | $2.12 \pm 0.0$                    | $0.0 \pm 0.0$                      | $1.44 \pm 0.0$                    |
| <b>Rumor Content (<math>R</math>)</b> |                                   |                                    |                                   |
| SVR+Post                              | $1.62 \pm 0.0$                    | $0.594 \pm 0.0$                    | $1.13 \pm 0.0$                    |
| EMB+BiLSTM                            | $1.67 \pm 0.01$                   | $0.539 \pm 0.01$                   | $1.22 \pm 0.02$                   |
| Chinese BERT                          | $1.59 \pm 0.03$                   | $0.599 \pm .019$                   | $1.15 \pm 0.03$                   |
| Chinese-BERT-WWM                      | $1.60 \pm 0.02$                   | $0.598 \pm .013$                   | $1.13 \pm 0.01$                   |
| ERNIE                                 | $1.61 \pm 0.01$                   | $0.585 \pm .001$                   | $1.14 \pm 0.02$                   |
| MacBERT                               | $1.58 \pm 0.00$                   | $0.605 \pm 0.02$                   | $1.13 \pm 0.02$                   |
| BERT-Weibo-Rumor (Ours)               | $1.57 \pm 0.01$                   | $0.610 \pm 0.02$                   | $1.11 \pm 0.01$                   |
| <b><math>R + P + U</math></b>         |                                   |                                    |                                   |
| Concat ‡                              | <b><math>1.55 \pm 0.01</math></b> | <b><math>0.633 \pm .004</math></b> | <b><math>1.10 \pm 0.01</math></b> |
| Max Pooling                           | $1.56 \pm 0.01$                   | $0.625 \pm 0.007$                  | $1.12 \pm 0.02$                   |
| Mean Pooling                          | $1.57 \pm 0.02$                   | $0.621 \pm 0.006$                  | $1.12 \pm 0.02$                   |
| Attention                             | $1.55 \pm 0.03$                   | $0.630 \pm 0.004$                  | $1.11 \pm 0.01$                   |

Our proposed neural fusion model (i.e.,  $R + P + U$  Concat) achieves the lowest RMSE score (1.55), and highest Pearson’s  $r$  correlation (0.633) surpassing all the other models. The best averaged MAE (1.11) is obtained by the Max Pooling and Concat models. Moreover, the Concat model performs significantly better (t-test;  $p < .05$ ) than the best transformer model ‘BERT-Weibo-Rumor’ (i.e., RMSE 1.57, Pearson’s  $r$  0.610 and MAE=1.11) fine-tuned using only text from the post. This demonstrates that user-related information is complementary to the content of a false rumor for inferring its popularity score. The Concat model performs the best RMSE (1.55) and Pearson’s  $r$  (0.633) than the other three fusion methods. This indicates that the high-dimensional representation obtained by concatenating  $H_1$ ,  $H_2$  and  $H_3$  (see Figure 5.2) is more informative than the low-dimensional representations from Attention, Max and Mean Pooling.

In general, the majority of the post-level transformer models (i.e., Chinese BERT, Chinese BERT-WWM, MacBERT) using the rumor’s text as input achieve similar performance with the exception of ERNIE (i.e., RMSE 1.61, Pearson’s  $r$  0.585 and MAE 1.14). Our BERT-

Weibo-Rumor achieves the RMSE (1.57), Pearson’s r (0.610), and MAE (1.11) overall slightly surpassing all other post-level transformer models. Finally, we observe that the two models that are trained from scratch i.e., SVR+BOW (RMSE 1.62, Pearson’s r 0.594 and MAE 1.16) and EMB+BiLSTM (RMSE 1.67, Pearson’s r 0.539 and MAE 1.14) achieve poorer RMSE and Pearson’s r than all transformer-based models except the MAE. These two simpler models have a significantly lower number of parameters and simpler structures than the BERT-style model, suggesting that competitive results can be achieved with models that do not require high computational resources.

## 5.6 Analysis

### 5.6.1 Ablation Study

We perform an ablation study to explore the predictive power of different feature combinations, i.e., rumor content ( $R$ ), user profile description ( $P$ ), and user attributes ( $U$ ). We evaluate five variants: (1) rumor content and user profile description ( $R + P$ ), (2) rumor content and user attributes ( $R + U$ ), (3) user profile description and user attributes ( $P + U$ ), (4) user profile description only ( $P$ ), and (5) user attributes only ( $U$ ). Besides, we also employ a linear model (i.e., Ridge Regression) to test the predictive performance of each user attribute from  $U_1$  to  $U_{12}$  except the ‘Verified Status’ (i.e., Boolean Value). To make a fair comparison, we test these combinations by using the same experimental setup (i.e., running it three times with different seeds). Table 5.6 shows the average performance (RMSE, MAE, and Pearson’s r).

We first observe that  $R+P$  (RMSE 1.56, Pearson’s r 0.620 and MAE 1.13) and  $R+U$  (RMSE 1.56, Pearson’s r 0.625 and MAE 1.12) have better predictive performance compared to the BERT-style models that use only  $R$ . This suggests that solely rumor content  $R$  contains limited information in terms of inferring its future popularity. The remaining three variants (i.e.,  $P + U$ ,  $P$ , and  $U$ ) without using the rumor content ( $R$ ) perform worse than SVR-BOW and EMB-BiLSTM (see Table 5.5, suggesting that the textual information of the rumor plays the most important role in predicting its future popularity. Given that the results for all single user attributes used are only just higher than the two weak baseline models (i.e., mean and median), we can infer that individual user attributes are not sufficiently informative in predicting the popularity of false rumors on Weibo.

**Table 5.6:** Average performance (RMSE, MAE, and Pearson’s r) for the ablation study of rumor popularity prediction. All Pearson’s r values are statistically significant ( $p < .001$ ). We list the results of the Concat model (i.e.,  $R+P+U$ ) at the top for reference.

| Model                    | RMSE                              | Pearson’s r                         | MAE                               |
|--------------------------|-----------------------------------|-------------------------------------|-----------------------------------|
| <b>Input</b>             |                                   |                                     |                                   |
| $R + P + U$              | <b><math>1.55 \pm 0.01</math></b> | <b><math>0.629 \pm 0.004</math></b> | <b><math>1.11 \pm 0.00</math></b> |
| $R + P$                  | $1.59 \pm 0.00$                   | $0.606 \pm 0.002$                   | $1.14 \pm 0.01$                   |
| $R + U$                  | $1.59 \pm 0.01$                   | $0.608 \pm 0.005$                   | $1.13 \pm 0.00$                   |
| $P + U$                  | $1.76 \pm 0.01$                   | $0.456 \pm 0.005$                   | $1.29 \pm 0.02$                   |
| $P$                      | $1.81 \pm 0.01$                   | $0.421 \pm 0.005$                   | $1.33 \pm 0.02$                   |
| $U$                      | $1.81 \pm 0.0$                    | $0.470 \pm 0.0$                     | $1.22 \pm 0.0$                    |
| <b>Linear Regression</b> |                                   |                                     |                                   |
| $U_1$                    | $1.91 \pm 0.0$                    | $0.278 \pm 0.0$                     | $1.49 \pm 0.0$                    |
| $U_2$                    | $1.96 \pm 0.0$                    | $0.143 \pm 0.0$                     | $1.52 \pm 0.0$                    |
| $U_3$                    | $1.92 \pm 0.0$                    | $0.223 \pm 0.0$                     | $1.49 \pm 0.0$                    |
| $U_4$                    | $1.92 \pm 0.0$                    | $0.309 \pm 0.0$                     | $1.50 \pm 0.0$                    |
| $U_5$                    | $1.97 \pm 0.0$                    | $0.102 \pm 0.0$                     | $1.55 \pm 0.0$                    |
| $U_6$                    | $1.98 \pm 0.0$                    | $0.016 \pm 0.0$                     | $1.56 \pm 0.0$                    |
| $U_8$                    | $1.95 \pm 0.0$                    | $0.168 \pm 0.0$                     | $1.53 \pm 0.0$                    |
| $U_9$                    | $1.95 \pm 0.0$                    | $0.164 \pm 0.0$                     | $1.53 \pm 0.0$                    |
| $U_{10}$                 | $1.94 \pm 0.0$                    | $0.182 \pm 0.0$                     | $1.52 \pm 0.0$                    |
| $U_{11}$                 | $1.97 \pm 0.0$                    | $0.098 \pm 0.0$                     | $1.55 \pm 0.0$                    |
| $U_{12}$                 | $1.95 \pm 0.0$                    | $0.187 \pm 0.0$                     | $1.52 \pm 0.0$                    |

Overall, all variants are inferior to the best  $R+P+U$  Concat model, which suggests that rumor content and user information are complementary to each other. Finally, linear regression models using individual user attributes  $U_1-U_2$  as input yield results that are close to the mean and median baselines.

### 5.6.2 Qualitative Analysis of Model Predictions

To uncover the main limitations of our best model (i.e.,  $R + P + U$  Concat), we perform an error analysis of false rumor cases where the model predicted a low popularity score for highly popular rumors (Cases Low 1, 2, 3) and vice versa (Cases High 1, 2, 3). Moreover, we analyze two cases where the model correctly predicted a popularity score almost identical

**Table 5.7:** Examples of prediction actual and predicted popularity scores made by our best performing  $R+P+U$  Concat model. For each example, we list the original Chinese false rumor  $R$ , its English Translation, and a link to the corresponding fact-checking page. Note that Weibo requires users to log in to access its fact-checking platform.

| False Rumors Content (in Chinese) and Explanations (in English) |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Pred. | Truth | Diff. |
|-----------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|-------|
| False rumors are incorrectly predicted to be low popularity     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |       |       |       |
| Low-1                                                           | <p>杨澜昨天终于承认了自己的美国籍身份。她理直气壮地说：“虽然我入了美国籍，但我出身于中国，所以从原产地角度而言，我不出席美国的两会而出席中国的两会是天经地义的” ...</p> <p>A false rumor about an official Chinese media host (named ‘杨澜’) has taken U.S. citizenship...</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaJ6wph6Kog">https://service.account.weibo.com/show?rid=K1CaJ6wph6Kog</a></p> <p>六小龄童，昨天因病去世，送“猴哥”最后一程...</p>                                                                                                 | 2.76  | 7.64  | -4.88 |
| Low-2                                                           | <p>A false rumor about the death of a famous actor (named 六小龄童)</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1Ca06Axk7ac1">https://service.account.weibo.com/show?rid=K1Ca06Axk7ac1</a></p> <p>必转！今天一个河北香河的朋友，香河华联超市前拍的照片，这个孩子一看就是被拐卖的！</p>                                                                                                                                                                                                             | 1.59  | 8.05  | -6.46 |
| Low-3                                                           | <p>Please share! This is a photo taken today by a friend in front of a supermarket in Xianghe, Hebei, this child is obviously being trafficked! (This is a false rumor about missing people.)</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaL7Ax166s">https://service.account.weibo.com/show?rid=K1CaL7Ax166s</a></p>                                                                                                                                 | 4.33  | 10.12 | -5.79 |
| False rumors are incorrectly predicted to be high popularity    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |       |       |       |
| High-1                                                          | <p>日本首相安倍晋三正式宣布辞职...</p> <p>Japanese Prime Minister Shinzo Abe officially announced his resignation...</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaN6Qtg66gk">https://service.account.weibo.com/show?rid=K1CaN6Qtg66gk</a></p> <p>求助大家帮忙，黄傲雪、7岁、身高，1.2米左右！3月11日中午12:30广州汇豪天下附近 丢失！求大家转发帮忙寻找...</p>                                                                                                                                                  | 4.36  | 0.69  | 3.67  |
| High-2                                                          | <p>Please help, [Girl’s Name], 7 years old, about 1.2 meters tall! She disappeared on March 11 at 12:30 p.m. in the [Location]! Please forward to help find this girl...</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaJ6gxc8aw1">https://service.account.weibo.com/show?rid=K1CaJ6gxc8aw1</a></p> <p>#成都提个醒#[太活跃的鱼千万别买] 去买鱼，结果看到摊贩往水盆内加入一种白色粉末，迅速用手搅拌，一会功夫白色粉末溶解，将半死不活的鱼虾倒入其中，一会儿就活蹦乱跳开，仿佛刚从河中捕回来的。这是一种能够致癌的催化剂，俗称鱼浮灵，也对智力有影响。相互转告一下，有必要让更多的人知道！</p> | 4.05  | 0.69  | 3.36  |
| High-3                                                          | <p>Do not buy live fish that are too active in the supermarket because they contain artificial additives, such as some carcinogens. [Emoji] Share it with each other, it is necessary to let more people know!</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaJ6wtc764d">https://service.account.weibo.com/show?rid=K1CaJ6wtc764d</a></p>                                                                                                              | 5.86  | 2.07  | 3.79  |
| False rumors that can be accurately predicted by the model      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |       |       |       |
| Acc-1                                                           | <p>一妇女喝了罐饮料，被送进医院，离开了世界。验尸死于於细螺旋体病，追踪她喝的饮料，是直接罐对嘴饮用。实验证明罐头受到鼠尿感染细螺旋体病毒。</p> <p>A woman died after drinking a can of drink contaminated with bacteria carried by rats.</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaJ6wt18qkj">https://service.account.weibo.com/show?rid=K1CaJ6wt18qkj</a></p> <p>Mr.Bean自杀了[泪]I love you, Mr.Bean![伤心][鲜花][蜡烛]</p>                                                                                                  | 4.56  | 4.70  | -0.14 |
| Acc-2                                                           | <p>Mr. Bean committed suicide. [Emoji] I love you Mr.Bean! [Emoji]</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaK6wNk8qwj">https://service.account.weibo.com/show?rid=K1CaK6wNk8qwj</a></p> <p>太恐怖了，我以后一定要拔掉！看看！！这个小女孩不幸死于电话充电器，就是因为大人平时充完电后没有及时把充电器插头部位拔出，小女孩拿起充电器另一头来玩含在嘴里不幸触电身亡..</p>                                                                                                                                                             | 4.78  | 4.64  | 0.14  |
| Acc-3                                                           | <p>Look! This little girl died tragically due to the smartphone charger. Her parents didn’t unplug the phone from charging, and the little girl picked up the charging head and put it in her mouth to play with it leading to her death.</p> <p><b>Fact-checking link:</b> <a href="https://service.account.weibo.com/show?rid=K1CaM6g1f8akh">https://service.account.weibo.com/show?rid=K1CaM6g1f8akh</a></p>                                                                                   | 1.02  | 1.09  | -0.07 |

to the actual score (Cases Acc 1, 2, 3). These cases are related to the most common topics discussed on Weibo (i.e., ‘Politics’, ‘Social Life’ and ‘Scientific’) (Liu et al., 2015). The false rumor cases (i.e., rumor content  $R$ , English translation, and fact-checking hyper-link) together with the actual and predicted popularity scores are shown in Table 5.7.

**Cases Low-1, High-1** We first observe that our model has difficulty in accurately predicting the popularity of false rumors related to politics (see Case Low-1 and Case High-3). Given that both cases were posted by unverified users who have a limited number of followers. However, Case Low-1 eventually became a highly popular false rumor, but Case High-3 received no engagement. This may be due to the fact that the content of the rumor in Case Low-1 is related to populism, which has been on the rise in the past decade on Weibo (Zhang et al., 2018; Zhang, 2020). Such rumors may attract a lot of engagement by other users.

**Cases Low-3, High-2** Cases Low-3 and High-2 are both false rumors related to missing people that were published in 2013 and 2014, respectively. In 2012, Professor Yu Jianrong (a famous Chinese sociologist), launched an event called ‘Saving children that beg on the streets’, calling for people to post and share information about begging children on Weibo to help them find their families (Zhang and Negro, 2013). Since then, Weibo has been widely used to find missing people in case of social emergencies, and many related rumors have emerged (Shan et al., 2019; Liu et al., 2020).

We observe that our model can identify Case Low-3 as a highly popular rumor, but it cannot accurately predict the popularity score (i.e., 10.12). We further explore the corresponding fact-checking link provided in Case Low-3 and discover that the false rumor (with 19k shares and 5k replies) was posted by a famous actor with millions of followers. The first three tokens of the rumor content is ‘必转!’ (‘i.e., Please Share!’) which reflects the fact that celebrities have high influence on the popularity of false rumors. Moreover, compared against other topics of false rumors such as ‘Politics’ and ‘Science’, the missing persons false rumors tend to be more difficult to verify as true or false in a short period of time after they are posted, as they usually require investigation by the authorities.

**Cases Acc-1, Acc-3** The last two cases are related to ‘Science’ and are accurately predicted by our model. We select accurate prediction cases by considering an absolute error less than 0.15, which is the 10-th percentile in the test set. These two cases are quite different (see Fact-checking links in Table 5.7). Case Acc-1 was posted by a verified user and has become highly popular. On the contrary, case Acc-3 was posted by an average Weibo user and has very low popularity (i.e, no shares and replies). This shows that our best model performs well in learning textual information and user attributes of false rumors about general life knowledge and junk science.



**Table 5.8:** Top 5 LIWC features associated with rumor popularity sorted by Pearson’s correlation ( $r$ ) between the normalized LIWC features frequency and the popularity scores of rumors ( $p < 0.001$ ). A positive  $r$  indicates a positive correlation, and vice versa.

| Rumor Content ( $R$ ) |               | User Profile Description ( $P$ ) |               |
|-----------------------|---------------|----------------------------------|---------------|
| LIWC Category         | Pearson’s $r$ | LIWC Category                    | Pearson’s $r$ |
| <b>conj</b>           | 0.153         | <b>WC</b>                        | 0.175         |
| <b>achieve</b>        | 0.121         | <b>WPS</b>                       | 0.166         |
| <b>cause</b>          | 0.112         | <b>work</b>                      | 0.110         |
| <b>AllPunc</b>        | 0.104         | <b>affiliation</b>               | 0.090         |
| <b>OtherP</b>         | 0.097         | <b>drives</b>                    | 0.082         |
| LIWC Category         | Pearson’s $r$ | LIWC Category                    | Pearson’s $r$ |
| <b>affiliation</b>    | -0.181        | <b>i</b>                         | -0.044        |
| <b>family</b>         | -0.180        | <b>ppron</b>                     | -0.041        |
| <b>social</b>         | -0.142        | <b>pronoun</b>                   | -0.037        |
| <b>male</b>           | -0.133        | <b>female</b>                    | -0.031        |
| <b>prep</b>           | -0.121        | <b>reward</b>                    | -0.029        |

### 5.6.3 Characterizing Highly Popular False Rumors

**Linguistic Analysis** Rumors with high popularity typically use stylistic features and sentiment to attract users’ attention (Alkhodair et al., 2020). We perform a linguistic analysis to uncover differences in linguistic patterns used between high and low popular false rumors. To this end, we use a standard psycho-linguistic analysis method, i.e. Linguistic Inquiry and Word Count<sup>12</sup> (LIWC) (Pennebaker et al., 2001), to represent the textual information (i.e., rumor content  $R$  and user profile description  $P$ ) into 95 different psycho-linguistic categories. Table 5.8 shows the univariate Pearson’s correlation test results between the popularity scores of rumors and LIWC features following Schwartz et al. (2013).

For the rumor content ( $R$ ), we observe that LIWC categories such as **Conjunctions** (e.g. and, but, etc.), **Cause** (e.g. because, effect, etc.), and **Achievement** (e.g. win, succeed, better, etc.) are most positively associated with false rumors of high popularity. In addition, rumor content of false rumors with high popularity contain more punctuation marks, i.e. **AllPunc** (all types of punctuation) and **OtherP** (other uncommon punctuation). We sample some cases with high popularity discovering that punctuation can be used as ‘Emoticon’ (e.g.,

<sup>12</sup>We use a Chinese LIWC version developed by (Huang et al., 2012) - <https://cliwceg.weebly.com/>

‘: (’, ‘@\_@’, ‘:P’ ) to express emotions. Similarly, these ‘Emoticons’ have been used to detect Twitter rumors with high engagement rate (Alkhodair et al., 2020).

Besides, we observe some LIWC categories related to emotional expression (e.g. **anger** and **negemo**) are positively correlated with high popularity Weibo rumors. They are also high-frequency words found in false Weibo rumors that can be detected early Song et al. (2019). Note that some common Emoticons (e.g., ‘: )’ and ‘: (’) also belong to the emotion categories in the LIWC dictionary. On the other hand, LIWC categories such as social environment related words (e.g., **social**, **family**, and **male referents** are more common in false rumors with lower popularity scores.

For user profile descriptions ( $P$ ), the LIWC categories **Words Count** and **Words per Sentence** show that Weibo users with longer descriptions are likely to share false rumors with high popularity. By exploring descriptive statistics of our dataset, we discover verified users usually have longer descriptions introducing themselves (i.e., average 34 tokens) than unverified users (i.e., average 15 tokens). These verified users (i.e., Verified Status  $U_7$ ) are also found to have a higher probability of spreading high popularity rumors in the future (see Table 5.9).

The analysis also shows that LIWC categories such as **Work** and **Affiliation** are positively correlated with high popularity rumors, which is the opposite of the negatively correlated LIWC categories discovered in the post. On the other hand, we observe that most negatively correlated LIWC categories (e.g., **i**, **Personal pronouns**, **Total pronouns**) are pronoun-related, however, they do not have high Pearson’s  $r$  values as they are common words in Weibo user profile descriptions.

**User Attributes ( $U$ )** Table 5.9 displays the sorted Pearson’s correlation  $r$  between user attributes (from  $U_1$  to  $U_{12}$ ) and popularity scores of false rumors ( $p < 0.001$ ). We first observe that all user profile attributes are positively correlated with the prevalence of rumors except the ‘Credit Score’ ( $U_6$ ),. This suggests that the Weibo Credit Score ( $U_6$ ) is actually a good indicator of user credibility. Thus posts from users with low Credit Scores may need to be prioritized for debunking.

We also observe that the Verified Status’ ( $U_7$ ) of user accounts has the highest Pearson’s correlation  $r$ , suggesting that false rumors posted by verified Weibo accounts are more likely to receive a larger number of reactions in the future. In social networks, verified accounts are

**Table 5.9:** Pearson’s Correlation  $r$  between the user attributes ( $U$ ) and the future popularity of Rumors ( $p < 0.001$ ), sorted in descending order.

|          | User Attributes ( $U$ )        | Pearson’s $r$ |
|----------|--------------------------------|---------------|
| $U_7$    | Verified Status                | 0.380         |
| $U_1$    | # of Followers                 | 0.255         |
| $U_4$    | # of Statuses                  | 0.254         |
| $U_3$    | # of Bi_Followers              | 0.243         |
| $U_{10}$ | # of Replies Received          | 0.214         |
| $U_{12}$ | # of All Reactions Received    | 0.213         |
| $U_8$    | # of shares Received           | 0.203         |
| $U_9$    | # of Likes Received            | 0.196         |
| $U_2$    | # of Followees                 | 0.181         |
| $U_{11}$ | # of Likes Received in Replies | 0.108         |
| $U_5$    | # of Favorites                 | 0.099         |
| $U_6$    | Credit Score                   | -0.057        |

generally considered more credible than average users, and these verified users are significantly more visible in online debates in case of public events (e.g., political events) (Hentschel et al., 2014; González-Bailón and De Domenico, 2021). Prior research (Liu et al., 2015) demonstrated that user Verified status ( $U_7$ ) and number of followers ( $U_1$ ) can be used as a proxy for the trustworthiness of the Weibo users. Similar to our dataset, high-popularity rumors are more likely to be shared by users with a larger number of followers ( $U_1$ ) and bi\_followers’ ( $U_3$ ) as social media users are more inclined to trust posts that were shared by their friends rather than strangers (Margolin et al., 2018). Other positively correlated factors are the reactions including the number of shares ( $U_8$ ), replies ( $U_{10}$ ), likes ( $U_9$ ) that users receive, which reflects the number of interactions they have in the social network.

## 5.7 Implications and Ethics Considerations of Our Study

In this section, we introduce the ethics considerations, theoretical and practical implications of our research.

### 5.7.1 Theoretical Implications

The theoretical implications of this work are as follows:

- We define a novel task of predicting future popularity of false rumors which has not been addressed in previous work. We introduce a new direction, and our task can be extended in multilingual and multi-platform settings.
- We provide extensive analyses (see Section 5.6) including qualitative analysis, psycholinguistic analysis (via LIWC), and user attributes analysis which can be used by social scientists and psychologists to complement studies on analyzing the characteristics of false rumors with high impact (Bronstein et al., 2019; Pennycook and Rand, 2019, 2021).

### 5.7.2 Practical implications

We believe that our work has several potential practical implications:

- First, our new dataset (including meta-features), pre-trained language model (i.e., Weibo-BERT-Rumor), and rumor popularity prediction system can be easily re-purposed by fact-checking platforms, professional journalists, researchers, and social media companies. Note that these resources will be released via user-friendly platforms such as HuggingFace and Github.
- Besides, our open source fusion network takes into account both post and user level meta-features to achieve the best predictive performance, and can be used as a strong baseline model for further research.
- Finally, our system can be combined with existing rumor detection models. For example, in some cases, social media platforms can obtain potential impact immediately upon discovery of a **false rumour**, which can prevent the spread of high-impact malicious posts at an early stage.

### 5.7.3 Ethics Considerations

Our work complies with Weibo’s API policy and has received approval from the Ethics Committee of our institution (Reference Number: 025470). Note that we have also submitted our research proposal to Weibo since we had to apply for the permission for accessing the Weibo official API. All false rumors were debunked and made public by the Weibo fact-checking platform. After anonymization, we will share the dataset splits and source code for research purposes.

## 5.8 Conclusion and Future Work

This paper presented the first study on future popularity prediction of false rumors on Weibo, based on both post and user-level information. This task is important for the timely detection of high-popularity rumors and complements existing methods for early rumor detection. A key contribution is a new Weibo dataset which includes 19,256 cases of false rumors and their associated popularity score, which is based on the engagement received. To predict the popularity of false rumors, we train a neural model that combines information from the rumor content, user profile description and user attributes, which outperforms strong baselines. Our proposed models and follow-up analysis would enable the prioritization of rumors for moderation and debunking, as well as be beneficial in computational linguistics for analyzing the main characteristics of popular false rumors.

However, we acknowledge that our current contributions, such as our new dataset and the Bert-Weibo-Rumor model, are limited to a mono-lingual setup. Furthermore, we believe that conducting further experiments on feature engineering, such as handcrafting features based on the rumor content, can help improve the model predictive performance. In addition, a detailed analysis of the hidden characteristics of rumor content can be used to study the features of rumors with high impact. In the future, we plan to extend this work towards studying the popularity of false rumors on different social media platforms and in a multi-lingual setting.

---

## BIBLIOGRAPHY

- Sarah A Alkhodair, Benjamin CM Fung, Steven HH Ding, William K Cheung, and Shih-Chia Huang. 2020. Detecting high-engaging breaking news rumors in social media. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–16.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.
- Jackie Ayoub, X Jessie Yang, and Feng Zhou. 2021. Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4):102569.
- Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. 2013. Popularity prediction in microblogging network: a case study on Sina Weibo. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 177–178.
- Parisa Bazmi, Masoud Asadpour, and Azadeh Shakery. 2023. Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Information Processing & Management*, 60(1):103146.
- Michael V Bronstein, Gordon Pennycook, Adam Bear, David G Rand, and Tyrone D Cannon. 2019. Belief in fake news is associated with delusionality, dogmatism, religious fundamental-

- ism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1):108–117.
- Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.
- Xueqin Chen, Fan Zhou, Fengli Zhang, and Marcello Bonsangue. 2021. Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning. *Information Processing & Management*, 58(5):102678.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Effective and effortless features for popularity prediction in microblogging network. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 269–270.
- Xiaofeng Gao, Zuowu Zheng, Quanquan Chu, Shaojie Tang, Guihai Chen, and Qianni Deng. 2021. Popularity prediction for single tweet based on heterogeneous bass model. *IEEE Transactions on Knowledge & Data Engineering*, 33(05):2165–2178.
- Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 907–910.

- Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. Identifying and understanding user reactions to deceptive and trusted social news sources. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–181.
- Sandra González-Bailón and Manlio De Domenico. 2021. Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences*, 118(11).
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Martin Hentschel, Omar Alonso, Scott Counts, and Vasileios Kandylas. 2014. Finding users we trust: Scaling up verified twitter users using their communication patterns. In *Eighth International AAAI Conference on Weblogs and Social Media*, volume 8, pages 591–594.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.
- Chin-Lan Huang, Cindy Chung, N. Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben Lam, and James Pennebaker. 2012. Development of the chinese linguistic inquiry and word count dictionary. *Chinese J Psychol*, 54:185–201.
- Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38.
- Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.



- Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Elena Kochkina, Tamanna Hossain, Robert L Logan IV, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116.
- Quyu Kong, Marian-Andrei Rizoiu, Siqi Wu, and Lexing Xie. 2018. Will this video go viral: Explaining and predicting the popularity of youtube videos. In *Companion Proceedings of the The Web Conference 2018*, pages 175–178.
- Vasileios Lampos, Nikolaos Aletras, Daniel Preoțiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on twitter. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413.
- Liuwu Li, Runwei Situ, Junyan Gao, Zhenguo Yang, and Wenyin Liu. 2017. A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1912–1917.
- Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.
- Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047.
- Yang Liu, Osamu Uchida, and Keisuke Utsu. 2020. A proposal on disaster information and rescue request sharing application using Sina Weibo. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pages 419–423. IEEE.

- Zhiyuan Liu, Le Zhang, CunChao Tu, and MaoSong Sun. 2015. Statistical and semantic analysis of rumors in chinese social media. *Scientia Sinica Informationis*, 45(12):1536–1546.
- Heng-yang Lu, Chenyou Fan, Xiaoning Song, and Wei Fang. 2021. A novel few-shot learning based multi-modality fusion model for covid-19 rumor detection from online social media. *PeerJ Computer Science*, 7:e688.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717.
- Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, 35(2):196–219.
- Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. ” nobody comes here anymore, it’s too crowded”; predicting image popularity on flickr. In *Proceedings of International Conference on Multimedia Retrieval*, pages 385–391.
- Stuart E Middleton, Lee Middleton, and Stefano Modafferi. 2013. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17.
- Yida Mu, Pu Niu, and Nikolaos Aletras. 2022. Identifying and characterizing active citizens who refute misinformation in social media. In *14th ACM Web Science Conference, WebSci ’22*, page 401–410.
- Gabriel Peres Nobre, Carlos HG Ferreira, and Jussara M Almeida. 2022. A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*, 59(1):102757.
- Shivam B Parikh, Vikram Patil, Ravi Makawana, and Pradeep K Atrey. 2019. Towards impact scoring of fake news. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 529–533. IEEE.

- Tanja Pavleska, Andrej Školka, Bissera Zankova, Nelson Ribeiro, and Anja Bechmann. 2018. Performance analysis of fact-checking organizations and initiatives in europe: a critical overview of online platforms fighting fake news. *Social Media and Convergence*, 29.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.
- Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.
- Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 365–374.
- Alicja Piotrkowicz, Vania Dimitrova, Jahna Otterbacher, and Katja Markert. 2017. Headlines matter: Using headlines to predict the popularity of news articles on twitter and facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 656–659.
- Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 518–527. IEEE.
- Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell,

- Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.
- Siqing Shan, Feng Zhao, Yigang Wei, and Mengni Liu. 2019. Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—a case study of Weibo (Chinese twitter). *Safety Science*, 115:393–413.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.
- Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.
- John H Smith and Nathaniel D Bastian. 2022. A ranked solution for social media fact checking using epidemic spread modeling. *Information Sciences*, 589:550–563.
- Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. CED: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.
- Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2021. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge & Data Engineering*, 33(08):3035–3047.
- Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency matters: A knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1412–1423.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

- Tomasz Trzciński and Przemysław Rokita. 2017. Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia*, 19(11):2561–2570.
- Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284.
- Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344.
- Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.
- Yandong Wang, Teng Wang, Xinyue Ye, Jianqi Zhu, and Jay Lee. 2016. Using social media for emergency response and urban sustainability: A case study of the 2012 beijing rainstorm. *Sustainability*, 8(1):25.
- Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3845–3854.

- Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twiterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*, pages 261–270.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network with applications to early rumor detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9042–9051.
- Yan Yan, Zhaowei Tan, Xiaofeng Gao, Shaojie Tang, and Guihai Chen. 2016. Sth-bass: a spatial-temporal heterogeneous bass model to predict single-tweet popularity. In *International Conference on Database Systems for Advanced Applications*, pages 18–32. Springer.
- Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454.
- Tauhid Zaman, Emily B Fox, and Eric T Bradlow. 2014. A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3):1583–1611.
- D Zhang. 2020. Digital nationalism on Weibo on the 70th chinese national day. *The Journal of Communication and Media Studies*, 6(1):1–19.
- Yinxian Zhang, Jiajun Liu, and Ji-Rong Wen. 2018. Nationalism on Weibo: Towards a multifaceted understanding of chinese nationalism. *The China Quarterly*, 235:758–783.
- Zhan Zhang and Gianluigi Negro. 2013. Weibo in China: Understanding its development through communication analysis and cultural studies. *Communication, Politics & Culture*, 46(2):199–216.

- Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522.
- Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623.
- Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

# CONCLUSION

This thesis presented four studies in the field of computational misinformation analysis. In this chapter, we summarise tasks, contributions, implications of the thesis and discuss potential future directions.

## 6.1 Summary of Thesis

In this thesis, we shed light on four research aims related to combating online misinformation:

- (i) To explore whether temporal factors can affect the model's predictive performance in rumour detection tasks.
- (ii & iii) To understand user behaviours towards online misinformation, i.e., achieving the automatic detection of users who diffuse or refute unreliable news stories on social media.
- (iv) To achieve the early detection of false rumours with higher popularity.



We address these aims using a standard research pipeline and summarise our findings in four separate papers. Our main contributions in each paper are as follows:

Paper I introduces the first study on examining the impact of *temporal concept drift* on four popular rumour detection classifiers (i.e., Twitter15 & 16, Weibo and PHEME). We propose a battery of control experiments to compare two evaluation methods i.e., using chronological splits and standard random splits, respectively. Our experimental results uncover that the use of both standard and stratified chronological splits leads to a substantially degradation of model performance on four popular rumour detection datasets. We argue that the role of *temporality* needs to be considered when evaluating modern rumour detection systems.

Paper II presents a new task of early detecting Twitter users who will propagate news items from unreliable news outlets in the future (e.g., Russia Today and Infowars). To this end, we develop a new publicly available annotated dataset consisting of 6.2k Twitter users. We evaluate a battery of supersized models achieves up to 79 macro F1-score for the binary classification task. Further linguistic analysis uncover the differences between language patterns (i.e., N-grams, Word Clusters and LIWC) in the two categories of Twitter users. For example, we observe that Twitter users spread news stories from unreliable news sources are more prevalent in posting tweets about topics related to politics and religion. On the other hand, users who only disseminate news from reliable news sources use more words related to self-disclosure in their social network.

Paper III is the first extensive work on identifying and characterising active citizens who refute misinformation across two stoical media networks (i.e., Weibo and Twitter) and languages (i.e., Chinese and English). We develop a publicly available collection of 47k Weibo users annotated into the two categories, i.e., misinformation spreaders and debunkers, respectively. Our proposed hierarchical transformer network considering users' historical timeline achieves the highest macro F1 scores (85.1 and 80.2) on the Weibo and Twitter datasets, respectively. Finally, we provide an extensive linguistic analysis uncovering the main differences in language use between the two categories of social media users across two platforms.

Paper IV introduces a large scale research on predicting the future popularity of false rumours given user and post-level features. We develop a new publicly available dataset based on the Weibo official fact-checking platform and a new pretrained language model (i.e., BERT-Weibo-Rumour). Besides, we propose a fusion network (considering source posts, users profile information and user attributes) that achieves the lowest RMSE score (1.55) and best Pearson’s correlation  $r$  (0.63). Our proposed quantitative analysis unveils that some user profile attributes (such as *Verified Status* and *Number of Followers*) have the highest Pearson’s correlation  $r$  with the high-impact false rumours, showing that false rumours diffused by verified users are most likely to receive more responses (e.g., shares, replies, etc.) in the future.

In general, we make several contributions:

- (i) We uncover that the use of temporal splits results in a substantial degradation of model predictive performance on four standard rumour detection benchmarks.
- (ii & iii) We develop two new datasets and hierarchical transformer-based models for two novel tasks: automatically detecting users who spread and refute misinformation, respectively.
- (iv) Our new model (i.e., a fusion neural network using both post and user-level information) achieves the best predictive performance on our new publicly available dataset

## 6.2 Future Work

This thesis can be extended in several directions:

- Initially, we plan to further examine the impact of temporal conceptual drift (i.e., Paper I) on interpretable misinformation detection systems which are developed with multiple modalities (e.g., text, images, and videos) (Chrysostomou and Aletras, 2022; Lin et al., 2021; Rao et al., 2021). In addition, our proposed

*chronological evaluation protocol* can be applied to more misinformation detection benchmarks in multilingual and multi-platform settings (Mu et al., 2023b; Thorne et al., 2018; Wang, 2017).

- We introduce two novel tasks in Papers II and III, i.e., predicting and characterising the behaviour of users who are most likely to diffuse or debunk misinformation in social media. Future research can improve the performance of the model by exploiting *network information*, e.g., user-level connections (such as follow and retweet) and profile information (Aletras and Chamberlain, 2018; Li and Aletras, 2022; Pan et al., 2019).
- We also plan to improve Papers II and III with psycho-linguistic research methods, e.g., using a data-driven approach to analyse the relationship between individuals' behaviours (e.g., debunk or diffuse misinformation in social media) and personality traits (Bronstein et al., 2019; Lin et al., 2023; Pennycook et al., 2018; Pennycook and Rand, 2019, 2020).
- Given the pipeline for developing the Weibo dataset (i.e., future popularity of false rumours) presented in Paper IV, we plan to improve this task to predict the future impact of unreliable posts on different social media platforms (e.g., Twitter and Facebook) and to analyse the characteristics of these posts in a multilingual setting (e.g., English and Spanish) (Ma et al., 2016, 2017; Zubiaga et al., 2016).
- Finally, we consider the integration of our proposed tasks and methods in this thesis with real-world applications of automatic misinformation detection. For example, the work in Papers II, III and IV can be combined with existing rumour detection systems to not only early debunk rumours, but also to flag such false rumours with potential high-impact in the future for the Weibo platform.

---

## BIBLIOGRAPHY

- Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.
- Ahmet Aker, Kevin Vincentius, and Kalina Bontcheva. 2019. Credibility and transparency of news sources: Data collection and feature analysis. In *NewsIR@ SIGIR*, pages 15–20.
- Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, pages 20–24.
- Michael V Bronstein, Gordon Pennycook, Adam Bear, David G Rand, and Tyrone D Cannon. 2019. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of applied research in memory and cognition*, 8(1):108–117.
- Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.
- Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Trends and*

- Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*, pages 40–52. Springer.
- George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10545–10553.
- Nicholas DiFonzo and Prashant Bordia. 2011. Rumors influence: Toward a dynamic social impact theory of rumor. In *The science of social influence*, pages 271–295. Psychology Press.
- Sabine A Einwiller and Michael A Kamins. 2008. Rumor has it: The moderating effect of identification on rumor impact and the effectiveness of rumor refutation 1. *Journal of applied social psychology*, 38(9):2248–2272.
- Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia computer science*, 121:817–825.
- Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.
- Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.
- Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.
- Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2018. Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018*, pages 507–511.

- Wenzhe Li and Nikolaos Aletras. 2022. Improving graph-based text representations with character and word level n-grams. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 228–233.
- Hause Lin, Gordon Pennycook, and David G Rand. 2023. Thinking more or thinking differently? using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, 230:105312.
- Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on Twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.
- Zhiming Mao, Xingshan Zeng, and Kam-Fai Wong. 2021. Neural news recommendation with collaborative news encoding and structural user encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 46–55.
- Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6:e325.
- Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023a. It’s about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 724–731.

- Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2023b. Examining temporalities on stance detection towards covid-19 vaccination. *arXiv preprint arXiv:2304.04806*.
- Yida Mu, Pu Niu, and Nikolaos Aletras. 2022. Identifying and characterizing active citizens who refute misinformation in social media. In *14th ACM Web Science Conference 2022*, pages 401–410.
- Jiaqi Pan, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. 2019. Twitter homophily: Network based prediction of user’s occupation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2633–2638.
- Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.
- Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.
- Gordon Pennycook and David G Rand. 2020. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200.
- Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Pp-rec: News recommendation with personalized user interest and time-aware news popularity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5457–5467.
- Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking network based on level-grained attention-masked BERT for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.
- James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 275–284.
- Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *ACL*, volume 2, pages 647–653.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.



- William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.
- Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9042–9051.
- Yungeng Xie, Rui Qiao, Guosong Shao, and Hong Chen. 2017. Research on chinese social media users’ communication behaviors during public emergency events. *Telematics and Informatics*, 34(3):740–754.
- Zheng Xu, Yunhuai Liu, Junyu Xuan, Haiyan Chen, and Lin Mei. 2017. Crowdsourcing based social media data analysis of urban emergency events. *Multimedia Tools and Applications*, 76:11567–11584.
- Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.
- Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623.
- Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.
- Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

---

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.