

SHAKSHI SHARMA

Fighting Misinformation in the Digital Age:
A Comprehensive Strategy for Characterizing,
Identifying, and Mitigating Misinformation on
Online Social Media Platforms



SHAKSHI SHARMA

Fighting Misinformation in the Digital Age:
A Comprehensive Strategy for Characterizing,
Identifying, and Mitigating Misinformation on
Online Social Media Platforms



UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on August 24, 2023 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Assoc. Prof. Rajesh Sharma
Institute of Computer Science
University of Tartu, Tartu, Estonia

Opponents

Prof. Joydeep Chandra
Associate Professor
Department of Computer Science
Indian Institute of Technology (IIT), Patna, India

Prof. Kiran Garimella
Assistant Professor
School of Communication and Information
Rutgers University, USA

The public defense will take place on the 6th of October, 2023 at 17:00 in Narva mnt 18-2049.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

Copyright © 2023 by Shakshi Sharma

ISSN 2613-5906 (print)

ISSN 2806-2345 (PDF)

ISBN 978-9916-27-342-5 (print)

ISBN 978-9916-27-343-2 (PDF)

University of Tartu Press

<http://www.tyk.ee/>

To my parents

ABSTRACT

With the development of Online Social Media (OSM) platforms like Twitter and Facebook, the way individuals connect and exchange information has changed dramatically. These platforms have, however, also led to the problem of misinformation, which can quickly spread on a global scale, and once it has been disseminated, it can be hard to stop. It is to be noted that misinformation is an umbrella term that can include rumors, fake news, etc. This thesis examines the multifaceted approach for combating misinformation in the digital era, focusing on three key dimensions: identifying misinformation content, developing a framework for identifying misinformation spreaders, and carrying out effective counter misinformation measures.

First, our proposed *characterization* approach aims to understand the traits of misinformation posts to identify cognitive behavior and motivations behind spreading misinformation. Scrutinizing social media posts' characteristics helps the research community detect and prevent misinformation. Our findings suggest that the misinformation and non-misinformation posts are linguistically different. Thus, these linguistic differences can be utilized to provide efficient automated methods for identifying and limiting the dissemination of misinformation on social media platforms.

Second, previous techniques for detecting suspicious or malicious users and identifying misinformation on platforms like Twitter have not considered user-level detection adequately. Categorizing a user as a misinformation spreader based on one post is insufficient. In this dimension, we develop an automated classification framework for *misinformation detection* that combines multiple posts and network information achieving up to 0.864 F1-Score (the highest among the other baseline models).

Third, existing approaches for mitigating the spread of misinformation on social media have limitations, such as the lack of external moderation and reliance on strict assumptions. We propose a *plan of action* approach which is an automated rebuttal of misinformation at scale by harnessing social media data and curated fact-checked data repositories. Particularly, this dimension focuses on the Twitter platform and COVID-19 misinformation, exploring two complementary approaches. It is a more practical approach to follow in the identification of misinformation posts that can identify the posts at an early stage as well as the later stage of misinformation spread compared to the existing approaches that can only provide solutions either at an early stage or at the later stage of the misinformation spread but not both. Thus, limiting the efficacy of previous approaches in preventing the spread of misinformation.

CONTENTS

List of original publications	14
1. Introduction	16
1.1. A Comprehensive Study on the Characterization of Misinformation Content	17
1.2. The Sentinel System: Framework for Misinformation Spreaders .	18
1.3. Plan of Action to Combat the Misinformation Menace	19
2. Literature Review	20
2.1. Misinformation Characterization	20
2.2. Misinformation Spreaders	22
2.3. Misinformation Mitigation Strategies	23
3. A Comprehensive Study on the Characterization of Misinformation Content	26
3.1. Dataset	27
3.2. Methodology	28
3.2.1. Pre-processing and Feature Extraction	28
3.2.2. Statistical Test and Machine Learning	29
3.3. Analyses	30
3.3.1. LIWC	30
3.3.2. Readability	34
3.3.3. SenticNet	35
3.3.4. Emotions	38
3.4. Prediction and Explainability	38
3.5. General Discussion	42
3.6. Summary	44
4. The Sentinel System: Framework for Misinformation Spreaders	45
4.1. Dataset Description and Methodology	47
4.1.1. Original Dataset	47
4.1.2. Transformation of Tweets Dataset into Rumor Spreaders Dataset	47
4.1.3. Extraction of three sets of features	51
4.2. Experiments and Result Section	52
4.2.1. Hyper-Parameters tuning	52
4.2.2. Optimization Techniques	54
4.2.3. Models Performance	54
4.3. Summary	57

5. Plan of Action to Combat the Misinformation Menace	59
5.1. Misinformation Social Media Corpus	60
5.1.1. Data Collection	60
5.1.2. Data Annotation Process	61
5.2. FaCov: Fact-Checked Articles Corpus	63
5.2.1. Data Collection	63
5.2.2. Extracted Features	65
5.2.3. COVID-19 Myths	68
5.2.4. Discussions around COVID-19	68
5.3. Automated MisInformation Rebuttal (AMIR)	72
5.3.1. Data Preparation: Discerning (sub-)topics	73
5.3.2. Data Preparation: Relationships Among Topics in Fact-check Article Corpus	75
5.3.3. Repurposing Social Media Content	78
5.3.4. Leveraging Fact-Checked Articles	81
5.4. Summary	86
6. Conclusions	88
7. Future Scope	90
Bibliography	93
Appendix A. Misleading Social Media Corpus Analysis	133
A.1. Language Exploration	133
A.1.1. Uncover Syntax	133
A.2. What are the Principal Topics of Discussion?	134
A.2.1. Opinion Study	136
A.2.2. Intense Emotions	138
A.3. The Influence of Visibility	139
A.3.1. The merry words of Twitter	140
A.3.2. Much ado about Hashtags	140
A.3.3. As you Like it	141
A.3.4. What's in a Name?	141
Acknowledgements	144
Sisukokkuvõte (Summary in Estonian)	145
Curriculum Vitae	147
Elulookirjeldus (Curriculum Vitae in Estonian)	150

LIST OF FIGURES

1. SHAP plot to determine feature importance. The SHAP values and feature names are represented by the x- and y-axes, respectively. Each data point represents a single instance. The red color represents a higher value for the feature than its average value, while the blue color denotes a lower value. A positive impact on the prediction is indicated by red values on the right side of the x-axis, and vice versa. The features are listed in order of decreasing importance (best seen in color).	40
2. Undirected Weighted Ego-Network of a Twitter User	46
3. Sentiments of Reply Tweets with respect to Rumor and Non-Rumor Tweets for all the five incidents	50
4. Framework	51
5. Performance of GCN model under five-fold cross validation for all the incidents	56
6. Performance of AUC-ROC Curve on all the five incidents	56
7. Key themes in True News	70
8. Key themes in Mix News	71
9. Key themes in False News	72
10. Automated Misinformation Rebuttal Pipeline.	73
11. Co-occurrence of prominent topics in the corpus of fact-checking articles.	77
12. Syntactic analysis of the Misleading and Non-Misleading tweets. The x-axis represents the counts of the specific Syntactic attribute (for example, noun) and the y-axis represents the density of the distribution. TTR is an abbreviation for Type-Token Ratio. Orange and blue colors represent the Misleading and Non-Misleading labels, respectively.	133
13. Sentiment analysis with respect to Misleading and Non-Misleading tweets. The x-axis and y-axis denote the labels of the tweets and percentages, respectively. Negative , Neutral , and Positive are different categories of the sentiments.	137
14. Emotion analysis of the Misleading and Non-Misleading tweets. Each axis denotes the emotion. The five emotions depicted are Happiness, Fear, Anger, Surprise, and Sadness. Non-Misleading tweets are represented in green color, and Misleading in red color (best seen in color).	139
15. Word Clouds with respect to Non-Misleading and Misleading tweets. The size of the word is proportional to its frequency.	140

16. Aggregation of Visibility Counts. Each cell represents the median values of the Retweets, Replies, and Likes with respect to the labels of tweets. NM and M denote the Non-Misleading and Misleading tweets, respectively. 142
17. Count of Vaccine Names used in a Tweet. X-axis and y-axis in (a) represents the count of the vaccines' names and percentages, respectively. The **green** and **red** colors indicate the Non-misleading and Misleading classes. Value 0 on the x-axis corresponds to no mention of the vaccine name in the tweet, Value 1 denotes mention of one vaccine, and so on. (b) shows the percentage of the tweets with respect to the count of the vaccine name (best seen in color). . . . 142

LIST OF TABLES

1. Data distribution of PHEME-9 among non-rumor and rumor tweets in source and reactions based on each individual case. NR and R represents non-rumor and rumor respectively	29
2. The table shows the significant difference of all LIWC features between the non-rumour and rumour source tweets	34
3. The table shows the significant difference of all LIWC features between the reactions of non-rumour and rumour tweets	34
4. The table shows the significant difference of all LIWC features between the aggregated non-rumour and aggregated rumour source (Src) and reaction tweets	35
5. The table shows the significant difference of all Readability features between the non-rumour and rumour source tweets	36
6. The table shows the significant difference of all Readability features between the reactions of non-rumour and rumour tweets	36
7. The table shows the significant difference of all SenticNet features between the aggregated non-rumour and aggregated rumour source and reaction tweets	37
8. The table shows the significant difference of all SenticNet features between the reactions of non-rumour and rumour tweets	37
9. The table shows the significant difference of all SenticNet features between the non-rumour and rumour source tweets	37
10. The table shows the percentage distribution of all the emotions in the non-rumour and rumour classes. Src, and Re represents the source and reply tweets respectively.	38
11. The table shows the partition of features according to categories and higher mean . PP and CP denotes Personal Pronouns and Cognitive Processes, respectively	39
12. Prediction Results of Random Forest model on five events. Src, and Re represents the source and reply tweets respectively. Evaluation metrics are denoted as Acc (Accuracy), Pr (Precision), Rec (Recall), and F1 (F1 Score).	39
13. Dataset Description	48
14. Distribution of Tweets and Spreaders for each of the incidents in the dataset. The two columns on the left show the number of tweets (rumor or non-rumor) posted by the source users only. The two columns on the right include the total number of users (source users and reply users) labeled as either rumor spreaders or not.	48
15. Dimensions of the PHEME dataset at various levels	49
16. An Example of the Tweets on Ferguson incident	49
17. Hyper-Parameters Tuning of models	53

18. Feature Selection Techniques (values in the cells indicates their corresponding p-values)	54
19. Metrics performance of different models	55
20. Selection of various sample sizes to fine-tune the pre-trained XLNet model. The values in the cells denote the average performance of the five-fold cross-validation along with the standard deviation. ACC, PR, RC, F1, and AUC represent Accuracy, Precision, Recall, F1-Score, and AUC ROC Score, respectively.	62
21. Evaluation metrics of the fine-tuned XLNet model. ACC, PR, RC, F1, and AUC represent Accuracy, Precision, Recall, F1-Score, and AUC ROC Score, respectively.	63
22. Ranking of the selected websites based on the Tranco list (sorted in descending order). Also, columns w/o Pre and w/ Pre represent number of articles without (before) pre-processing and with (after) pre-processing, respectively.	65
23. Attribute names with their absolute numbers of non-null values in the FaCov dataset.	66
24. Truth labels retrieved from the Fact-Check websites after pre-processing under the 3-class labeling, and the number of labels.	67
25. Fact-checks present in the dataset that are related to widely spread myths according to WHO, extracted from Title column. The myths are then further categorized based on the labels.	69
26. Absolute number (and %) of Tweets per topic.	74
27. Topics and Subtopics identified from the Misleading Tweets dataset [SSD22].	74
28. Topics and subtopics identified from the FaCov dataset [Sha+22]. Furthermore, items highlighted in Blue color represents the mapped topics from the misleading tweets dataset [SSD22].	75
29. (Multilabel topics) Topmost topics distribution in fact-checked articles FaCov dataset. The Count column indicates the total number of fact-checked articles that contains the specific topic.	76
30. (Multilabel topics) Secondmost topics distribution in fact-checked articles FaCov dataset. The Count column indicates the total number of fact-checked articles that contains the specific topic.	76
31. Comparison of the Spacy NER model with and without fine-tuning the model on the manually labeled data.	79
32. Top entities mislabeled by the unaugmented NER model. The left and right sides represent vaccine names and non-vaccine names entities.	79
33. Evaluation of the complementary approaches. AMIR_SM and AMIR_FC represent social media non-misleading posts recommendations (first approach) and fact-checked articles recommendations (second approach), respectively.	81

34. Not mapped topic: Operation Warp Speed	83
35. Not mapped topic: Real Side-effects	83
36. P-values of the Kolmogorov Smirnov Test for each of the syntactic characteristics extracted from the dataset.	134
37. Top 5 most discussed topics in Misleading and Non-Misleading COVID-19 vaccination tweets.	136
38. Topic Modeling with respect to each sentiment. M and NM represent Misleading and Non-Misleading. OWS and VaEf denote Operation Warp Speed and Vaccine Efficacy, respectively.	138
39. Topic modeling with respect to each five emotions. M and NM represent Misleading and Non-Misleading, respectively. OWS and VaEf are abbreviations for Operation Warp Speed and Vaccine Efficacy, respectively.	138
40. Use of 10 different Hashtags (left-side represents the hashtags that are mostly present in Non-Misleading tweets but not in Misleading tweets, and vice-versa on the right).	141

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

- I **Shakshi Sharma**, and Rajesh Sharma. "Identifying possible rumor spreaders on twitter: A weak supervised learning approach." In 2021 **International Joint Conference on Neural Networks (IJCNN)**, pp. 1-8. IEEE, 2021.
Lead author: The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.
- II **Shakshi Sharma**, Rajesh Sharma, and Anwitaman Datta. "(Mis) leading the COVID-19 Vaccination Discourse on Twitter: An Exploratory Study of Infodemic Around the Pandemic." **IEEE Transactions on Computational Social Systems**, 2022.
Lead author: The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.
- III Sabur Butt, **Shakshi Sharma (Equal Contribution)**, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. "What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses." **Computers in Human Behavior** 135, p.107345, 2022.
Lead author: The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.
- IV **Shakshi Sharma**, Ekanshi Agrawal, Rajesh Sharma, and Anwitaman Datta. "FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset." In **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, Vol. 16, pp. 1312-1321, 2022.
Lead author: The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.

Publications not included in the thesis

- I Mudit Dhawan, **Shakshi Sharma (Equal Contribution)**, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. "GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection." ArXiv preprint (arXiv:2202.12478), 2022.
Lead author: The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.
- II Mohit Mayank, **Shakshi Sharma**, and Rajesh Sharma. "DEAP-FAKED: Knowledge graph based approach for fake news detection." In **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, (pp. 47-51). IEEE, 2022.
- III Raj Jagtap, Abhinav Kumar, Rahul Goel, **Shakshi Sharma**, and Rajesh Sharma, and Clint P George. "Misinformation Detection on YouTube Using Video

Captions.", ArXiv Preprint (arXiv:2107.00941), 2021.

The author performed the implementation and the analysis of the experiments and contributed substantially in writing.

1. INTRODUCTION

Online Social Media (OSM) platforms such as Twitter and Facebook were initially created to promote interpersonal interactions, but they have now evolved into a haven for those who frequently spread false information by taking advantage of the connectivity and reach of the platforms on a worldwide scale [SS21]. Specifically, these platforms allow for the open exchange of ideas, but they haven't been able to stop the propagation of false information because the content on these sites is not third-party verified, and people frequently (whether knowingly or unwittingly) propagate false information. On these platforms, studies have shown that false information travels six times more quickly than true news, thereby leading to societal panic, anxiety, and financial harm [SS21].

Academic literature [CXK22] covers numerous concepts about disseminating false information. The key differences between these concepts are facticity and intention. The degree to which the author intentionally misleads the audience is referred to as intention. Facticity is a term used to describe the extent to which the content is based on actual events. These concepts are important in understanding information disorder, which refers to spreading false information to others. Misinformation, Disinformation, and Malinformation are the three primary forms of information disorder. Misinformation is the term for unintentionally disseminating inaccurate information, such as an article containing out-of-date facts. Contrarily, disinformation is false information that is purposefully spread in order to inflict harm, such as news reports that employ fabricated figures. True information that is misused, such as when words are taken out of context to deceive an audience, is known as malinformation.

Other similar concepts in the literature include fake news, which generates incorrect news that imitates reliable news to trick the audience, and rumors, which are unconfirmed reports or assertions that may be partially or wholly incorrect. Many scholars use the word "misinformation" as a collective phrase to refer to any false information since it can be challenging to determine the purpose behind information without clearly separating it from other related notions [CXK22]. Thus, in this thesis, we follow the same terminology, that is, we refer to each false piece of information as misinformation.

Numerous research has demonstrated the adverse impacts of misinformation on society. According to a 2019 Pew Research Center study¹, 64% of American respondents said that "made-up news" significantly complicated their understanding of current events. MIT found in 2018² that fake stories posted on Twitter had a 70% higher chance of being retweeted than true news reports. In its 2021 research³, the Reuters Institute for the Research of Journalism at the University

¹<https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/>

²<https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308>

³<https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/dnr-executive-summary>

of Oxford drew attention to widespread concerns about the proliferation of incorrect and misleading information, notably with reference to the Coronavirus. The World Economic Forum⁴ predicted that the cost of misinformation to the US economy in 2020 was almost \$78 billion annually.

Misinformation on social media has been the subject of research from a variety of perspectives, including content [FBC12; Jin+17; KT19; FJP51; Zho+20], context [SWL19; VMR17; WYN21; Wan+20; RMS21; Hu+21; DKF20; Cui+20; CWL19], and propagation-based [Del+16; Gor+19; LW18] approaches. The mentioned studies show the grave effects of misinformation on people, society, and the economy. However, there is still much to be discovered to fully understand the phenomenon of misinformation on social media. Additionally, the rapidly evolving nature of social media platforms and the constantly changing ways in which people use them present ongoing challenges for researchers in this field. Despite these challenges, continued research is crucial in order to develop effective strategies for combating the spread of false information on social media and promoting greater digital literacy among users. As the impact of misinformation continues to be felt in areas such as politics, public health, and beyond, understanding its causes and effects is more important than ever.

In order to better understand the mechanisms underlying the spread of misinformation and create potent countermeasures, additional research must be conducted. This could involve identifying the traits of social media misinformation content, identifying the users who spread misinformation, and looking into measures to control the misinformation, which constitutes the primary objective of this thesis.

To conclude, this thesis explores the multifaceted approach for countering misinformation in the digital age, focusing on three main dimensions: characterizing misinformation content, developing a framework for detecting misinformation spreaders, and carrying out effective counter misinformation strategies. To achieve these goals, state-of-the-art Artificial Intelligence (AI) techniques, such as Machine Learning (ML), Deep Learning (DL), Graph Neural Networks (GNN), and Natural Language Processing (NLP), have been utilized to analyze social media datasets in order to accomplish the objectives mentioned in each dimension. The following sections provide a brief explanation of these dimensions.

1.1. A Comprehensive Study on the Characterization of Misinformation Content

The first dimension of work is the fusion of descriptive and predictive tasks, which is understanding the characteristics (traits) of online social media content thus, users with respect to misinformation, especially rumors in this case. The rumor is a piece of information whose validity is in doubt at the time of posting. In

⁴<https://www.zdnet.com/article/online-fake-news-costing-us-78-billion-globally-each-year/>

other words, there is doubt about whether the information being posted is true or false. Through comprehensive psycho-linguistic analyses of rumor and non-rumor posts, we aim to uncover patterns and markers that can aid in identifying misinformation.

Research can benefit from the characterization of misinformation users by understanding more about the behaviors of both non-misinformation and misinformation users on social media sites. By focusing on the content of misinformation, these insights can aid in the creation of tactics that effectively halt the spread of misinformation. Additionally, it can aid in the development of improved detection algorithms that can detect misinformation accurately. Researchers can create more precise algorithms to detect and report misinformation in real-time by better knowing the traits of these individuals. This will help to provide an educated and healthy online environment.

Contribution. In our proposed *characterization* approach, we propose an in-depth understanding of the traits of the rumors and non-rumors posts. This work is inspired by the idea that understanding the writing styles of such users can help understand cognitive behavior or identify the main motivations behind posting misinformation. Therefore, analyzing the characteristics of such users on social media helps the research community detect and prevent misinformation. Our approach is explained in detail in Chapter 3.

1.2. The Sentinel System: Framework for Misinformation Spreaders

The second dimension of work is predictive in nature, where we develop a framework to detect and prevent the spread of misinformation. Particularly, we propose user-level rumor detection, that is, classifying the user as a possible rumor spreader or not.

While researchers have proposed various techniques for identifying misinformation spreaders and detecting social media posts involved in spreading misinformation on platforms like Twitter, these approaches have not adequately considered user-level detection. In addition, we found that previous studies classify the user as a misinformation spreader if a post is detected as a misinformation post [Mas+19; KNL21]. However, we believe that one post is not sufficient to categorize the user as a misinformation spreader or not. This is due to the reason that it is highly possible that one post shared by the user might be due to a lack of knowledge or shared not intentionally. Other studies [Mas+19; LW20; Dou+21] identify the user as a misinformation spreader or not by utilizing their social network structure, such as followers network, reply networks, etc., along with the content. However, previous studies either focus on one piece of article (or post) or network structure with one piece of article (or post) but not the combination of posts posted by the user along with their network structure.

Given the alarming concern about the spread of misinformation, it is crucial to

prioritize such research studies that address these issues.

Contribution. In our *misinformation detection* approach, inspired by the above mentioned points, we develop a classification framework that employs more than one post of a user and its network structure in combination. The primary goal is to develop a better framework by combining multiple posts with the network information of a user. Refer to Chapter 4 *misinformation spreaders* research that provides detailed insights into how we overcome these challenges to develop a more effective framework in misinformation detection.

1.3. Plan of Action to Combat the Misinformation Menace

The third dimension of work is descriptive in nature, where we employ techniques to counter misinformation. Cost-effective, scalable solutions for combating misinformation are the need of the hour. The menace of misinformation, with accelerated dissemination through social media and amplified in echo chambers, poses grave risks. A mix of technical, social, and regulatory policies together is needed to thwart these. In this context, it's important to have technical mechanisms that can quickly and effectively refute misinformation on a large scale. These tools should be tailored to the specifics of the misinformation and concerned users have a crucial role to play.

In the previous works, there are various network-based, deep learning based approaches for mitigating the spread of misinformation on social media [Sha+19; Sax+20; SBL19; XDZ22; Wan+22]. However, these methods have limitations such as a lack of external moderation and reliance on strict assumptions that misinformation has already been identified and its propagation is tracked. In addition, there are works that utilize fact-checked articles for recommendations. While fact-checked articles are a reliable source of information, they are not always immediately available, leading to not being able to stop the spread of misinformation at the initial phase when users are most vulnerable and also likely to be searching for information about it.

Contribution. This *plan of action* chapter explores how existing information obtained from social media and augmented with more curated fact-checked data repositories can be harnessed to facilitate automated rebuttal of misinformation at scale. While the ideas herein can be generalized and reapplied in the broader context of misinformation mitigation using a multitude of information sources and catering to the spectrum of social media platforms, this chapter is confined in its scope to the only rebuttal of tweets, and in the specific context of misinformation regarding COVID-19. In this chapter, we explore two complementary approaches to carry out a rebuttal of misinformation in the context of the Twitter platform. For detailed analysis, please refer to Chapter 5.

2. LITERATURE REVIEW

We present the misinformation literature from three perspectives. First, *Misinformation Characterization* where researchers focus on identifying the traits of the misinformation users, for example, by analyzing their writing styles to gain a deeper understanding of their cognitive behavior to identify the root causes of misinformation and create efficient countermeasures [UC22; Sak+21]. Second, *Misinformation Detection* where researchers develop automated frameworks to classify the content of news as misinformation or not utilizing traditional to advanced machine learning (ML), deep learning (DL), and Natural Language Processing (NLP) techniques. One such use case of developing these automated frameworks is to categorize the news when the misinformation starts spreading to the masses on social media, and we don't have fact-checked articles or such to validate the news. Third, in *Misinformation Recommendations* where the recommendation system is created to combat misinformation employing fact-checked articles. This type of recommender system can be useful when fact-checked articles are present to validate the news coming from social media. Please note that as discussed in Chapter 1, *Introduction*, our proposed solutions can be generalized to any form of misinformation. Therefore, we used the term misinformation in place of fake news, rumor, and misleading while discussing the broader context of the problem.

2.1. Misinformation Characterization

In this section, we first focus on the psychological perspective on conspiracy theories, followed by works of rumor detection and the necessity to understand the explainability aspect in the rumor domain.

1. **Psychological perspective:** Several studies used psychological perspectives to find a correlation with conspiracy theories. However, conspiracies and misinformation can have different characteristics. The focus of the conspiracies is especially to set a narrative by a set of people in some organizations, while misinformation can spread due to misunderstanding or inaccurate information as well without any intention to manipulate the users [MP22]. But the overall aim of misinformation or conspiracy theories is the same, that is, it spreads lies everywhere. [DSC17] wrote extensively on the psychological factors and divided them into existential (e.g. desire for control), social (e.g. desire to maintain a positive image of the self or group), and epistemic (e.g. desire for understanding) factors. The authors explained that these factors contribute to the popularity of conspiracy theories. Another study discovered [Lan+17] that conspiracies urge from the need for uniqueness. Similarly, in light of the COVID-19 pandemic, researchers [Cal+19] have correlated a high level of conspiracy thinking of parents with the delay of vaccination among children. While there are studies that have attempted to discover the correlation of psycholinguistics

with fake news [Shu+19] and conspiracy [Rai+21] in general, we narrowed it down to the understanding of rumor. Various NLP studies have been published related to rumor detection in text. These studies mostly revolve around the data collection, techniques, and features suitable for identifying rumor text. Other than PHEME which is the benchmark dataset for rumor detection and is used for this study, the popular text-based datasets include RUMDECT [Ma+16], SNAP data [YL11], CrisisLexT26 [OVC15], MULTI [Jin+17], KWON [Kwo+13] and RUMOUREVAL [Der+17].

2. **Rumor detection:** Among the components of rumor detection, rumor has been explored [VV20] in the context of rumor tracking, rumor stance classification, rumor detection, and rumor veracity classification. In the study [ZLP16] authors introduced context-aware rumor detection using a sequential classifier to detect rumors from the tweets divided into five news events. In an outbreak of stories, identifying the emergency is also important for the timely detection of credible information. The work [Xia+12] used an unsupervised algorithm to label the tweets as credible and incredible and identified the urgency of the news verification using supervised machine learning methods and multiple features (content, author, and diffusion). The rumor detection task can be topic-level [Yan+15] or post-level [Der+17], where the task is to identify if the topic is relevant to the text or if the post information has rumor respectively. The trend of rumor detection on topic-level has shifted quickly towards the post-level understanding of rumor and, hence, needs insightful analyses for understanding linguistic and psycholinguistic attributes of rumor. The need to understand the explainability aspect of rumors stems from the need for early detection. Early detection of the rumor is essential to mitigate the harm and the study [MGW18] segregated the rumors using a propagation tree. They used recursive neural networks to classify the tweets into false rumor, non-rumor, unverified rumor and true rumor. A Recurrent Neural Network (RNN) based aware provenance approach was proposed [Duo+17] combining the textual information and provenance information to enhance the results. To tackle the cases where provenance details were missing, they used a fusion of text and provenance information. Among the machine learning models [Ash+21; GSZ19; Jin+17; Sam+16] Support vector machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and XGBoost have been used repeatedly. Deep learning models (Convolutional neural network (CNN), Long short-term memory (LSTM), Recurrent neural networks (RNN)) [SB20; KDQ20; Ngu17] and transformer-based models (XLNet, BERT, RoBERTa, DistilBERT, and ALBERT) [Sch+21; Che+21] have often given the state of the art results in multiple rumor tasks.
3. **Misinformation detection employing psycholinguistics features:** Psycholinguistic features on rumor detection studies have involved the use of

Long Short-Term Memory (LSTM) with LIWC features i.e swear words and personal pronouns [TP10] and emotions [VRA18]. The study [VRA18] showed the existence of false rumors that triggered fear, disgust, and surprise in the rumor replies, while, [GRC19] proposed an LSTM-based model with emotions to check the credibility of articles. Similarly, the authors [GGR21] gave a comparative analysis over various profile, linguistic and psychological characteristics (Big-five personality traits, LIWC, emotions) and used CNN-based classifier with word embeddings and psycho-linguistic characteristics for classification. In fake news tasks, the role of user profiles was evaluated by [SWL18], which showed implicit (age, location) and explicit features (follower count, status count) that contributed to fake news. In addition, they showed how the combination of these features with the psycho-linguistic characteristics (LIWC, Writing styles) can be effective. Another study [Gia+20] proposed a CNN-based model combined with personality traits (Big-five) and LIWC to distinguish between fake users.

Contrary to these studies, our main focus of the first contribution is to provide insights into the source as well as reaction tweets comparing different events, as every event triggers a different insight. Moreover, we discuss the explainability part in the context of which of these features highly contribute to the rumor or non-rumor classes as well as their reactions. To the best of our knowledge, none of these papers indulges in the explainability aspects of psycholinguistics in rumor tweets and reaction tweets.

2.2. Misinformation Spreaders

Researchers have proposed numerous techniques for identifying suspicious or fake user profiles across various online social media (OSM) platforms such as Facebook [GK17], Twitter [Ben+10; WZG18], Tuenti [Cao+12], to name a few. In this section, we discuss past studies across two dimensions. First, we discuss work related to the identification of fake or suspicious users. Next, we present works about the detection of rumors.

1. **Suspicious profile detection:** Identification of suspicious (or fake) profiles on Twitter, a representative of OSM was initially analyzed using apparent features such as the *number of followers* and the *number of followings* [Ben+10], *user-profile-name*, *screen-name*, and *email parameters* [Tho+13]. Some other works have focused on identifying suspicious followers by using multi-modal information [Jia+14], which is out of the scope of this work. It is a well observed fact that mobile phones have been playing a vital role in the popularity of OSM platforms. Nonetheless, it has facilitated the rise of fake profiles as well [SN13]. Therefore, mobile phones too have been used as a medium in tracking fake profiles. For instance, in [PLB13], the authors observed the daily behavior of users using mobile phone activity for detecting fake profiles. Besides researchers have also

used camera-based sensors in detecting fake profiles within or across multi-social networks [Ber+15].

2. **User profiles in rumor tasks:** We assume that users who are often involved in disseminating rumors are more likely to be *rumor spreaders* in comparison to users whose involvement in spreading rumors is less. Our approach for identifying the possible *rumor spreaders* exploits (rumor and non-rumor) tweets along with other features present in the dataset. Therefore, we further present related literature with respect to user profiles in rumor tasks. User profiles have been analyzed as an important aspect in detecting rumor propagation. Analysis of the user profile [Shu+19], identification of source of the rumor [Dev+20], and genuineness score of the users in the social network which are spreading the rumors [Rat+17] have been utilized in past works.

Few approaches in rumor detection have also explored graph-based methodologies. Specifically, in [RSP20], the authors constructed a Twitter follower graph and exploited diffusion patterns of (mis) information for rumor detection. Some of the papers [Bia+20; WXM19; Hua+19] have employed Graph Convolutional Network (GCN) based approaches for detecting rumors. Unlike these works, we study the rumors to identify possible *rumor spreaders*.

Previous literature for misinformation (rumor) spreaders classify the user as a misinformation user based on one misinformation post by the user. Specifically, if the post is detected as a misinformation post then the user of that post is also considered a misinformation user or misinformation spreader. However, we argue that one post is not enough to categorize the user as a misinformation user as there may be multiple reasons for sharing a misinformation post, such as lack of knowledge or not sharing intentionally. Thus, a more comprehensive approach is needed to categorize the misinformation spreaders.

Based on the above mentioned points, our second contribution focuses on identifying possible *rumor spreaders* by using not only the textual data that is being spread by these *rumor spreaders* (source and reply posts of a user) considering the multiple posts of a user but also exploring their ego networks. To accomplish this objective, we used the Graph Convolution Network (GCN) approach which has been mainly used in the past for identifying rumors and not for identifying possible *rumor spreaders*.

2.3. Misinformation Mitigation Strategies

Previous studies [Bal+22] have identified various factors that contribute to the spread of misinformation, such as low factual understanding and difficulty recognizing misinformation. To combat misinformation, providing correct information to users is crucial. While many studies have focused on detecting misinformation

and interpreting black-box models [But+22], combating misinformation in an automated manner has received relatively less attention so far. We present literature on combating misinformation. Our study uses COVID-19 as a case, so we also present literature on COVID-19 misinformation.

1. **Misinformation Detection:** There are various network-based approaches for mitigating the spread of misinformation on social media [KKT03; Ngu+12; Sha+19; ZZS13; Far+17]. However, these methods have limitations such as a lack of external moderation. Additionally, these methods are difficult to deploy on real-world social networks due to the dynamic and volatile nature of information diffusion and user behavior. Deep learning based approaches utilize advanced neural networks architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to model the text and context of news articles [Sha+22; SS21]. These models can be trained on large datasets of labeled news articles to learn the underlying patterns and features that distinguish fake news from real news. Once trained, these models can be used to classify new articles as fake or real with varying degrees of accuracy. Additionally, these models can also be used to extract features from news articles that can be used as inputs for other mitigation strategies, such as URLs or news recommendations. However, these approaches also have limitations, such as the need for large amounts of labeled training data, and are not robust against biases in the training data. There have been some efforts to use recommender systems to combat fake news, primarily focused on recommending fact-checked URLs to a small specific group of users, e.g., fact-checkers [VL18; You+19].
2. **COVID-19 Misinformation & Mitigation:** The studies [Ska+22; Lie+22] show that misinformation about COVID-19 vaccines on social media platforms can lead to hesitancy and a decrease in vaccine uptake. Strategies to combat misinformation include tracking vaccine misinformation in real-time and engaging with social media to disseminate correct information [Isl+21; Sha+23], semantic network analysis [MOS21] which revealed that the most common topics of vaccine misinformation were related to safety, efficacy, and conspiracy theories and that misinformation was spread by a small number of users who were highly active in the anti-vaccination movement.

The work that is closest to ours [Wan+22] focuses on utilizing the user’s reading history to identify the event or topic they are interested in. However, it may struggle to identify new or breaking events that are not yet a part of the user’s reading history. This means that the approach may not be able to provide the user with timely and relevant information about new or unexpected events. Another limitation of the approach is that it relies solely on fact-checked articles for recommendations. While fact-checked articles are a reliable source of information, they are not always immediately available. Misinformation can spread quickly,

often before fact-checking organizations have a chance to verify the information and publish their findings. This means that the approach may not be able to provide recommendations for news that is being widely shared and discussed in real-time, which is when users are most vulnerable and also likely to be searching for information about it.

Inspired by these works, our third contribution focuses on developing a pipeline to automate the rebuttal of misinformation on social media. The pipeline combines two complementary approaches, the first approach is identifying and reusing relevant existing but factual/non-misinformation tweets as counter tweets. This is the baseline approach that can be used even when fact-checking websites are yet to carry out the vetting exercise for a given misinformation, or while such articles are identified and matched by our second approach. The second approach is recommending fact-checking articles from various fact-checking websites when they are available. This pipeline creates a more practical and scalable rebuttal system to combat misinformation on social media, by providing users with verified, reliable, and more comprehensive information from reputable sources.

3. A COMPREHENSIVE STUDY ON THE CHARACTERIZATION OF MISINFORMATION CONTENT

The credibility of information is the most decisive aspect on social media as the unmoderated nature of social media text has resulted in several cases of misinformation spreading [Der+17]. In this work, we focus on rumor, which is a specific kind of misinformation, whose authenticity has not been verified [Zub+18]. In the past few years, researchers have analysed rumors from two different directions which can be divided into descriptive analyses of rumors, and the detection of rumors using a variety of machine learning and deep learning techniques [MGW18; GSZ19; SS21].

Despite these apparent robust techniques, increasing rumors motivates the development of systems that, by gathering and analysing the collective judgements of users [Luk+16], are able to reduce the spread of rumors by accelerating the sense-making process [DB14]. In particular, linguistics and natural language processing researchers have taken the onus to study how users have discussed rumors and to understand the psycho-linguistic attributes connected to rumor spreading and detection [GGR21; Shu+19]. Scientific studies aim to understand the malicious intentions of spreading rumors through the psychological processes involved in the use of language can help in textual classification and behaviour analyses of users.

Research Question. To aid in the mitigation of misinformation, in this work, we performed the analysis of rumor vs. non-rumor tweets using psycholinguistic approaches, which is the study of the interrelation between linguistic factors and psychological aspects. In this chapter, we address the following research question: *Can we use psycholinguistics features to conduct an in-depth analysis of rumor and non-rumor tweets?*

It should be noted that we are not considering the user-level features and are solely focusing on textual information. To be specific, we used psycholinguistics attributes that tend to convey the latent (hidden) meaning of the text. However, the patterns of these attributes cannot be determined in individual instances and need to have aggregated supervised data processed by computing psycholinguistic methods and statistical evidence. To the best of our knowledge, this is the first study to use psycholinguistics features to conduct an in-depth analysis of rumor and non-rumor tweets. Specifically, we investigate the following research questions by dividing the main research question mentioned above:

RQ 1: Is there a difference in psycho-linguistic characteristics between rumor and non-rumor source tweets?

RQ 2: How can psycho-linguistic features be used to differentiate between the reactions that rumor and non-rumor tweets attract?

RQ 3: Does the contribution of psycho-linguistic features vary from event to

event or do they remain consistent for all events?

RQ 4: Can we exploit these features for classifying rumor and non-rumor tweets using Machine Learning models?

RQ 5: Which psycho-linguistic features are highly discriminative for identifying rumor and non-rumor class for the classification task?

To answer these research questions, we used the PHEME-9 dataset, consisting of 9 different events (Section 3.1). We performed psycho-linguistic studies to address RQ1, RQ2 and RQ3, and extracted four types of features from the dataset: Linguistic Inquiry and Word Count (LIWC), Readability, SenticNet, and Emotions. In computerized text analyses, Linguistic Inquiry and Word Count [Pen+15] is a gold standard in understanding linguistic aspects of motivations, thoughts, feelings, and personality. SenticNet is used to derive concept-level sentiment analyses from the text. Readability features indicate the easiness to interpret a text depending on its unique attributes. Finally, Emotions show us the true nature of the feelings which cannot be interpreted by plain polarity and sentiment detection. All of these features provided us with more insight and perspective into the rumor and non-rumor tweets. We calculated the statistical significance and mean values of every psycho-linguistic feature for rumor source tweets, non-rumor source tweets, reactions of rumor tweets, and reactions of non-rumor tweets, respectively. The statistical significance tests helped us to assess the difference between the rumor and non-rumor psycho-linguistic features.

The rest of the chapter is organized as follows. We explain the methodology in Section 3.2 and the results of our analysis in Section 3.3. For RQ4 and RQ5, we further used machine learning classifiers (classical as well as ensemble-based) on every event to evaluate the effectiveness of the psycholinguistic features in classifying the tweets into rumor vs. non-rumor classes. Lastly, we use SHAP, an AI Explainability tool based on shapley values to identify the measure of contributions each feature has in the model (Section 3.4) and provide a general discussion (Section 3.5). Finally, we conclude this article with some future works in Section 3.6.

3.1. Dataset

This study has been carried out using the PHEME dataset [KLZ18], which consists of nine events, wherein all the events are breaking news. To be specific, each event contains the source tweets which are divided into rumor and non-rumor source tweets. Similarly, every source tweet has a set of reaction tweets which are again divided into rumored reaction tweets and non-rumor reaction tweets. The reactions are triggered by either a rumored source tweets and the non-rumored source tweets and the reactions triggered by rumor and non-rumor tweets are set into rumored reaction tweets and non-rumored reaction tweets, respectively. However, it should be noted that the reactions have no ground truth to be labelled as the rumor or non-rumor and are just a reaction in the form of replies. The dataset ini-

tially reported five events [ZLP16], but was later extended to nine events to bring more variety in context. Understanding the events for rumors in the text also helps us understand the linguistic differences since these incidences were rife with rumors and gained significant media attention. The nine incidences mentioned are explained below:

- Ferguson unrest: The incident refers to the protest in Missouri, USA after an 18 year old African-American was shot by a white police officer.
- Ottawa shooting: The incident occurred in Canada’s Ottawa Parliament Hill, which resulted in the death of a Canadian soldier.
- Sydney siege: Lindt cafe siege in Sydney was a terrorist attack when a gunman held several people, hostage, in a cafe.
- Charlie Hebdo shooting: The weekly Charlie Hebdo satirical newspaper office was invaded by two brothers in Paris, which resulted in 12 people being killed and 11 injured.
- Germanwings plane crash: The plane enroute from Barcelona to Dsseldorf was crashed due to the suicidal tendencies of the co-pilot. All passengers and crew were killed in the crash.
- Gurlitt: The incident relates to the Gurlitt art collection. Hildebrand Gurlitt, was an art merchant who acted for the Nazis, who died with no direct descendants. People started spreading rumors about the acceptance of the artwork by Berns Museum.
- Prince-Toronto: The rumor started with a deleted tweet of a "pop-up show" by the Prince’s band. People suspected a surprise secret concert.
- Putin missing: The incident refers to the public disappearance of the Russian president.
- Ebola: The soccer star Michael Essien was thought to have contracted Ebola and the rumor went viral.

Table 1 shows the statistical distribution among source and reaction tweets. The Ebola case study had no list of rumor tweets, so we disregarded the case study and performed the experiments with the remaining eight case studies.

3.2. Methodology

We combined various psycholinguistic features such as LIWC, SenticNet [Cam+14], readability indexes, and emotions to bring out true insights about user patterns. The features were selected based on their previous impact [LC22; Pér+17] and usage in the text classification tasks and their individual contribution in the fake news/ rumor detection tasks.

3.2.1. Pre-processing and Feature Extraction

We first thoroughly pre-processed the noisy data of Twitter and then extracted previously mentioned features from it. The LIWC features were extracted through

	Source tweets		Reaction tweets	
	NR	R	NR	R
Charlie	1621	458	29302	68887
German	231	238	1764	2256
Sydney	699	522	14621	8154
Putin	112	126	236	361
Prince	4	229	3	666
Ottawa	420	470	5428	5966
Gurlitt	77	61	15	26
Ferguson	859	284	16837	6195

Table 1: Data distribution of PHEME-9 among non-rumor and rumor tweets in source and reactions based on each individual case. NR and R represents non-rumor and rumor respectively

the LIWC (2015) program which records all punctuations and words. We used Textstatistic ¹ python library to extract readability features including Gunning Fog, Flesch Reading Ease, Flesch-Kincaid, Simple Measure of Gobbledygook (SMOG) and Dale-Chall. It is important to keep the period in a sentence to extract readability features, hence, we only removed hashtags, user mentions, emojis and URL's. SenticNet features were calculated using the SenticNet API ² and the sentic words were combined to achieve the phrase features. For SenticNet features, we used word lemmatization and removed all punctuations, URL's, user mentions, hashtags, and custom stopwords (without negation words) for pre-processing. Emotions were calculated using DistilRoBERTa base model [San+19] trained on a combination of multiple emotion datasets for English predicting Ekman's six basic emotions, plus a neutral class which can be tested on the HuggingFace API ³.

3.2.2. Statistical Test and Machine Learning

To verify that the difference between the features (indicated in Section 3.2.1) of rumor and non-rumor classes are significant, we employed Kolmogorov Smirnov (KS) test⁴. This is a non-parametric test used to compare the two distributions. In our case, the two distributions correspond to rumor and non-rumor features, as explained in Section 3.3.

Next in Section 3.4, we used these features as inputs to machine learning models to evaluate whether they are good enough to classify the tweet as a rumor or non-rumor. Furthermore, we utilized the SHAP explainability tool⁵, indicating which features are more relevant than others when making predictions. We

¹<https://pypi.org/project/textatistic/>

²<https://sentic.net/api>

³<https://huggingface.co/j-hartmann/emotion-english-distilroberta-base?>

⁴https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

⁵<https://shap.readthedocs.io/en/latest/overviews.html>

also performed Holm-Bonferroni Method [Abd10] for testing multiple corrections. Multiple correction testing reveals the false positives in the p-values derived from the Kolmogorov Smirnov (KS) test. We performed the tests on all features and found no false positives.

3.3. Analyses

In this section, using psycholinguistic analysis we studied the characteristics of rumor and non-rumor source tweets and the corresponding reactions to the tweets. Different events show us insightful information on the nature of rumor categories and how the social and psychological meaning of words can change in every scenario. In addition, we calculated the statistical significance test five times to validate that the difference in both classes (rumor and non-rumor) was not due to random chance. We aggregated the mean value to evaluate the overall influence of the individual features for every class.

3.3.1. LIWC

LIWC features tell us about the psychometric properties of the tweets. Table 2 and 3 show the statistical difference (p-values of the KS Test as discussed in the Methodology Section) between the rumor and non-rumor tweets' LIWC feature in each individual event for source tweets and reaction tweets, respectively. The Tables use red and green colors to show the absence or presence of a statistically significant difference, respectively, between rumor and non-rumor tweets for each LIWC category and event. Table 4, on the other hand, is an aggregated representation of all the events to find the significant differences for LIWC features in source tweets and their reactions. Specifically, we concatenated the LIWC feature of the rumor and non-rumor source (reaction) tweets for all the events together. In order to determine if there was a statistically significant difference between the rumor and non-rumor cases after aggregation, we used the KS test.

Linguistic Processes. These processes include text-related analysis. In particular, **Word Count** (WC) (Table 2 and 3, Row 1) tells us about the engagement and domination of a user in a conversation. The word count needs to be balanced in deceptive scenarios where the descriptiveness of the scenario needs to be balanced, and too many words can reveal inaccuracies. In some events, word count was proven to be statistically significant in either the source tweets or the reaction tweets. Based on the WC retrieved from the LIWC program, we calculated the mean value (μ) of the WC for all events of rumor source tweets (RS), non-rumor source tweets (NRS), rumor reply tweets (RR), and non-rumor reply tweets (NRR), separately.

The mean WC (μ) of rumor source tweets (RS) and rumor reply tweets (RR) was 20 and 14.63, respectively. *One possible reason for this can be that source tweets needed more convincing and engagement. Where the mean WC of rumor source tweets (RS) and non-rumor source tweets (NRS) was almost the same,*

mean WC in non-rumor reply tweets (NRR) was found to be higher than rumor reply tweets (RR). It could be due to non-rumor reply tweets (NRR) needs to have more engagement to deny the rumor claims.

The **function words** (Table 2 and 3, Row 2) in LIWC include total pronouns, impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, and negations. Among the personal pronouns (I, s/he, they, we), we identify that the mean (μ) in rumor source tweets (RS) is 1 against the 3.20 in non-rumor source tweets (NRS). Pronouns give us a lot of insight into the personality of the users as they indicate how users are communicating with each other [BS08; Arg+06; SCG08] and what is the intent [Ber60] of the conversation. We saw a means score of 1st person singular (I), 1st person plural (we), 2nd person (you), 3rd person singular (S/he) and third-person plural (they) all to be higher in non-rumor source tweets (NRS). The same trend was seen in the rumor (RR) and non-rumor reaction tweets (NRR) where the use of personal pronouns was significantly more. Although the collective significance of function words in reaction tweets (Table 4, Row 2) was not seen, when we divided the significance based on events (Table 3), we saw 5 out of 8 events showing significant difference between non-rumor and rumor reaction tweets. Use of **prepositions** (to, with above) shows us concern with precision [PS03; New+03] and was found to be higher in rumor source tweets ($\mu = 11$ in RS) compared to rumor reaction tweets ($\mu = 7.81$ in RR). **Negation words** psychologically correlate with inhibition [TT08; PMF97] and was seen to be higher in rumor reaction tweets (RR) with the mean value of 2.02 compared to 1 in rumor source tweets (RS). However, non-rumor tweets has higher negation words than rumor tweets. Similarly, other attributes such as **conjunction** ($\mu = 2$ in RS, $\mu = 2.39$ in NRS), **common adverbs** ($\mu = 2$ in RS, $\mu = 2.35$ in NRS), **auxiliary verbs** ($\mu = 4$ in RS, $\mu = 4.71$ in NRS) and **impersonal pronouns** ($\mu = 1$ in RS, $\mu = 2.34$ in NRS) all saw lower mean in rumor source tweets (RS). Use of Impersonal pronouns, auxiliary verbs, conjunctions and negation words was seen to be more in reaction tweets where non-rumor reactions had a higher mean than rumor reactions.

Psychological Processes. Among the psychological processes we analyzed **affective process** (positive emotions, negative emotions), **social processes** (female references, male references, family, friends), **cognitive processes** (insight, causation, discrepancy, tentative, certainty, differentiation), **perceptual processes** (see, hear, feel), **biological processes** (body, sexual, ingestion, health), **drives** (affiliation, achievement, power, reward, risk), **time orientations** (past focus, future focus, present focus), **relativity** (motion, space, time) and **personal concerns** (work, leisure, home, money, religion).

Though **affective processes** are discussed extensively in Section 3.3.4, LIWC stats showed positive emotions words ($\mu = 1$ in RS, $\mu = 2.16$ in NRS) and negative emotions ($\mu = 2$ in RS, $\mu = 3.19$ in NRS) breaking down to more anxiety, anger and sadness in the non-rumor source tweets (NRS). The rumor ($\mu = 7.034$) and non-rumor ($\mu = 7.624$) reaction stats show us a similar story with higher affective

processes in non-rumor tweets. One can observe (Table 2 and 3, Row 3) a huge disparity between the positive and negative mean of words in the rumor source tweets showing the imbalance of emotions. On the other hand, reaction tweets show a very similar mean of negative words as reactions to rumors are meant to trigger negative emotions.

Social words correlate with social concerns [New+03; New+08; TP10], and social support [Les+07], we observed (Table 2 and 3, Row 4) a variety of observations in both source and reaction tweets. Type of rumor can draw many such references and one can see a pattern of more female ($\mu = 0.0598$ in RS, $\mu = 0.1196$ in NRS) and friends ($\mu = 0.0695$ in RS, $\mu = 0.1465$ in NRS) reference words in non-rumor source tweets and more family ($\mu = 0.2274$ in RS, $\mu = 0.1493$ in NRS) and male ($\mu = 0.4901$ in RS, $\mu = 0.3109$ in NRS) average words in rumor source tweets.

Cognitive Processes gives us the insight of the reasoning and difference in the thought process of the authors [Arg+06; Bat+02; Lie+02; TP10]. Cognitive processes showed a significant difference in the events of both reaction and source tweets (Table 2 and 3, Row 5) The event of Putin and Prince had no significant difference in cognitive processes due to the nature of the rumor (no supporting claim). *Cognitive processes were also used higher by the users in non-rumor source and reaction tweets. Except for the tentative words (maybe, perhaps) that were used more in the rumor source tweets and reaction tweets, all subcategories of cognitive processes had a higher value of mean in the non-rumor source and reaction tweets.*

Perceptual Processes tell us about the sensory experiences in the text, including seeing, hearing, and feeling related words. *We saw that the perceptual processes (Table 2 and 3, Row 6) did not create a lot of impact in the individual scenarios except in Putin’s case in the source tweets where the category was very relevant to the scenario of Putin being absent since his last sighting.* Similarly, **biological processes** (Body, health, sexual, ingestion) was also very scenario-specific and Sydney’s case was about a hostage scenario and an act of terrorism. *We conclude that biological processes (Table 2 and 3, Row 7) become significant where the incident is directly related to health, body, sex etc.*

Time orientations in the psychological processes gave us good insight about the rumor and non-rumor scenarios where *rumor source tweets (Table 2, Row 13) were more past focused ($\mu = 3$ in RS, $\mu = 1.52$ in NRS) and non-rumor source tweets were more present focused ($\mu = 6.42$ in NRS, $\mu = 5$ in RS).* Reactions (Table 3, Row 13) to non-rumor and rumor tweets followed the same trend.

Relativity (area, bend, exit) includes motion (arrive, car, go), space (down, in, thin), time (end, until, season) related words. Relativity, in general, proved to be less significant in reactions (Table 3, Row 9) of the scenarios and significant in five out of eight scenarios in the source tweets (Table 2, Row 9). *Though relativity is more heavily used in rumor source tweets and their reactions, however, an important observation to make is that motion related words were present more*

in the non-rumor source tweets and the reactions of the rumor tweets. A combination of higher mean in the negation words and motion related words in the non-rumor source tweets shows us the attempted correction in the direction of the conversation.

People show their personal concerns in the reactions and sources of the tweet where we can see an emphasis on work, money, leisure, home religion and death related words. **Personal concerns** can be seen to be significant in some events (Table 2 and 3, Row 12) i.e Charlie, Sydney and Ferguson in the source tweets. However, this can be case dependent as all these events were linked to violence and reported abuse. *In the incidences that mattered, we saw a higher mean of work, death, and leisure-related words in the rumored cases whereas more use of home, money and religion related words in the non-rumored source tweets.* **Informal Language** (Table 2 and 3, Row 10) on social media is expected in general and the data presented a similar scenario where the *mean of non-rumored source tweet* ($\mu = 5.32$ in NRS, $\mu = 5$ in RS) *was higher in mean.* In general, the reaction of the tweets had a high mean ($\mu = 4.024$ in NRS, $\mu = 4.051$ in RS) of informal words as well.

Drives (Table 2 and 3, Row 8) are the motivational factors and in rumor detection it can identify a lot about the authors motives/agendas behind the tweets. Drives might include affiliation (ally, friend), power (superior, bully), reward (prize, benefit), risk (danger, doubt), and achievement (win, success) related words. Drives had a significant impact on the differentiation between non-rumor and rumor resources. Breaking down the reaction on rumor and non-rumor tweets, it also played a vital role in differentiating how drives of users were different in every scenario impacting through the rumor and non-rumor sources. Further narrowing revealed that *rumored sources had higher mean in reward, risk and power related words and non-rumor source tweets had more mean of affiliation and achievement related words.*

Punctuation. Punctuations (Table 2 and 3, Row 11) such as question marks and assents show how people are communicating with each other. Punctuations can also show us attempted effort for explanation or emphasis through indicators such as apostrophes and parentheses. *Punctuations did not have a lot of impact in the reactions of rumor and non-rumor; however, played some part in differentiating the rumor and non-rumor source tweets. We observed that non-rumor sources of tweets have relatively higher mean periods, commas, semi-colons, question marks, apostrophes and parenthesis.* The reactions, in general, had more mean of punctuations for both rumor and non-rumor.

Other Grammar. The other grammar category (Table 2 and 3, Row 14) is the only category that is statistically significant in both rumor and non-rumor source tweets and their reactions (Table 4, Row 14). The grammar category of LIWC includes common verbs, common adjectives, comparisons, interrogatives and quantifiers. Common verbs can explain the temporal focus of the tweets along with common adjectives that identify the actions. In the non-rumor and rumor source

tweets, common adjectives almost show similar means and hence do not contribute to differentiating. *The reactions to the rumor and non-rumor tweets however used much more common adjectives and verbs. Non-rumor source tweets and the reactions to non-rumor source tweets had a higher mean of interrogative and comparison words compared to rumor source tweets and their reactions, signifying that users were questioning more about the non-rumor tweets compared to rumor tweets. Reaction of both categories also showed a greater mean of quantifiers.*

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
WC	0.95	0.006	0.98	0.129	0.04	7.02E-05	0.0002	8.04E-05
Function words	2.44E-15	0.07	7.45E-18	0.012	0.33	1.10E-18	0.001	2.88E-10
Affect Words	1.75E-09	0.15	1.22E-33	0.39	0.51	0.0001	7.51E-11	0.001
Social Words	3.15E-07	0.78	5.93E-17	0.94	0.36	3.30E-19	0.002	0.0006
Cognitive Processes	1.50E-07	4.37E-05	1.69E-21	0.36	0.99	1.65E-09	6.87E-05	4.22E-09
Perpetual Processes	0.99	0.96	0.11	0.01	0.06	0.98	0.79	0.31
Biological Processes	0.99	0.28	6.66E-08	0.999	0.54	0.59	1	0.46
Core Drives and Needs	0.84	0.09	5.98E-06	0.0005	0.38	0.03	0.003	0.0002
Relativity	3.86E-11	0.63	0.005	0.39	0.07	0.002	0.009	0.038
Informal Speech	0.04	3.68E-07	0.001	0.01	0.12	0.0005	0.33	5.09E-13
All Punctuation	0.0003	0.44	0.033	0.34	0.8	0.24	0.079	7.77E-16
Personal Concerns	3.75E-07	0.2	8.28E-05	0.2	0.22	0.18	0.057	0.0001
Time Orientation	0.001	0.21	1.56E-05	0.07	0.69	4.12E-05	0.0791	0.024
Grammar Other	0.039	0.002	0.01	0.03	0.86	0.014	0.03	0.01
Language Metrics	0.0005	0.77	3.57E-14	0.23	0.66	1.70E-06	0.25	9.81E-14
Summary Variable	0	0.87	2.75E-20	0.36	1.95E-08	0.012	0.006	1.13E-182

Table 2: The table shows the significant difference of all LIWC features between the non-rumour and rumour source tweets

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
WC	3.36E-05	0.06	0.04	0.048	0.5	0.34	0.93	0.55
Function words	1.41E-32	0.28	3.60E-28	7.30E-01	0.67	2.38E-05	0.5	1.36E-22
Affect Words	3.01E-07	0.052	2.17E-50	0.64	0.79	0.014	0.52	1.40E-06
Social Words	2.42E-17	0.0006	6.76E-26	0.83	0.62	0.005	0.77	1.34E-06
Cognitive Processes	4.93E-17	0.03	7.42E-27	0.98	0.25	0.001	0.77	3.04E-28
Perpetual Processes	0.01	0.94	0.99	0.95	0.97	0.015	0.94	6.80E-08
Biological Processes	0.66	0.94	0.17	0.99	0.38	0.89	1	0.194
Core Drives and Needs	0.01	0.004	0.004	0.41	0.32	0.006	0.5	0.14
Relativity	1.16E-42	0.51	1.80E-10	0.92	0.98	0.9	0.81	0.008
Informal Speech	1.33E-05	0.23	1.17E-05	0.016	0.68	0.21	0.99	4.11E-19
All Punctuation	0.19	0.48	0.031	0.33	0.1	0.15	0.57	8.91E-10
Personal Concerns	0.25	0.001	0.002	0.99	0.99	0.0014	0.57	0.47
Time Orientation	4.72E-15	5.13E-05	3.17E-10	0.68	0.4	1.17E-10	0.965	1.49E-08
Grammar Other	5.04E-08	1.13E-08	5.69E-14	0.29	0.28	1.26E-05	9.93E-07	9.93E-07
Language Metrics	2.51E-12	0.0005	1.22E-42	0.27	0.09	0.132	0.75	7.91E-18
Summary Variable	0	2.47E-44	0	5.09E-17	8.36E-09	3.66E-15	0.004	0

Table 3: The table shows the significant difference of all LIWC features between the reactions of non-rumour and rumour tweets

3.3.2. Readability

Readability allows differentiating between a text that is easy to comprehend compared to a text that is complicated and requires a high level of education or intelligence for understanding. There are many readability scores used to evaluate the text, we considered the most popular tests to evaluate tweets. Table 5

	Src tweets	Reaction
WC	4.70E-05	9.99E-01
Function words	4.70E-05	9.99E-01
Affect Words	4.70E-05	9.99E-01
Social Words	4.70E-05	9.99E-01
Cognitive Processes	4.70E-05	9.99E-01
Perpetual Processes	4.70E-05	9.99E-01
Biological Processes	4.70E-05	9.99E-01
Core Drives and Needs	4.70E-05	9.99E-01
Relativity	4.70E-05	9.99E-01
Informal Speech	4.70E-05	9.99E-01
All Punctuation	4.70E-05	9.99E-01
Personal Concerns	3.75E-07	2.50E-01
Time Orientation	4.20E-05	9.99E-01
Grammar Other	0.03	5.05E-08
Language Metrics	1	1
Summary Variable	4.70E-05	1.00E+00

Table 4: The table shows the significant difference of all LIWC features between the aggregated non-rumour and aggregated rumour source (Src) and reaction tweets

and 6 shows the significant difference in readability scores between rumor and non-rumor tweets and their reactions to various events. The Table presents Flesch score [FJP51], Flesch-Kincaid score [Kin+75], Gunning-Fog score [Gun69], Smog score [Mc 69], Dale-Chall score [Fle48]. The Flesch score in source tweets ($\mu = 72.26$ in RS, $\mu = 74.12$ in NRS) indicated "fairly easy to use" text and reactions ($\mu = 80.04$ in R, $\mu = 79.57$ in NR) indicated the same trend, where the high number of the score (0-100) indicates more easiness to read. If we translate that score to US academic grade level (0-18) using Flesch-Kincaid we see that the source tweets ($\mu = 7.22$ in RS, $\mu = 7.76$ in NRS) and their reactions ($\mu = 4.66$ in R, $\mu = 4.89$ in NR) show a 7th grade of understanding to read the text. The Gunning-Fog score (grades 1-20) points the source tweets ($\mu = 8.9$ in RS, $\mu = 9.41$ in NRS) and their reactions ($\mu = 7.2$ in R, $\mu = 7.51$ in NR) in the direction of Smog score ($\mu = 7.6$ in RS, $\mu = 8.11$ in NRS) where the easiness of reading is suitable for a 7th-9th grader. Lastly, Dale-Chall score for readability indicated the text required the reading comprehension of 11th to 15th grade student. Although we found no significant difference collectively of readability scores, however, we found many incidences of significance individually when divided per event. *It would be fair to say that Twitter rumor spreaders and their responses engage casual conversations and are designed to target masses for rapid spreading of news which is different from formal news platforms.*

3.3.3. SenticNet

SenticNet features $\in [-1, +1]$ give us a commonsense understanding of the text by translating the hourglass wheel of emotions [Plu01; CLH12] into statistical values. We considered the sentic values (Aptitude, Pleasantness, Attention, and Sensitivity) and polarity associated with the concept. We observed a shift of emo-

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
flesch_score	0.065	4.35E-09	0.11	0.005	0.0007	1.61E-06	0.378	1.00E-04
fleschkincaid_score	1.90E-01	6.42E-09	1.00E-03	0.039	0.035	1.00E-04	0.31	3.10E-07
gunningfog_score	1.00E-02	0.273	1.07E-09	0.065	0.1074	0.001	1.00E-02	7.57E-05
smog_score	7.00E-01	0.25	1.21E-07	0.055	0.32	2.20E-01	0.03	0.294
dalechall_score	6.58E-06	6.24E-06	2.61E-05	5.70E-07	0.14	7.30E-08	9.30E-01	2.90E-01

Table 5: The table shows the significant difference of all **Readability** features between the non-rumour and rumour **source tweets**

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
flesch_score	5.49E-01	0.005	0.01	0.86	0.66	2.04E-10	0.784	0.009
fleschkincaid_score	2.00E-03	3.82E-06	1.35E-08	7.63E-01	0.24	2.00E-03	0.69	3.66E-07
gunningfog_score	3.94E-11	0.00024	1.62E-10	0.47	0.44	0.02	0.797	8.36E-05
smog_score	3.27E-06	0.01	2.58E-06	0.53	0.91	0.001	1	6.61E-01
dalechall_score	6.00E-04	0.0003	7.14E-05	0.39	0.85	0.06	0.17	4.40E-02

Table 6: The table shows the significant difference of all **Readability** features between the **reactions** of non-rumour and rumour tweets

tions throughout the tweets giving a mix of SenticNet values. The aggregated statistical significance can be seen in Table 7 which shows all SenticNet values significant for rumor and non-rumor source tweets and their reactions. *rumor source tweets were weighted more towards pleasantness ($\mu = 0.07872$ in RS, $\mu = 0.03275$ in NRS), attention ($\mu = 0.11747$ in RS, $\mu = 0.08301$ in NRS) and polarity ($\mu = 0.10902$ in RS, $\mu = 0.08646$ in NRS) while non-rumor tweets had more mean of sensitivity ($\mu = 0.03$ in RS, $\mu = 0.05351$ in NRS) related emotions. The mean value of aptitude ($\mu = 0.09925$ in RS, $\mu = 0.09879$ in NRS) on the other hand can be seen to be very close although towards positive aptitude emotions (trust and acceptance). Among the reaction tweets we saw no aggregate statistical significance for any of SenticNet values, however, Table 8 shows significance in many individual scenarios such as the Sydney case where all SenticNet features had significant differences among reactions. Reactions of rumor and non-rumor tweets gave us negative mean pleasantness ($\mu = -0.01907$ in R, $\mu = -0.0258$ in NR) weighing more towards non-rumor tweets along with sensitivity ($\mu = 0.0756$ in R, $\mu = 0.0853$ in NR), polarity ($\mu = 0.0373$ in R, $\mu = 0.0383$ in NR) and aptitude ($\mu = 0.05901$ in R, $\mu = 0.0681$ in NR). *Attention related emotions (interest, anticipation and vigilance) were seen more in rumor reactions ($\mu = 0.04763$ in R, $\mu = 0.0389$ in NR) clearly showing how people are more attentive towards rumored content that is designed to draw more attention.* The statistics identify that Sentic net values for non-rumor reactions draw more attention to emotions like annoyance, anger, acceptance, trust, grief and sadness compared to rumor reactions. It should be noted that the emotions are triggered depending on the scenario and drive of rumor spreaders. Table 9 clearly shows a significant difference of Sentic-Net values in source tweets showing incidences like Gurliit in which sentic values had no significance and Sydneys case where sentic values plays important role in differentiation rumor and non-rumor sources.*

	Src tweets	Reaction
pleasantness	5.82E-05	0.9999999846
attention	5.82E-05	0.9999999846
sensitivity	4.64E-05	0.9999999846
aptitude	5.82E-05	0.9999999846
polarity	5.82E-05	0.9999999846

Table 7: The table shows the significant difference of all [SenticNet](#) features between the [aggregated](#) non-rumour and aggregated rumour source and reaction tweets

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
pleasantness	7.78E-10	0.05	0.001	0.29	0.309	5.33E-05	0.82	0.725
attention	3.84E-09	2.20E-01	1.00E-04	8.71E-01	0.45	1.60E-04	0.71	1.10E-02
sensitivity	9.00E-03	0.91	3.00E-03	0.56	0.29	0.036	0.6	1.83E-07
aptitude	1.19E-24	0.839	1.00E-03	0.87	0.95	1.38E-06	0.56	7.00E-02
polarity	1.01E-20	0.66	8.00E-04	0.271	0.33	6.42E-05	0.56	9.00E-03

Table 8: The table shows the significant difference of all [SenticNet](#) features between the [reactions](#) of non-rumour and rumour tweets

	Charlie	German	Sydney	Putin	Prince	Ottawa	Gurliit	Ferguson
pleasantness	1.84E-06	4.30E-01	3.51E-06	3.37E-07	0.975	8.50E-09	0.81	1.43E-01
attention	4.40E-07	3.40E-01	2.00E-04	0.0039	0.03	9.87E-08	0.92	8.30E-01
sensitivity	7.00E-04	0.42	3.72E-05	5.22E-07	0.37	0.38	9.00E-01	1.51E-07
aptitude	1.77E-11	0.03	1.10E-04	0.0039	0.5	2.31E-08	0.4431	0.26
polarity	5.54E-11	5.50E-01	2.19E-07	2.31E-05	0.93	2.11E-08	3.70E-01	3.40E-02

Table 9: The table shows the significant difference of all [SenticNet](#) features between the non-rumour and rumour [source tweets](#)

3.3.4. Emotions

Table 10 shows the emotion percentages across the rumor and non-rumor categories. *Fear and Sadness are the two most instigated emotions in the rumor tweets. The reactions to the rumor showed that the percentage of fear and sadness was converted into anger and surprise while the highest instigated emotion being neutral. Non-rumor source tweets had the highest percentage of neutrality, followed by fear, anger and sadness. The reactions to non-rumor source tweets had less percentage of fear and greater percentages of neutrality, anger and surprise.* We can see the patterns of rumored tweets trying to use negative emotions to instil fear among people and as a reaction, many people felt fear compared to reactions to non-rumor sources in general. It should be noted that the majority of the rumored incidences in the study were related to some sort of tragedy, however, the distribution among the same news in rumored and non-rumored forms shows the extent of negative and positive emotions used to achieve the potential motives.

	Rumour Src	Non-rumour Src	Rumour Re	Non-rumour Re
Anger	7.83%	18.52%	15.64%	20.60%
Disgust	1.59%	3.13%	3.74%	4.49%
Fear	27.64%	19.41%	9.70%	7.29%
Joy	4.15%	8.20%	6.55%	9.03%
Neutral	19.30%	28.11%	35.85%	36.24%
Sadness	27.14%	15.16%	12.99%	9.33%
Surprise	12.35%	7.46%	15.54%	13.02%

Table 10: The table shows the percentage distribution of [all the emotions](#) in the non-rumour and rumour classes. Src, and Re represents the source and reply tweets respectively.

The Table 11 gives a bird’s eye view to the analyses section. We partitioned the rumor and non-rumor source tweets and their reactions based on feature impact and mean. The analysis is further strengthened by the SHAP plots and explanation given in Section 3.4.

3.4. Prediction and Explainability

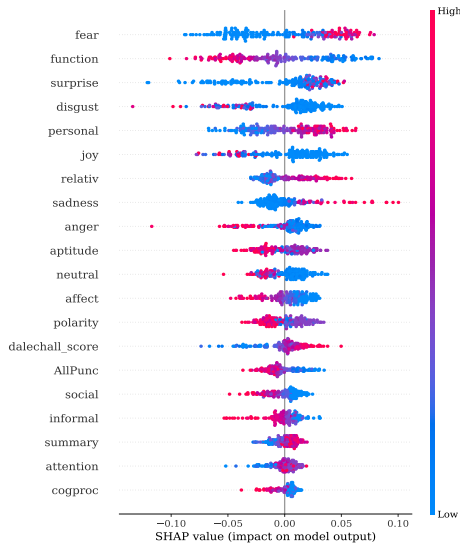
In this section, we investigate how effective are the four types of features (discussed in the previous section) distinguishing between the rumor and non-rumor categories (classes). We use these features as inputs to various Machine Learning (ML) models, both classical and ensemble-based, to categorise each tweet as rumor or non-rumor. Each event in the dataset is passed through 30 ML models individually. Moreover, we train two separate models based on two types of inputs, that is, source tweets features and reply tweets features. As previously stated, if a source tweet is classified as a rumor, all of their reply tweets are also classified as rumors and vice versa. This aids in determining whether the individual features of

	Higher mean in source tweets	Higher mean in source tweets reply tweets
rumor	-Prepositions -Social words (family) -Cognitive Processes (tentative) -Time (past) -Relativity -Personal (work, death, leisure) -Drives (reward, risk, power) -SenticNet (pleasantness, attention, polarity) -Emotions (fear, sadness)	-Negation words -Cognitive Processes (tentative) -Time (past) -Relativity -Emotions (anger, surprise)
Non-rumor	-Average word count -PP (1st person [singular (I), plural (we)], 2nd person (you), 3rd person [singular (S/he), plural (they)]) -Conjunction, common adverbs, auxiliary verbs -Affective processes -CP (insight, causation, discrepancy, certainty, differentiation) -Time (present) -Personal (home, money, religion) -Informal language -Drives (affiliation, achievement) -Grammar (interrogative, comparison) -SenticNet (sensitivity) -Emotions (neutral, fear, anger)	-Negation words -Affective processes -CP (insight, causation, discrepancy, certainty, differentiation) -Impersonal pronouns, auxiliary verbs, conjunctions -Time (present) -Grammar (interrogative, comparison) -SenticNet (pleasantness, sensitivity) -Emotions (fear) -PP

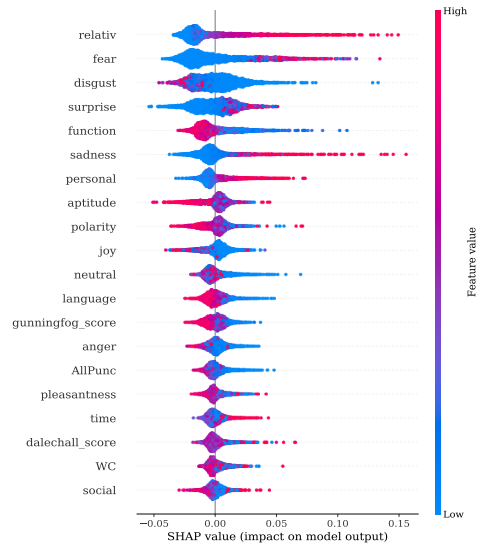
Table 11: The table shows the partition of features according to categories and **higher mean**. PP and CP denotes Personal Pronouns and Cognitive Processes, respectively

	Charlie		German		Sydney		Putin		Prince		Ottawa		Gurlitt		Ferguson	
	Src	Re	Src	Re	Src	Re	Src	Re	Src	Re	Src	Re	Src	Re	Src	Re
Acc	0.87	0.82	0.81	0.76	0.77	0.77	0.92	0.67	1	1	0.85	0.83	0.92	0.85	0.84	0.73
Pr	0.87	0.83	0.81	0.76	0.77	0.77	0.93	0.67	1	1	0.85	0.85	0.92	0.85	0.84	0.74
Rec	0.87	0.82	0.81	0.76	0.77	0.77	0.92	0.67	1	1	0.85	0.85	0.92	0.85	0.84	0.73
F1	0.87	0.76	0.81	0.76	0.77	0.77	0.91	0.67	1	1	0.85	0.85	0.92	0.85	0.84	0.65

Table 12: Prediction Results of **Random Forest** model on five events. Src, and Re represents the source and reply tweets respectively. Evaluation metrics are denoted as Acc (Accuracy), Pr (Precision), Rec (Recall), and F1 (F1 Score).



(a) Charlie hebdo (source tweets)



(b) Charlie hebdo (reply tweets)

Figure 1: SHAP plot to determine feature importance. The SHAP values and feature names are represented by the x- and y-axes, respectively. Each data point represents a single instance. The **red** color represents a higher value for the feature than its average value, while the **blue** color denotes a lower value. A positive impact on the prediction is indicated by **red** values on the right side of the x-axis, and vice versa. The features are listed in order of decreasing importance (best seen in color).

source and reply tweets are distinct enough to classify the tweet into one of two classes.

We split the source tweets (psycho-linguistic features) dataset into 80% train and 20% test set for each event. Further, we applied a ten-fold cross validation technique on the train set, which resulted in 10 train-fold sets. Following that, each ML model is trained on each train-fold set and then evaluated on the 20% test set. Basically, we used ten-fold for training, and the final evaluation was done on the test data. We also apply the oversampling technique where the dataset is imbalanced. The same procedure is followed for reply tweets. We used default hyper-parameters of the ML models provided by scikit-learn⁶ package.

Due to space constraints, we only report the results of the best performing model (refer to Supplementary Table 1-4 for more details. Please note that in total, we train and tested 30 models but we are providing the results of the next best models only as providing the results of all the 30 models is unnecessary.), which is the Random Forest model, in Table 12. Each of the eight events is evaluated based on four metrics: Accuracy, Precision, Recall, and F1 Score. The Table shows that the results of source tweets are better (# of cases - 6) or equal (# of cases - 2) to the results of reply tweets, implying that source tweets are more important to classify rumor and non-rumor tweets than reply tweets.

Next, the SHAP Explainability AI tool is utilized to evaluate the contributions of each feature in the classification task. Specifically, by computing the average marginal contributions of each feature, this tool assists in determining the significant features. Figure 1 shows the significant features for the rumor class in descending order. Due to space limits, we are only showing the SHAP plot (refer to Supplementary Discussion 1-7 for more details) of *Charlie hebdo* event for both source and reply tweets in Figures 1a and 1b respectively. It can be noted that for the source tweets, *fear* is the highest contributing attribute, whereas for reply tweets *relative* attributes the most. We also illustrate the features that have a positive and negative impact on the class. Particularly, in *fear* attribute, as the red color samples, which are on the right side of the x-axis, are more than the blue color samples, this indicates a positive impact on the rumor class. This means that the higher the value of the *fear* attribute, the better the chances of predicting the rumor class. This is intuitive as the *fear* emotion is high in rumor tweets.

The features: *surprise*, *personal*, *sadness*, *relativ*, *summary*, *attention*, *grammar*, *WC*, *dalechall_score*, *drives*, *fleschkincaid_score*, *time*, and *percept* all have a positive impact on the rumor class, as seen in Figure 1a. We also noticed that the SHAP plot of reply tweets, Figure 1b has the same but few different features than the SHAP plot of source tweets, Figure 1a: *sensitivity*, *affect*, *informal*, *bio*, all of which have a positive influence on the rumor class. The remaining attributes in both the Figures have a negative impact, meaning that as the value of these attributes decreases, the likelihood of correctly predicting rumor class increases.

⁶<https://scikit-learn.org/stable/>

We noticed a similar trend in other terrorist-related or killing people events such as *Sydney siege*, *Germanwings crash*. In comparison, in non-terrorist events such as *Ferguson*, which is an event about protest, we observed that (not shown due to space limits, refer to Supplementary Discussion 1-7), these events contains more positive influenced features with respect to rumor class than the *Charlie* event. One possible explanation could be that the nature of the event was about speculation, future occurrence, and was driven by fear.

3.5. General Discussion

The practical implications of the current findings will help enable alerts for rumors based both on the source tweets and the reactions to those tweets. Reactions showing a high proportion of the psycholinguistic attributes relating to rumor reaction class can indicate the uncertainty of people towards the tweets. It also enables the psychological understanding of rumors in depth. Furthermore, each event in the dataset leads to a different context, such as some events are related to serious rumors, such as violence and killing, whereas other events are benign rumors, such as missing rumors. The divergent results suggest that analyzing or understanding the dataset is quite significant as each event could provide a different aspect. This highlights the adaptability of our methodology to capture the distinct patterns and dynamics within each event in the dataset, leading to more accurate and contextually relevant findings. By acknowledging the differences observed across events, we can enhance the robustness and comprehensiveness of our approach through further investigation and refinement. We validated our approach by utilizing them as features of the machine learning models. Results show that these features are significant in detection models.

We noticed some unexpected results obtained from the analysis. For instance, in relation to the punctuation results presented, it is surprising to observe that non-rumor source tweets exhibit relatively higher mean usage of punctuation marks such as periods, commas, semi-colons, question marks, apostrophes, and parentheses. These findings may suggest that non-rumor source tweets employ more syntactic complexity or expressive writing styles when compared to rumor source tweets. However, it is crucial to note that the interpretation of these results should be made with caution, as several factors may influence the usage of punctuation in tweets, including individual writing habits, cultural differences, and contextual factors. Another instance is that pleasantness is higher in rumors. This finding, although intriguing, needs to be carefully analyzed and interpreted. The measurement of pleasantness in rumors might be influenced by various factors, such as the specific language used in rumor dissemination or the emotional impact of rumors on the recipients. It is crucial to acknowledge that the determination of pleasantness in the text is a complex task, often subject to interpretation and subjectivity. Besides the contexts of the dataset, another factor that can affect the analysis is how and in what circumstances the dataset has been collected. For in-

stance, which platform has been used to collect the data and why. These questions for data collection have a great impact on the generalizability of the results.

Limitations. It is important to mention that the dataset used has a fixed number of words in the Twitter text, which is a limitation of the platform. Another limitation of a social media text is the extravagant use of informal text, including slang, elongated texts, and emojis. However, these limitations have little or no effect on our studies and hence, can be replicated to study regional differences in psycho-linguistics rumor text. All the psycholinguistic features played an impact on the results. The Shapley values (based on every incident) resonated with the statistical findings of both the source tweets and their reactions. Based on different scenarios, the ranking of the features in the feature selection algorithms changed. However, the generalizability derived from the algorithms using these features remained consistent.

Furthermore, we highlight the specifications of the best performing model, Random Forest. We utilized the scikit-learn software, in particular, version 1.0.2. The Random Forest Classifier has several important parameters, including a total of 100 trees, an entropy-based split criterion, and two minimum samples for the split. We also investigated the trained models for both source and reply tweets for all the events. According to our observations, both sorts of best-performing models have the same parameters. This suggests that the same type of features are sufficient to train the source and reply tweets models well when it comes to classifying tweets as rumor or non-rumor. One could argue why there are distinct models for the source and reply tweets when the same parameters of the best-performing model give the best performance. The reason for this is that we want to make sure that the extracted features from the source (or reply) tweets are adequate to train the model without the requirement for the features from the reply (or source) tweets. This suggests that psycho-linguistic features extracted from texts alone can be used to train machine learning models, which could be significant in rumor categorization tasks.

The purpose of this research was to perform an in-depth analysis of the psycholinguistics side of the rumor task. We discovered a substantial difference between rumor and non-rumor psycho-linguistics source features, as well as between reply features.

We discovered that rumor source tweets used more past-related words, prepositions and contain drives (motivation) related to reward, risk, and power. Similarly, non-rumor source tweets had more mean features such as affective processes, cognitive processes (insight, causation, discrepancy, certainty, differentiation), present-related words, informal language, and had drives related to affiliation and achievement (RQ 1). The highest percentage of neutrality was found in non-rumor source tweets and non-rumor reactions, whereas rumor source tweets were driven by fear and grief, and their reactions invited anger and fear. Attention-related emotions (interest, anticipation, and vigilance) were more prevalent in rumor reactions, according to SenticNet (RQ 2). Readability has a considerable

impact on the majority of events. All the psycholinguistic features played a considerable impact and remain consistent on the events (RQ 3). We also explored the effectiveness of these features in predicting rumors. Specifically, we discovered that the ensemble-based Random Forest model, for all events, outperformed the other used models (RQ 4). As machine learning models are black-box in nature, we utilized the SHAP AI Explainability tool to look for the features that are more important than the other features. This helped in understanding which features contribute the most in classifying the tweets into one of two categories (RQ 5).

3.6. Summary

The previous research does not focus much on investigating the characteristics of rumor and non-rumor posts especially psycholinguistics analyses. The purpose of this research was to perform an in-depth analysis of the psycholinguistics side of the rumor task. We discovered a substantial difference between rumor and non-rumor psycho-linguistics source features, as well as between reply features.

We discovered that rumor source tweets used more past related words, prepositions and contain drives (motivation) related to reward, risk, and power. Similarly, non-rumor source tweets had more mean of features such as affective processes, cognitive processes (insight, causation, discrepancy, certainty, differentiation), present-related words, informal language, and had drives related to affiliation and achievement. The highest percentage of neutrality was found in non-rumor source tweets and non-rumor reactions, whereas rumor source tweets were driven by fear and grief, and their reactions invited anger and fear. Attention-related emotions (interest, anticipation, and vigilance) were more prevalent in rumor reactions, according to SenticNet. Readability has a considerable impact on the majority of events. We also explored the effectiveness of these features in predicting rumors. Specifically, we discovered that the ensemble-based, Random Forest model, for all events outperformed the other used models. As machine learning models are black-box in nature, we utilised the SHAP AI Explainability tool to look for the features that are more important than the other features. This helped in understanding which features contribute the most in classifying the tweets into one of two categories.

4. THE SENTINEL SYSTEM: FRAMEWORK FOR MISINFORMATION SPREADERS

Online Social Media (OSM) platforms, initially developed for connecting individuals, have become a hotbed for users who spread misinformation regularly by exploiting the connectivity and globality of these platforms [Kwa+10]. On one side, these platforms offer a place for the expression of views. However, on the flip side, these platforms have not been able to regulate the spread of misinformation. According to [Diz18], false information propagates six times faster than true news on these platforms, which can result in panic, fear, and financial loss to society [Zub+18]. Thus, it is important to control misinformation propagation before it spreads to the masses. However, it is not a trivial task to identify such users as they blend very well by creating many connections with users who are not involved in spreading misinformation activities [Fir+14].

In this chapter, we focus on rumors, which is one type of misinformation (other types are fake news, hoaxes, etc.). One way to control the spread of misinformation is by identifying users who are possibly the misinformation spreaders, that is, users who are often involved in spreading the misinformation. We discovered that if a post is classified as a misinformation post [Mas+19; KNL21], prior studies claim that the user is a misinformation spreader. However, we don't think one post is enough to label a user as a misinformation spreader or not. This is owing to the fact that it is quite likely that one post posted by the user was done so unintentionally or due to ignorance. Other studies [Mas+19; LW20; Dou+21] label the user as a misinformation spreader or not by utilizing their social network structure, such as followers network, reply networks, etc., along with the content. However, previous studies either focus on one piece of article (or post) or network structure with one piece of article (or post) but not the combination of posts posted by the user along with their network structure.

Research Question. Inspired by the above limitations, our research question is as follows, *Can we build a framework that utilizes both the network and posts attributes of the users in order to better identify the misinformation¹ (rumor) spreaders?*

This chapter focuses on identifying the “possible” *rumor spreaders*. We define *rumor spreaders* as those users who are often engaged in spreading the rumors [Zub+18], and the term “possible” in our work points to the fact that it is very likely that the user could be a *rumor spreader*. Identifying possible *rumor spreaders* is important, as they could be the potential source of misinformation propagation. Curbing on such users means controlling the misinformation diffusion as well. However, this problem has not acquired significant attention in contrast to detecting *rumors*. This is primarily due to the lack of an annotated dataset

¹Misinformation is a broader term including fake news, rumor, etc. and our framework can be applied to any form of misinformation.

about *rumor spreaders* [HGC19]. Thus, identifying possible *rumor spreaders* is challenging in many aspects, which is the theme of this research work.

In brief, we use the **PHEME** dataset, which contains *rumor* and *non-rumor* tweets about five incidents, that is, i) *Charlie hebdo*, ii) *German wings crash*, iii) *Ottawa shooting*, iv) *Sydney siege*, and v) *Ferguson* occurred between 2014 and 2015. In order to transform the tweets dataset into rumor spreaders dataset, one of the techniques we utilize is the weak supervised learning technique, which is a branch of machine learning used to label the unannotated data using few or noisy labels in order to avoid the expensive task of manual annotation of the data [RVH19]. Please note that the labels that are generated using this approach are the *near ground-truth* labels. Thus, we use the term ‘possible’ for rumor spreaders dataset.

After the data transformation step, we leverage the following three distinct features for classifying possible *rumor spreaders*:

1. **User Features:** This includes users’ features such as *number of followers*, *number of favorites*.
2. **Text Features:** We exploit users’ tweets as a second set of features.
3. **Ego-Network Features:** We also explore the network of users who posted the tweets and the users who respond to those tweets. We illustrate this using Figure 2, wherein the central node **I** is a node who has posted a tweet, which we refer to as initiator user and the nodes **R1**, **R2**, **R3**, and **R4** are the responder users who have replied to the user **I**’s tweet. The weights on the edges correspond to the number of times user **I**, and responder user interacted with each other. Based on our dataset, we are only able to create a one-hop network (ego-network) comprising of the initiator user and its responders.

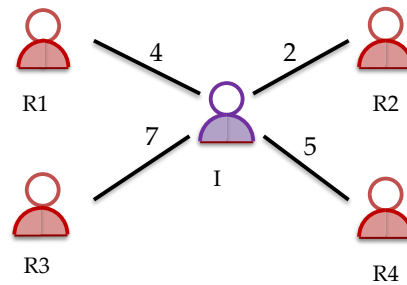


Figure 2: Undirected Weighted Ego-Network of a Twitter User

It is important to note that the network formed between the initiator of the tweets and their corresponding responders naturally calls for a technique that can exploit network features as well. Thus, in order to capture the network properties, we employ Graph Neural Network (GNN) based approach for identifying possible *rumor spreaders*. To be specific, we explore Graph Convolutional Networks

(GCN), a particular type of GNN for our analysis. In contrast to the baseline approaches: SVM, RF, and LSTM, the GCN approach exhibit better performance and is able to achieve a value of 0.864 for F1-Score and 0.720 for AUC-ROC Score.

The rest of the paper is organized as follows. Section 4.1 covers dataset description and methodology used for identifying possible *rumor spreaders*. Section 4.2 discusses experiments and their results, and Section 4.3 summarizes the chapter.

4.1. Dataset Description and Methodology

In this section, we first discuss the **PHEME** dataset (Section 4.1.1). Next, in Section 4.1.2, we describe how we transform the **PHEME** tweets dataset into the rumor spreaders dataset. Finally, we explain three different types of features extracted from the dataset, which are provided as input to the machine learning algorithms (Section 4.1.3).

4.1.1. Original Dataset

This paper utilizes the **PHEME**¹ dataset, which is a collection of *rumor* and *non-rumor* tweets that have been extensively used in previous works [KLZ18; Zub+18]. The dataset comprises five events (or incidents) - *Charlie hebdo*, *German wings crash*, *Ottawa shooting*, *Sydney siege* and *Ferguson*. For the rest of this work, we refer to them as *Charlie*, *German*, *Ottawa*, *Sydney*, and *Ferguson*, respectively. The dataset contains information about the tweets pertaining to these incidents that have been posted as breaking news during the year 2014 – 2015. To be specific, data is provided in the form of five files, where each file is related to five particular incidents. The data in each file is stored in JSON format, having information about the source (or initiator’s) tweet and its corresponding information. Table 13 provides detailed information about various fields.

Furthermore, each source tweet has the ground-truth regarding whether the tweet is a *rumor* or *non-rumor*. Table 14, column ‘# of Tweets (%)’ provides information about the number of *rumor* and *non-rumor* tweets for each of the incidents for our analysis.

4.1.2. Transformation of Tweets Dataset into Rumor Spreaders Dataset

Due to the lack of an annotated dataset of users who are spreading the *rumors* on OSM platforms, we first transform the **PHEME** dataset (which carries information pertaining to the initiator’s tweets, such as users who replied to the initiator’s tweets, the *followers count* of the initiator user, etc.) into the rumor spreaders dataset. Table 15, column ‘Tweets’, shows the original dimensions of

¹https://figshare.com/articles/PHEME_dataset_of_rumors_and_non-rumours/4010619

Table 13: Dataset Description

No	Fields	Description
1	user id	unique id of the initiator user
2	tweet	tweet posted by initiator
3	# of followers	of the initiator
4	# of favorites	of the initiator
5	verified user	source user has verified account or not
6	reply user id	unique id of the reply user
7	reply tweet	tweet posted by reply user
8	# of reply followers	of the reply user
9	# of reply favorites	of the reply user
10	verified reply user	reply user has verified account or not
11	label	initiator's tweet is rumor or not

Table 14: Distribution of Tweets and Spreaders for each of the incidents in the dataset. The two columns on the left show the number of tweets (rumor or non-rumor) posted by the source users only. The two columns on the right include the total number of users (source users and reply users) labeled as either rumor spreaders or not.

Incidents	# of Tweets (%)		# of Spreaders (%)	
	Rumor	Non-rumor	Rumor	Non-rumor
Charlie	458 (22%)	1621 (78%)	13879 (74.2%)	4821 (25.8%)
German	238 (50.7%)	231 (49.3%)	1464 (50.3%)	1442 (49.7%)
Ottawa	470 (52.8%)	420 (47.2%)	3978 (51.1%)	3794 (48.9%)
Sydney	522 (42.8%)	699 (57.2%)	7545 (61.8%)	4658 (38.2%)
Ferguson	284 (24.8%)	859 (75.2%)	3792 (35%)	7001 (65%)

the incidents, wherein each cell value represents the total number of initiator’s tweets (36189 rows in case of *Charlie*) and the total number of features, including ground-truth labels (11 columns for all the incidents).

Table 15: Dimensions of the **PHEME** dataset at various levels

Incidents	Tweets	Rumor Spreaders	Adjacency Matrix
Charlie	36189, 11	18700, 304	18700, 18700
German	4020, 11	2906, 304	2906, 2906
Ottawa	11394, 11	7772, 304	7772, 7772
Sydney	22775, 11	12203, 304	12203, 12203
Ferguson	46064, 11	10793, 304	10793, 10793

Table 16: An Example of the Tweets on Ferguson incident

No	Initiator Tweet	Reply Tweet 1	Reply Tweet 2
1	The mother of the boy killed in #Ferguson speaking to media about the loss of her son. http://t.co/YlxEDKoeBB	[USER] guess the cops were protecting and serving the community again.	[USER] my heart aches for her! This was so wrong!
2	Police in #Ferguson once charged a man w/ destruction of property for bleeding on their uniforms after they beat him http://t.co/MRVP76sdUP	[USER] that’s not true Dudeee!! please go and read good newspapers.	[USER] not at all, but they need to change some things at #ferguson PD. [USER]

In order to identify possible *rumor spreaders*, we start by placing each user with its corresponding tweets followed by its responders and their respective reply tweets. As part of the data cleaning process, we remove non-alphanumeric characters, URLs, stopwords, punctuations, lowercase all the words and perform additional Natural Language Processing operations as well, for instance, Porter Stemming of the words. We cover the steps taken for the conversion:

Preliminary Investigation: Sentiment Analysis of the Reply tweets: It is highly likely that a tweet may have attracted multiple responses (or replies). To exemplify, Table 16 shows an example of the two tweets from *Ferguson* incident. The first row corresponds to the *non-rumor* tweet along with its replies (in this Table, we have shown only two replies, but a tweet can have any number of replies), whereas the second row corresponds to the *rumor* tweet. It can be seen that the reply tweets possess sentiments with respect to the posted tweet, which can help in identifying the *rumors*, and thus, *rumor spreaders*. For instance, *non-rumor* tweets are in support of the initiator’s tweet, hence, represents a positive sentiment. Whereas in the case of *rumor* tweets, reply tweets do not show support of the initiator’s tweet, indicating negative sentiment. This could be a key indication that sentiments of the reply tweets play a key role in identifying whether the tweets posted by the user is *rumor* or not.

In this regard, we first try to analyze the sentiments of the reply tweets using *TextBlob API*². Figure 3 displays the sentiments of the reply tweets with respect to *rumor* and *non-rumor* tweets for all five incidents. Specifically, the x-axis

²https://textblob.readthedocs.io/en/dev/api_reference.html

represents the positive and negative sentiments with respect to *rumor* and *non-rumor* reply tweets for all the incidents and y-axis corresponds to its percentage. To capture the same, we consider the reply tweets under the *rumor* category if the initiator’s tweet is a *rumor* otherwise *non-rumor*. It is clear from Figure 3 that reply tweets under the *rumor* category have mostly negative sentiments and vice-versa for all the incidents (we have excluded the # of neutral sentiments, which are very few in number to avoid confusion). Thus, it can be validated that the sentiments of the reply tweets can be utilized to identify *rumor* tweets, and hence, possible *rumor spreaders*.

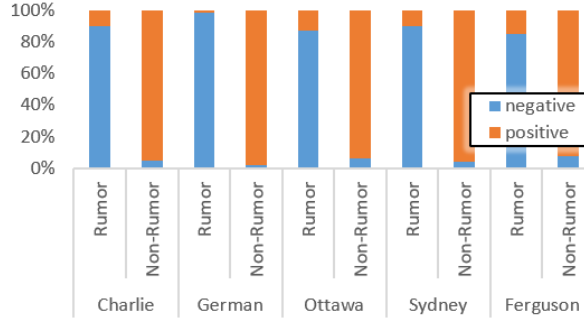


Figure 3: Sentiments of Reply Tweets with respect to Rumor and Non-Rumor Tweets for all the five incidents

Step 1 of Transformation: Labeling Reply tweets Using Weak Supervised Learning Approach:

In order to identify the possible rumor spreaders, we would like to utilize the stance of reply tweets in our approach. However, no such information is present in the dataset. Thus, in order to label each reply tweet, we apply the *MinHash*³ algorithm in line with [Nil+19]. The *MinHash* finds the similarities between each pair of the initiator’s tweet and the reply tweet. Specifically, if both these tweets are similar, then we assign the same label to the reply tweet as the initiator’s tweet, indicating that the reply tweet is in support of the initiator’s tweet. Otherwise, we assign the opposite label to the reply tweet. We considered two tweets to be similar if their similarity score is greater than or equal to 85% (this threshold is validated manually). This approach of labeling the reply tweets is what we call as weak supervised learning approach due to the fact that we do not have the manual annotation of these tweets.

Step 2 of Transformation: Calculating *rumor spreaders’ intensity score*: Tweets’ labels only indicate whether a particular tweet is a *rumor* or *non-rumor*. Therefore, in order to identify possible *rumor spreaders*, we calculate a score that indicates the intensity (high intensity means user posting more number of rumor tweets) with which users spread *rumors*, which we term as *rumor spreaders’ in-*

³snaPy API: <https://pypi.org/project/snappy/>

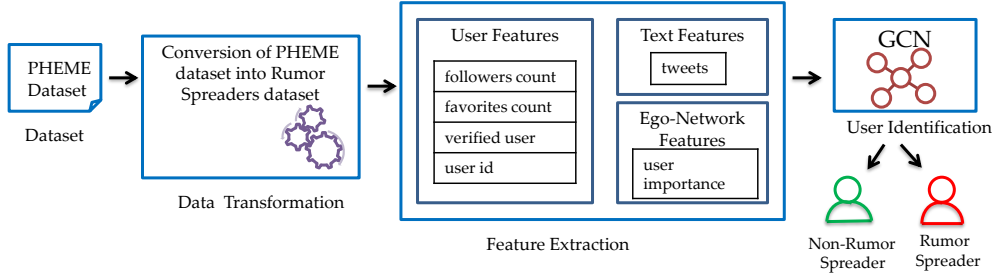


Figure 4: Framework

tensity score. We compute this score for each user by using the following formula:

$$score = \frac{\# of times user tweets rumor}{Total \# of times user tweets} \quad (4.1)$$

where the denominator is calculated by counting the total number of tweets posted by a user, whereas the numerator is calculated by counting the total number of *rumor* tweets posted by a user. The score range lies between $[0, 1]$, where 0 means not a *rumor spreader*, and 1, indicating possibly a *rumor spreader*.

In order to validate the effectiveness of this score, we calculate the degree (number of connections) from the user-user reply graph, as shown in Figure 2. We observe that the nodes (or users) who are connected to many other nodes (or users), that is, high degree, are more involved in posting *rumor* tweets as compared to nodes (or users) who have a low degree. We then manually check the users with their *rumor spreaders' intensity score* calculated using Equation 1. The score is in line with the degree, which verifies our approach of identifying possible *rumor spreaders*.

To model the problem as a binary classification problem, we put a threshold of 0.5 to create two classes of users. That is, ***if the rumor spreaders' intensity score is < 0.5, then we assign 0 label (indicating non-rumor spreader class); otherwise, we assign 1 representing the possible rumor spreader class.*** The reason for choosing this threshold is based on the observation that the sentiments of the tweets as discussed in Step 1 are positive when the *rumor spreaders' intensity score* is less than 0.5, which is indicative of *non-rumor spreaders* class and vice-versa. Based on this threshold conversion, Table 14, column '# of Spreaders (%)' shows the number of possible *rumor spreaders* and *non-rumor spreaders* for each of the five incidents in the dataset.

4.1.3. Extraction of three sets of features

In this section, we describe three distinct sets of features that we extract, to be utilized by our machine learning models for predicting possible *rumor spreaders*.

1. **User Features:** This set of features represents user's profile-based information that includes *followers count*, *favorites count*, and *verified users*.

2. **Text Features:** This feature represents text-related features, such as the *tweet* column in our dataset. We employ popular *Word2Vec* embedding (a specific type of Vector Space Models) [Mik+13] to convert the tweets into a numeric vector. Specifically, for each unique word in the *tweet*, *Word2Vec* generates its corresponding 300-dimension numeric vector, which is then aggregated in such a way that each sentence represents a 300-dimension vector. Besides, we remove few noisy words as well, for instance, aaand, aand, aaaaand using English vocabulary.
3. **Ego-Network Features:** To capture the network properties in the data, we created a feature which we call as *User Importance* (*user_imp*) feature. This feature helps in adding the network properties in the dataset by calculating the importance of each user with respect to replies it has received. The formula for calculating this feature is as follows:

$$user_imp = \frac{\# \text{ of replies a user gets}}{\# \text{ of replies of all users}} \quad (4.2)$$

Figure 4 summarizes the framework, which we discussed in this Section. Section 4.2 discusses the parameters used by GCN and the results of various approaches.

4.2. Experiments and Result Section

In this section, we discuss few of the main hyper-parameter tuning used by our models (Section 4.2.1). Next, we discuss various optimization techniques (Section 4.2.2), and lastly, we discuss the results of our approaches (Section 4.2.3).

As mentioned previously in Figure 2, there is a network formation between initiators and responders in terms of tweets. Thus, we use GCN [KW16], a type of GNN technique, which can exploit both the network structure and features (such as *tweets*, *followers count*). In particular, we perform a binary classification for identifying if a specific Twitter user is possibly involved in *rumor* spreading or not. In addition, we compare GCN with other baseline approaches (SVM, RF, and LSTM).

4.2.1. Hyper-Parameters tuning

Table 17 shows the values for various hyper-parameter settings that are chosen for fine-tuning the models for improving the performance. Here, we have specified few of the main hyper-parameters only. For instance, SVM has a hyper-parameter, *kernel*, which is a function that finds the similarity score between two data points even from the high dimensional input space in order to find the optimal hyper-plane. RF has a hyper-parameter known as, *number of trees in forest* that specifies how many trees should be formed so that they can be used as a parallel estimators in order to make final prediction. The rest of the hyper-parameters Table 17 are

specific to the neural networks (for both LSTM and GCN). The *number of layers in a network* represents the total number of layers used in a model, *number of channels in each layer* represents the total number of output neurons to be used in a layer, *drop out layer* is used to avoid overfitting of the data. The *activation function* decides which neuron to activate for the next layer in the network, *number of epochs* refers to the number of iterations the neural network model is to be trained for. The *loss function* is used for refining the model after every epoch. The NA values in the table refer to Not Applicable.

Table 17: Hyper-Parameters Tuning of models

Parameters\ Models	SVM	RF	LSTM	GCN
Kernel	radial basis function	NA	NA	NA
# of trees	NA	100	2	2
# of layers	NA	NA	2	2
# of channels in first layer	NA	NA	32	32
# of channels in second layer	NA	NA	2	2
Drop out layer	NA	NA	0	2
# of Epochs	NA	NA	300	300
Activation function	NA	NA	sigmoid	sigmoid
Loss function	average	hinge	binary cross entropy	binary cross entropy

Apart from the Table 17 hyper-parameters, the GCN model takes two additional inputs - graph adjacency matrix and nodes' features matrix. The graph adjacency matrix stores the nodes' neighbors information in a ZXZ matrix, where Z represents the total number of nodes (users) in the dataset. The nodes' features matrix ZXF is the final matrix of the preprocessing step, where Z represents the number of nodes and F is the size of total features. As already mentioned, Table 15, column 'Rumor Spreaders', shows the dimensions of the nodes' features matrix for each of the incidents. Specifically, each cell value represents the nodes, Z and the features, F where F represents the features (attributes). The nodes' features are *followers count*, word embedding of the tweet, *verified users*, *favorites count*, and *user_imp*. In addition, the dimensions of the adjacency matrix are shown in Table 15, column 'Adjacency Matrix', wherein each cell value represents the total number of users in a dataset. Utilizing these two inputs, in simple terms, GCN aggregates the neighboring nodes' features by multiplying the hidden state by the adjacency matrix [Wu+20]. Furthermore, the graph size depends upon the number of nodes in the adjacency matrix for each event (Table 15). Specifi-

cally, following the machine learning approach, the graph has been further divided into the train, validation, and test sets. After providing required inputs to GCN, the model is trained to predict possible *rumor* and *non-rumor spreaders*.

In order to predict possible *rumor spreaders*, the nodes have labels as 0 (*non-rumor spreader*) or 1 (possibly a *rumor spreader*). In Section 4.2.3, we discuss results of all the machine learning approaches.

4.2.2. Optimization Techniques

As part of the optimization, we perform following steps -

1. **Cross Validation:** To ensure the effectiveness of our model and to avoid overfitting of data, we perform K-Fold cross-validation on our dataset where $K = 5$. To avoid the class imbalance problem in *Charlie* incident, we use Stratified K-Fold.
2. **Standardization:** All the features are standardized before training the machine learning model.
3. **Feature Importance:** In addition to the above two techniques, three feature selection techniques, namely, *Chi-Square*, *Information Gain*, *Gain Ratio* are applied to each of the five incidents of rumor spreaders dataset to check whether each feature is correlated with the target variable. Table 18 depicts the p-values of three feature selection techniques. For all the features, p-value < 0.05 which shows that these features are important in predicting possible *rumor spreaders*. Thus, we consider all these features in our experiments.

Table 18: Feature Selection Techniques (values in the cells indicates their corresponding p-values)

Features	Chi-Square	Information Gain	Gain Ratio
followers count	5.28e-05	4.62e-07	8.56e-09
favorites count	8.66e-17	11.65e-22	5.56e-11
verified users	2.26e-10	1.44e-05	3.67e-07
user importance	7.87e-20	9.93e-23	7.83e-26

4.2.3. Models Performance

In this section, we discuss the micro and macro-analysis of our evaluation for all the incidents using five metrics.

1. Macro-Analysis: Table 19 provides the macro-averaged results of our machine learning models for all the five metrics for *Charlie*, *German*, *Ottawa*, *Sydney*, and *Ferguson* respectively. In general, GCN outperforms other classifiers in all the five metrics. SVM and RF perform better than LSTM in most of the metrics. Considering per incident evaluations, the results of *German* incident outperforms with a significant margin, whereas *Charlie* results exceed with a small margin. In

Table 19: Metrics performance of different models

S.No.	Metrics	SVM	RF	LSTM	GCN
Charlie					
1	Accuracy	0.748	0.760	0.671	0.790
2	Precision	0.748	0.758	0.571	0.790
3	Recall	0.748	0.628	0.571	0.790
4	F1-Score	0.853	0.840	0.778	0.864
5	AUC-ROC	0.600	0.600	0.570	0.690
German					
6	Accuracy	0.552	0.567	0.541	0.715
7	Precision	0.553	0.567	0.541	0.717
8	Recall	0.552	0.567	0.541	0.716
9	F1-Score	0.572	0.567	0.546	0.709
10	AUC-ROC	0.552	0.566	0.540	0.720
Ottawa					
11	Accuracy	0.567	0.565	0.552	0.675
12	Precision	0.567	0.565	0.552	0.681
13	Recall	0.566	0.565	0.552	0.677
14	F1-Score	0.578	0.569	0.559	0.655
15	AUC-ROC	0.566	0.566	0.550	0.680
Sydney					
16	Accuracy	0.618	0.639	0.561	0.655
17	Precision	0.565	0.606	0.541	0.655
18	Recall	0.520	0.579	0.542	0.664
19	F1-Score	0.751	0.740	0.638	0.690
20	AUC-ROC	0.618	0.638	0.540	0.660
Ferguson					
21	Accuracy	0.652	0.675	0.585	0.705
22	Precision	0.598	0.634	0.549	0.671
23	Recall	0.519	0.589	0.549	0.658
24	F1-Score	0.782	0.778	0.676	0.783
25	AUC-ROC	0.652	0.674	0.550	0.660

spite of using Stratified K-Fold, the class imbalance problem might have affected the *Charlie* results.

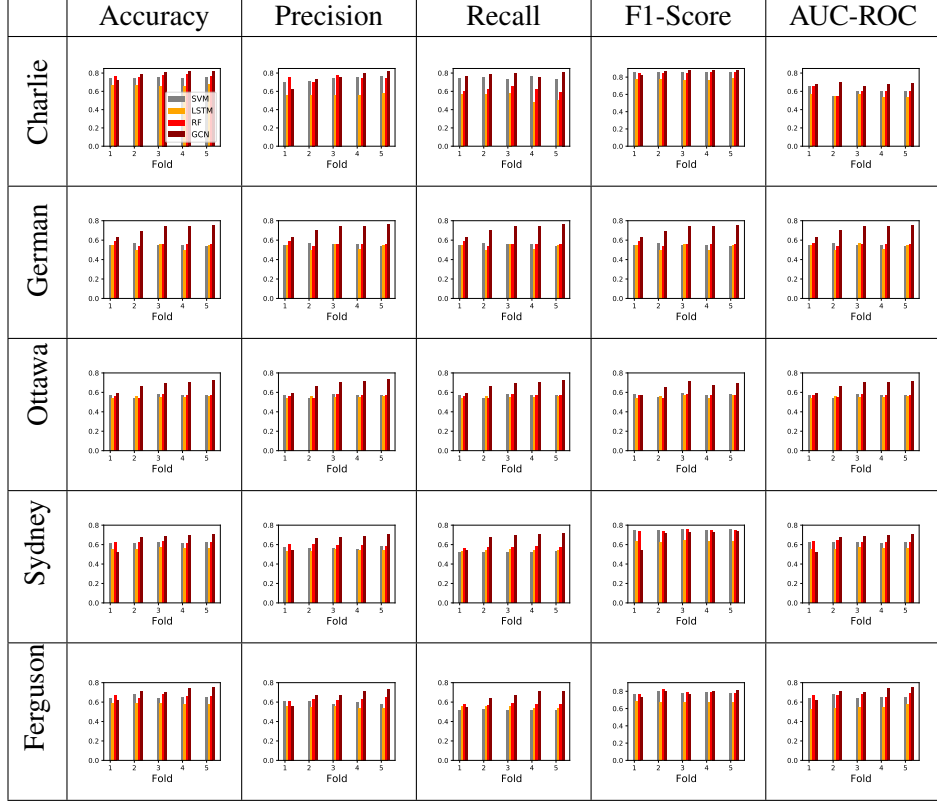


Figure 5: Performance of GCN model under five-fold cross validation for all the incidents

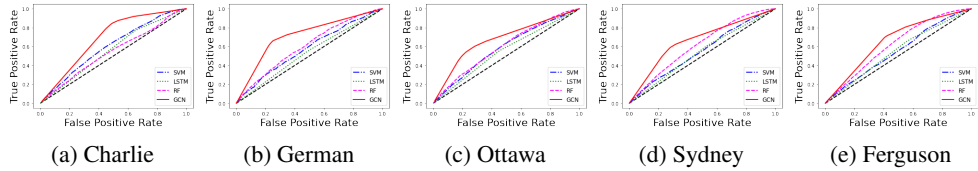


Figure 6: Performance of AUC-ROC Curve on all the five incidents

2. Micro-Analysis: Figure 5 shows the micro-performance of each of the approaches under five metrics for all the incidents in our dataset. Specifically, the x-axis represents the fold number under five-fold cross-validation, whereas the y-axis represents the metrics used for evaluation. To summarize, each incident consisting of five plots depicting Accuracy, Precision, Recall, F1-Score, and AUC-ROC Score for each fold. It can be seen from the plots that, in general, the GCN approach is performing better than the baseline models. However, GCN approach on *German* and *Ottawa* incidents performed significantly better compared to the

other two approaches. This clearly implies that the GCN approach is a natural fit for our problem statement.

Furthermore, we plot the AUC-ROC plot for each of the incidents to understand the micro-performance of the algorithm at different thresholds. Figure 15 shows the AUC-ROC Curve for *Charlie*, *German*, *Ottawa*, *Sydney*, and *Ferguson* respectively. It can be noticed that SVM, RF, and LSTM perform little better than the random model, whereas GCN is better with a good margin. However, RF performs better than the other two baselines in the case of the *Ferguson* incident. Besides, the GCN approach performs well on lower thresholds for *German*, *Ottawa*, and *Sydney*. In contrast, the reverse is the case for *Charlie*, and *Ferguson* indicating the reasons, higher values of Accuracy, Precision, Recall, and F1-Score than AUC-ROC values for the *Charlie* and higher values of Accuracy, Precision, and F1-Score than AUC-ROC values for the *Ferguson* incident. In all cases, the results are indicative that GCN is able to exploit the relation among the initiator’s tweet and its responders, which helped it to perform better.

4.3. Summary

This work utilized the **PHEME** dataset to identify possible *rumor spreaders*. Our first task was to utilize existing techniques to transform the rumor detection dataset into the rumor spreaders dataset which is otherwise challenging to obtain the user-level data wherein more than one post is used for the classification of users. The second task involved building a better framework for the classification task by employing both network and post attributes of the user. Specifically, using a weak supervised learning approach, we model this problem as a binary classification task by applying various machine learning models to the transformed rumor spreaders dataset. In the classification task, our results show that GCN (compared to baseline models) is able to perform better (to raise red flags for possible *rumor spreaders*) by exploiting relationships of possible *rumor spreaders* who could blend well with *non-rumor spreaders*. The overall performance of the GCN shows the effectiveness of this approach. This work is also feasible even if we have few replies. To be specific, this approach is specifically designed to examine the network structure created by the initial post and the replies that followed. Though it is true that fake news articles often receive fewer or the same number of replies than real news [Shu+20; SSD22; Jag+21], social media users do reply to misinformation (or non-misinformation) posts, especially if it concerns any breaking news or trending topics. Therefore, even if there aren’t lots of replies to a post or a few posts, our approach still works. In case there is only one reply to each post, our method can still operate, although the network formation will make a lesser impact on the detection mechanism.

Limitations. One of the limitations of this study is that the topics are related to crime and terrorist attacks. Diverse topics could help in understanding the pattern more efficiently. Another limitation is the formal evaluation of the rumor spread-

ers' annotations. As the manual validation of the rumor spreaders' annotations is not feasible, so, we used a semi-manual approach (*degree and sentiments*) to validate that the annotations are at par with our approach and understanding of the problem. Still, the formal evaluation is considered to be a good measure to assess the quality of the annotations.

5. PLAN OF ACTION TO COMBAT THE MISINFORMATION MENACE

Misinformation has emerged as a major societal threat in recent years; specifically in the COVID-19 pandemic, it has wreaked havoc, for instance, by fuelling vaccine hesitancy. The fears, uncertainties, and doubts (FUD) around COVID-19 and the available vaccines have been amplified by the ongoing discourse on social media - leading to an infodemic (a portmanteau of information and epidemic, referring to the spread of possibly accurate and inaccurate information about a disease, itself spreading like an epidemic). The infodemic has been fuelled by prolific misinformation spreaders, e.g., the ‘disinformation dozen’ [The21], who may have special interests and agendas in doing so. [MW21] studied vaccine hesitancy in several low and middle-income countries (LMIC) and compared that with the US and Russia, which were some of the countries at the forefront of COVID-19 vaccine research. The average vaccine acceptance rate in LMICs was reported to be 80.3%, compared to 64.6% in the United States and 30.4% in Russia. Such vaccination hesitancy has made the impact of a harsh pandemic much worse than necessary, e.g., a study estimated a factor of up to 8 more fatalities [Mes21] attributed to this.

Given the pivotal role of social media both as a medium of propagation of misleading information as well as an easy source to probe the prevailing sentiments at scale, we chose to use openly accessible COVID-19 vaccine-related tweets to drive our research. Unlike traditional surveys, for example, [Bos+21], which are able to collect and analyze detailed demographic breakdowns such as age, race, gender, education, income, geography, etc., our approach is not granular. However, the advantages of sensing information from open social media are its sheer scale and the possibility to carry out continuous and real-time monitoring once the data pipeline is set up¹.

Research Question. Based on the above mentioned points, this chapter explored how existing information obtained from social media and augmented with more curated fact-checked data repositories can be harnessed to facilitate automated rebuttal of misinformation at scale. While the ideas herein can be generalized and reapplied in the broader context of misinformation mitigation using a multitude of information sources and catering to the spectrum of social media platforms, this work serves as a proof of concept, and as such, it is confined in its scope to the only rebuttal of tweets, and in the specific context of misinformation regarding COVID-19. It leverages two publicly available datasets, viz. FaCov (fact-checked articles) [Sha+22] and misleading (social media Twitter) [SSD22] data on COVID-19 Vaccination. This leads to our research question, that is, *How accurately can we provide recommendations to social media users in order to mitigate misinformation?*

¹This work presents only a retrospective analysis.

In this chapter, we first focus on the data collection followed by the analysis of the Twitter social media (Section 5.1) and fact-checked articles (Section 5.2) corpora. Next, utilizing these corpora for providing recommendations to misinformation corpus afterward to combat misinformation (Section 5.3).

5.1. Misinformation Social Media Corpus

In this section, we discuss the collection process of the Twitter corpus to be used for the recommendations later on. Please note that our approach can be generalized to any form of misinformation (such as fake news, rumors, misleading, etc.), therefore, we interchangeably use the terms misleading and misinformation.

We collect a moderate-sized representative corpus of tweets (over 200,000) pertaining to COVID-19 vaccination spanning a period of seven months (September 2020 - March 2021). Following a Transfer Learning approach, we utilize a pre-trained Transformer-based XLNet model to classify tweets as Misleading or Non-Misleading and manually validate the results with random subsets of samples.

Furthermore, in order to validate that the data collected is sound enough to be used for the task of rebuttal of misinformation, we performed the following analysis. We leverage the data to study and contrast the characteristics of tweets in the corpus that are misleading in nature against non-misleading ones. This exploratory analysis enables us to design features such as sentiments, hashtags, nouns, and pronouns which can, in turn, be exploited for classifying tweets as (Non-)Misleading using various Machine Learning (ML) models in an explainable manner. Specifically, several ML models are employed for prediction, with up to 90% accuracy, with the importance of each feature explained using SHAP Explainable AI (XAI) tool.

5.1.1. Data Collection

The first news of the world’s COVID-19 vaccine registration² in August 2020, as well as Trump’s order to carry out vaccination even before it had been thoroughly tested and approved³, signaled a vaccine rush. Naturally, this amplified the discussions around Covid-19 vaccines both offline and online. We examine the type of discussions about COVID-19 vaccination that spread over Twitter since Twitter provides an easy to mine source of information representing all the important narratives. We collected tweets related to COVID-19 vaccination from September 2020 to March 2021. Most countries had begun vaccine rollouts⁴ as of March 2021. A significant mass of anti-vaxxers remain. Nevertheless, gradually, people

²<https://www.theguardian.com/society/2020/apr/18/coronavirus-vaccine-trials-could-be-completed-by-mid-august>

³<https://www.bbc.com/news/world-us-canada-53899908>

⁴<https://www.cnn.com/2021/03/25/covid-live-updates.html>

are less hesitant to get vaccinated⁵, and most of the current narratives and discussion points were already formed and matured in the aforementioned period. As such, we focus on the period leading up to large-scale vaccine roll-out, which is critical for studying and understanding the nature of misinformation and its spread.

Given the Twitter API restrictions, collecting an exhaustive dataset was impractical. As such, we aimed for a representative sample instead. We queried the Twitter Streaming API with a variety of relevant keywords: *vaccine*, *covid-vaccine*, *chinesevirus*, *VaccinesWork*, *NoMasks*, *antivaxxer*, *antivaccine*, *antivax*, *COVID19*, *covax*. Since there are no precise keywords for the Misleading category, we also considered anti-vaccine tweets to cover a broader range of aspects even though not all Misleading tweets are anti-vaccine and vice-versa. This resulted in over 200,000 tweets. After filtering the tweets⁶, the final dataset used in this study has 114,635 tweets in English.

5.1.2. Data Annotation Process

In order to perform the analysis, we further process and label the tweets as Misleading and Non-Misleading.

Misleading, a type of misinformation, is defined as fabricating an issue or an individual that is not true [Mol+21; Zub+18]. Specifically, after applying standard NLP techniques to clean the tweets, we designate a tweet to be Misleading when the content of a tweet deviates from the evidence shared by news media or reputable sources such as the WHO (World Health Organization), and even if it uses facts in parts, it might add connotations to it that encourages vaccine hesitancy. Otherwise, we consider the tweet as Non-Misleading. For instance, the tweet, **“I know its so bloody weird that Gates funded every single vaccine moderna Pfizer Oxford etc. Something my gut says dont have the vaccine.”** spins a narrative to increase people’s distrust. Thus, in this work we consider it as a Misleading tweet. The partial truth in the Misleading tweets could be the result of the widespread COVID-19 vaccination half-truths and myths⁷. These narratives are designed to prompt vaccine hesitancy [Cos+20] and eventually become the root cause of further Misleading information.

Since data annotation is a costly and labor-intensive method, thus, first manual labeling of a sample subset of tweets is performed by human annotators, which is used for model parameter tuning and validation when we explored the **Transfer Learning** [PY09] approach to annotate each tweet in the complete dataset as Misleading or Non-Misleading. To that end, we first took a random sample of 1,500 tweets from the dataset. This sampling was agnostic of the topics these tweets might be about. We manually annotated them as either Misleading or

⁵<https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021>

⁶We removed tweets that aren’t relevant to COVID-19 or are not in English.

⁷<https://edition.cnn.com/2020/12/18/health/myths-covid-vaccine-debunked/index.html>

Non-Misleading tweets. Three annotators labeled the sample tweets. Then, using Fleiss’ kappa score [McH12] (k) to test the quality of the dataset annotation, an **inter-annotator** agreement amongst three annotators was performed, producing $k = 0.84$, validating a high quality of agreement. The resulting labels had the ratio 47.7:52.3 among Misleading and Non-Misleading tweets.

We experimented with the sampling of tweets by selecting various sample sizes and utilizing five-fold cross-validation. Initially, we fine-tuned the pre-trained XLNet transformer model⁸ with a sample size of 100. The results, as can be seen in Table 20, are less promising and show the model’s underfitting. Next, we increased the sample size to 500 and observed an improvement over the previous performance. We continued with increasing the sample size. We observe that the performance dropped when the sample size was 2,000. This is because while the training set’s accuracy, precision, recall, and other metrics all reached 0.92, the test set’s results were noticeably poorer. Consequently, we used 1,500 samples to fine-tune the pre-trained XLNet model. In order to demonstrate the stability of the results, we repeated these experiments five times but with different random seeds (1, 124, 2012, 2022, 1000).

Table 20: Selection of various sample sizes to fine-tune the pre-trained XLNet model. The values in the cells denote the average performance of the five-fold cross-validation along with the standard deviation. ACC, PR, RC, F1, and AUC represent Accuracy, Precision, Recall, F1-Score, and AUC ROC Score, respectively.

Sample size	ACC	PR	RC	F1	AUC
100	0.25 $\pm$ 0.012	0.24 $\pm$ 0.005	0.24 $\pm$ 0.022	0.24 $\pm$ 0.010	0.26 $\pm$ 0.100
500	0.42 $\pm$ 0.23	0.45 $\pm$ 0.023	0.44 $\pm$ 0.012	0.45 $\pm$ 0.003	0.44 $\pm$ 0.302
1000	0.72 $\pm$ 0.009	0.77 $\pm$ 0.120	0.77 $\pm$ 0.231	0.77 $\pm$ 0.011	0.73 $\pm$ 0.003
1,500	0.86 $\pm$ 0.001	0.86 $\pm$ 0.002	0.88 $\pm$ 0.001	0.86 $\pm$ 0.111	0.86 $\pm$ 0.012
2,000	0.68 $\pm$ 0.113	0.73 $\pm$ 0.023	0.72 $\pm$ 0.013	0.72 $\pm$ 0.112	0.72 $\pm$ 0.302

After choosing 1,500 as the sample size for the tweets, we then finally fine-tune the XLNet language model using the manually labeled tweets. XLNet is an extension of the Transformer-XL model and learns the bidirectional contexts. BERT and RoBERTa were also tested. XLNet outperformed all other models in our case. The 1,500 manually annotated sampled tweets were split into an 80:20 train and test set utilizing a five-fold cross-validation technique. Specifically, 300 out of 1,500 tweets were used as a test set, and the rest were used as a training set. Table 21, row 1 shows the evaluation metrics calculated on the test set. The results shown are the average value on the five test sets obtained using the five-fold cross-validation technique. Subsequently, we used the fine-tuned XLNet model to annotate the entire dataset. To validate the efficiency of the annotated tweets, we manually verified 1,000 random tweets. This sample was also balanced in terms of the predicted labels. Table 21, row 2 shows the evaluation metrics of this

⁸https://huggingface.co/transformers/model_doc/xlnet.html

validation. The accuracy for the sample was 0.86, indicating that our model had been well-trained, and we can rely on these labels to get further insights.

Table 21: Evaluation metrics of the fine-tuned XLNet model. ACC, PR, RC, F1, and AUC represent Accuracy, Precision, Recall, F1-Score, and AUC ROC Score, respectively.

Experiment	# of samples	PR	RC	F1	ACC	AUC
Test set	300	0.86	0.88	0.86	0.86	0.86
Validate	1,000	0.88	0.88	0.89	0.88	0.88

For a detailed analysis of the misleading corpus, please refer to Appendix Appendix A.

5.2. FaCov: Fact-Checked Articles Corpus

In this section, we present FaCov⁹ – a collection of full-length articles that discuss and determine the truth value of widespread news and claims around the coronavirus. The articles are collected from 13 prominent fact-checking websites that discuss such news in detail and contain accompanying information such as authors, dates, truth labels, summary, links to the articles, and the full-length fact-check article itself, as discussed in the 5.2.1 section. Also discussed in this section, we provide two types of truth classifications to each of the articles: 3-class, with labels like *True*, *Mix*, *False*, and a stricter 2-class, with just *True* and *False* labels.

This dataset spans the time frame of the first two years of the pandemic and addresses limitations of existing data collections on the Covid-19 infodemic in the following manners:

1. We cumulate COVID-19 articles from multiple fact-checking websites, as such websites discuss and validate posts from various social media networks, so diversifying our dataset.
2. Our dataset is rich enough to include a variety of information about COVID-19 and its variants, along with insights such as named entities cited in the dataset, frequently appearing words and phrases, social media platforms mentioned in the articles, and prevalent myths surrounding COVID-19 according to the WHO.

5.2.1. Data Collection

We present the construction of the dataset in this section by going over aspects such as the discovery of sources, collecting articles, and extracting features to generate some insights into the data, which later help in examining the COVID-19 discourse and finally, the recommendations.

⁹<https://www.doi.org/10.5281/zenodo.5854656>

Discovering Sources: . For the collection of fact-check articles on COVID-19, domains of various fact-checking internet sources were collected from Google Fact Check Tools¹⁰. This was accomplished by querying terms and phrases such as “coronavirus,” “COVID-19,” “pandemic,” “2019-21 coronavirus pandemic,” and their variations, as well as any similar terms provided by the tool. The results comprised of fact-check articles from several websites, and the domains of these websites were noted. This helped in narrowing our search down to those websites that contain relevant news pertaining to COVID-19. We shortlisted 13 such anglophone fact-check websites based on their popularity as well as the quality and quantum of relevant content on those sites, and their domains are listed in Table 22, column 1. The reason behind working with these 13 websites was that COVID fact-check articles from these sites were the ones that surfaced in fact-check searches on Google Fact Check Tools. Here, by surface, we mean an appearance in the first 100 search results. This signaled that these websites are popular and carry fact-check articles on widespread claims. Moreover, the Tranco list¹¹ generated on 30 November 2021 is also used to determine whether or not the websites are genuinely popular. We also looked at a few additional fact-checking websites. We were unable to scrape them, however. As a result, we chose only these 13 websites, and their ranking as per the Tranco list is shown in Table 22.

So, we conclude that the articles from these 13 sites are indeed important and thus, can provide insight into the subjects being addressed in viral social media and news items, as well as in online social discussions on COVID-19.

Collecting Articles: . The articles were collected by web-scraping pages from the websites collected earlier, using the Web Scraper¹² browser extension. More specifically, the sections of these websites that dealt exclusively with COVID-19 related content were scraped. In cases where the website did not have such a specified section, the search functionality within the website was used to query terms related to COVID-19, and the articles in the search results were scraped. Also, in some cases, all articles were scraped, and those unrelated to COVID-19 were filtered out in the pre-processing stage. All the articles collected were then put together into one CSV. The final total number of rows in the dataset is shown in Table 22.

The following information was extracted along with the articles:

- Title of the fact-check article
- URL of the fact-check article
- Claim being discussed in the article (if available)
- Summary of the fact-check article (if available)
- Content of the fact-check article

¹⁰<https://toolbox.google.com/factcheck/explorer>

¹¹ Available at <https://tranco-list.eu/list/X74N>

¹²<https://webscraper.io/>

Websites	Rank	w/o Pre	w/ Pre
reuters.com	120	267	116
usatoday.com	177	100	24
indianexpress.com	947	11	5
rappler.com	3044	160	100
afp.com	3210	454	451
politifact.com	3332	1,170	1,169
factcheck.org	5287	735	337
indiatvnews.com	7310	79	25
thelogicalindian.com	53104	1,211	215
boomlive.in	108250	774	474
polygraph.info	244920	449	112
factchecker.in	430705	181	41
covid19factcheck.com	-	19	19
# of articles	-	5,610	3,088

Table 22: Ranking of the selected websites based on the Tranco list (sorted in descending order). Also, columns w/o Pre and w/ Pre represent number of articles without (before) pre-processing and with (after) pre-processing, respectively.

- Label assigned by the article to the claim
- Author of the fact-check article (if available)
- Date of publication of the article (if available)

Pre-processing and data-cleaning: . Pre-processing involved searching the CSV for COVID-related terms and manually reviewing the articles that contained those terms. Irrelevant entries were thus removed, and any entries that contained empty or null values were also eliminated. Furthermore, for articles from some websites (for instance, FactCheck.org), the labels were not explicitly available or accessible for scraping but rather had to be retrieved from the article’s content. In such cases, we went over each article’s content and labeled them manually. We initially gathered a total of 5,610 fact-check articles and the information that accompanied them, which after the pre-processing totaled **3,088** which became a part of the final dataset, as represented in Table 22. Additionally, Table 23 includes the absolute numbers of non-null values in the dataset. In addition, we found no duplicate claims in the dataset.

Time frame:. The fact-check articles in the final dataset cover the period from the start of COVID-19, December 2019, till the first week of December 2021, spanning a period of roughly two years.

5.2.2. Extracted Features

Article Summarization:. One of the columns present in the dataset is the *original summary*. However, as this field had 26% null values, we construct the extractive summary [MG14] programmatically using Term Frequency-Inverse Doc-

Attribute names	Absolute number of the non-null values (out of 3,088)
title	3,088
URL	3,088
claim	2,540
summary	2,286
content	3,088
label	3,088
author	2,956
date	2,929

Table 23: Attribute names with their absolute numbers of non-null values in the FaCov dataset.

ument Frequency (TF-IDF) approach, an NLP-based technique [CAS16]. Deep learning-based approaches were also investigated. However, the input size for the majority of the models is limited to 512 words. On average, our *Content* has 5000 words. Other deep learning models that allow for longer text sizes, such as the popular BigBird [Zah+20] transformer model, did not perform well on our dataset compared to the NLP-based technique.

The TF-IDF is a numerical statistic that indicates how important a term is to a document in a corpus. Its value increases in proportion to the number of times a word appears in a document but is offset by the frequency of the term in the corpus, which helps regulate the fact that certain words are more frequent than others. Following the TF-IDF approach, one required parameter is the number of important sentences to include in the summarization. Thus, we initially compute the mean (μ) and standard deviation (σ) of the number of sentences in each article. Next, we used ($\mu + \sigma$) value as the total number of sentences to include for summarizing every article, which was five. The programmatically generated summary is with the column name as *generated summary*.

Data Annotations: . Table 24 refers to scraped websites that have a variety of labels from multiple fact-checking websites such as *False*, *True*, *Mostly True*, *Partly True (Half True)*, *Mostly False*, *Pants on Fire*, *Yes*, *No*, *Misleading*, *Altered*, *Missing Context*, *Misrepresented*, and *Manipulated*.

We simply aggregate these raw labels into three classes to further analyze data based on the labels and avoid ambiguity in the diversity of labels issued to each article by various fact-checking websites. In addition, two-class labels are included for simplicity, though the three-class labels provide richer context and are used for the analysis of the data presented in this work.

1. 3-class (*True*, *Mix*, *False*): In this grouping, the articles that fall under *True*, *Yes* class are put into *True* class, *False*, *No*, *Fake*, and *Pants on Fire (PoF)* articles are put in the *False* class, and the rest of the labels mentioned come

Website	True	Mix	False	#
BoomLive	–	Misleading	False	2
FactChecker.in	True	Partly True Misleading	False	4
IndiaTV News	True	Partly True	False	3
Reuters	–	Mostly Fake Half True Misleading	Fake	4
Indian Express	True	–	False	2
Logical Indian	True	Half True Mostly True Misleading	Fake, False	6
Covid19 Fact Check	Yes	–	No	2
USA Today	–	Altered Missing Context	False	3
AFP FactCheck	True	Misleading Misrepresented Manipulated	False	5
Politifact	True	Mostly False Half True Mostly True	PoF False	6
FactCheck.org	True	Mostly Fake Half True Mostly True Misleading	Fake	6
Polygraph	True	Likely False Misleading Unclear Unsubstantiated Disputed Uncertain Dangerous	False	9
Rappler	–	Missing Context Altered Photo Altered Video	False	4

Table 24: Truth labels retrieved from the Fact-Check websites after pre-processing under the 3-class labeling, and the number of labels.

under *Mix* class. The *Mix* class label signifies that the statements are somewhat true, but either they lack additional knowledge, context, or important facts, among others. This grouping has been decided as per the labels’

definitions in the paper [SN20]. The data distribution for the 3-class is as follows: The *True* class has 72 articles, *Mix* has 818 articles, and the *False* class has 2198 articles. Table 24 shows the distribution of labels for each of the 13 websites in the 3-class annotation.

2. 2-class (*True*, *False*): In this case, we have a *True* class for all the articles labeled as *True* and *Yes*. The rest of the articles fall under the *False* class. The *True* class has 72 articles, whereas the *False* class has 3016 articles.

5.2.3. COVID-19 Myths

There have been several myths about the COVID-19 epidemic. To explore such information in our dataset, we look into the articles that verify the veracity of these myths. We determined the myths to study based on the myth busters list on the WHO website¹³. Table 25 illustrate the myths we discovered in the *Title*, as well as additional categorization of the myths using the *True*, *Mix*, *False* labels. We only include myths that have at least one article addressing them. According to our data, the topmost myth is Hydroxychloroquine which is used to treat malaria and autoimmune diseases and can cure COVID-19. Particularly, the two fact-check articles, that is, *no, ivermectin & Hydroxychloroquine do not treat covid-19*, and *US medical association did not change stance on Hydroxychloroquine as covid-19 treatment* indicate the myths busted surrounding Hydroxychloroquine. Furthermore, unsurprisingly, the *False* label comprises most of the myths-related articles.

5.2.4. Discussions around COVID-19

We discuss the key themes surrounding the COVID-19 pandemic present in our dataset. The 3-class labels *True*, *Mix*, and *False* discussed in the *Dataset* section are used to discover the primary themes (topics). The generative model (non-transformer based models), the LDA¹⁴ approach, has been used to choose the topics. To consider a different approach from LDA, we also looked into a transformer-based model, BerTopic modeling¹⁵. However, we noticed that the LDA approach yields better results in our case. Particularly, we manually look for the topic name obtained from LDA and BerTopic for all three labels in *Content*. We observed that the LDA topics are generalized and cover the articles' whole agenda more than the BerTopic approach. Next, in LDA, we utilize the Grid Search strategy to investigate the optimal number of topics (themes) in the data.

¹³<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters>

¹⁴<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>

¹⁵<https://maartengr.github.io/BERTopic/index.html>

Myths about COVID-19, its Prevention, and Cure	# of myths in Title column	Labels		
		True	Mix	False
Ineffectiveness of Alcohol-based hand sanitizers	12	0	4	8
Hydroxychloroquine	23	0	8	15
Vitamins & Minerals supplements	7	1	1	5
Usage of Masks while exercising	5	0	2	3
Infection via Water and Swimming	19	0	0	19
Bacteria as a cause	7	0	1	6
Use of Oxygen cylinders and related news	15	0	2	13
Pepper in soup as a cure	1	0	0	1
Spread of infection through houseflies	1	0	1	0
Use of disinfectants on the human body	8	0	1	7
5G networks and relation to COVID	15	1	1	13
Exposure to sun for protection from infection	6	0	1	5
Changes in Life-insurance policies	7	0	2	5
Holding breath as a test for COVID infection	4	0	1	3
Snowy weather as protection	1	0	0	1
Hand dryers as a preventative measure	1	0	0	1
Vaccines against pneumonia as a measure for COVID	3	0	0	3
Rinsing nose with saline water to flush out the infection	7	0	0	7
Consuming garlic	5	0	0	5
Use of antibiotics against the virus	2	0	0	2
# of articles discussing the myths	149	2	25	122

Table 25: Fact-checks present in the dataset that are related to widely spread myths according to WHO, extracted from Title column. The myths are then further categorized based on the labels.

As a result, we identified three significant themes. Furthermore, each primary theme has been investigated further to uncover sub-topics.

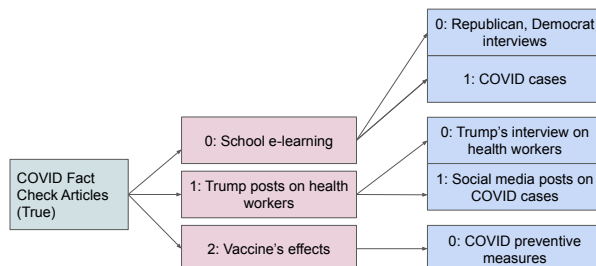


Figure 7: Key themes in True News

True News. The first primary theme is *e-learning in Schools*. It has been discussed and fact-checked whether in-person learning helps reduce COVID cases or which states in a certain country have implemented a whole e-learning system. The two sub-topics found under this theme are *Republican and Democrats' interviews* and *COVID cases*. Interviews with Republicans and Democrats discuss the economic impact of e-learning during the epidemic. The other theme discusses the graph of COVID cases after the implementation of e-learning in schools.

The second primary theme is *Donald Trump's social media posts on health-care workers*. Basically, Trump's message on social media about free vaccination for healthcare employees is accurate and is one of the primary themes in True news. Under this theme, there are two sub-topics; *Trump's interviews on health-care employees* and *his social media posts about the amount of COVID-19 cases in various states*. In addition, the same news, along with the description of an increasing number of COVID-19 cases in several states, has been popular in his conversations with health care personnel.

The third primary theme is *Vaccination's effects*. There is a well-known debate and confusion among people about the true risks and effects of vaccination. As a result, there are multiple fact-checked articles concerning vaccination. *COVID prevention measures* are the only sub-topic under this area. With getting vaccinated or not as the main issue, the importance of administering the vaccine to patients as soon as feasible has also been emphasized to prevent and mitigate the COVID-19 pandemic.

Mix News. The first primary topic in Mix news is *COVID-19 deaths in states*. Some posts claim to have counted the total number of deaths caused by COVID-19 in different states of a country, which may or may not be accurate but lacks sufficient or further information. As a result, it is always a good idea to double-check such information by studying the full article, as these forms of news might lead us astray by creating a false context. *Misleading posts, photos, and videos* is one sub-topic. As the name suggests, these posts include images or videos in addition to text. For example, photos or videos may be misleading or provide incorrect context in relation to the content. The image could be from an event that

occurred a long time ago but is being staged as a recent one.

The second primary topic is *Trump and Biden's statements about China*. There are articles about Trump and Biden regarding China; for example, they talk about Trump's claims that Biden wrote him an apology letter after criticizing travel restrictions on China. However, it was later revealed that Biden never wrote an apology, and Trump put a layer of exaggeration on top of that. Two sub-themes emerged from this theme; *Democrats' debate of COVID reports* and *Trump and Biden's elections in the midst of the pandemic*. Due to the upcoming elections, both parties became heavily active in discussions of COVID-related problems in order to gain a larger share of the vote.

The third primary topic, *Masks* (common in False news' topic as well). Wearing masks even after being jabbed has been misinterpreted by the majority of people, resulting in the dissemination of fake information. The three sub-topics are; *Misleading posts, pictures, and videos*, *Drug trials, and Immunity*, and *China's COVID reports and research*. The articles under this theme claim on misleading social media posts on wearing masks, spreading the wrong context on the effectiveness of drug trials, and generating false reports about China's cases and research indicating mask is not necessary to wear.

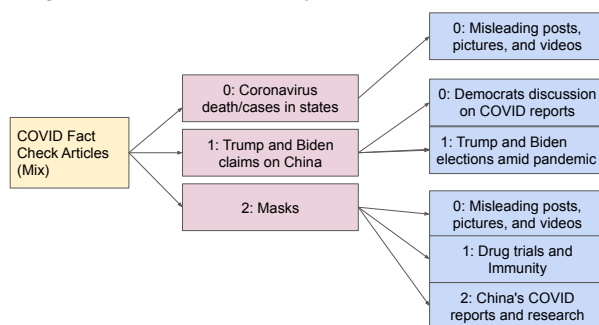


Figure 8: Key themes in Mix News

False News. The first primary theme is the *Spread of COVID-19*. There is a lot of fake news out there about how COVID-19 spreads, and people spread incorrect data records on social media. Despite WHO's efforts to dispel COVID-19 falsehoods, people continue to make erroneous claims. This kind of False news has had the greatest impact in all countries. There are three sub-topics under this theme; *False death reports*, *Masks*, and *False videos on vaccine*. The main sources of propagating the virus include false reports of celebrity deaths on social media claims that wearing masks does not help prevent COVID-19 and posting bogus vaccine videos on social media.

The second primary theme is *Misleading posts, pictures, and videos* (same topic as in Mix news). Social media has been a major source of spreading misleading information, especially in COVID-19 times. Because pictures (images) and videos increase social media visibility in terms of likes and shares, there are posts that contain either false pictures or videos that are irrelevant to the COVID-

19 issue. Following the reverse image search, the picture or video was discovered to have occurred in the past on any other occasions. This has been really concerning, and mitigating such misleading information in such a short period of time has proven tough. One sub-theme, *Lockdown news on social media*, falls under this umbrella. The news about the lockdown, such as the length of the lockdown and the directions to follow, is incorrect and lacks any accurate reporting or official links. This emphasizes that we should be cautious of any news until we receive official confirmation.

The third primary theme is *Masks*. Wearing or not wearing masks after vaccination proved to be a substantial point of discussion and the propagation of false news. This frequently discussed *China/Wuhan-related* news quoting that people in China (Wuhan) wore masks even before the COVID-19 outbreak.

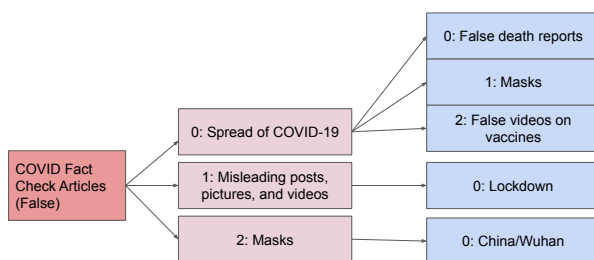


Figure 9: Key themes in False News

Furthermore, we performed additional exploratory analysis, such as which social media platforms are being used most in the articles, mentions of COVID-19 variants found in the dataset, and so on. Please refer to the original paper for a detailed analysis of this work [Sha+22].

5.3. Automated Misinformation Rebuttal (AMIR)

In this section, we now present the recommendations approach, that is, providing a rebuttal system to counter misinformation utilizing the two corpora discussed in the previous sections. This rebuttal system can be applied to any form of misinformation, thus, we use the terms misleading, fake news, and misinformation interchangeably.

Combating Misinformation: A Two-Pronged Approach

In order to automate the rebuttal of misinformation on social media, we have developed a pipeline that utilizes two complementary approaches (as shown in Figure 10). The first identifies and reuses existing related but factual/non-misleading tweets by other users as recommended counter tweets. This is the baseline approach that can be used even when fact-checking websites are yet to carry out the vetting exercise for a given misinformation, or while such articles are identified and matched by our second approach.

The second recommends fact-checking articles (subject to availability) from

various fact-checking sites. This provides users with verified, reliable, and more comprehensive information from reputable sources to counter the misinformation. By utilizing both of these automation approaches in tandem, we create a scalable rebuttal pipeline to combat misinformation on social media.

There are COVID-19 datasets available in the literature that have limitations [CL20; Che+21; Dha+22; Li+20; Pat+21; Hao+20; Kim+21; IK20; Wah+22; SN20; Hos+20]. For instance, the timeframe is limited, full-length fact-checked articles are unavailable or articles are written in non-English. Therefore, we opt to utilize the misleading social media and the FaCov fact-checked articles corpora that cover all mentioned limitations.

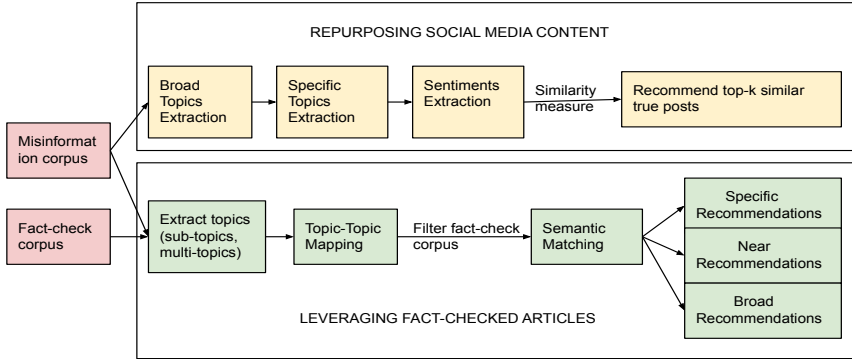


Figure 10: Automated Misinformation Rebuttal Pipeline.

Though instantiated with COVID-19 specific (mis-)information on Twitter, the applied methodology and principles are universal and generalizable, subject to training the models with suitable datasets. Likewise, the pipeline (see Figure 10) is inherently modular in nature, such that, for a given individual step or task, alternative approaches can be applied separately or even in conjunction.

About the datasets, while both datasets identify key themes surrounding the COVID-19 pandemic, the FaCov dataset also uncovers sub-topics, which we leverage in our analysis. Next, we delve into the information extracted from the datasets, which serves as the foundation for our recommendation approaches.

5.3.1. Data Preparation: Discerning (sub-)topics

We extend the analysis of the data from our own papers [Sha+22; SSD22] (discussed in the previous sections of this chapter) to examine topics at a finer granularity and study the relationships among them.

Social media (Twitter) corpus [SSD22]. We use Latent Dirichlet Allocation (LDA) topic modeling to assign topics to each tweet in the misleading corpus [SSD22]. This refines the prior work in [SSD22] by assigning specific topic(s) to individual tweets rather than simply identifying broad topics within the dataset. Specifically, if a tweet contains a specific topic, we label it accordingly. This resulted in 12 identified topics, with a small number of tweets remaining unlabeled.

To improve the labeling process for these tweets, we also checked for the presence of synonyms for each topic in the unlabeled tweets and were able to assign some tweets to the previously identified topics. Despite this, a small fraction (2.98%) of tweets remain unlabeled (labeled as **Unknown**) shown in Table 37.

Table 26: Absolute number (and %) of Tweets per topic.

Topics	Tweets: number (%)
Choices	33,150 (28.9%)
Politics	25,276 (22%)
Vaccine Efficacy	21,936 (19.1%)
Shots	9,568 (8.34%)
Trump	8,432 (7.35%)
Data & Facts	3,601 (3.14%)
Unknown	3,426 (2.98%)
Trials	3,217 (2.8%)
Myths	2,376 (2%)
Operation Warp Speed	1,369 (1.19%)
Real Side-Effects	1,216 (1.06%)
Approval	883 (0.77%)
Availability	185 (0.16%)

Table 27: Topics and Subtopics identified from the Misleading Tweets dataset [SSD22].

Main topic	Sub-topic 1	Sub-topic 2	Sub-topic 3
Politics	Operation Warp Speed	Vaccine Efficacy	Vaccine in countries
Vaccine Efficacy	Vaccine Effects on pregnant women	Trust & risk of the vaccines	vaccine-related data (deaths, illness, shot, dose)
Vaccine Choices	MRNA vaccines (allergies)	Orders of vaccines	Individual's interest on vaccines
Vaccine Shots	Second doses received in time	Single doses received in time	-
Operation Warp Speed	White house providing funding to manufacturing companies	Money involved in providing doses of various vaccines	Administration, distribution of approved vaccines
Real Side-effects	Fever & soreness after Pfizer shots	Soreness and pain after shots	Experiencing fatigue & headache
Trump	Involvement of Trump in the Operation Warp Speed	-	-
Data & Facts	Report shots of vaccines	-	-
Myths	Bill gates conspiracy theory	Deaths & diseases caused by vaccines	Severe allergies due to vaccines
Trials	Phase trials of various vaccines	Placebo effects	Participants facing illnesses
Vaccine Approval	-	-	-
Vaccine Availability	-	-	-

We similarly identify up to three sub-topics of the main topics in the misleading corpus (shown in Table 27). The number of topics and the sub-topics have been chosen based on the high coherence value in the LDA technique. There are a couple of topics, viz. **Trump** and **Data & Facts**, wherein only one sub-topic is identified. Two topics, namely, **Approval** and **Availability** have no further sub-topics.

Fact-checked articles corpus [Sha+22]. Table 28 shows the most discussed topics and corresponding sub-topics in the fact-checked articles dataset [Sha+22], capturing a variety of issues including posts involving Trump, vaccine-related

concerns, the number of deaths due to COVID-19, falsely relating images or videos to COVID-19.

It’s interesting to note that the topics identified in the misleading Twitter corpus are more fine-grained and nuanced as compared to fact-checked articles FaCov corpus. For instance, on the topic of **Vaccine Efficacy**, the three sub-topics delve into its effect on pregnant women, concerns about the vaccines, and statistics about the vaccine. Whereas in the case of the FaCov dataset, the only sub-topic of the **Vaccine’s effects** is COVID preventive measures. This is likely because Twitter is a platform where users openly express their concerns and perspectives, whereas, in the fact-checked articles FaCov dataset, only specific posts flagged by many people are fact-checked and posted on fact-checking websites. This is in line with observations drawn from other social media platforms such as Reddit [TD23], where it has also been noted that there are more nuances and diversity in user generated content in the form of comments and discussions, with respect to website article subjects.

Table 28: Topics and subtopics identified from the FaCov dataset [Sha+22]. Furthermore, items highlighted in **Blue** color represents the mapped topics from the misleading tweets dataset [SSD22].

Main topic	Sub 1	Sub 2	Sub 3
School e-learning	Republican, Democrat interviews	COVID cases	-
Trump posts on health Workers (Trump)	Trump’s interview on health workers	Social media posts on COVID cases	-
Vaccine’s effects (Vaccine Efficacy) (Vaccine Shots) (Vaccine Choices) (Myths) (Vaccine Availability)	COVID preventive measures	-	-
Coronavirus death/cases in states (Politics)	Misleading posts, pictures, and videos	-	-
Trump and Biden claims on China (Politics) (Trump)	Democrats discussion on COVID reports	Trump and Biden elections amid pandemic	-
Masks (Myths) (Trials) (Data & Facts)	Misleading posts, pictures, and videos	Drug trials and Immunity	China’s COVID reports and research
Spread of COVID-19 (Politics) (Myths)	False death reports	Masks	False videos on vaccines
Misleading posts, pictures, and videos (Myths)	Lockdown	-	-

5.3.2. Data Preparation: Relationships Among Topics in Fact-check Article Corpus

We investigate the co-occurrence of topics for each fact-checked article, considering the second most suitable topic among the same topics using the probability distribution from the LDA approach. In Tables 29 and 30 we show the distribution of the top two topics assigned to each article. The *Count* column indicates the total number of fact-checked articles that contains the specific topic. The widely discussed topics are the **Vaccine Effects** and **Trump posts on health Workers** in the topmost and second most topics distribution, indicating that the majority of

Table 29: (Multilabel topics) Topmost topics distribution in fact-checked articles FaCov dataset. The Count column indicates the total number of fact-checked articles that contains the specific topic.

Topics	Count
Vaccine Effects	967
Trump posts on health Workers	664
Misleading posts, pictures, and videos	577
Coronavirus death/cases in states	461
School e-learning	306
Spread of COVID-19	81
Masks	32

Table 30: (Multilabel topics) Secondmost topics distribution in fact-checked articles FaCov dataset. The Count column indicates the total number of fact-checked articles that contains the specific topic.

Topics	Count
Trump posts on health Workers	834
Misleading posts, pictures, and videos	635
Coronavirus death/cases in states	558
Vaccine Effects	458
School e-learning	315
Masks	137
Spread of COVID-19	130
Unknown	21

the fact-checked articles are related to clarifying the efficiency and symptoms of the various COVID-19 vaccines. The **Trump posts on health workers** topic also gained popular traction in false claims.

Figure 11 shows the relation between prominent topics. The nodes and edges of the graph represent the topics' names and if both the topics have been assigned to the same article, respectively. The edges' weight (indicated with thickness in the figure) encapsulates the number of times the same two topics are assigned to an article.

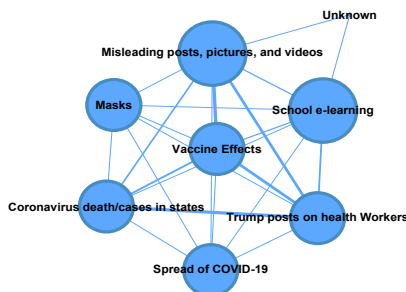


Figure 11: Co-occurrence of prominent topics in the corpus of fact-checking articles.

We identify four relatively strong relations, namely, **Coronavirus death/cases in states — Trump posts on health workers**; **Vaccine Effects — Misleading posts, pictures, and videos**; **Vaccine Effects — Trump posts on health workers**; **Trump posts on health workers — Misleading posts, pictures, and videos**. These show that the fact-checked articles mostly present the number of COVID-19 (death) cases within the topics with Trump's posts on health workers. Vaccine-related concerns have been widely linked with misleading posts, images, or videos which are wrongly linked to COVID-19. In addition, Trump's posts on health workers are linked with misleading posts, images, or videos.

Including the second most suitable topic ensures that a wider choice of relevant fact-checked articles are considered for recommendation, while the combination of top topics helps improve the accuracy of the recommendations to users with relevant and accurate information to counter misinformation (refer to the third category of recommendation in Section 5.3.4).

Knowing the Unknown: The **Unknown** topic label is linked with the top-most topics of articles, with the second most topic being the **Unknown**. There were a total of 21 fact-checked articles that were not assigned a second topic (thus **Unknown**), as reported in Table 30. We explored those 21 **Unknown** articles manually. They discuss an assortment of themes such as throwing away God's idols into the river due to COVID-19, lockdown concerns, questioning the existence of COVID-19, training drills wrongly linked to COVID-19 havoc, mutation of human vs. man-made viruses, and black fungus correlated with COVID-19. Though these **Unknown** topics may not be as popular as the other topics in the

dataset, they are still vital, since they add diversity and coverage of fact-checking of a wider variety of misinformation.

5.3.3. Repurposing Social Media Content

Our first approach to combating misinformation on social media is to utilize content from the platform itself to recommend relevant non-misinformation posts in response to a misinformation post (refer to the top box in Figure 10). This is particularly useful in situations where fact-checking articles might not (yet) be available.

To that end, we employ a combination of three Natural Language Processing (NLP) techniques. The first technique is Latent Dirichlet Allocation (LDA) topic modeling, which is used to extract general themes and topics from the posts (as elaborated previously in Section 5.3.1). We also use Named Entity Recognition¹⁶ (NER) to identify specific entities mentioned in the posts (Section 5.3.3). Finally, we match the sentiments of the posts to ensure that the recommended posts provide a balanced perspective on the topic (Section 5.3.3).

Specific Topic Extraction. We evaluated a variety of Named Entity Recognition (NER) models, including the latest pretrained roberta-base NER model, which successfully identified named entities in 57% of our tweets data. However, after careful examination, we found that the `en_core_web_sm` model performed best, identifying entities in 88% of our data with at least one entity.

Upon further investigation, we discovered a few issues with pre-existing models. Firstly, certain words specific to COVID-19, such as Pfizer, Shots, and Johnson & Johnson, were incorrectly labeled as PERSON, GPE (Geopolitical entity). Secondly, there were still 12% of entries that contained zero or null entities.

To address these issues, we fine-tuned the NER model on a subset of 100 manually labeled random instances from our dataset. We added a new entity type called VAC_TYPE for vaccine names using a list of manually labeled VAC_TYPE entities: [pfizer, astrazeneca, mrna, astrazenca, jnj, oxford, sputnik, modern, variants, #pfizer, booster, #astrazeneca, biontech, Covidshield]. Additionally, we also included spelling errors in the entities and did not remove hashtags from the dataset as they play a crucial role in identifying misinformation. We trained the augmented model for 30 epochs with a dropout rate of 0.3, using five-fold cross-validation.

We then compared the performance of the augmented model (fine-tuned) to the unaugmented model on the same subset of 100 examples, as shown in Table 31. We observed that the augmented model performed better in all metrics by a significant margin. Finally, we labeled all tweets in the dataset using the augmented model. We obtained the list of VAC_TYPE entities that are labeled by the augmented model, which included new entries beside the ones we had provided during training, viz. [phizer, myrna, zenca, novavax, johnsonandjohnson,

¹⁶<https://spacy.io/api/entityrecognizer>

johnson, mirna].

Through this process, we were able to improve the detection of entities and entity types in the dataset. In particular, for 112,994 (98.5%) of our records, at least one entity was identified (up from 88% with the unaugmented one). Besides the identification of new entities across the dataset, 36,829 previously identified entities were reclassified by the augmented model: 35,480 (96.3%) of these were of VAC_TYPE and 1,349 (3.7%) were of non-VAC_TYPE. Table 32 represents the top entities divided into VAC_TYPE and non-VAC_TYPE entities. For example, the entity Pfizer has been incorrectly labeled (in the present context) as ORG by the unaugmented model but as VAC_TYPE by the augmented model. Since most of the dataset contains vaccine names in their tweets, it clearly indicates why VAC_TYPE entities detected by the model are in such a big number.

Table 31: Comparison of the Spacy NER model with and without fine-tuning the model on the manually labeled data.

Spacy model	Accuracy	Precision	Recall	F1 Score
w/o training	0.27	0.25	0.26	0.25
w/ training	0.89	0.90	0.88	0.87

While the augmented model generally performed better, there were a few instances where it failed to identify entities accurately. For example, in a few cases, geographic places were mislabeled as VAC_TYPE by the augmented model but correctly labeled as GPE by the unaugmented model. Additionally, there were instances where both models failed to identify the entity accurately, for instance, the un/augmented models have incorrectly labeled Ohio as MONEY and VAC_TYPE, respectively. Despite these limitations, our approach resulted in a substantial improvement in the detection of named entities and entity types within the tweets.

Table 32: Top entities mislabeled by the unaugmented NER model. The left and right sides represent vaccine names and non-vaccine names entities.

VAC_TYPE			Others		
Entity	Mislabeled	Correct	Entity	Mislabeled	Correct
pfizer	ORG	VAC_TYPE	millions	CARDINAL	MONEY
moderna	GPE	VAC_TYPE	billions	CARDINAL	MONEY
astrazeneca	ORG	VAC_TYPE	trump	ORG	PERSON
johnson and johnson	PERSON	VAC_TYPE	biden	ORG	PERSON
novavax	ORG	VAC_TYPE	lock down	NORP	EVENT

Sentiments Extraction. To counter confirmation bias, we use the VADER API [HG14] to detect the sentiment of tweets as positive, negative, or neutral. To recommend related non-misleading tweets to counter misleading tweets, we employ two approaches for determining similarity: direct string matching and vector dimension matching. We found that vector dimension matching with GloVe embedding [PSM14] and cosine similarity worked best for our data.

Top-K non-misleading relevant tweets. In a nutshell, we employed three criteria - general topic, entities extracted by our augmented NER model, and sentiment

- to match misleading tweets with equivalent non-misleading tweets. The top-K (where K can be a user-defined parameter) most similar non-misleading tweets are then identified using the GloVe embedding and cosine similarity approach to be recommended as counters.

Criterion for Evaluation of First Approach. To evaluate our approaches, we employ the Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) metrics, which are frequently employed in recommendations tasks [Wan+22; Lia+16; Har+17], particularly in settings where ground truth is absent. This is particularly pertinent since it is impractical to manually evaluate the recommendations' quality. Hence, we stick to the two widely used metrics that are suitable in our case. Particularly, we adapted the definition of relevant items considered in both the mentioned metrics. To that end, we define Precision@k (P@k) as follows:

$$P@k = \frac{\# \text{ relevant recommendations to misleading tweet}}{\# \text{ recommendations to misleading tweet}} \quad (5.1)$$

Here, a recommendation is considered relevant if, after similarity matching of misleading and non-misleading tweet pairs, the non-misleading tweets satisfy all three criteria: matching topics, entities, and sentiments with the misleading tweet. The total number of recommendations involves non-misleading tweets recommendations utilizing similarity measure from the whole corpus (that is, without any criteria). The rest of the MAP calculation is the same as per the predefined formula[Lia+16].

$$MAP@k = \frac{1}{N} \sum_{i=1}^N \frac{1}{m} \sum_{k=1}^K P@k \cdot rel(k) \quad (5.2)$$

where $rel(k)$ represents whether the recommendation is relevant or not. The MRR metric remains the same as in [Xu+16].

$$MRR@k = \frac{1}{R} \sum_{r=1}^R \frac{1}{k_r} \quad (5.3)$$

where k_r is the rank of the first relevant recommendations. R is the number of misleading tweets that we seek recommendations for.

Table 33, row 1 represents the evaluation metrics MRR@k and MAP@k performances on this social media non-misleading (true) posts recommendations approach. It can be observed that at first, with an increasing value of k, performance starts to increase and then decreases. MRR@10 and MAP@10 achieve better performance meaning that ten non-misleading posts recommendations are good enough to be presented as a counter for the misleading post. Furthermore, giving users too many recommendations (though "too many" is a subjective word here) may cause them to lose interest in reading further.

Table 33: Evaluation of the complementary approaches. AMIR_SM and AMIR_FC represent social media non-misleading posts recommendations (first approach) and fact-checked articles recommendations (second approach), respectively.

	MRR@3	MRR@5	MRR@10	MRR@15	MRR@20	MAP@3	MAP@5	MAP@10	MAP@15	MAP@20
AMIR_SM	0.541	0.616	0.689	0.615	0.612	0.528	0.567	0.746	0.703	0.677
AMIR_FC	0.513	0.553	0.586	0.663	0.617	0.586	0.683	0.729	0.757	0.738

5.3.4. Leveraging Fact-Checked Articles

In this section, we discuss our second approach, that is, the four steps used for providing customized fact-checked article(s) based recommendations for a misinformation post (refer to the bottom box in Figure 10).

Association of topics (topic-topic mapping). First, we map the topics identified across both corpora to be able to identify a shortlist of the fact-checked article(s) dataset which may share relevance to a given misleading tweet.

Distance Metric Approach: We use the Pyldavis API¹⁷ which considers the top two principal components of topic distances using the Jensen-Shannon divergence as a way to quantify the difference between two distributions.

For each topic in the misleading Twitter corpus, we find its coordinates among the topics in the FaCov corpus. To determine the closest match between one misleading topic and all the FaCov topics, we calculate the Euclidean distance between them. The closest match (that is, the smallest distance) between a misleading topic and a FaCov topic is then mapped together for topic-topic mapping. This analysis is performed for all misleading Twitter corpus topics, resulting in 15 topic-topic mappings (shown in Table 28).

We also tried a few variants to directly match strings of the keywords extracted from LDA in both corpora, which are as follows:

1. **Naive approach:** We first removed stop words, lowercasing, and stemming of all the keywords extracted from LDA in both datasets. Then, examining the keywords extracted from the LDA topic modeling analysis for each topic, if there were any exact matching words between two topics, we considered their frequency of occurrence and sort them in descending order. We repeated this analysis with all the topics' keywords in the second dataset by fixing the topic of one dataset. The final decision for mapping is based on the number of matched words and their frequency of occurrence, resulting in 15 topic-topic mappings which turned out to be identical to the mappings obtained through the measurement of the distance of distributions.
2. **Term Frequency–Inverse Document Frequency (TF-IDF) weighted approach:** To find the mapping of the two topics from the datasets, the TF-IDF¹⁸ of each keyword in the topics is identified and used to find the cosine

¹⁷<https://pyldavis.readthedocs.io/en/latest/>

¹⁸https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

similarity of the topics. The problem with this approach was that it returned quite low similarity scores when the matching number of keywords is few, which was often the case, rendering it inapplicable.

3. **Without TF-IDF weighted approach:** Cosine similarity scores without weighting using the spacy API¹⁹ yielded marginally better results but failed to accurately reflect the similarity of COVID-19 related terms, such as a 0 score between covid and mask.

We manually validated all the topic-topic mapping to ensure the correctness of the results, and for the given data sets, we observed an accuracy of 100% for the mappings. We also evaluated the quality of the topic-topic mapping using the **Rank-k** approach. Particularly, for the naive approach which had the identical results as the distance metric approach, we investigated the rank of each exact word (sorted in descending order if the matched words are more than one) that is mapped in all the topic-topic mappings. The value of k was determined by averaging all the ranks retrieved from repeating the approach for all the mappings. We observed that the exact words that are used to map the two topics appeared in the top 5 occurrences, so the rank was 5. Since the rank is quite well, we conclude that the naive approach is able to map the topics fast and hence, does not degrade the quality of the approach.

Given the disparity in the number of topics in each corpus, multiple topics of the misleading corpus mapped to one topic in the FaCov corpus, as shown in Table 28. **Blue** color corresponds to the topics of the Twitter dataset being mapped to the topics (black color) of the FaCov dataset. For instance, **Vaccine's effects** has been mapped to multiple topics such as **Vaccine Efficacy, Shots, Choices**, etc.

Three topics of the Twitter dataset have not been mapped to any of the topics in the FaCov dataset, namely, **Operation Warp Speed, Real Side-effects, and Vaccine Approval**. We further dig deep into each of the three topics, as can be seen in Tables 34 and 35.

In the fact-checked articles of these three topics, we discovered that the **Operation Warp Speed, and Real Side-effects** are mentioned very infrequently. As such, these are not the main topics of the fact-checked articles, hence, these topics are not detected in the FaCov dataset even though these words are sub-parts that are being used to explain the main topic of the articles. We then looked into these articles' topmost and second most topics.

In the 14 Operation Warp Speed articles (Table 34), the most discussed topics are vaccine effects and trump-related posts. Similarly, in the 12 articles of the Real side-effects keyword, the most discussed topic is the vaccine effects. However, in Vaccine Availability, we found a total of 185 fact-checked articles, including false positive cases (not enumerated due to space constraints).

Filtering the fact-checked articles. After topic-topic mapping, for a given misleading tweet from a specific topic, its mapped with the topic of the FaCov dataset

¹⁹<https://spacy.io/api/doc>

Table 34: Not mapped topic: Operation Warp Speed

No.	Topic 1	Topic 2
1.	Vaccine Effects	Masks
2.	Masks	Trump posts on health Workers
3.	Trump posts on health Workers	Vaccine Effects
4.	Trump posts on health Workers	Vaccine Effects
5.	Vaccine Effects	Trump posts on health Workers
6.	Trump posts on health Workers	Spread of COVID-19
7.	Trump posts on health Workers	Vaccine Effects
8.	Vaccine Effects	Trump posts on health Workers
9.	Trump posts on health Workers	Masks
10.	Vaccine Effects	Masks
11.	Vaccine Effects	Masks
12.	Vaccine Effects	Masks
13.	Masks	Vaccine Effects
14.	Coronavirus death/cases in states	Vaccine Effects

Table 35: Not mapped topic: Real Side-effects

No.	Topic 1	Topic 2
1.	Vaccine Effects	Masks
2.	Vaccine Effects	Masks
3.	Vaccine Effects	-
4.	Vaccine Effects	Misleading posts, pictures, and videos
5.	Vaccine Effects	Trump posts on health Workers
6.	Vaccine Effects	-
7.	Vaccine Effects	Coronavirus death/cases in states
8.	Vaccine Effects	Misleading posts, pictures, and videos
9.	Vaccine Effects	Trump posts on health Workers
10.	Vaccine Effects	Masks
11.	Vaccine Effects	Masks
12.	Vaccine Effects	Trump posts on health Workers

utilizing Table 28 to filter out irrelevant fact-checked articles and shortlist only those articles that are specific to the misleading tweet’s topic.

Semantic Matching. The filtered fact-checked articles for a given misinformation tweet are then used to carry out semantic matching using a pre-trained Sentence Textual Similarity transformer model [RG19]. Specifically, each [tweet, article] pair is fed to the pre-trained model which outputs their cosine similarity scores. These scores are sorted in descending order to analyze whether the most relevant fact-checked article(s) from the FaCov corpus for a given misleading tweet can be recommended or not.

We used two pre-trained similarity transformer models that are popular and trained on large and multiple datasets, including social media datasets, namely, *all-MiniLM-L6-v2* and *paraphrase-MiniLM-L6-v2* from the API²⁰. In our case, the *all-MiniLM-L6-v2* model is the best performing model as per manual validation. One big advantage of these text similarity models is that they also take care of the contextual (or synonymous) words while comparing the pair. Specifically, employing these pre-trained transformer models also take into account the context of the sentences rather than just focusing on finding similar words. Next, we investigated the following questions while performing the similarity task.

1. **Title or content?** We performed the text similarity on both types of pairs, that is, [tweet, title column of the article] and [tweet, content column of the article]. The *title* and *content* columns are the title of the fact-checked article and the whole content of the fact-checked article, respectively. We found that the *title* column performs better than the *content* column. This could be because tweets are short in length, so comparing the tweet and the title makes more sense as compared to the tweet and content, which has many words, and that could hinder the matching process. We performed our analysis with all the different combinations of topics. From the Twitter corpus, we picked a misleading tweet from a specific topic.
2. **How to pick a misleading tweet?** We considered the engagement analysis, which includes the replies count, retweets count and likes count of a tweet. The premise is that if the engagement analysis of a misleading tweet is high, it is more likely to find relevant fact-checked articles in the FaCov corpus. We checked by picking the highest and lowest engagements of the same topic’s misleading tweets and found no such correlation between high engagement and the presence of relevant fact-checked articles. Thus, the misleading tweet of a specific topic has been chosen randomly.

The Three Customized Recommendations. After mapping the topics of the two corpora (Section 5.3.4), we performed the text similarity among a misleading tweet from a topic with all the fact-checked articles from the mapped topic using semantic matching to find the cosine similarity scores between each pair of [misleading tweet, fact-checked article] (Section 5.3.4). For all the various topic

²⁰<https://www.sbert.net/index.html>

combinations, we observe that the similarity scores do not exceed 69% in the corpora. By analyzing these combinations, we surmise the following categories of recommendations:

1. **Specific recommendations:** If the similarity scores are equal to or greater than 62%, then the specific recommendations of the article(s) are provided to a misleading tweet. The threshold of 62% has been decided by manual validation. The process involved validating with different threshold values and analyzing the results to see if the recommendations provided were specific and relevant to the misleading tweet. Based on these evaluations, the threshold of 62% was deemed to provide the best results and was chosen as the final threshold. When we refer to specific recommendations, we mean that the recommended articles should have the exact same topic as the given misleading tweet. The user will be able to access articles that are closely relevant to the information stated in the misleading tweet in this way, giving them access to more reliable and correct information. For instance, the top recommended article for the misleading tweet *“they all used aborted fetus either in development or testing. Especially, in the vaccine astrazeneca johnson and johnson have fetal cell lines and is being said the mrna is something never tested on humans”* from the topic **Vaccine Efficacy** is *“Johnson & Johnson’s COVID-19 vaccine does not contain aborted fetal cells”* with 65% similarity score.
2. **Almost near recommendations:** If the similarity score is less than 62%, then we recommend such articles that have an almost near topic discussed in the given misleading tweet. For instance, the top recommended article for the misleading tweet *“i am proud of you for refusing it. Astrazeneca vaccine is dubious too many bad cases than goods in results of taking that, so whether it works or not it won’t be any of my portion”* from the topic **Choices** is *“Experts say the Oxford AstraZeneca COVID-19 vaccine is safe and that its benefits far outweigh possible risks”* with 59% similarity score.
3. **General/Broad recommendations:** In a case where there is no mapping of the tweet’s topic with the fact-checked topic, that is, zero articles have been retrieved from the fact-checked corpus, the second best prominent topic (discussed in Section 3.3) can be used to find the similarity score which can further be employed for broad recommendations. Using the second most suitable topic ensure that all relevant fact-checked articles are considered for recommendation. This can help to improve the accuracy of the recommendations and ensure that users are provided with the most relevant and accurate information to counter misinformation. For instance, the **School e-learning** topic of FaCov has no mapping with the **Operation Warp Speed** of the misleading corpus. In this case, the **School e-learning** topic in the FaCov corpus is associated/connected with the topic, **Trump posts on health workers** (as per Figure 11). So, **Trump posts on health workers** can be

utilized to perform the semantic matching with the **Operation Warp Speed** topic in the misleading corpus. For instance, the top recommended article for the misleading tweet “*maybe he kept some so he can sell them to the black market remember he is in debt or to punish BioNTech pfizer for letting the world know that operation warp speed money was not involved in their vaccine development. This mean guy is capable of heinous crimes*” from the topic **Operation Warp Speed** is “*No, Trump didn’t tweet his blood is a vaccine*” with a 37% similarity score.

Evaluation Criterion of Second Approach. To evaluate our second approach, which is recommending fact-checked articles to a misleading post, we employed the MRR@k and MAP@k metrics as discussed in Section 5.3.3. Other widely used metrics such as NDCG@k require the availability of ground truth availability which is not the case for our problem. Table 33, row 2 shows the performance of the MRR@k and MAP@k metrics for different values of k. It can be seen from the table that 15 fact-checked articles’ recommendations for a misleading post provide good performance. Further, it is to be noted the higher value of k does not necessarily mean better performance as we may not have the relevant recommendations to provide to the social media user. Additionally, giving the user too many recommendations can make them less effective due to limited attention. Thus, the focus here is on the quality rather than the quantity of the recommendations.

5.4. Summary

In this chapter, we propose the practical approach of automated misinformation rebuttal systems to battle misinformation on social media, amortizing user-generated content from the same platform in conjunction with articles from websites reputed and specializing in fact-checking popular misinformation.

The whole pipeline involves several complex tasks, e.g., detecting and distinguishing social media content carrying misinformation from those that do not, understanding the broad and precise topic of the content, and correlating with the content of fact-checked articles when applicable, in order to determine the relevance of recommendations.

While we have used manual inspections and human interpretation of the intermediate results, since all the actual steps in the pipeline are algorithmic in nature, our approach can eliminate human-in-the-loop and in that sense, in principle, it can function at scale, and in near real-time. Nevertheless, a real live deployment dealing with a plethora of sources would need to optimize the various algorithms for computational efficiency in addition to any work to improve the accuracy of the results. For example, right now, in Section 5.3.4 we use what can be deemed as an exhaustive search for establishing topic-topic maps for data from the two different data sources, which is undesirable, more so when we would need to accommodate a wider set of topics and data sources.

Given the niche and novelty of this work, there is also no well-established benchmark data to evaluate the quality of recommendations. Hence, in this work, due to the unavailability of the ground truth of recommendations which is generally the case in real-world, we focus on the two widely used metrics by adapting the definition to fit our scenarios to evaluate our complementary approaches. This, along with longitudinal user case studies employing A/B tests on the efficacy of even the premise itself, namely timely and precise rebuttals that might thwart misinformation (and to which extent) are open-ended issues that need to be investigated.

Limitations. The approach described in this contribution primarily tackles COVID-19 misinformation on the Twitter platform, which could be one of the limitations of this work. Another area of improvement is the lack of ground truth availability for the recommendations, which limits our possibility of employing a variety of evaluation metrics to evaluate our approaches.

6. CONCLUSIONS

As posts on online social media platforms are not independently checked by a third party, users continuously spread misinformation. Studies have demonstrated that on these platforms, misinformation spreads six times as rapidly as actual news, causing society to worry, anxiety, and financial loss. It is of utmost importance to develop various strategies to fight misinformation.

In this thesis, we focus on three multifaceted approaches to combat misinformation. The first dimension of work is the fusion of descriptive and predictive tasks, which is understanding the characteristics (traits) of online social media content thus, users with respect to misinformation, especially rumors in this case. We discovered a substantial difference between rumor and non-rumor psycholinguistics source features, as well as between reply features. We discovered that rumor source tweets used more past-related words, prepositions, and contained drives (motivation) related to reward, risk, and power. Similarly, non-rumor source tweets had more mean features such as affective processes, cognitive processes (insight, causation, discrepancy, certainty, differentiation), present-related words, informal language, and had drives related to affiliation and achievement. We also explored the effectiveness of these features in predicting rumors. Specifically, we discovered that the ensemble-based Random Forest model, for all events, outperformed the other used models. As machine learning models are black-box in nature, we utilized the SHAP AI Explainability tool to look for the features that are more important than the other features. The research community can leverage the language variations between posts containing misinformation and those not containing misinformation to develop automated systems for identifying and preventing the spread of misinformation on social media platforms.

The second dimension of work is predictive in nature, where we develop a framework to detect and prevent the spread of misinformation. Particularly, we propose user-level rumor detection, that is, classifying the user as a possible rumor spreader or not. Previous techniques for detecting suspicious or malicious users and identifying misinformation on platforms like Twitter have not considered user-level detection adequately. Categorizing a user as a misinformation spreader based on one post is insufficient. To combat this, a system for automatic misinformation detection is created that integrates network data and many postings to get a high F1-Score of 0.864, outperforming previous baseline models. Particularly, the PHEME dataset was used to identify possible rumor spreaders using a weak supervised learning approach. We modeled this problem as a binary classification task by applying various machine learning models to the transformed rumor spreaders dataset. Our results show that GCN (compared to baseline models) is able to perform better (to raise red flags for possible rumor spreaders) by exploiting relationships of possible rumor spreaders who could blend well with non-rumor spreaders.

The third work is descriptive in nature, where we recommend articles/posts

to users who spread misinformation. The drawbacks of earlier strategies for reducing the spread of misinformation on social media platforms include a lack of external moderation. An action plan that combines social media data with curated fact-checked data repositories to deploy automated debunking of misinformation at scale is suggested as a solution to this problem. This strategy focuses on the Twitter platform and COVID-19 misinformation, examining two complementing strategies that can spot posts that disseminate misinformation both early and late in the process, outperforming earlier strategies that could only target one step. This strategy could be more successful in stopping the spread of misinformation. Specifically, it leverages two publicly available datasets, viz. FaCov (fact-checked articles) and misleading (social media Twitter) data on COVID-19. The first identifies and reuses existing related but factual/non-misleading tweets by other users as recommended counter tweets. This is the baseline approach that can be used even when fact-checking websites are yet to carry out the vetting exercise for a given misinformation or while such articles are identified and matched by our second approach. The second recommends fact-checking articles (subject to availability) from various fact-checking sites. This provides users with verified, reliable, and more comprehensive information from reputable sources to counter the misinformation. By utilizing both of these automation approaches in tandem, we create a scalable rebuttal pipeline to combat misinformation on social media.

7. FUTURE SCOPE

In this thesis, we investigated three dimensions to combat misinformation in the digital era. This section discusses potential future directions to improve the effectiveness of combating misinformation, building upon the findings and contributions of this thesis.

There are multiple directions to extend the *characterization* work discussed in Chapter 3. The first potential direction for future improvement is to extend the scope of the characterization work by adding the user-level analysis. Understanding the human psychology of rumor spreaders can offer insightful knowledge about the causes and tendencies of the propagation of misinformation. By examining user-level aspects such as demographic data, social network traits, and prior online behavior, researchers may develop a more thorough understanding of the factors that lead to rumor spreading. To acquire qualitative information on the beliefs, attitudes, and motives of those who have been recognized as rumor spreaders, surveys or interviews with them may be conducted. The second future direction could be incorporating additional psycho-linguistic features. There are still unexplored territories to investigate, even if the current study has concentrated on verbal clues and patterns. For instance, morphological characteristics, such as word construction and structure, may shed light on the language used by those who disseminate rumors. Investigating the relationships between various morphological traits and the propagation of rumors may reveal language tactics used in the rumor posts. As a third direction, expanding this research to take into account the anti-rumor tweets dataset may offer a more thorough picture of the interactions between rumor users and those who actively debunk them. Researchers can learn more about the efficacy of various anti-rumor techniques and the counter-narratives used to refute misinformation by examining the interactions and reactions between rumor users and those involved in debunking efforts. This might entail examining the language and content of anti-rumor tweets, finding convincing arguments and reasoning tactics, and examining the dynamics of rumor spreaders and debunkers in social networks.

The second dimension of work is where we discuss our framework utilizing graph neural networks technique to classify the user as a misinformation spreader in Chapter 4. We would like to improve this work through the following multiple plans. First, modeling the problem as a multiclass prediction problem would be an intriguing expansion to consider the problem as a binary classification work. We can capture a more detailed depiction of various misinformation spreaders by taking into account many classifications. This strategy would enable us to categorize the users as those who disseminate rumors, users who might convey misinformation, and users who actively debunk or do not spread misinformation. We may reduce the loss in information translation and better understand the intricacies of user behavior in the propagation of misinformation by framing the problem as a multiclass prediction. Second, it would be beneficial to explore other graph neu-

ral network approaches, also known as inductive learning, which are designed for handling unseen graph structures in order to predict the rumor spreaders before the misinformation spread. Graph Attention Networks (GAT) and GraphSage are two approaches in this field that show promising performance. GAT makes use of attention processes to evaluate the significance of nearby nodes as messages are being sent, allowing it to identify more intricate linkages in the graph. On the other hand, GraphSage generates embeddings for nodes using a neighborhood aggregation method, making it appropriate for inductive learning on networks with different architectures. By combining these methods into the framework, we may be able to increase the model's reliability and accuracy in detecting those who propagate misinformation. Third, it would be worthwhile to expand the research beyond the existing range of topics by investigating cross-topic analysis in rumors. The banking sector, for instance, is an intriguing field to research. We can learn more about the particular dynamics and traits of misinformation within this sector by examining rumors that are connected to financial markets, investments, or economic events. This investigation may entail gathering pertinent datasets, modifying the framework to account for the special characteristics of financial rumors, and researching the dissemination and effects of misinformation in this particular situation. This extension would enable the creation of domain-specific detection and mitigation measures and offer a complete understanding of the propagation of misinformation across several domains.

The proposed action plan for COVID-19 corpora to combat misinformation utilizing two complementary approaches is a practical and feasible approach, as discussed in Chapter 5. Thus, it is interesting to generalize our work across several dimensions. First, it would be beneficial to broaden the system's scope to include topics other than COVID-19. While pandemic-specific misinformation is the main focus of the present study, broadening the focus to cover other crucial topics might offer a more comprehensive analysis. The work can be used to fight misinformation on a variety of topics by merging corpora and statistics from other areas, including politics, health, science, and social concerns. This would entail gathering and studying pertinent datasets, modifying the algorithms and models to fit the new research areas, and modifying the current framework. Second, in addition to the sites covered by the existing approach, it is important to consider content from a wider variety of open social media networks, including Reddit, into account. Each social media has different user dynamics and characteristics that contribute to the spread of misinformation. A more comprehensive knowledge of how misinformation spreads and the ability of the system to handle a wider range of sources would be provided by including data from sites like Reddit, which have their own unique communities and content patterns. To support this expansion, it would be necessary to modify the data collecting and analysis procedures to account for the unique features and APIs of other platforms. The third crucial factor to take into account is how non-English information is addressed. A thorough rebuttal system should be able to handle content in several languages

because misinformation circulates around the world in many different languages. The system would need to be modified to process and analyze non-English content by adding language-specific models and resources, training multilingual models, and increasing the system's language coverage. This would increase the system's utility on a worldwide scale by allowing it to offer effective rebuttals and recommendations for misinformation in a variety of languages. The fourth direction could be the evaluation of the approach's practical efficacy, with a focus on its capacity to convince users and lessen the impact of misinformation, which is a key area for future research. The influence and results of the strategy on users' beliefs, attitudes, and behaviors can be assessed via user studies, surveys, or experimental designs.

BIBLIOGRAPHY

- [AA05] Lada A. Adamic and Eytan Adar. “How To Search a Social Network”. In: *Social Networks* 27 (2005), p. 2005.
- [AA20] Manal Almalki and Fahad Azeez. “Health chatbots for fighting COVID-19: a scoping review”. In: *Acta Informatica Medica* 28.4 (2020), p. 241.
- [Abd10] Hervé Abdi. “Holm’s sequential Bonferroni procedure”. In: *Encyclopedia of research design* 1.8 (2010), pp. 1–8.
- [Aga+18] Rishav Raj Agarwal et al. “Predicting financial trouble using call data—On social capital, phone logs, and financial trouble”. In: *PloS one* 13.2 (2018), e0191863.
- [Ahm17] Sami Ahmed. “Cryptocurrency & Robots: How to Tax and Pay Tax on Them”. In: *SCL Rev.* 69 (2017), p. 697.
- [AIS16] Kumar Arun, Garg Ishan, and Kaur Sanmeet. “Loan Approval Prediction based on Machine Learning Approach”. In: *IOSR J. Comput. Eng* 18.3 (2016), pp. 18–21.
- [Akh+18] Md Shad Akhtar et al. “No, that never happened!! Investigating rumors on Twitter”. In: *IEEE Intelligent Systems* 33.5 (2018), pp. 8–15.
- [Ako+12] Leman Akoglu et al. “PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs”. In: *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012*. 2012, pp. 439–450. DOI: 10.1137/1.9781611972825.38. URL: <http://dx.doi.org/10.1137/1.9781611972825.38>.
- [Alt13] Edward I Altman. “Predicting financial distress of companies: revisiting the Z-score and ZETA® models”. In: *Handbook of research methods and applications in empirical finance*. Edward Elgar Publishing, 2013.
- [Alv+15] Ivan Alves et al. *Network analysis of the EU insurance sector*. ESRB Occasional Paper Series 07. European Systemic Risk Board, July 2015. URL: <https://ideas.repec.org/p/srk/srkops/201507.html>.
- [AM03] A. Kermarrec A. Ganesh and L. Massoulie. “Peer-to-peer membership management for gossip-based protocols”. In: *IEEE Transactions on Computers*, vol. 52. 2003.
- [AM18] Amelia Acker and Dhiraj Murthy. “Venmo: Understanding Mobile Payments As Social Media”. In: *Proceedings of the 9th International Conference on Social Media and Society*. SMSociety ’18. Copenhagen, Denmark: ACM, 2018, pp. 5–12. ISBN: 978-1-4503-

- 6334-1. DOI: 10.1145/3217804.3217892. URL: <http://doi.acm.org/10.1145/3217804.3217892>.
- [Ant+18] Nino Antulov-Fantulin et al. “Inferring short-term volatility indicators from the Bitcoin blockchain”. In: *International Conference on Complex Networks and their Applications*. Springer. 2018, pp. 508–520.
- [AOT15] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. “Systemic Risk and Stability in Financial Networks”. In: *American Economic Review* 105.2 (Feb. 2015), pp. 564–608. DOI: 10.1257/aer.20130456. URL: <http://www.aeaweb.org/articles?id=10.1257/aer.20130456>.
- [APL18] Ángel Francisco Agudo-Peregrina, Diego Perez, and Martin Langberg. “What Banking and Phone Data Tell us about the Socioeconomic Groups and Their Consumption Patterns?”. In: *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*. 2018, pp. 595–597. DOI: 10.1109/ASONAM.2018.8508330. URL: <https://doi.org/10.1109/ASONAM.2018.8508330>.
- [Apo+09] Andrea Apolloni et al. “A Study of Information Diffusion over a Realistic Social Network Model”. In: *International Conference on Computational Science and Engineering*. 2009.
- [Arg+06] Jaime Arguello et al. “Talk to me: Foundations for successful individual-group interactions in online communities”. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2006, pp. 959–968.
- [ARS18] Israa Alqassem, Iyad Rahwan, and Davor Svetinovic. “The anti-social system properties: Bitcoin network data analysis”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2018).
- [Ash+21] Noman Ashraf et al. “CIC at CheckThat! 2021: Fake news detection using machine learning and data augmentation”. In: CLEF. 2021.
- [AU18] Kristen M Altenburger and Johan Ugander. “Monophily in social networks introduces similarity among friends-of-friends”. In: *Nature human behaviour* 2.4 (2018), pp. 284–290.
- [AX21] Jon Agle and Yunyu Xiao. “Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science”. In: *BMC Public Health* 21.1 (2021), pp. 1–12.
- [Azo94] E Michael Azoff. *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc., 1994.
- [BA10] Melek Acar Boyacioglu and Derya Avci. “An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock

- market return: the case of the Istanbul stock exchange”. In: *Expert Systems with Applications* 37.12 (2010), pp. 7908–7912.
- [BA99] A. L. Barabasi and R. Albert. “Emergence of scaling in random networks”. In: *Science (New York, N.Y.)* 286 (Oct. 1999), pp. 509–512.
- [Bac+12a] Lars Backstrom et al. “Four Degrees of Separation”. In: *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci ’12. New York, NY, USA: ACM, 2012, pp. 33–42.
- [Bac+12b] Lars Backstrom et al. “Four degrees of separation”. In: *Proceedings of the 4th Annual ACM Web Science Conference*. 2012, pp. 33–42.
- [Bad+09] Randy Baden et al. “Persona: An Online Social Network with User Defined Privacy”. In: *ACM SIGCOMM*. 2009.
- [Bae12] Jae Kwon Bae. “Predicting financial distress of the South Korean manufacturing industries”. In: *Expert Systems with Applications* 39.10 (2012), pp. 9159–9165.
- [Bal+07] Roberto Baldoni et al. “TERA: topic-based event routing for peer-to-peer architectures”. In: *Proceedings of the 2007 inaugural international conference on Distributed event-based systems*. DEBS ’07. Toronto, Ontario, Canada: ACM, 2007, pp. 2–13. DOI: 10.1145/1266894.1266898. URL: <http://doi.acm.org/10.1145/1266894.1266898>.
- [Bal+22] Vimala Balakrishnan et al. “Infodemic and fake news—A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review”. In: *International Journal of Disaster Risk Reduction* (2022), p. 103144.
- [Bat+02] Sonja V Batten et al. “Physical and psychological effects of written disclosure among sexual abuse survivors”. In: *Behavior Therapy* 33.1 (2002), pp. 107–122.
- [BBE07] Chris Baumann, Suzan Burton, and Gregory Elliott. “Predicting consumer behavior in retail banking”. In: *Journal of Business and Management* 13.1 (2007), pp. 79–96.
- [BBM12] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. “Topic-Aware Social Influence Propagation Models”. In: *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*. ICDM ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 81–90. ISBN: 978-0-7695-4905-7. DOI: 10.1109/ICDM.2012.122. URL: <http://dx.doi.org/10.1109/ICDM.2012.122>.
- [Ben+10] Fabricio Benevenuto et al. “Detecting spammers on twitter”. In: *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. Vol. 6. 2010, p. 12.
- [Ber+15] Flavio Bertini et al. “Profile resolution across multilayer networks through smartphone camera fingerprint”. In: *Proceedings of the 19th*

- International Database Engineering & Applications Symposium*. 2015, pp. 23–32.
- [Ber+21] George Berry et al. “Estimating Homophily in Social Networks Using Dyadic Predictions”. In: *Sociological Science* 8 (2021), pp. 285–307.
- [Ber60] Daniel E Berlyne. “Conflict, arousal, and curiosity.” In: (1960).
- [BFL14] Annika Baumann, Benjamin Fabian, and Matthias Lischke. “Exploring the Bitcoin Network”. In: vol. 1. Apr. 2014, pp. 369–374. DOI: 10.5220/0004937303690374.
- [BG18] Amirhosein Bodaghi and Sama Goliaei. “A novel model for rumor spreading on social networks with considering the influence of dissenting opinions”. In: *Advances in Complex Systems* 21.06n07 (2018), p. 1850011.
- [BGL11] Albert-Laszlo Barabasi, Natali Gulbahce, and Joseph Loscalzo. “Network Medicine: A Network-based Approach to Human Disease”. In: *Nature reviews. Genetics* 12 (Jan. 2011), pp. 56–68. DOI: 10.1038/nrg2918.
- [Bia+20] Tian Bian et al. “Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks”. In: *arXiv preprint arXiv:2001.06362* (2020).
- [Bir+99] Kenneth P. Birman et al. “Bimodal multicast”. In: *ACM Trans. Comput. Syst.* 17.2 (1999).
- [BJ19] Venkateswarlu Bonta and Nandhini Kumares2and N Janardhan. “A comprehensive study on lexicon based approaches for sentiment analysis”. In: *Asian Journal of Computer Science and Technology* 8.S2 (2019), pp. 1–6.
- [Blo+08] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.
- [BLS15] Yazan Boshmaf, Dionysios Logothetis, and Georgios Siganos. *Method and system for predicting victim users and detecting fake user accounts in online social networks*. US Patent App. 14/140,965. July 2015.
- [Böh+15] Rainer Böhme et al. “Bitcoin: Economics, technology, and governance”. In: *Journal of Economic Perspectives* 29.2 (2015), pp. 213–38.
- [Bos+21a] J Bosman et al. “Who are the unvaccinated in America? There’s no one answer”. In: *The New York Times* (2021).
- [Bos+21b] Julie Bosman et al. *Who Are the Unvaccinated in America? There’s No One Answer*. 2021.
- [Bot+ar] Cecile Bothorel et al. “Clustering attributed graphs: models, measures and methods”. In: *Network Science* (to appear).

- [BP98] Sergey Brin and Lawrence Page. “The anatomy of a large-scale hypertextual Web search engine”. In: *Comput. Netw. ISDN Syst.* 30.1-7 (1998), pp. 107–117.
- [BPA22] Shovan Bhowmik, Priyo Ranjan Kundu Prosun, and Kazi Saeed Alam. “A novel three-level voting model for detecting misleading information on COVID-19”. In: *Advanced Techniques for IoT Applications: Proceedings of EAIT 2020*. Springer. 2022, pp. 374–383.
- [Bri06] Michael Brinkmeier. “PageRank revisited”. In: *ACM Trans. Internet Technol.* 6 (3 Aug. 2006), pp. 282–301. ISSN: 1533-5399. DOI: <http://doi.acm.org/10.1145/1151087.1151090>. URL: <http://doi.acm.org/10.1145/1151087.1151090>.
- [BS08] Jenna L Baddeley and Jefferson A Singer. “Telling losses: Personality correlates and functions of bereavement narratives”. In: *Journal of Research in Personality* 42.2 (2008), pp. 421–438.
- [BSV03] R. Bhagwan, S. Savage, and G. Voelker. “Understanding availability”. In: *IPTPS '03: The 2nd International Workshop on P2P Systems*. 2003.
- [Buc+] Sonja Buchegger et al. “PeerSoN: P2P Social Networking - Early Experiences and Insights”. In.
- [Bur19] Talha Burki. “Vaccine misinformation and social media”. In: *The Lancet Digital Health* 1.6 (2019), e258–e259.
- [Bur20] Talha Burki. “The online anti-vaccine movement in the age of COVID-19”. In: *The Lancet Digital Health* 2.10 (2020), e504–e505.
- [But+22a] Sabur Butt et al. “What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses”. In: *Computers in Human Behavior* (2022).
- [But+22b] Sabur Butt et al. “What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses”. In: *Computers in Human Behavior* (2022), p. 107345.
- [Cal+19] Timothy Callaghan et al. “Parent psychology and the decision to delay childhood vaccination”. In: *Social science & medicine* 238 (2019), p. 112407.
- [Cam+14] Erik Cambria et al. “Sentic API: a common-sense based API for concept-level sentiment analysis”. In: *CEUR Workshop Proceedings*. Vol. 144. 2014, pp. 19–24.
- [Cao+12] Qiang Cao et al. “Aiding the detection of fake accounts in large scale social online services”. In: *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*. 2012, pp. 197–210.

- [Car16] Jon Carrick. “Bitcoin as a complement to emerging market currencies”. In: *Emerging Markets Finance and Trade* 52.10 (2016), pp. 2321–2334.
- [CAS16] Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. “Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)”. In: *ComTech: Computer, Mathematics and Engineering Applications 7.4* (2016), pp. 285–294.
- [Caz+16] Lorenzo Cazzoli et al. “A large scale study to understand the relation between twitter and financial market”. In: *2016 third European network intelligence conference (ENIC)*. IEEE. 2016, pp. 98–105.
- [CB20] Wen-Ying Sylvia Chou and Alexandra Budenz. “Considering Emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence”. In: *Health communication* 35.14 (2020), pp. 1718–1722.
- [CB21] Charles Alfred A Cruz and Francis F Balahadia. “Analyzing Public Concern Responses for Formulating Ordinances and Laws using Sentiment Analysis through VADER Application”. In: *International Journal of Computing Sciences Research* 6 (2021), pp. 842–856.
- [CF02] P. Gibbons C. Palmer and C. Faloutsos. “ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs”. In: *KDD*. 2002.
- [CFB04] Vicent Cholvi, Pascal Felber, and Ernst W. Biersack. “Efficient search in unstructured peer-to-peer networks”. In: *SPAA*. 2004.
- [Cha+21] Mohit Chandra et al. ““ A Virus Has No Religion”: Analyzing Islamophobia on Twitter During the COVID-19 Outbreak”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 2021, pp. 67–77.
- [Cha+99] Philip K Chan et al. “Distributed data mining in credit card fraud detection”. In: *IEEE Intelligent Systems and Their Applications* 14.6 (1999), pp. 67–74.
- [Che+18] Weiling Chen et al. “Unsupervised rumor detection based on users’ behaviors using neural networks”. In: *Pattern Recognition Letters* 105 (2018), pp. 226–233.
- [Che+21a] Ben Chen et al. “Transformer-Based Language Model Fine-Tuning Methods for COVID-19 Fake News Detection”. In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by Tanmoy Chakraborty et al. Cham: Springer International Publishing, 2021, pp. 83–92.
- [Che+21b] Mingxi Cheng et al. “A COVID-19 rumor dataset”. In: *Frontiers in Psychology* 12 (2021), p. 644801.

- [Che+22] Xueqin Chen et al. “Multi-view learning with distinguishable feature fusion for rumor detection”. In: *Knowledge-Based Systems* 240 (2022), p. 108085.
- [Chi+11] A. Chiuso et al. “Gossip algorithms for distributed ranking”. In: *American Control Conference ACC11 (submitted)*. 2011.
- [Cho10] Seong Eun Cho. “Cross-cultural comparison of Korean and American social network sites: exploring cultural differences in social relationships and self-presentation”. In: *Unpublished doctoral dissertation, Rutgers University, New Brunswick, New Jersey*. (2010).
- [CHP17] Eunsuk Chong, Chulwoo Han, and Frank C Park. “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies”. In: *Expert Systems with Applications* 83 (2017), pp. 187–205.
- [CKT06] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. “Evolutionary Clustering”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: ACM, 2006, pp. 554–560. ISBN: 1-59593-339-5. DOI: 10 . 1145 / 1150402 . 1150467. URL: <http://doi.acm.org/10.1145/1150402.1150467>.
- [CL20] Limeng Cui and Dongwon Lee. “Coaid: Covid-19 healthcare misinformation dataset”. In: *arXiv preprint arXiv:2006.00885* (2020).
- [CLF+20] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set”. In: *JMIR Public Health and Surveillance* 6.2 (2020), e19273.
- [CLH12] Erik Cambria, Andrew Livingstone, and Amir Hussain. “The hourglass of emotions”. In: *Cognitive behavioural systems*. Springer, 2012, pp. 144–157.
- [CLP11] Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. “Rumor spreading in social networks”. In: *Theoretical Computer Science* 412.24 (2011), pp. 2602–2610.
- [CMG09] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. “A Measurement-driven Analysis of Information Propagation in the Flickr Social Network”. In: *WWW*. 2009.
- [CMN04] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. “Finding community structure in very large networks”. In: *Physical Review* 70 (2004).
- [CMN08] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. “Hierarchical structure and the prediction of missing links in networks”. In: *Nature* 453 (2008), pp. 98–101.
- [CMW15] Simone Caschili, Francesca Romana Medda, and Alan Wilson. “An Interdependent Multi-Layer Model: Resilience of International Net-

- works". In: *Networks and Spatial Economics* 15.2 (June 2015), pp. 313–335. ISSN: 1572-9427. DOI: 10.1007/s11067-014-9274-2. URL: <https://doi.org/10.1007/s11067-014-9274-2>.
- [CMY14] Brett W Cantrell, John M McInnis, and Christopher G Yust. "Predicting credit losses: Loan fair values versus historical costs". In: *The accounting review* 89.1 (2014), pp. 147–176.
- [Com+12] David Combe et al. "Getting Clusters from Structure Data and Attribute Data." In: *ASONAM*. IEEE Computer Society, 2012, pp. 710–712. ISBN: 978-0-7695-4799-2. URL: <http://dblp.uni-trier.de/db/conf/asunam/asonam2012.html#CombeLEG12>.
- [Cor20] Warren Cornwall. *Officials gird for a war on vaccine misinformation*. 2020.
- [Cos+08] P. Costa et al. "Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks". In: *Selected Areas in Communications, IEEE Journal on* 26.5 (June 2008).
- [Cos+20] Alessandro Cossard et al. "Falling into the Echo Chamber: The Italian Vaccination Debate on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 130–140. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/7285>.
- [CPB17] Francesco Calderoni, Carlo Piccardi, and Domenico Brunetto. "Communities in criminal networks: A case study". In: *Social Networks* 48 (Jan. 2017), pp. 116–125. DOI: 10.1016/j.socnet.2016.08.003.
- [Cro+16] Michael Crosby et al. "Blockchain technology: Beyond bitcoin". In: *Applied Innovation* 2.6-10 (2016), p. 71.
- [CS16] Tao-Hung Chang and Davor Svetinovic. "Data Analysis of Digital Currency Networks: Namecoin Case Study". In: *21st International Conference on Engineering of Complex Computer Systems, ICECCS 2016, Dubai, United Arab Emirates, November 6-8, 2016*. Ed. by Hai Wang and Mounir Mokhtari. IEEE Computer Society, 2016, pp. 122–125. DOI: 10.1109/ICECCS.2016.023. URL: <https://doi.org/10.1109/ICECCS.2016.023>.
- [CSN09] A Clauset, C. R. Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Rev.* 51 (4 Nov. 2009), pp. 661–703.
- [CSP12] Denzil Correa, Ashish Sureka, and Mayank Pundir. "iTop: Interaction Based Topic Centric Community Discovery on Twitter". In: *Proceedings of the 5th Ph.D. Workshop on Information and Knowledge*. PIKM '12. Maui, Hawaii, USA: ACM, 2012, pp. 51–58. ISBN: 978-1-4503-1719-1. DOI: 10.1145/2389686.2389698. URL: <http://doi.acm.org/10.1145/2389686.2389698>.

- [Cui+20] Limeng Cui et al. “Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 492–502.
- [CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.
- [CWL19] Limeng Cui, Suhang Wang, and Dongwon Lee. “Same: sentiment-aware multi-modal embedding for detecting fake news”. In: *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 2019, pp. 41–48.
- [CXK22] Sijing Chen, Lu Xiao, and Akit Kumar. “Spread of misinformation on social media: What contributes to it and how to combat it”. In: *Computers in Human Behavior* (2022), p. 107643.
- [DB14] Leon Derczynski and Kalina Bontcheva. “Pheme: Veracity in Digital Social Networks.” In: *UMAP workshops*. 2014.
- [DB20] Suparna Dhar and Indranil Bose. “Emotions in Twitter communication and stock prices of firms: the impact of Covid-19 pandemic”. In: *Decision* 47.4 (2020), pp. 385–399.
- [Del+16] Michela Del Vicario et al. “The spreading of misinformation online”. In: *Proceedings of the national academy of Sciences* 113.3 (2016), pp. 554–559.
- [Del06] Matteo Dell’amico. “Highly Clustered Networks with Preferential Attachment to Close Nodes”. In: *Proceedings of European Conference on Complex Systems 2006*. 2006.
- [Dem+87] Alan Demers et al. “Epidemic algorithms for replicated database maintenance”. In: *PODC ’87: Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*. ACM, 1987.
- [Der+17] Leon Derczynski et al. “SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours”. In: *arXiv preprint arXiv:1704.05972* (2017).
- [Dev+20] P Suthanthira Devi et al. “Veracity Analysis and Prediction in Social Big Data”. In: *Information and Communication Technology for Sustainable Development*. Springer, 2020, pp. 289–298.
- [DeV+21] Matthew DeVerna et al. “CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines”. In: *arXiv e-prints* (2021), arXiv–2101.
- [Dha+22a] Arkin Dharawat et al. “Drink Bleach or Do What Now? Covid-HeRA: A Study of Risk-Informed Health Decision Making in the Presence of COVID-19 Misinformation”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 1218–1227.

- [Dha+22b] Mudit Dhawan et al. “GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection”. In: *arXiv preprint arXiv:2202.12478* (2022).
- [DHA03] Anwitaman Datta, Manfred Hauswirth, and Karl Aberer. “Updates in Highly Unreliable, Replicated Peer-to-Peer Systems”. In: *ICDCS '03: Proceedings of the 23rd International Conference on Distributed Computing Systems*. IEEE Computer Society, 2003.
- [Diz18] Peter Dizikes. *Study: On Twitter, false news travels faster than true stories*. <http://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308>. 2018.
- [DKB17] Matthew Dixon, Diego Klabjan, and Jin Hoon Bang. “Classification-based financial markets prediction using deep neural networks”. In: *Algorithmic Finance* 6.3-4 (2017), pp. 67–77.
- [DKF20] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel. “Cascade-LSTM: A tree-structured neural classifier for detecting misinformation cascades”. In: *proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*. 2020, pp. 2666–2676.
- [DM04] Nicolas Ducheneaut and Robert J. Moore. “The Social Side of Gaming: A Study of Interaction Patterns in a Massively Multiplayer Online Game”. In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. Chicago, Illinois, USA: ACM, 2004, pp. 360–369. ISBN: 1-58113-810-5. DOI: 10.1145/1031607.1031667. URL: <http://doi.acm.org/10.1145/1031607.1031667>.
- [Dou+21] Yingtong Dou et al. “User preference-aware fake news detection”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2051–2055.
- [DQA04] Anwitaman Datta, Silvia Quarteroni, and Karl Aberer. “Autonomous Gossiping: A Self-Organizing Epidemic Algorithm for Selective Information Dissemination in Wireless Mobile Ad-Hoc Networks”. In: *Semantics of a Networked World* (2004).
- [DRF04] Yanlei Diao, Shariq Rizvi, and Michael J. Franklin. “Towards an Internet-Scale XML Dissemination Service”. In: *Proceedings of the 26th International Conference on Very Large Data Bases*. 2004.
- [DS11] Anwitaman Datta and Rajesh Sharma. “GoDisco: Selective Gossip based Dissemination of Information in Social Community based Overlays”. In: *ICDCN*. 2011.
- [DSC17] Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. “The psychology of conspiracy theories”. In: *Current directions in psychological science* 26.6 (2017), pp. 538–542.

- [Duo+17] Chi Thang Duong et al. “Provenance-based rumor detection”. In: *Australasian Database Conference*. Springer. 2017, pp. 125–137.
- [Dut+03] D. Dutta et al. *The Design of a Distributed Rating Scheme for Peer-to-Peer Systems*. 2003.
- [Dwi+20] Yogesh K Dwivedi et al. “Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life”. In: *International Journal of Information Management* 55 (2020), p. 102211.
- [DZ06] Zhidong Deng and Yi Zhang. “Complex Systems Modeling Using Scale-Free Highly-Clustered Echo State Network”. In: *Proceedings of the International Conference on Neural Networks 2006*. 2006, pp. 3128–3135.
- [ED10a] Frank Emmert-Streib and Matthias Dehmer. “Identifying critical financial networks of the DJIA: Toward a network-based index”. In: *Complexity* 16.1 (2010), pp. 24–33. DOI: 10.1002/cplx.20315. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.20315>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cplx.20315>.
- [ED10b] Frank Emmert-Streib and Matthias Dehmer. “Influence of the Time Scale on the Construction of Financial Networks.” In: *PLoS ONE* 5.9 (2010).
- [ELG19] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. “Fake news detection on social media: a systematic survey”. In: *2019 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM)*. IEEE. 2019, pp. 1–8.
- [Eug+03] Patrick Th. Eugster et al. “The many faces of publish/subscribe”. In: *ACM Comput. Surv.* 35.2 (2003).
- [Eug+04] Patrick T. Eugster et al. “Epidemic Information Dissemination in Distributed Systems”. In: *Computer* 37.5 (May 2004), pp. 60–67. ISSN: 0018-9162.
- [Eva+20] Sarah Evanega et al. “Coronavirus misinformation: quantifying sources and themes in the COVID-19 ‘infodemic’”. In: *JMIR Preprints* 19.10 (2020), p. 2020.
- [Far+17] Mehrdad Farajtabar et al. “Fake news mitigation via point process based intervention”. In: *International conference on machine learning*. PMLR. 2017, pp. 1097–1106.
- [FBC12] Song Feng, Ritwik Banerjee, and Yejin Choi. “Syntactic stylometry for deception detection”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012, pp. 171–175.
- [Fer20] Emilio Ferrara. “# covid-19 on twitter: Bots, conspiracies, and social media activism”. In: *arXiv preprint arXiv: 2004.09531* (2020).

- [Fir+14] Michael Fire et al. "Friend or foe? Fake profile identification in online social networks". In: *Social Network Analysis and Mining* 4.1 (2014), p. 194.
- [FJ15] Seth Freedman and Ginger Zhe Jin. "The Information Value of On-line Social Networks : Lessons from Peer-to-Peer Lending". In: 2015.
- [FJP51] James N Farr, James J Jenkins, and Donald G Paterson. "Simplification of Flesch reading ease formula." In: *Journal of applied psychology* 35.5 (1951), p. 333.
- [FKE12] Michael Fire, Gilad Katz, and Yuval Elovici. "Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies". In: *Human Journal* 1.1 (2012), pp. 26–39.
- [FL03] Marcel Fafchamps and Susan Lund. "Risk Sharing Networks in Rural Philippines". In: *Journal of Development Economics* 1 (2003), pp. 261–87.
- [Fle48] Rudolph Flesch. "A new readability yardstick." In: *Journal of applied psychology* 32.3 (1948), p. 221.
- [For10] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (2010), pp. 75–174. ISSN: 0370-1573. DOI: DOI : 10 . 1016 / j . physrep . 2009 . 11 . 002. URL: <http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1>.
- [Fur+17] Angelo Furno et al. "Two-level clustering fast betweenness centrality computation for requirement-driven approximation". In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 1289–1294.
- [GA15] Alex Greaves and Benjamin Au. "Using the bitcoin transaction graph to predict the price of bitcoin". In: *No Data* (2015).
- [GB21] Federico Germani and Nikola Biller-Andorno. "The anti-vaccination infodemic on social media: A behavioral analysis". In: *PloS one* 16.3 (2021), e0247642.
- [GDJ06] Saikat Guha, Neil Daswani, and Ravi Jain. "An Experimental Study of the Skype Peer-to-Peer VoIP System". In: *IPTPS'06: The 5th International Workshop on Peer-to-Peer Systems*. 2006.
- [GG10] Marco Galbiati and Simone Giansante. *Emergence of networks in large value payment systems (LVPSs)*. Department of Economic Policy, Finance and Development (DEPFID) University of Siena 0110. Jan. 2010.
- [GGR21] Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. "Detection of conspiracy propagators using psycho-linguistic characteristics". In: *Journal of Information Science* (2021), p. 0165551520985486.

- [Gia+20] Anastasia Giachanou et al. “The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers”. In: *International Conference on Applications of Natural Language to Information Systems*. Springer. 2020, pp. 181–192.
- [GK17] Aditi Gupta and Rishabh Kaushal. “Towards detecting fake user accounts in Facebook”. In: *ISEA Asia Security and Privacy (ISEASP)*. IEEE. 2017, pp. 1–6.
- [GL16] Aditya Grover and Jure Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.
- [GN02] M. Girvan and M. E. J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826.
- [GN06] M. T. Gastner and M. E. J. Newman. “The spatial structure of networks”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 49.2 (Jan. 2006), pp. 247–252.
- [Gor+19] Genevieve Gorrell et al. “SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*. Association for Computational Linguistics. 2019, pp. 845–854.
- [Gra73a] M. S. Granovetter. “The Strength of Weak Ties”. In: *The American Journal of Sociology* 78.6 (1973), pp. 1360–1380.
- [Gra73b] M.S. Granovetter. “The Strength of Weak Ties”. In: *The American Journal of Sociology* 78.6 (1973), pp. 1360–1380.
- [Gra78] M. Granovetter. “Threshold Models of Collective Behavior”. In: *The American Journal of Sociology* 83.6 (1978), pp. 1420–1443.
- [Gra82] Mark Granovetter. “The Strength of Weak Ties: A Network Theory Revisited”. In: *Sociological Theory*. 1982, pp. 105–130.
- [GRC19] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. “Leveraging emotional signals for credibility detection”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 877–880.
- [Gre+16] Klaus Greff et al. “LSTM: A search space odyssey”. In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.
- [Gri+19] Nir Grinberg et al. “Fake news on Twitter during the 2016 US presidential election”. In: *Science* 363.6425 (2019), pp. 374–378.
- [Gru+04] Daniel Gruhl et al. “Information diffusion through blogspace”. In: *WWW*. 2004.

- [GSZ19] Yixin Geng, Jie Sui, and Qian Zhu. “Rumor detection of Sina Weibo based on SDSMOTE and feature selection”. In: *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE. 2019, pp. 120–125.
- [Gun69] Robert Gunning. “The fog index after twenty years”. In: *Journal of Business Communication* 6.2 (1969), pp. 3–13.
- [Guo+06] Qiang Guo et al. “Growing scale-free small-world networks with tunable assortative coefficient”. In: *Physica A-statistical Mechanics and Its Applications* 371 (2006), pp. 814–822. DOI: 10.1016/j.physa.2006.03.055.
- [Gup+14] Aditi Gupta et al. “Tweetcred: Real-time credibility assessment of content on twitter”. In: *International Conference on Social Informatics*. Springer. 2014, pp. 228–243.
- [Gur+15] Supraja Gurajala et al. “Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach”. In: *Proceedings of the 2015 International Conference on Social Media & Society*. 2015, pp. 1–7.
- [GZ06] Anna Goldenberg and Alice Zheng. “Exploratory study of a new model of evolving networks”. In: *In ICML workshop on statistical network analysis*. SpringerVerlag, 2006.
- [HA03] Bernardo A. Huberman and Lada A. Adamic. “Information dynamics in a networked world”. In: *Complex Networks, Lecture Notes in Physics*. Springer, 2003, pp. 371–398.
- [Han+12] Lilia Hannachi et al. “Community Extraction Based on Topic-Driven-Model for Clustering Users Tweets”. English. In: *Advanced Data Mining and Applications*. Ed. by Shuigeng Zhou, Songmao Zhang, and George Karypis. Vol. 7713. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 39–51. ISBN: 978-3-642-35526-4. DOI: 10.1007/978-3-642-35527-1_4. URL: http://dx.doi.org/10.1007/978-3-642-35527-1_4.
- [Hao+20] Fatima Haouari et al. “ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection”. In: *arXiv preprint arXiv:2010.08768* (2020).
- [Har+17] Khalid Haruna et al. “A collaborative approach for research paper recommender system”. In: *PloS one* 12.10 (2017), e0184516.
- [He+14] Benlan He et al. “Prediction of Customer Attrition of Commercial Banks based on SVM Model”. In: *Procedia Computer Science* 31 (2014). 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014, pp. 423–430. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2014.05.286>. URL: <http://www.sciencedirect.com/science/article/pii/S1877050914004633>.

- [HFN13] Ben Holtz, Julie Fortuna, and Jocelyn Neff. “Evolutionary Structural Analysis of the Bitcoin Network”. In: 2013.
- [HG14] Clayton Hutto and Eric Gilbert. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.
- [HGC19] Sooji Han, Jie Gao, and Fabio Ciravegna. “Neural language model based training data augmentation for weakly supervised early rumor detection”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, pp. 105–112.
- [HK02a] P. Holme and B. J. Kim. “Growing scale-free networks with tunable clustering”. In: *Physical Review E* 65 (2002).
- [HK02b] Petter Holme and Beom J. Kim. “Growing scale-free networks with tunable clustering”. In: *Physical Review E* 65.2 (Jan. 2002).
- [Hof+21] Beth L Hoffman et al. “# DoctorsSpeakUp: Lessons learned from a pro-vaccine Twitter event”. In: *Vaccine* (2021).
- [Hor20] Richard Horton. “Offline: managing the COVID-19 vaccine infodemic”. In: *Lancet (London, England)* 396.10261 (2020), p. 1474.
- [Hos+07] M. Hosseini et al. “A Survey of Application-Layer Multicast Protocols”. In: 9.3 (2007).
- [Hos+20] Tamanna Hossain et al. “COVIDLies: Detecting COVID-19 misinformation on social media”. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. 2020.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [HSL15] Daning Hu, Gerhard Schwabe, and Xiao Li. “Systemic risk management and investment analysis with financial network analytics: research opportunities and challenges”. In: *Financial Innovation* 1.1 (June 2015), p. 2. ISSN: 2199-4730. DOI: 10.1186/s40854-015-0001-x. URL: <https://doi.org/10.1186/s40854-015-0001-x>.
- [Hu+21] Linmei Hu et al. “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)*. 2021, pp. 754–763.
- [Hua+19] Qi Huang et al. “Deep Structure Learning for Rumor Detection on Twitter”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

- [Hue08] Maik Huettinger. “Cultural dimensions in business life: Hofstede’s indices for Latvia and Lithuania”. In: *Baltic Journal of Management* 3.3 (2008), pp. 359–376.
- [HYL17] William L Hamilton, Rex Ying, and Jure Leskovec. “Representation learning on graphs: Methods and applications”. In: *arXiv preprint arXiv:1709.05584* (2017).
- [Iam+] Adriana Iamnitchi et al. “The Small World of File Sharing”. In: *IEEE Transactions on Parallel and Distributed Systems* ().
- [IK20] Isa Inuwa-Dutse and Ioannis Korkontzelos. “A curated collection of covid-19 online datasets”. In: *arXiv preprint arXiv:2007.09703* (2020).
- [IMR19] Mohammad Raihanul Islam, Sathappan Muthiah, and Naren Ramakrishnan. “RumorSleuth: joint detection of rumor veracity and user stance”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, pp. 131–136.
- [Isl+21] Md Saiful Islam et al. “COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence”. In: *PloS one* 16.5 (2021), e0251605.
- [Jag+21] Raj Jagtap et al. “Misinformation Detection on YouTube Using Video Captions”. In: *arXiv preprint arXiv:2107.00941* (2021).
- [JG19] Andrej Janchevski and Sonja Gievska. “AndrejJan at SemEval-2019 Task 7: A Fusion Approach for Exploring the Key Factors pertaining to Rumour Analysis”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019, pp. 1083–1089.
- [Jia+14] Meng Jiang et al. “Detecting suspicious following behavior in multi-millionnode social networks”. In: *Proceedings of the 23rd International Conference on World Wide Web*. 2014, pp. 305–306.
- [Jin+17] Zhiwei Jin et al. “Multimodal fusion with recurrent neural networks for rumor detection on microblogs”. In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 795–816.
- [JJH14] Soufiene Jaffali, Salma Jamoussi, and AbdelmajidBen Hamadou. “Grouping Like-Minded Users Based on Text and Sentiment Analysis”. English. In: *Computational Collective Intelligence. Technologies and Applications*. Ed. by Dosam Hwang, JasonJ. Jung, and Ngoc-Thanh Nguyen. Vol. 8733. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 83–93. ISBN: 978-3-319-11288-6. DOI: 10 . 1007 / 978 - 3 - 319 - 11289 - 3 _ 9. URL: http://dx.doi.org/10.1007/978-3-319-11289-3_9.

- [JK21] Dariusz Jemielniak and Yaroslav Krempovych. “An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter”. In: *Public Health* 200 (2021), pp. 4–6.
- [JL17] Huisu Jang and Jaewook Lee. “An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information”. In: *Ieee Access* 6 (2017), pp. 5427–5437.
- [JM21] Perna Juneja and Tanushree Mitra. “Auditing e-commerce platforms for algorithmically curated vaccine misinformation”. In: *arXiv preprint arXiv:2101.08419* (2021).
- [JMB09] Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. “T-Man: Gossip-based fast overlay topology construction”. In: *Comput. Netw.* 53.13 (2009).
- [JRT10] Matthew O. Jackson, Tomas Rodriguez-barraquer, and Xu Tan. *Social Capital and Social Quilts: Network Patterns of Favor Exchange* *. 2010.
- [JSL99] Anders Johansen, Didier Sornette, and Olivier Ledoit. “Predicting financial crashes using discrete scale invariance”. In: *arXiv preprint cond-mat/9903321* (1999).
- [JTO08] Wolfgang John, Sven Tafvelin, and Tomas Olovsson. “Trends and differences in connection-behavior within classes of internet backbone traffic”. In: *Proceedings of the 9th international conference on Passive and active network measurement. PAM’08*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 192–201.
- [JW02] Glen Jeh and Jennifer Widom. “SimRank: a measure of structural-context similarity”. In: *KDD ’02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002.
- [Kal+15] Harry A. Kalodner et al. “An Empirical Study of Namecoin and Lessons for Decentralized Namespace Design”. In: *14th Annual Workshop on the Economics of Information Security, WEIS 2015, Delft, The Netherlands, 22-23 June, 2015*. 2015. URL: http://www.econinfosec.org/archive/weis2015/papers/WEIS%5C_2015%5C_kalodner.pdf.
- [Kar+00] R. Karp et al. “Randomized rumor spreading”. In: *FOCS ’00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 2000.
- [KDQ20] Chandra Mouli Madhav Kotteti, Xishuang Dong, and Lijun Qian. “Ensemble Deep Learning on Time-Series Representation of Tweets for Rumor Detection in Social Media”. In: *Applied Sciences* 10.21 (2020), p. 7541.

- [KF13] Natascha A Karlova and Karen E Fisher. “A social diffusion model of misinformation and disinformation for understanding human information behaviour”. In: (2013).
- [KG04] James S. H. Kwok and S. Gao. “Knowledge sharing community in P2P network: a study of motivational perspective.” In: *J. Knowledge Management* 8.1 (2004), pp. 94–102. URL: <http://dblp.uni-trier.de/db/journals/jkm/jkm8.html#KwokG04>.
- [Khe+02] Abdelmajid Khelil et al. “An epidemic model for information diffusion in MANETs”. In: *Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems*. MSWiM ’02. ACM, 2002, pp. 54–60.
- [Kim+21] Jisu Kim et al. “FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period”. In: *Telematics and Informatics* 64 (2021), p. 101688.
- [Kim03] Kyoung-jae Kim. “Financial time series forecasting using support vector machines”. In: *Neurocomputing* 55.1-2 (2003), pp. 307–319.
- [Kin+75] J Peter Kincaid et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch, 1975.
- [KKT03] David Kempe, Jon Kleinberg, and Éva Tardos. “Maximizing the spread of influence through a social network”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 137–146.
- [Kle00a] J. Kleinberg. “The small-world phenomenon: an algorithm perspective”. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. STOC ’00. ACM, 2000, pp. 163–170.
- [Kle00b] Jon Kleinberg. “The Small-World Phenomenon: An Algorithmic Perspective”. In: *in Proceedings of the 32nd ACM Symposium on Theory of Computing*. 2000, pp. 163–170.
- [Kle99] Jon M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment”. In: *Journal of the ACM* 46 (1999), pp. 668–677.
- [KLZ18a] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. “All-in-one: Multi-task learning for rumour verification”. In: *arXiv preprint arXiv:1806.03713* (2018).
- [KLZ18b] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. *PHEME dataset for Rumour Detection and Veracity Classification*. June 2018. DOI: 10.6084/m9.figshare.6392078.v1. URL: https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078/1.
- [KM18] Werner Kristjanpoller and Marcel C Minutolo. “A hybrid volatility forecasting framework integrating GARCH, artificial neural net-

- work, technical analysis and principal components analysis”. In: *Expert Systems with Applications* 109 (2018), pp. 1–11.
- [KM20] Nisha Kundu and Yogesh Kumar Meena. “Detecting Suspicious Users in Social Networks Using Text Analysis”. In: *Smart Systems and IoT: Innovations in Computing*. Springer, 2020, pp. 463–473.
- [KNL21] Mansoor Karami, Tahora H Nazer, and Huan Liu. “Profiling fake news spreaders on social media through psychological and motivational factors”. In: *Proceedings of the 32nd ACM conference on hypertext and social media*. 2021, pp. 225–230.
- [Kol+22] Boshko Koloski et al. “Knowledge graph informed fake news classification via heterogeneous representation ensembles”. In: *Neurocomputing* 496 (2022), pp. 208–226.
- [Kou+20] Ramez Kouzy et al. “Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter”. In: *Cureus* 12.3 (2020).
- [KR20] Deepak Kumar and SK Rath. “Predicting the Trends of Price for Ethereum Using Deep Learning Techniques”. In: *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Springer, 2020, pp. 103–114.
- [KRD04] Mujtaba Khambatti, KyungDong Ryu, and Partha Dasgupta. “Structuring Peer-to-Peer Networks Using Interest-Based Communities”. English. In: *Databases, Information Systems, and Peer-to-Peer Computing*. Ed. by Karl Aberer, Manolis Koubarakis, and Vana Kalogeraki. Vol. 2944. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 48–63. ISBN: 978-3-540-20968-3. DOI: 10.1007/978-3-540-24629-9_5. URL: http://dx.doi.org/10.1007/978-3-540-24629-9_5.
- [Kre+04] Marc van Kreveld et al. “Distributed Ranking Methods for Geographic Information Retrieval”. In: *In 20th European Workshop on Computational Geometry*. Springer, 2004, pp. 231–243.
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.
- [KSN19] Akshi Kumar, Saurabh Raj Sangwan, and Anand Nayyar. “Rumour veracity detection on twitter using particle swarm optimized shallow classifiers”. In: *Multimedia Tools and Applications* 78.17 (2019), pp. 24083–24101.
- [KT19] Hamid Karimi and Jiliang Tang. “Learning hierarchical discourse-level structure for fake news detection”. In: *arXiv preprint arXiv:1903.07389* (2019).

- [Kul09] Santosh Kulkarni. “Badumna Network Suite: A Decentralized Network Engine for Massively Multiplayer Online Applications”. In: *Peer-to-Peer Computing*. 2009.
- [Kum+05a] Ravi Kumar et al. “On the Bursty Evolution of Blogspace”. In: *World Wide Web* 8.2 (2005).
- [Kum+05b] Ravi Kumar et al. *Theoretical Analysis of Geographic Routing in Social Networks*. Tech. rep. June 2005. URL: <http://publications.csail.mit.edu/tmp/MIT-CSAIL-TR-2005-040.pdf>.
- [Kum18] Suni Ajay Kumar. “A regression model for crypto-currency price”. In: (2018).
- [KW16] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).
- [Kwa+10] Haewoon Kwak et al. “What is Twitter, a social network or a news media?” In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 591–600.
- [Kwo+13] Sejeong Kwon et al. “Prominent features of rumor propagation in online social media”. In: *2013 IEEE 13th international conference on data mining*. IEEE. 2013, pp. 1103–1108.
- [KYC15] Sibel B. Kusimba, Yang Yang, and Nitesh V. Chawla. “Family Networks of Mobile Money in Kenya.” In: *Information Technology in International Development* 11 (2015), pp. 1–21.
- [Kyr+09] F. Kyriakopoulos et al. “Network and eigenvalue analysis of financial transaction networks”. In: *The European Physical Journal B* 71.4 (July 2009), p. 523. ISSN: 1434-6036. DOI: 10.1140/epjb/e2009-00255-7. URL: <https://doi.org/10.1140/epjb/e2009-00255-7>.
- [Lah18] Salim Lahmiri. “Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression”. In: *Applied Mathematics and Computation* 320 (2018), pp. 444–451.
- [Lam+08a] Renaud Lambiotte et al. “Geographical dispersal of mobile communication networks”. In: *PHYSICA A* 387 (2008), p. 5317.
- [Lam+08b] Renaud Lambiotte et al. “Geographical dispersal of mobile communication networks.” In: vol. 387. 2008, pp. 5317–5325.
- [Lam21] Rabindra Lamsal. “Design and analysis of a large-scale COVID-19 tweets dataset”. In: *Applied Intelligence* 51.5 (2021), pp. 2790–2804.
- [Lan+17] Anthony Lantian et al. “I know things they don’t know!” In: *Social Psychology* (2017).
- [Laz+02] Michael A Lazarus et al. *Predictive modeling of consumer financial behavior*. US Patent 6,430,539. Aug. 2002.

- [Laz+07] Michael A Lazarus et al. *Predictive modeling of consumer financial behavior using supervised segmentation and nearest-neighbor matching*. US Patent 7,165,037. Jan. 2007.
- [LB+95] Yann LeCun, Yoshua Bengio, et al. “Convolutional networks for images, speech, and time series”. In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.
- [LB19] Salim Lahmiri and Stelios Bekiros. “Cryptocurrency forecasting with deep learning chaotic neural networks”. In: *Chaos, Solitons & Fractals* 118 (2019), pp. 35–40.
- [LBS10] Hui Li, Sourav S. Bhowmick, and Aixin Sun. “Affinity-driven Prediction and Ranking of Products in Online Product Review Sites”. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM ’10. ACM, 2010.
- [LC22] Colin Jingwei Lee and Hui Na Chua. “Using Linguistics and Psycholinguistics Features in Machine Learning for Fake News Classification Through Twitter”. In: *Proceedings of International Conference on Data Science and Applications*. Springer. 2022, pp. 717–730.
- [LD12] Kwan Hui Lim and A. Datta. “Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter Using Tweeting Links”. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*. Vol. 1. Dec. 2012, pp. 214–221. DOI: 10.1109/WI-IAT.2012.53.
- [LD13] Kwan Hui Lim and Amitava Datta. “A Topological Approach for Detecting Twitter Communities with Common Interests.” In: vol. 8329. *Ubiquitous Social Media Analysis.*, 2013, pp. 23–43.
- [Leo+16] Yannick Leo et al. “Correlations of Consumption Patterns in Social-economic Networks”. In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’16. Davis, California: IEEE Press, 2016, pp. 493–500. ISBN: 978-1-5090-2846-7. URL: <http://dl.acm.org/citation.cfm?id=3192424.3192516>.
- [Les+07] Gilly Leshed et al. “Feedback for guiding reflection on teamwork practices”. In: *Proceedings of the 2007 international ACM conference on Supporting group work*. 2007, pp. 217–220.
- [LHu+11] Gaston L’Huillier et al. “Topic-based Social Network Analysis for Virtual Communities of Interests in the Dark Web”. In: *SIGKDD Explor. Newsl.* 12.2 (Mar. 2011), pp. 66–73. ISSN: 1931-0145. DOI: 10.1145/1964897.1964917. URL: <http://doi.acm.org/10.1145/1964897.1964917>.

- [Li+20] Yichuan Li et al. “MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation”. In: *arXiv preprint arXiv:2011.04088* (2020).
- [Li+22] Ke Li et al. “Dynamic probabilistic graphical model for progressive fake news detection on social media platform”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 13.5 (2022), pp. 1–24.
- [LI18] CHANG M LIY and LYUS InIctuOculi. “Exposingaigcreated fake-videosbydetectingeyebinking”. In: *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China*. 2018, pp. 11–13.
- [Lia+16] Dawen Liang et al. “Modeling user exposure in recommendation”. In: *Proceedings of the 25th international conference on World Wide Web*. 2016, pp. 951–961.
- [Lib+05] D. Liben-Nowell et al. “Geographic routing in social networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.33 (2005), pp. 11623–11628.
- [Lie+02] Patricia Liehr et al. “Expressing health experience through embodied language”. In: *Journal of Nursing Scholarship* 34.1 (2002), pp. 27–32.
- [Lie+22] Cristian Lieneck et al. “Facilitators and Barriers of COVID-19 Vaccine Promotion on Social Media in the United States: A Systematic Review”. In: *Healthcare*. Vol. 10. 2. MDPI. 2022, p. 321.
- [LKF07] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. “Graph evolution: Densification and shrinking diameters”. In: *ACM Trans. Knowl. Discov. Data* 1.1 (Mar. 2007), pp. 2+. ISSN: 1556-4681.
- [LLZ18] Jiaqi Liang, Linjing Li, and Daniel Zeng. “Evolutionary dynamics of cryptocurrency transaction networks: An empirical study”. In: (2018).
- [Loo+21] Sahil Loomba et al. “Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA”. In: *Nature human behaviour* 5.3 (2021), pp. 337–348.
- [Loo94] Siew Lan Loo. “Neural networks for financial forecasting”. PhD thesis. University of London, 1994.
- [LPV13] Mingfeng Lin, Nagpurnanand R. Prabhala, and Siva Viswanathan. “Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending”. In: *Management Science* 59 (2013), pp. 17–35.
- [LR06] Michael Levi and Peter Reuter. “Money laundering”. In: *Crime and Justice* 34.1 (2006), pp. 289–375.
- [LS] Monero Logo and Nicolas van Saberhagen. “Monero (cryptocurrency)”. In: ().

- [LT17] PICHL Lukáš and KAIZOJI Taisei. “Volatility analysis of Bitcoin price time series”. In: *Quantitative Finance and Economics* 1 (2017), pp. 474–485.
- [Lu+12] Yong Lu et al. “Social Influence and Defaults in Peer-to-Peer Lending Networks”. In: *ICIS*. 2012.
- [Luk+16] Michal Lukasik et al. “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)*. 2016, pp. 393–398.
- [Lüt13] Helmut Lütkepohl. “Vector autoregressive models”. In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing, 2013.
- [LW18] Yang Liu and Yi-Fang Wu. “Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.
- [LW20] Yang Liu and Yi-Fang Brook Wu. “Fned: a deep network for fake news early detection on social media”. In: *ACM Transactions on Information Systems (TOIS)* 38.3 (2020), pp. 1–33.
- [M01] Batty M. “Cities as small worlds”. In: *Environment and Planning B: Planning and Design* 28.5 (2001), 637 –638.
- [Ma+16] Jing Ma et al. “Detecting rumors from microblogs with recurrent neural networks”. In: (2016).
- [Mab18] Sagwadi Mabunda. “Cryptocurrency: The new face of cyber money laundering”. In: *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. IEEE. 2018, pp. 1–6.
- [Mal+21] Larissa G Malagoli et al. “A Look into COVID-19 Vaccination Debate on Twitter”. In: *13th ACM Web Science Conference 2021*. 2021, pp. 225–233.
- [Mar+21] Julio Emilio Marco-Franco et al. “COVID-19, Fake News, and Vaccines: Should Regulation Be Implemented?” In: *International Journal of Environmental Research and Public Health* 18.2 (2021), p. 744.
- [Mas+19] Faiza Masood et al. “Spammer detection and fake user identification on social networks”. In: *IEEE Access* 7 (2019), pp. 68140–68152.
- [MB05] Edward C Malthouse and Robert C Blattberg. “Can we predict customer lifetime value?” In: *Journal of interactive marketing* 19.1 (2005), pp. 2–16.
- [Mc 69] G Harry Mc Laughlin. “SMOG grading-a new readability formula”. In: *Journal of reading* 12.8 (1969), pp. 639–646.

- [MC20] Shahan Ali Memon and Kathleen M Carley. “Characterizing covid-19 misinformation communities using a novel twitter dataset”. In: *arXiv preprint arXiv:2008.00791* (2020).
- [MCB11] Huina Mao, Scott Counts, and Johan Bollen. “Predicting financial markets: Comparing survey, news, twitter and search engine data”. In: *arXiv preprint arXiv:1112.1051* (2011).
- [McH12] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pp. 276–282.
- [MCP16] Tanushree Mitra, Scott Counts, and James Pennebaker. “Understanding anti-vaccination attitudes in social media”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10. 2016.
- [Mes21] Daniela Olivera Mesa. *Report 43 - Quantifying the impact of vaccine hesitancy in prolonging the need for Non-Pharmaceutical Interventions to control the COVID-19 pandemic*. <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-43-vaccine-hesitancy/>. 2021.
- [MG14] Nikita Munot and Sharvari S Govilkar. “Comparative study of text summarization methods”. In: *International Journal of Computer Applications* 102.12 (2014).
- [MGW18] Jing Ma, Wei Gao, and Kam-Fai Wong. “Rumor detection on twitter with tree-structured recursive neural networks”. In: Association for Computational Linguistics. 2018.
- [MGW19] Jing Ma, Wei Gao, and Kam-Fai Wong. “Detect rumors on Twitter by promoting information campaigns with generative adversarial learning”. In: *The World Wide Web Conference*. 2019, pp. 3049–3055.
- [Mik+13] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
- [Mil67] S. Milgram. “The Small World Problem”. In: *Psychology Today* 2 (1967), pp. 60–67.
- [Mis+07] Alan Mislove et al. “Measurement and Analysis of Online Social Networks”. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. San Diego, California, USA: ACM, 2007, pp. 29–42.
- [MK20] Yelena Mejova and Kyriaki Kalimeri. “Advertisers jump on coronavirus bandwagon: Politics, news, and business”. In: *arXiv preprint arXiv:2003.00923* (2020).
- [ML09] Zhien Ma and Jia Li. “Dynamical Modeling and Analysis of Epidemics”. In: (2009).

- [ML19] Bjarke Mønsted and Sune Lehmann. “Algorithmic detection and analysis of vaccine-denialist sentiment clusters in social networks”. In: *arXiv preprint arXiv:1905.12908* (2019).
- [Mol+21] Maria D Molina et al. ““Fake news” is not simply false information: a concept explication and taxonomy of online content”. In: *American behavioral scientist* 65.2 (2021), pp. 180–212.
- [Mon+10] Géraldine Del Mondo et al. “A Graph Model for Spatio-temporal Evolution”. In: 16.11 (June 1, 2010), pp. 1452–1477.
- [Mon+19] Federico Monti et al. “Fake news detection on social media using geometric deep learning”. In: *arXiv preprint arXiv:1902.06673* (2019).
- [Mos+22] Ahmadreza Mosallanezhad et al. “Domain adaptive fake news detection via reinforcement learning”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 3632–3640.
- [MOS21] Chad Melton, Olufunto A Olusanya, and Arash Shaban-Nejad. “Network analysis of COVID-19 vaccine misinformation on social media”. In: *Stud Health Technol Inform* 287 (2021), pp. 165–166.
- [MP22] Antoine Marie and Michael Bang Petersen. “Political conspiracy theories as tools for mobilization and signaling”. In: *Current Opinion in Psychology* (2022), p. 101440.
- [MS11] John Markoff and Somini Sengupta. *Separating You and Me? 4.74 Degrees*. <https://www.nytimes.com/2011/11/22/technology/between-you-and-me-4-74-degrees.html>. [Online; accessed 21-November-2011]. 2011.
- [MS17] Jinxuan Ma and Lynne Stahl. “A multimodal critical discourse analysis of anti-vaccination information on Facebook”. In: *Library & Information Science Research* 39.4 (2017), pp. 303–310.
- [MSC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a Feather: Homophily in Social Networks”. In: *Annual Review of Sociology* 27.1 (2001), pp. 415–444.
- [MSS22] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. “DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection”. In: *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2022.
- [Mur+21] Jamie Murphy et al. “Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom”. In: *Nature communications* 12.1 (2021), pp. 1–15.
- [MW21] S. Machingaidze and C.S. Wiysonge. “Understanding COVID-19 vaccine hesitancy”. In: *Nature Medicine* (2021).
- [Nak+08] Satoshi Nakamoto et al. “Bitcoin: A peer-to-peer electronic cash system”. In: (2008), p. 1.

- [Nak09] Satoshi Nakamoto. *Bitcoin: A peer-to-peer electronic cash system*. 2009. URL: <http://www.bitcoin.org/bitcoin.pdf>.
- [New+03] Matthew L Newman et al. “Lying words: Predicting deception from linguistic styles”. In: *Personality and social psychology bulletin* 29.5 (2003), pp. 665–675.
- [New+08] Matthew L Newman et al. “Gender differences in language use: An analysis of 14,000 text samples”. In: *Discourse processes* 45.3 (2008), pp. 211–236.
- [New01a] M. Newman. “Scientific collaboration networks: I. Network construction and fundamental results”. In: *Phys. Rev. E*, vol. 64, 2001.
- [New01b] M. E. J. Newman. “Clustering and preferential attachment in growing networks”. In: *Physical Review E* 64.2 (July 2001), pp. 025102+.
- [Ngu+12] Nam P Nguyen et al. “Containment of misinformation spread in online social networks”. In: *Proceedings of the 4th Annual ACM Web Science Conference*. 2012, pp. 213–222.
- [Ngu17] Tu Ngoc Nguyen. “A comprehensive low and high-level feature analysis for early rumor detection on twitter”. In: *arXiv preprint arXiv:1711.00726* (2017).
- [Nil+19] Shirin Nilizadeh et al. “Think outside the dataset: Finding fraudulent reviews using cross-dataset analysis”. In: *The World Wide Web Conference*. 2019, pp. 3108–3115.
- [NSY22] Cindy Sing Bik Ngai, Rita Gill Singh, and Le Yao. “Impact of COVID-19 vaccine misinformation on social media virality: content analysis of message themes and writing strategies”. In: *Journal of medical Internet research* 24.7 (2022), e37806.
- [NW99] M. E. J. Newman and D. J. Watts. “Scaling and percolation in the small-world network model”. In: *PHYS.REVE* 60 (1999).
- [OVC15] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. “What to expect when the unexpected happens: Social media communications across crises”. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 2015, pp. 994–1009.
- [Pal+12] Diana Palsetia et al. “User-Interest based Community Extraction in Social Networks”. In: *The 6th SNA-KDD Workshop*. 2012.
- [Pan+18] Dibakar Raj Pant et al. “Recurrent neural network based bitcoin price prediction by twitter sentiment analysis”. In: *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*. IEEE. 2018, pp. 128–132.
- [Par+20] Aldo Pareja et al. “EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs.” In: *AAAI*. 2020, pp. 5363–5370.

- [Par19] Nikita Nitin Parab. “Twitter Rumour Detection using Temporal Property of Tweets”. PhD thesis. Dublin, National College of Ireland, 2019.
- [PAS14] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. “Deepwalk: Online learning of social representations”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 701–710.
- [Pat+20] Parth Patwa et al. “Fighting an infodemic: Covid-19 fake news dataset”. In: *arXiv preprint arXiv:2011.03327* (2020).
- [Pat+21] Parth Patwa et al. “Fighting an infodemic: Covid-19 fake news dataset”. In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CON-STRANT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer. 2021, pp. 21–29.
- [PDA13] F. Pozzi, T. Di Matteo, and T. Aste. “Spread of risk across financial markets: better to invest in the peripheries”. English. In: *Scientific Reports* 3 (Apr. 2013). ISSN: 2045-2322. DOI: 10.1038/srep01665.
- [Pen+15] James W Pennebaker et al. *The development and psychometric properties of LIWC2015*. Tech. rep. 2015.
- [Pér+17] Verónica Pérez-Rosas et al. “Automatic detection of fake news”. In: *arXiv preprint arXiv:1708.07104* (2017).
- [PGC06] J. A. Patel, I. Gupta, and N. Contractor. “JetStream: Achieving Predictable Gossip Dissemination by Leveraging Social Network Principles”. In: *Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications*. 2006.
- [Pie+21] Francesco Pierri et al. “VaccinItaly: monitoring Italian conversations around vaccines on Twitter”. In: *arXiv preprint arXiv:2101.03757* (2021).
- [PLB13] Charles Perez, Marc Lemercier, and Babiga Birregah. “A dynamic approach to detecting suspicious profiles on social platforms”. In: *2013 IEEE International Conference on Communications Workshops (ICC)*. IEEE. 2013, pp. 174–178.
- [Plu01] Robert Plutchik. “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”. In: *American scientist* 89.4 (2001), pp. 344–350.
- [PMF97] James W Pennebaker, Tracy J Mayne, and Martha E Francis. “Linguistic predictors of adaptive bereavement.” In: *Journal of personality and social psychology* 72.4 (1997), p. 863.
- [Pot+17] Martin Potthast et al. “A stylometric inquiry into hyperpartisan and fake news”. In: *arXiv preprint arXiv:1702.05638* (2017).

- [PP71] Vilfredo Pareto and Alfred N. Page. “Translation of Manuale di economia politica”. In: *Manual of political economy* (1971).
- [PR04] Gorman S P and Kulkarni R. “Spatial small worlds: new geographic patterns for an information economy”. In: *Environment and Planning B: Planning and Design* 31.2 (2004), 273 –296.
- [PS03] James W Pennebaker and Lori D Stone. “Words of wisdom: language use over the life span.” In: *Journal of personality and social psychology* 85.2 (2003), p. 291.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.
- [Pur+20] Neha Puri et al. “Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases”. In: *Human vaccines & immunotherapeutics* 16.11 (2020), pp. 2586–2593.
- [PY09] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [Qaz+11] Vahed Qazvinian et al. “Rumor has it: Identifying misinformation in microblogs”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 1589–1599.
- [Rai+21] Stephen A Rains et al. “Psycholinguistic Markers of COVID-19 Conspiracy Tweets and Predictors of Tweet Dissemination”. In: *Health Communication* (2021), pp. 1–10.
- [Raj18] Mladen Rajko. “Digital Currencies as a Substitute Means of Payment”. In: *ECONOMIC INTEGRATIONS, THEORY, COOPERATION AND PERSPECTIVES Integrations reforms and business challenges*. 2018.
- [Rap53] Anatol Rapoport. “Spread of information through a population with sociostructural bias: I. Assumption of transitivity.” In: *Bulletin of Mathematical Biophysics* 15 (1953).
- [Rat+17] Bhavtosh Rath et al. “From retweet to believability: Utilizing trust to identify rumor spreaders on twitter”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2017, pp. 179–186.
- [RB02] Albert Reka and Barabási. “Statistical mechanics of complex networks”. In: *Rev. Mod. Phys.* 74 (June 2002), pp. 47–97.
- [Ren+17] Stephanie Rendón de la Torre et al. “Fractal and multifractal analysis of complex networks: Estonian network of payments”. In: *The European Physical Journal B* 90.12 (Dec. 2017), p. 234. ISSN: 1434-

6036. DOI: 10.1140/epjb/e2017-80214-5. URL: <https://doi.org/10.1140/epjb/e2017-80214-5>.
- [RG19] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).
- [RMS21] Bhavtosh Rath, Xavier Morales, and Jaideep Srivastava. “SCARLET: explainable attention based graph neural network for fake news spreader prediction”. In: *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I*. Springer. 2021, pp. 714–727.
- [Roo+20] Jon Roozenbeek et al. “Susceptibility to misinformation about COVID-19 around the world”. In: *Royal Society open science* 7.10 (2020), p. 201199.
- [Ros+14] Arad Rostampour et al. *Method for detecting suspicious individuals in a friend list*. US Patent 8,725,672. May 2014.
- [Ros+15] Giulio Rossetti et al. “Community-centric Analysis of User Engagement in Skype Social Network”. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ASONAM ’15. Paris, France: ACM, 2015, pp. 547–552. ISBN: 978-1-4503-3854-7. DOI: 10.1145/2808797.2809384. URL: <http://doi.acm.org/10.1145/2808797.2809384>.
- [RRR20] Viju Raghupathi, Jie Ren, and Wullianallur Raghupathi. “Studying public perception about vaccination: A sentiment analysis of tweets”. In: *International journal of environmental research and public health* 17.10 (2020), p. 3464.
- [RS13] Dorit Ron and Adi Shamir. “Quantitative Analysis of the Full Bitcoin Transaction Graph”. In: *Financial Cryptography and Data Security*. Ed. by Ahmad-Reza Sadeghi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 6–24. ISBN: 978-3-642-39884-1.
- [RSP20] Nir Rosenfeld, Aron Szanto, and David C Parkes. “A Kernel of Truth: Determining Rumor Veracity on Twitter by Diffusion Pattern Alone”. In: *arXiv* (2020), arXiv–2002.
- [RSR20] Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. “The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic”. In: *Canadian journal of emergency medicine* 22.4 (2020), pp. 418–421.
- [RVH19] Alex Ratner, Paroma Varma, and Braden Hancock. “Weak supervision: A new programming paradigm for machine learning—sail blog”. In: *Visited on* 6.26 (2019), p. 2020.

- [RY18] Zuherman Rustam and Frederica Yaurita. “Insolvency prediction in insurance companies using support vector machines and fuzzy kernel c-means”. In: *Journal of Physics: Conference Series*. Vol. 1028. 1. IOP Publishing Ltd. 2018, p. 012118.
- [SA13] Rajesh Sharma and Datta Anwitaman. “GoDisco++: A Gossip Algorithm for Information Dissemination in Multi-dimensional Community Networks”. In: *Pervasive Mob. Comput.* 9.2 (Apr. 2013), pp. 324–335. ISSN: 1574-1192.
- [Sak+21] Ahmed Shahriar Sakib et al. “Identifying insomnia from social media posts: psycholinguistic analyses of user tweets”. In: *Journal of medical Internet research* 23.12 (2021), e27613.
- [Sal21] Malik Sallam. “COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates”. In: *Vaccines* 9.2 (2021), p. 160.
- [Sam+16] Justin Sampson et al. “Leveraging the implicit structure within social media for emergent rumor detection”. In: *Proceedings of the 25th ACM international on conference on information and knowledge management*. 2016, pp. 2377–2382.
- [San+19] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *ArXiv abs/1910.01108* (2019).
- [San+21] Abraham C Sanders et al. “Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse”. In: *medRxiv* (2021), pp. 2020–08.
- [Sax+20] Akraati Saxena et al. “Mitigating misinformation in online social network with top-k debunkers and evolving user opinions”. In: *Companion proceedings of the web conference 2020*. 2020, pp. 363–370.
- [SB20] S Santhoshkumar and LD Dhinesh Babu. “Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks”. In: *Social Network Analysis and Mining* 10.1 (2020), pp. 1–17.
- [SBL19] Kai Shu, H Russell Bernard, and Huan Liu. “Studying fake news via network analysis: detection and mitigation”. In: *Emerging research challenges and opportunities in computational social network analysis and mining* (2019), pp. 43–65.
- [SC09] Robert P Schumaker and Hsinchun Chen. “Textual analysis of stock market prediction using breaking financial news: The AZFin text system”. In: *ACM Transactions on Information Systems (TOIS)* 27.2 (2009), pp. 1–19.
- [SC20] Pardeep Singh and Satish Chand. “Predicting the Popularity of Rumors in Social Media Using Machine Learning”. In: *Social Net-*

- working and Computational Intelligence*. Springer, 2020, pp. 775–789.
- [Sce+11] Salvatore Scellato et al. “Socio-spatial Properties of Online Location-based Social Networks”. In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2011.
 - [SCG08] Rachel A Simmons, Dianne L Chambless, and Peter C Gordon. “How do hostile and emotionally overinvolved relatives view relationships?: What relatives’ pronoun use tells us”. In: *Family Process* 47.3 (2008), pp. 405–419.
 - [Sch+20] Leonard Schild et al. “" go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19”. In: *arXiv preprint arXiv:2004.04046* (2020).
 - [Sch+21] Mina Schütz et al. “Automatic Fake News Detection with Pre-trained Transformer Models”. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by Alberto Del Bimbo et al. Cham: Springer International Publishing, 2021, pp. 627–641.
 - [SD12a] Rajesh Sharma and Anwitaman Datta. “Decentralized Information Dissemination in Multidimensional Semantic Social Overlays”. In: *ICDCN*. 2012.
 - [SD12b] Rajesh Sharma and Anwitaman Datta. “SuperNova: Super-peers Based Architecture for Decentralized Online Social Networks”. In: *COMSNETS* (2012).
 - [Sea+20] Richard F Sear et al. “Quantifying covid-19 content in the online health opinion war using machine learning”. In: *Ieee Access* 8 (2020), pp. 91886–91893.
 - [Seg04] Mark R Segal. “Machine learning benchmarks and random forest regression”. In: (2004).
 - [SGM20] Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. “Foundations and modelling of dynamic networks using Dynamic Graph Neural Networks: A survey”. In: *arXiv preprint arXiv:2005.07496* (2020).
 - [Sha+19] Karishma Sharma et al. “Combating fake news: A survey on identification and mitigation techniques”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.3 (2019), pp. 1–42.
 - [Sha+22a] Wajiha Shahid et al. “Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities”. In: *IEEE Transactions on Computational Social Systems* (2022).
 - [Sha+22b] Shakshi Sharma et al. “FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset”. In: *ICWSM* (2022).
 - [Sha+23] Shakshi Sharma et al. “Misinformation Concierge: A Proof-of-Concept with Curated Twitter Dataset on COVID-19 Vaccination”. In: *CIKM*. 2023.

- [Sha09] Devavrat Shah. “Gossip Algorithms”. In: *Found. Trends Netw.* 3.1 (2009).
- [Shi+03] Shuming Shi et al. “Distributed page ranking in structured p2p networks”. In: *In ICPP*. 2003, pp. 179–186.
- [Shu+19a] Kai Shu et al. “defend: Explainable fake news detection”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 395–405.
- [Shu+19b] Kai Shu et al. “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*. 2019, pp. 436–439.
- [Shu+20] Kai Shu et al. “Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media”. In: *Big data* 8.3 (2020), pp. 171–188.
- [Sic+18] Rosa Sicilia et al. “Twitter rumour detection in the health domain”. In: *Expert Systems with Applications* 110 (2018), pp. 33–40.
- [Sin+19] Shivangi Singhal et al. “SpotFake: A Multi-modal Framework for Fake News Detection”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2019, pp. 39–47.
- [ŞJ05] Özgür Şimşek and David Jensen. “Decentralized search in networks using homophily and degree disparity”. In: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland, 2005.
- [Ska+22] Ingjerd Skafle et al. “Misinformation about COVID-19 vaccines on social media: rapid review”. In: *Journal of medical Internet research* 24.8 (2022), e37367.
- [SL04] Dexter Chi Wai Siu and Tak Pang Lau. “Distributed ranking over peer-to-peer networks”. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. WWW Alt. '04. ACM, 2004.
- [SMM09] Julia Schwarz, Jennifer Mankoff, and H. Scott Matthews. “Reflections of Everyday Activities in Spending Data”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: ACM, 2009, pp. 1737–1740. ISBN: 978-1-60558-246-7. DOI: 10 . 1145 / 1518701 . 1518968. URL: <http://doi.acm.org/10.1145/1518701.1518968>.
- [SMÜ19] Rajesh Sharma, Artem Mateush, and Jaan Übi. “Tale of Three States: Analysis of Large Person-to-Person Online Financial Transactions in Three Baltic Countries”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 1497–1505.

- [SN13] Mohammad Salehan and Arash Negahban. “Social networking on smartphones: When mobile phones become addictive”. In: *Computers in human behavior* 29.6 (2013), pp. 2632–2639.
- [SN20] Gautam Kishore Shahi and Durgesh Nandini. “FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19”. In: *arXiv preprint arXiv:2006.11343* (2020).
- [SO10] Shashi Shekhar and Dev Oliver. “Computational Modeling of Spatio-temporal Social Networks: A Time-Aggregated Graph Approach”. In: *2010 Specialist Meeting Spatio-Temporal Constraints on Social Networks*. 2010.
- [Sol+19] R Sudhesh Solomon et al. “Understanding the psycho-sociological facets of homophily in social network communities”. In: *IEEE Computational Intelligence Magazine* 14.2 (2019), pp. 28–40.
- [Son+05] Xiaodan Song et al. “Modeling and predicting personal information dissemination behavior”. In: *KDD ’05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005.
- [Sov18] Yhlas Sovbetov. “Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, bitcoin, and monero”. In: *Journal of Economics and Financial Analysis* 2.2 (2018), pp. 1–27.
- [SS04] Alex J Smola and Bernhard Schölkopf. “A tutorial on support vector regression”. In: *Statistics and computing* 14.3 (2004), pp. 199–222.
- [SS21a] Shakshi Sharma and Rajesh Sharma. “Identifying possible rumor spreaders on twitter: A weak supervised learning approach”. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.
- [SS21b] Shakshi Sharma and Rajesh Sharma. “Identifying Possible Rumor Spreaders on Twitter: AWeak Supervised Learning Approach”. In: *Proceedings of the 19th international conference on World wide web*. 2021.
- [SSD22] Shakshi Sharma, Rajesh Sharma, and Anwitaman Datta. “(Mis) leading the COVID-19 Vaccination Discourse on Twitter: An Exploratory Study of Infodemic Around the Pandemic”. In: *IEEE Transactions on Computational Social Systems* (2022).
- [SWL18] Kai Shu, Suhang Wang, and Huan Liu. “Understanding user profiles on social media for fake news detection”. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 430–435.
- [SWL19] Kai Shu, Suhang Wang, and Huan Liu. “Beyond news contents: The role of social context for fake news detection”. In: *Proceedings of*

- the twelfth ACM international conference on web search and data mining*. 2019, pp. 312–320.
- [SYB+14] David Schwartz, Noah Youngs, Arthur Britto, et al. “The ripple protocol consensus algorithm”. In: *Ripple Labs Inc White Paper 5.8* (2014), p. 151.
- [Tam+19] Nguyen Thanh Tam et al. “From anomaly detection to rumour detection using data streams of social platforms”. In: *Proceedings of the VLDB Endowment* 12.9 (2019), pp. 1016–1029.
- [TCL08] Dinh Nguyen Tran, Frank Chiang, and Jinyang Li. “Friendstore: cooperative online backup using trusted nodes”. In: *SocialNets '08: Proceedings of the 1st Workshop on Social Network Systems*. 2008.
- [TD23] Zachary Tan and Anwitaman Datta. “The first year of the Covid-19 pandemic through the lens of r/Coronavirus subreddit: An exploratory study”. In: *Health and Technology* (2023).
- [TDL21] Lisa Tang, Sabrina Douglas, and Amar Laila. “Among sheeple and antivaxxers: Social media responses to COVID-19 vaccine news posted by Canadian news organizations, and recommendations to counter vaccine hesitancy”. In: *Canada communicable disease report= Relevé des maladies transmissibles au Canada* 47.12 (2021), pp. 524–533.
- [The21] The Center for Countering Digital Hate. *The Disinformation Dozen*. 2021.
- [Tho+13] Kurt Thomas et al. “Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse”. In: *Presented as part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13)*. 2013, pp. 195–210.
- [TKT21] Mike Thelwall, Kayvan Kousha, and Saheeda Thelwall. “Covid-19 vaccine hesitancy on English-language Twitter”. In: *Profesional de la información (EPI)* 30.2 (2021).
- [TP10] Yla R Tausczik and James W Pennebaker. “The psychological meaning of words: LIWC and computerized text analysis methods”. In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.
- [Tsa+17] Avraam Tsantekidis et al. “Using deep learning to detect price change indications in financial markets”. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 2511–2515.
- [TT08] Paul J Taylor and Sally Thomas. “Linguistic style matching and negotiation outcome”. In: *Negotiation and Conflict Management Research* 1.3 (2008), pp. 263–281.
- [UC22] Apoorva Upadhyaya and Joydeep Chandra. “Spotting flares: The vital signs of the viral spread of tweets made during communal incidents”. In: *ACM Transactions on the Web* 16.4 (2022), pp. 1–28.

- [Vaz03] Alexei Vazquez. “Growing networks with local rules: preferential attachment, clustering hierarchy and degree correlations”. In: *PHYS.REV.E* 67 (2003), p. 056104.
- [VD01] Peter C Verhoef and Bas Donkers. “Predicting customer potential value an application in the insurance industry”. In: *Decision support systems* 32.2 (2001), pp. 189–199.
- [VDE17] Manuel R Vargas, Beatriz SLP De Lima, and Alexandre G Evsukoff. “Deep learning for stock market prediction from financial news articles”. In: *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE. 2017, pp. 60–65.
- [Vel+17] Petar Veličković et al. “Graph attention networks”. In: *arXiv preprint arXiv:1710.10903* (2017).
- [VGB11] Stefania Vitali, James Glattfelder, and Stefano Battiston. “The Network of Global Corporate Control”. In: *PloS one* 6 (Oct. 2011), e25995. DOI: 10.1371/journal.pone.0025995.
- [Vig+09] Ymir Vigfusson et al. “GO: Platform Support For Gossip Applications”. In: *P2P* (2009).
- [VL18] Nguyen Vo and Kyumin Lee. “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*. 2018, pp. 275–284.
- [VMR17] Soroush Vosoughi, Mostafa ‘Neo’ Mohsenvand, and Deb Roy. “Rumor gauge: Predicting the veracity of rumors on Twitter”. In: *ACM transactions on knowledge discovery from data (TKDD)* 11.4 (2017), pp. 1–36.
- [Vol+17] Svitlana Volkova et al. “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*. 2017, pp. 647–653.
- [VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online”. In: *Science* 359.6380 (2018), pp. 1146–1151.
- [VV20] Deepika Varshney and Dinesh Kumar Vishwakarma. “A review on rumour prediction and veracity assessment in online social network”. In: *Expert Systems with Applications* (2020), p. 114208.
- [VVM18] Siddhi Velankar, Sakshi Valecha, and Shreya Maji. “Bitcoin price prediction using machine learning”. In: *2018 20th International Conference on Advanced Communication Technology (ICACT)*. IEEE. 2018, pp. 144–147.
- [Wah+22] Jan Philip Wahle et al. “Testing the generalization of neural language models for COVID-19 misinformation detection”. In: *Infor-*

mation for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I. Springer. 2022, pp. 381–392.

- [Wal01] Steven Walczak. “An empirical analysis of data requirements for financial forecasting with neural networks”. In: *Journal of management information systems* 17.4 (2001), pp. 203–222.
- [Wan+18] Yaqing Wang et al. “Eann: Event adversarial neural networks for multi-modal fake news detection”. In: *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 2018, pp. 849–857.
- [Wan+19] Chenxu Wang et al. “Modeling rumor propagation and mitigation across multiple social networks”. In: *Physica A: Statistical Mechanics and its Applications* 535 (2019), p. 122240.
- [Wan+20] Youze Wang et al. “Fake news detection via knowledge-driven multimodal graph convolutional networks”. In: *Proceedings of the 2020 international conference on multimedia retrieval*. 2020, pp. 540–547.
- [Wan+22a] Shoujin Wang et al. “Veracity-Aware and Event-Driven Personalized News Recommendation for Fake News Mitigation”. In: *Proceedings of the ACM Web Conference 2022*. WWW ’22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 3673–3684. ISBN: 9781450390965. DOI: 10.1145/3485447.3512263. URL: <https://doi.org/10.1145/3485447.3512263>.
- [Wan+22b] Shoujin Wang et al. “Veracity-aware and event-driven personalized news recommendation for fake news mitigation”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 3673–3684.
- [Wat+21] James Waters et al. “Exploring the sentiment of entrepreneurs on Twitter”. In: *Plos one* 16.7 (2021), e0254337.
- [WDA10] Christian von der Weth, Anwitaman Datta, and Sally Ang. “COBS: A tool for collaborative browsing and search on the web”. In: *ICME*. 2010.
- [WDN02] Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman. “Identity and search in social networks”. In: *Science* 296 (2002), pp. 1302–1305.
- [Web+19] Mark Weber et al. “Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics”. In: *arXiv preprint arXiv:1908.02591* (2019).
- [WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994. URL: <http://scholar.google.com/scholar.bib?q=info:>

gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0.

- [WH15] Ziyi Wang and Jingti Han. “Visualization of the UK Stock Market Based on Complex Networks for Company’s Revenue Forecast”. In: *Information and Knowledge Management in Complex Systems*. Ed. by Kecheng Liu et al. Cham: Springer International Publishing, 2015, pp. 186–194. ISBN: 978-3-319-16274-4.
- [Wil+16] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.
- [WLL21] Wei Wu, Hanjia Lyu, and Jiebo Luo. “Characterizing discourse about COVID-19 vaccines: A Reddit version of the pandemic story”. In: *Health Data Science 2021* (2021).
- [Woo+14] Gavin Wood et al. “Ethereum: A secure decentralised generalised transaction ledger”. In: *Ethereum project yellow paper* 151 (2014), pp. 1–32.
- [WS21] Claire Wardle and Eric Singerman. “Too little, too late: social media companies’ failure to tackle vaccine misinformation poses a real threat”. In: *bmj* 372 (2021).
- [WS98] D. Watts and S. Strogatz. “Collective dynamics of small-world networks”. In: *Nature*, vol. 393. 1998.
- [WT07] Ken Wakita and Toshiyuki Tsurumi. “Finding Community Structure in Mega-scale Social Networks: [Extended Abstract]”. In: *Proceedings of the 16th International Conference on World Wide Web. WWW ’07*. Banff, Alberta, Canada: ACM, 2007, pp. 1275–1276. ISBN: 978-1-59593-654-7. DOI: 10 . 1145 / 1242572 . 1242805. URL: <http://doi.acm.org/10.1145/1242572.1242805>.
- [Wu+20] Zonghan Wu et al. “A comprehensive survey on graph neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [WXM19] Penghui Wei, Nan Xu, and Wenji Mao. “Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity”. In: *arXiv preprint arXiv:1909.08211* (2019).
- [WYN21] Kun Wu, Xu Yuan, and Yue Ning. “Incorporating relational knowledge in explainable fake news detection”. In: *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part III*. Springer. 2021, pp. 403–415.
- [WZG18] Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. “Sybilblind: Detecting fake users in online social networks without manual labels”. In: *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer. 2018, pp. 228–249.

- [XDZ22] Xiaofei Xu, Ke Deng, and Xiuzhen Zhang. “Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 2022, pp. 1206–1214.
- [Xia+12] Xin Xia et al. “Information credibility on twitter in emergency situation”. In: *Pacific-Asia Workshop on Intelligence and Security Informatics*. Springer. 2012, pp. 45–59.
- [XKS13] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. “Overlapping community detection in networks: The state-of-the-art and comparative study.” In: *ACM Comput. Surv.* 45.4 (2013), p. 43. URL: <http://dblp.uni-trier.de/db/journals/csur/csur45.html#XieKS13>.
- [XLC15] Jennifer Jie Xu, Eric Lu, and Michael Chau. “The Effects of Lender-Borrower Communication on P2P Lending Outcomes”. In: *ICIS*. 2015.
- [Xu+16] Zhenghua Xu et al. “Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2016, pp. 1921–1924.
- [Xu+19] Kuai Xu et al. “Detecting fake news over online social media via domain reputations and content understanding”. In: *Tsinghua Science and Technology* 25.1 (2019), pp. 20–27.
- [Yan+11] Shuang-Hong Yang et al. “Like Like Alike: Joint Friendship and Interest Propagation in Social Networks”. In: *Proceedings of the 20th International Conference on World Wide Web*. WWW ’11. Hyderabad, India: ACM, 2011, pp. 537–546. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963481. URL: <http://doi.acm.org/10.1145/1963405.1963481>.
- [Yan+15] Zhifan Yang et al. “Emerging rumor identification for social media with hot topic detection”. In: *2015 12th Web Information System and Application Conference (WISA)*. IEEE. 2015, pp. 53–58.
- [Yan+21] Qi Yang et al. “Online Communication Shifts in the Midst of the Covid-19 Pandemic: A Case Study on Snapchat”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 830–840.
- [Yaq20] Ussama Yaqub. “Tweeting During the Covid-19 Pandemic: Sentiment Analysis of Twitter Messages by President Trump”. In: *Digital Government: Research and Practice* 2.1 (2020), pp. 1–7.
- [YJB02] Soon-Hyung Yook, Hawoong Jeong, and Albert-László Barabási. “Modeling the Internet’s large-scale topology”. In: *Proceedings of the National Academy of Sciences* 99.21 (Oct. 2002), pp. 13382–13386.

- [YL11] Jaewon Yang and Jure Leskovec. “Patterns of temporal variation in online media”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 177–186.
- [YL12] Jaewon Yang and J. Leskovec. “Community-Affiliation Graph Model for Overlapping Network Community Detection”. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. Dec. 2012, pp. 1170–1175. DOI: 10.1109/ICDM.2012.139.
- [YML14] Jaewon Yang, Julian J. McAuley, and Jure Leskovec. “Community Detection in Networks with Node Attributes”. In: *CoRR* abs/1401.7267 (2014). URL: <http://arxiv.org/abs/1401.7267>.
- [You+19] Di You et al. “Attributed multi-relational attention network for fact-checking url recommendation”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 1471–1480.
- [YSC19] Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. “Examining emergent communities and social bots within the polarized online vaccination debate in Twitter”. In: *Social Media+ Society* 5.3 (2019), p. 2056305119865465.
- [Yua+19] Chunyuan Yuan et al. “Jointly embedding the local and global relations of heterogeneous graph for rumor detection”. In: *arXiv preprint arXiv:1909.04465* (2019).
- [YYZ17] Bing Yu, Haoteng Yin, and Zhanxing Zhu. “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting”. In: *arXiv preprint arXiv:1709.04875* (2017).
- [Zah+20] Manzil Zaheer et al. “Big Bird: Transformers for Longer Sequences.” In: *NeurIPS*. 2020.
- [Zar+20] Koosha Zarei et al. “A first instagram dataset on covid-19”. In: *arXiv preprint arXiv:2004.12226* (2020).
- [ZG05] Alice X. Zheng and Anna Goldenberg. *A Generative Model for Dynamic Contextual Friendship Networks*. Tech. rep. Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA, July 2005. URL: http://research.microsoft.com/%5C~%7B%7Dalicez/papers/dcfm%5C_tr.pdf.
- [Zha+19] Ling Zhao et al. “T-gcn: A temporal graph convolutional network for traffic prediction”. In: *IEEE Transactions on Intelligent Transportation Systems* (2019).
- [Zho+20] Xinyi Zhou et al. “Fake news early detection: A theory-driven model”. In: *Digital Threats: Research and Practice* 1.2 (2020), pp. 1–25.
- [ZHS09] Peixiang Zhao, Jiawei Han, and Yizhou Sun. “P-Rank: a comprehensive structural similarity measure over information networks”. In: *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. 2009.

- [ZLP16] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. “Learning reporting dynamics during breaking news for rumour detection in social media”. In: *arXiv preprint arXiv:1610.07363* (2016).
- [Zub+18] Arkaitz Zubiaga et al. “Detection and resolution of rumours in social media: A survey”. In: *ACM Computing Surveys (CSUR)* 51.2 (2018), pp. 1–36.
- [ZZS13] Ke Zhou, Hongyuan Zha, and Le Song. “Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes”. In: *Artificial Intelligence and Statistics*. PMLR. 2013, pp. 641–649.

Appendix A. MISLEADING SOCIAL MEDIA CORPUS ANALYSIS

A.1. Language Exploration

Here, we uncover ten distinct Syntactic styles of Misleading and Non-Misleading tweets.

A.1.1. Uncover Syntax

After assigning the labels for each tweet as discussed in Section 5.1.2, we study ten Syntactic aspects (attributes) which were found to be more significant than others in our dataset to distinguish the structural patterns of both types of tweets. We used the NLTK library¹ to extract the part-of-speech tags. First, we visualize the Syntactic distributions of both Misleading and Non-Misleading tweets. Next, to validate that the difference in both the distributions is indeed significant, we use Kolmogorov Smirnov Test².

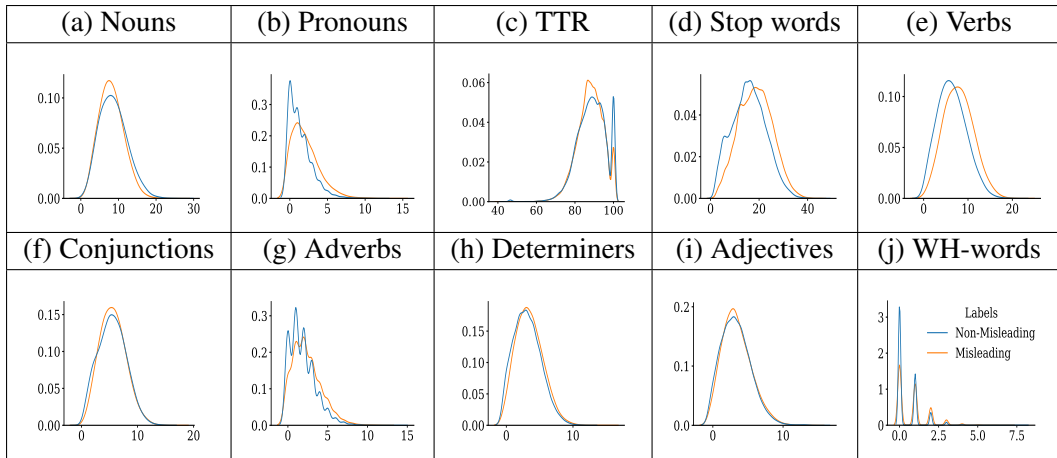


Figure 12: Syntactic analysis of the Misleading and Non-Misleading tweets. The x-axis represents the counts of the specific Syntactic attribute (for example, noun) and the y-axis represents the density of the distribution. TTR is an abbreviation for Type-Token Ratio. **Orange** and **blue** colors represent the Misleading and Non-Misleading labels, respectively.

First, we look at the **Nouns**, the main building blocks of any sentence. We observe from Figure 12(a) that visually there is a slight variation in both distributions. To determine whether this difference in the distributions is statistically significant, we calculated the p-value of Nouns (see Table 36, row 1), which is much lower than the significance level (5% as default value), which implies that

¹<https://www.nltk.org/book/ch05.html>

²<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>

the two distributions are in fact dissimilar. Second, **Pronouns** are the substitute for Nouns. Figure 12(b) shows that Pronouns are more used in Non-Misleading tweets than Misleading tweets. Third, **Type-Token Ratio (TTR)** measures the lexical diversity (quality) of the text. Specifically, it is the ratio between the total number of unique words (types) in the text and the total number of words in the text. The higher the value of this ratio, the higher the lexical diversity of the text.

$$TTR = \frac{\# of Types}{\# of Tokens} * 100 \quad (A.1)$$

We notice in Figure 12(c) that the distributions are right-skewed, indicating that the text is of good quality in terms of lexical diversity. However, we observe some differences between the distributions. When TTR is near to 100, the density of the Misleading distribution is lower in comparison to its mean. Whereas, Non-Misleading distribution has a similar density with respect to its mean. This implies that Misleading tweets are less lexically diverse in contrast to Non-Misleading tweets. In addition, the p-value present in Table 36, row 3 is lower, indicating the distributions are different.

Table 36: P-values of the Kolmogorov Smirnov Test for each of the syntactic characteristics extracted from the dataset.

Syntactic Attributes	P-values
Nouns	1.12e-109
Pronouns	4.29e-78
TTR	2.09e-96
Stop words	1.10e-44
Verbs	7.81e-50
Conjunctions	5.89e-57
Adverbs	1.84e-25
Determiners	0.0
Adjectives	0.0
WH-words	0.0

We investigate other Syntactic aspects namely, **Stop words** (Figure 12(d)), **Verbs** (Figure 12(e)), **Conjunctions** (Figure 12(f)), **Adverbs** (Figure 12(g)), **Determiners** (Figure 12(h)), **Adjectives** (Figure 12(i)), and **WH-words** (Figure 12(j)). Their p-values present in the Table 36. Although the values for **Determiners**, **Adjectives** and **WH-words** are near but less than the significance level, we consider them distinguishable attributes in finding Misleading and Non-Misleading tweets.

A.2. What are the Principal Topics of Discussion?

Next, we performed topic modeling using the LDA approach [Xu+19], we inspect the top five most talked-about topics among Misleading and Non-Misleading

tweets (see Table 37). We employed grid-search to identify the optimal number of topics, determining that to be five for both Misleading and Non-Misleading tweets for our data collection.

Following are the key themes of Misleading tweets -

1) *Politics*: The frequent targets of these tweets are politics; for example, the tweet - **“I don’t even trust this Govt to take a vaccine they are desperate to sell us This makes me feel sad But now if cellulitis is also a side effect of it And blood clots And Morrison brushes that aside and or lies to us about it What ?”**. Here, the key point of discussion is not to believe the government, and thus misleads the readers by bringing political angle into the vaccination debate and creating this type of vaccine prejudice.

2) *Myths and Side-Effects*: The false stories and myths about vaccines have the greatest effect on people’s minds; for instance, this tweet - **“IF you are allergic to eggs and chicken, you are not going to receive a dose of the 1.02 million doses of Oxford-Astrazeneca vaccine that is expected in Kenya on Tuesday 2 March, 2021.”** is attempting to steer readers away from the true story³, namely the delivery of said volume of vaccines under the COVAX initiative.

3) *Vaccine Efficacy*: There is a lot of confusion related to the COVID-19 vaccine efficacy in the discussions, for instance, - **“How can we trust the vaccine when the efficacy is not reported precisely??”**.

4) *Trump’s Role*: Trump is well-known for being heavily involved in many false reports⁴. This tweet confirms Trump’s involvement in the tweet - **“Why did the whitehouse turn down Pfizer offer of the vaccine as the 1st to receive it? Seems Mr. Trump joins hands with Operation Warp Speed and they are deliberately slowing the speed of vaccine rollout.”**

5) *Vaccines Choice*: There is skepticism about choosing the vaccines. For instance, this tweet tells one of the myths about altering DNA for the Johnson & Johnson vaccine - **“Do not get the Johnson & Johnson version. The MOA is different than that of Pfizer’s or Moderna’s vaccine. Picture a tennis ball inside a basketball. J&J’s enters the nucleus (tennis ball) that can permanently alter DNA.”**. The key themes of Non-Misleading tweets are:

1) *Operation Warp Speed*: The most discussed topic - Operation Warp Speed, initiated by the US government to facilitate the development and distribution of COVID vaccines. An example of a tweet with this theme - **“[username] Pfizer vaccine was self funded. Nothing to do with warp speed, the government or the lying ex president trump. TRUMP IS A LIAR”**.

2) *Shots & Real Side-Effects*: These two themes are about informing an individual’s actual experience after getting the vaccine shot. We provide an example of a tweet that is related to both the themes - **“I had the PFIZER shot(s), 30 days apart, and both times it didn’t hurt. My arm was sore later on for a day or**

³<https://www.unicef.org/kenya/press-releases/over-1-million-covid-19-vaccine-doses-arrive-nairobi-via-covax-facility>

⁴<https://www.cnn.com/2021/01/13/trump-tweets-legacy-of-lies-misinformation-distrust.html>

Table 37: Top 5 most discussed topics in Misleading and Non-Misleading COVID-19 vaccination tweets.

Misleading	Non-Misleading
Politics	Operation Warp Speed
Myths & Side-Effects	Shots
Vaccine Efficacy	Vaccine Efficacy
Role of Trump	Real Side-Effects
Vaccine Choices	Data & Facts

two. 24 hours after the second shot I got a fever and chills and promptly went to sleep. Woke up 10 hours later feeling great. No other side effects. GET THIS VACCINE!"

3) *Vaccine Efficacy*: Both Misleading and Non-Misleading have a shared theme. This indicates that they are both discussing the vaccine's efficacy. Misleading tweets, on the other hand, aim to cause uncertainty regarding vaccine efficacy, while Non-Misleading tweets emphasize the positive aspects of vaccine efficacy, such as - **"#JabMe: A single shot of either the Pfizer or Oxford vaccine provides about 80 percent protection against being treated in a hospital, according to the latest data from the UK vaccination program."**

4) *Data & Facts*: Non-Misleading tweets are more concerned with presenting accurate information, such as - **"BBC: Around 5 million Europeans have already received the AstraZeneca vaccine. Of this figure, about 30 cases had reported "thromboembolic events" - or developing blood clots. European medicines regulator said there was no indication the jab was causing the blood clots."**

The top five themes indicate the different subjects explored in both types of tweets. Precisely, Misleading tweets mostly mislead the reader using political dimension or raise fear among the people for vaccination. In comparison, Non-Misleading tweets discuss the real side-effects of the vaccination and try to bring the facts with evidence.

A.2.1. Opinion Study

Previously, we explored the Syntactic dimension. We now look into the second dimension, the role of opinion in relation to Misleading and Non-Misleading tweets.

Sentiment Matters. We first explore the impact of sentiments on Misleading and Non-Misleading tweets, considering three broad categories of sentiments: Positive, Negative, and Neutral. The sentiments are calculated using VADER API [Yaq20; Wat+21]. To detect a tweet as Positive, Negative, or Neutral, we utilized the compound score. The compound score is the sum of the valence scores of each word, yielding a range of [-1, 1]. A score of -1 indicates a strong Negative sentiment, whereas a score of +1 suggests a strong Positive sentiment. We utilized a 0.05 threshold value. The selection of the 0.05 threshold has been adopted from the previous literature [BJ19; CB21]. Additionally, to validate the choice

of threshold, we randomly sampled 150 tweets with a balanced combination of Positive, Negative, and Neutral sentiment labels obtained from the VADER API (50 random tweets from each label). Following that, we manually annotated the sentiments for 150 tweets and got an accuracy of 0.96, confirming the efficacy of the selected threshold. Thus, a tweet is considered to have a Positive sentiment if the compound score is greater than or equal to 0.05, neutral if the score is between 0.05 and -0.05; Negative otherwise.

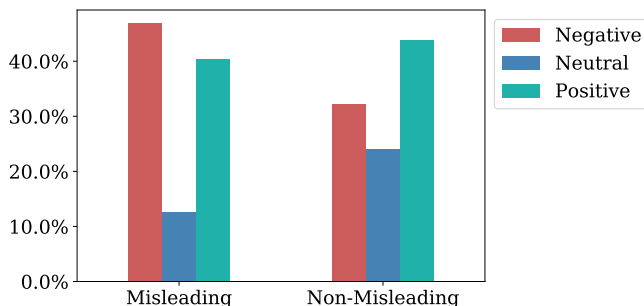


Figure 13: Sentiment analysis with respect to Misleading and Non-Misleading tweets. The x-axis and y-axis denote the labels of the tweets and percentages, respectively. **Negative**, **Neutral**, and **Positive** are different categories of the sentiments.

Figure 13 shows that Negative sentiments are more prevalent in Misleading tweets followed by Positive and Neutral sentiments. An example of a Misleading tweet with Positive sentiment is **“A little Angel in my dreams today told that our bodies will be developing antibodies on its own within a few days without vaccination”**. While a reading of it indicates vaccine prejudice and skepticism, the sentiment analysis tool latches upon the positive sounding phrases in there.

We then identify the topics which are being discussed with respect to sentiments. Apart from Vaccine Efficacy which confirms the above mentioned tweet’s theme, the related topics Operation Warp Speed and Trials are also identified within the Positive sentiments of Misleading tweets (See Table 38). We surmise that Misleading tweets with Positive sentiments inject the negativity by sugar-coating the tweets with positive words to easily trick people into either believing in their positive hypothetical situations or providing a new dimension to the topic.

Positive sentiments, on the other hand dominate in Non-Misleading tweets, though a substantial portion of them again have Negative or Neutral sentiments. An instance of a Non-Misleading tweet with Negative sentiment is **“Dr Kathrin Jansen, Pfizer’s head of vaccine development: We were never part of the Warp Speed ... We have never taken any money from the U.S. government, or from anyone. Trump is a liar”**. In this instance, the Negative sentiment of the Non-Misleading tweet is due to it counteracting the Misleading information. Many facts and news related to the pandemic have naturally Negative sentiments. Similar to this tweet argument, topics that are discovered in the Negative senti-

ments of Non-Misleading tweets are Operation Warp Speed and Vaccine Efficacy, in addition to, Trials and Data & Facts (See Table 38). These Non-Misleading tweets with Negative sentiments indicate that they are either attempting to clarify claims against COVID Vaccination’s Development Companies or myths against the vaccination process with their choice of negative words.

Table 38: Topic Modeling with respect to each sentiment. M and NM represent Misleading and Non-Misleading. OWS and VaEf denote Operation Warp Speed and Vaccine Efficacy, respectively.

	Positive	Negative	Neutral
M	Trials, OWS, VaEf	VaEf, OWS, Trials, Myths	VaEf, OWS, Trials
NM	Trump, Real side-effects, VaEf	Data & Facts, OWS, VaEf, Trials	Data & Facts, VaEf, Real side-effects

Overall, Negative sentiments are more common in Misleading tweets, whereas, Non-Misleading tweets have more Positive sentiments. We go through five emotions in detail in the next Section.

A.2.2. Intense Emotions

The sentiments serve as the foundation for analyzing the tweets. As a result, we dig deeper into the impact of emotions on tweets.

Table 39: Topic modeling with respect to each five emotions. M and NM represent Misleading and Non-Misleading, respectively. OWS and VaEf are abbreviations for Operation Warp Speed and Vaccine Efficacy, respectively.

	Fear	Surprise	Sadness	Anger	Happiness
M	Trials, OWS, VaEf	Trials, Politics, Trump, Myths	VaEf, OWS, Politics	VaEf, OWS, Availability, Trials	VaEf, approval
NM	Trump, Politics, VaEf	Shots, Data & Facts, OWS, VaEf	Data & Facts, VaEf, Politics	VaEf, Data & Facts, Real side-effects, Availability	VaEf, Real side-effects

Figure 14 displays the five different emotions - Anger, Fear, Happiness, Sadness, and Surprise retrieved by employing the text2emotion API [DB20]. In this API, for each tweet, we get a dictionary where keys denote emotion categories, and values indicate its score. Higher the score, the higher the emotion in that tweet. We classify the tweets as per the emotion that has the highest score. In the Figure, the value with respect to each emotion axis corresponds to the percentage of a particular emotion in a (*Non-*)*Misleading* class.

In Misleading tweets, the most common emotion is Fear, followed by Surprise, Sadness, Happiness, and, finally, Anger. Whereas, Non-Misleading tweets have a tie for the first place with Fear and Surprise, followed by Happiness, Sadness, and Anger.

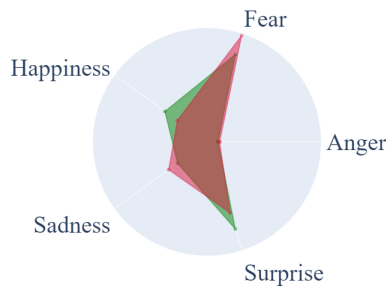


Figure 14: Emotion analysis of the Misleading and Non-Misleading tweets. Each axis denotes the emotion. The five emotions depicted are Happiness, Fear, Anger, Surprise, and Sadness. Non-Misleading tweets are represented in **green** color, and Misleading in **red** color (best seen in color).

Fear and Surprise are the two most popular emotions across both types of tweets. It is understandable given that 45% of unvaccinated people are afraid to get the vaccine because they are worried about the adverse side-effects⁵. To confirm this, we studied the topics around Fear and Surprise emotions. The topics which are similar to the above statement are Trials and Vaccine Efficacy in both Misleading and Non-Misleading categories (See Table 39). However, the emotion Fear is higher in Misleading tweets in contrast to Non-Misleading tweets, which is attributable to the observation that most Misleading tweets reference fake and fabricated vaccine side-effects and false narratives of Operation Warp Speed (Table 39). In contrast, the emotion Surprise is higher in Non-Misleading tweets than Misleading tweets. These often discuss the governments' fast response towards vaccination, fitting into the Data & Facts topic.

Furthermore, emotions of Anger and Sadness are higher in Misleading tweets. One of the possible reasons could be that these tweets often involve a political dimension and accuse the government of not making the right decisions. The topics covering under both emotions are Politics, Vaccine Availability, and Operation Warp Speed in the Misleading category. Non-Misleading tweets that have emotion Happiness discuss their experience about receiving the shot and facing no bogus side-effects spreading across the Internet. A related topic is Real side-effects in the Non-Misleading category. To summarize, the majority of the Misleading tweets have Fear emotions more than Non-Misleading tweets.

A.3. The Influence of Visibility

So far, our analysis was confined to the content of the tweets themselves. The focus of the third dimension is looking at information and meta-data in the tweets

⁵<https://www.vox.com/recode/22330018/covid-vaccine-hesitancy-misinformation-carnegie-mellon-facebook-survey>

#vaccinatedandproud represents that tweet is in support of the vaccination process. Thus, the choice of the hashtags can provide a clue about the Misleading tweets.

Co-hashtags: We also consider the combination of hashtags that frequently occurred together in a tweet, i.e., co-hashtags. In Non-Misleading tweets, we find 280 co-hashtags, while 86 co-hashtags are found in Misleading tweets. After filtering those co-hashtags that occurred more than once, we found that co-hashtags repeatedly occurred only in Non-Misleading tweets. There is no pattern (consistency) concerning co-hashtags in Misleading tweets, making their hashtags more random.

Table 40: Use of 10 different Hashtags (left-side represents the hashtags that are mostly present in Non-Misleading tweets but not in Misleading tweets, and vice-versa on the right).

Non-Misleading	Misleading
fullyvaccinated	saynotopoisonvaccines
savetheplanet	vaccineextortion
healthnews	pseudoscience
thisismyshot	trumpvirusdeath240k
covid19updates	untestedvaccine
scienceisreal	iwillnotgetvaccinated
publichealth	billgatesisevil
inthenews	abolishbigpharma
vaccinatedandproud	novaccine4me
2ndshot	astrazenecapoison

A.3.3. As you Like it

The number of Retweets, Replies, and Likes count are all essential visibility attributes. Figure 16 depicts the median (mean) values of the counts of Retweets, Replies, and Likes for both classes. In our case, the median and mean values are the same. On average, Replies to the tweets remain the same regardless of the type of information it contains. However, there is a variation in the Retweets count and Likes count. Relatively, the Misleading tweets get fewer Retweets and Likes than Non-Misleading tweets.

A.3.4. What's in a Name?

The names of the COVID-19 vaccinations are frequently mentioned in tweets. In this regard, we attempt to assess the influence of vaccine names in both Misleading and Non-Misleading tweets. In our collected data, we discovered that the majority of the discussions revolve around the five vaccinations: Pfizer, Moderna, AstraZeneca, Covaxin, and Johnson & Johnson. These names are used either in an individual or combined manner.

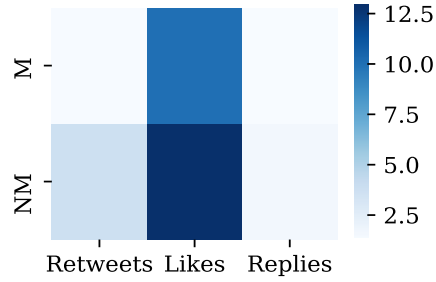


Figure 16: Aggregation of Visibility Counts. Each cell represents the median values of the Retweets, Replies, and Likes with respect to the labels of tweets. NM and M denote the Non-Misleading and Misleading tweets, respectively.

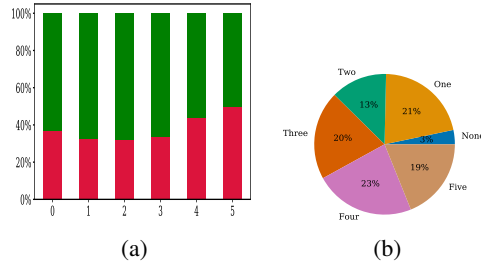


Figure 17: Count of Vaccine Names used in a Tweet. X-axis and y-axis in (a) represents the count of the vaccines' names and percentages, respectively. The green and red colors indicate the Non-misleading and Misleading classes. Value 0 on the x-axis corresponds to no mention of the vaccine name in the tweet, Value 1 denotes mention of one vaccine, and so on. (b) shows the percentage of the tweets with respect to the count of the vaccine name (best seen in color).

Figure 17 shows that the proportion of Misleading tweets is lower than the proportion of Non-Misleading tweets until the number of vaccine names is fewer than or equal to three. When the count reaches four or five, the number of Misleading tweets begins to rise, i.e., when the number of vaccine names in a tweet grows to more than three, the likelihood of a Misleading tweet also grows.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Rajesh Sharma for his invaluable, constant support and feedback throughout my Ph.D. Also, I would like to thank Dr. Anwitaman Datta for their guidance.

I would like to extend my appreciation to my family, friends, and teachers for their constant support throughout the journey.

SISUKOKKUVÕTE

Väärteabe vastu võitlemine digitaalajastul: väärteabe iseloomustamise, tuvastamise ja leevendamise põhjalik strateegia veebipõhistel sotsiaalmeediaplatvormidel

Veebipõhiste sotsiaalmeediaplatvormide, nagu Twitter ja Facebook, esilekerkimine on hõlbustanud valeteabe ülemaailmset levitamist, soodustades sotsiaalse hirmu, ärevuse ja majandusliku kahju kasvu. Lõputöö uurib mitmekülgset lähene misviisi desinformatsiooniga võitlemiseks digiajastul, keskendudes kolmele põhidimensioonile: valeinformatsiooni sisu tuvastamine, raamistiku väljatöötamine valeinformatsiooni levitajate tuvastamiseks, ja tõhusate desinformatsioonivastaste meetmete rakendamine.

Töö esimeseks mõõtmeks on kirjeldavate ja ennustavate ülesannete sulandumine, milleks on Interneti-sotsiaalmeedia sisu omaduste (tunnuste) mõistmine, seega kasutajad seoses desinformatsiooniga, antud juhul eriti kuulujuttudega. Kuulujutt on infokild, mille paikapidavus on postitamise hetkel kahtluse all. See töö on inspireeritud ideest, et selliste kasutajate kirjutamisstiilide mõistmine võib aidata mõista kognitiivset käitumist või tuvastada valeinfo postitamise peamised motiivid. Avastasime olulise erinevuse kuulujuttude ja mittekuulujuttude psühholingvistika allikatunnuste, aga ka vastusefunktsioonide vahel. Avastasime, et kuulujuttude allika säutsud kasutasid rohkem minevikuga seotud sõnu, eessõnu ja sisaldasid tasu, riski ja võimuga seotud tõuke (motivatsiooni). Samamoodi olid kuulujuttude allikatest mittekuuluvatel säutsidel rohkem alatuid jooni, nagu afektiivsed protsessid, kognitiivsed protsessid (mõistmine, põhjuslik seos, lahknevus, kindlus, eristamine), olevikuga seotud sõnad, mitteametlik keel ning neil oli seoseid ja saavutusi. Samuti uurisime nende funktsioonide tõhusust kuulujuttude ennustamisel. Täpsemalt avastasime, et ansamblipõhine Random Foresti mudel ületas kõigi sündmuste puhul teisi kasutatud mudeleid. Kuna masinõppemudelid on oma olemuselt mustad, kasutasime muudest funktsioonidest olulisemate funktsioonide otsimiseks tööriista SHAP AI seletatavus.

Töö teine mõõde on olemuselt ennustav, kus töötame välja raamistiku, mis võimaldab tuvastada ja tõkestada väärinfo levikut. Eelkõige pakume välja kasutajataseme kuulujuttude tuvastamise, st kasutaja klassifitseerimise võimalikuks kuulujuttude levitajaks või mitte. Varasemad meetodid kahtlaste või pahatahtlike kasutajate tuvastamiseks ja desinformatsiooni tuvastamiseks platvormidel nagu Twitter ei ole kasutajataseme tuvastamist piisavalt arvesse võtnud. Kasutaja liigitamine valeinfo levitajaks ühe postituse põhjal on ebapiisav. Selle vastu võitlemiseks luuakse süsteem automaatseks valeinformatsiooni tuvastamiseks, mis integreerub võrguandmed ja paljud postitused, et saada kõrge F1-skoor 0,864, mis ületab varasemaid baasmudeleid. Eelkõige kasutati PHEME andmestikku võimalike kuulujuttude levitajate tuvastamiseks, kasutades nõrka juhendatud õppimisviisi. Modelleerisime selle probleemi binaarse klassifitseerimise ülesandena, rakenda-

des erinevaid masinõppemudeleid muudetud kuulujutu levitajate andmekogumile. Meie tulemused näitavad, et GCN (võrreldes algmodelitega) suudab paremini toimida (tõsta punaseid lippe võimalike kuulujuttude levitajate jaoks), kasutades ära võimalike kuulujuttude levitajate suhteid, kes võiksid hästi seguneda mittekuulujuttude levitajatega.

Kolmas töö on oma olemuselt kirjeldav, kus soovitame artikleid/postitusi kasutajatele, kes levitavad väärinfot. Sotsiaalmeedia platvormidel valeinformatsiooni leviku vähendamise varasemate strateegiate puuduste hulka kuulub välise modereerimise puudumine. Selle probleemi lahendamiseks soovitatakse tegevuskava, mis ühendab sotsiaalmeedia andmed kureeritud faktidega kontrollitud andmehooldatega, et võtta kasutusele automaatne valeinformatsiooni paljastamine. See strateegia keskendub Twitteri platvormile ja COVID-19 desinformatsioonile, uurides kahte täiendavat strateegiat, mis võivad tuvastada postitusi, mis levitavad valeinformatsiooni nii protsessi alguses kui ka hilises etapis, edestades varasemaid strateegiaid, mis võiksid olla suunatud vaid ühele sammule. See kasutab kahte avalikult kättesaadavat andmekogumit, st. FaCov (faktidega kontrollitud artiklid) ja eksitavad (sotsiaalmeedia Twitter) andmed COVID-19 vaktsineerimise kohta. Esimene tuvastab ja taaskasutab teiste kasutajate olemasolevad seotud, kuid faktilised/mitteeksitavad säutsud kui soovitatavad vastusäutsid. See on põhiline lähenemisviis, mida saab kasutada isegi siis, kui faktide kontrollimise veebisaidid ei ole veel teatud valeinformatsiooni kontrollinud või kui sellised artiklid on tuvasutatud ja sobib meie teise lähenemisviisiga. Teine soovitab faktikontrolli artikleid (olenevalt saadavusest) erinevatelt faktikontrolli saitidelt. See annab kasutajatele usaldusväärsetest allikatest pärit kontrollitud, usaldusväärset ja põhjalikumat teavet, et desinformatsiooni tõrjuda. Kasutades mõlemat automatiseerimismeetodit koos, loome skaleeritava ümberlükkamise torustiku, et võidelda valeinformatsiooniga sotsiaalmeedias.

CURRICULUM VITAE

Personal data

Name: Shakshi Sharma
Email: shakshi.sharma@ut.ee
Date of Birth: 26-08-1991
Citizenship: Indian
Language: Hindi, and English

Education

2020–Present **Ph.D.** at Institute of Computer Science, University of Tartu, Estonia. Advised by Dr. Rajesh Sharma.

2015–2017 **Master of Technology** in Computer Science and Engineering, Analytics at National Institute of Technology Delhi, India. Advised by Dr. Akanksha Juneja.
Thesis: Deep Learning in Computer Vision for Real-world Scene Recognition.

2008–2012 **Bachelor of Technology** in Computer Science Engineering, Jan Nayak Ch. Devi Lal Vidyapeeth, Sirsa, Haryana, India.

Employment and Research Visit

Sep, 2020–Present Junior Research Fellow, University of Tartu, Estonia.

Mar, 2023–Apr, 2023 Research Intern, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom.

May, 2022–Jul, 2022 Research Intern, Department of Computer Science and Engineering, International Institute of Information Technology (IIIT), Hyderabad, India.

Jul, 2018–Jun, 2020 Assistant Professor, Department of Computer Science and Engineering, The NorthCap University, Gurugram, Haryana, India.

Jun, 2017–Jun, 2018 Graduate Programmer, Fidelity International Company, Gurugram, Haryana, India.

Scientific work

Main fields of interest:

- Data Science
- Machine Learning
- Deep Learning

- Natural Language Processing
- Graph Neural Networks
- AI Ethics

Publications

- I **Shakshi Sharma**, Ekanshi Agrawal, Rajesh Sharma, Anwitaman Datta. "Fa-Cov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset". **Proceedings of the International AAAI Conference on Web and Social (ICWSM)**, 2022.
- II Sabur Butt, **Shakshi Sharma (Equal Contribution)**, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. "What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses". **Computers in Human Behavior**, 2022.
- III **Shakshi Sharma**, Rajesh Sharma, and Anwitaman Datta. "Misleading the Covid-19 vaccination discourse on Twitter: An exploratory study of infodemic around the pandemic." **IEEE Transactions on Computational Social Systems**, 2022.
- IV Mudit Dhawan, **Shakshi Sharma (Equal Contribution)**, Aditya Kadam, Rajesh Sharma, Ponnurangam Kumaraguru. "GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection." arXiv preprint arXiv:2202.12478, 2022.
- V Mohit Mayank, **Shakshi Sharma**, Rajesh Sharma. "DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection". **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, 2022.

- VI Raj Jagtap, Abhinav Kumar, Rahul Goel, **Shakshi Sharma**, Rajesh Sharma, Clint P George “Misinformation Detection on YouTube Using Video Captions.” arXiv preprint arXiv:2107.00941, 2021.
- VII **Shakshi Sharma**, and Rajesh Sharma. “Identifying possible rumor spreaders on twitter: A weak supervised learning approach.” **International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2021.

ELULOOKIRJELDUS

Isiklikud andmed

Nimi: Shakshi Sharma
E-post: shakshi.sharma@ut.ee
Sünniaeg: 26-08-1991
Kodakondsus: Indialane
Keelteoskus: Hindi, inglise

Haridus

2019–hetkel Ph.D. arvutiteaduses, Tartu Ülikool, Eesti. Nõustab dr Rajesh Sharma.
2015–2017 Tehnoloogia magister arvutiteaduses ja -tehnoloogias, analüütika Delhi Riiklikus Tehnoloogiainstituudis, India. Nõustaja dr Akanksha Juneja.
Lõputöö: Arvutinägemise süvaõpe reaalses maailmas Stseeni äratundmine.
2008–2012 Tehnoloogia bakalaureus arvutiteaduse inseneri erialal, Jan Nayak Ch. Devi Lal Vidyapeeth, Sirsa, Haryana, India.

Teenistuskäik

Sept, 2020–hetkel Nooremteadur, Tartu Ülikool, Eesti.
Märts 2023–Aprill 2023 Teaduspraktikant, arvutiteaduse osakond, UniSheffieldi ülikool, Sheffield, Ühendkuningriik.
Mai 2022–Juuli 2022 Rahvusvahelise Infotehnoloogia Instituudi (IIIT) arvutiteaduse ja -tehnika osakonna teadurpraktik, Hyderabad, India.
juuli 2018–juuni 2020 Arvutiteaduse osakonna dotsent ja tehnika, NorthCapi ülikool, Gurugram, Haryana, India.
Juuni 2017– Juuni 2018 Programmeerija, Fidelity International Company, Gurugram, Haryana, India.

Teadustegevus

Peamised uurimisvaldkonnad:

- Andmeteadus
- Masinõpe
- Sügav õppimine
- Loomuliku keele töötlemine
- Graafika närvivõrgud
- AI eetika

Väljaanded

- I **Shakshi Sharma**, Ekanshi Agrawal, Rajesh Sharma, Anwitaman Datta. "FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset". **Proceedings of the International AAAI Conference on Web and Social (ICWSM)**, 2022.
- II Sabur Butt, **Shakshi Sharma (Equal Contribution)**, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. "What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses". **Computers in Human Behavior**, 2022.
- III **Shakshi Sharma**, Rajesh Sharma, and Anwitaman Datta. "Misleading the Covid-19 vaccination discourse on Twitter: An exploratory study of infodemic around the pandemic." **IEEE Transactions on Computational Social Systems**, 2022.
- IV Mudit Dhawan, **Shakshi Sharma (Equal Contribution)**, Aditya Kadam, Rajesh Sharma, Ponnurangam Kumaraguru. "GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection." *arXiv preprint arXiv:2202.12478*, 2022.
- V Mohit Mayank, **Shakshi Sharma**, Rajesh Sharma. "DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection". **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, 2022.

- VI Raj Jagtap, Abhinav Kumar, Rahul Goel, **Shakshi Sharma**, Rajesh Sharma, Clint P George “Misinformation Detection on YouTube Using Video Captions.”arXiv preprint arXiv:2107.00941, 2021.
- VII **Shakshi Sharma**, and Rajesh Sharma. “Identifying possible rumor spreaders on twitter: A weak supervised learning approach.”**International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2021.

**DISSERTATIONES INFORMATICAЕ
PREVIOUSLY PUBLISHED IN
DISSERTATIONES MATHEMATICAE
UNIVERSITATIS TARTUENSIS**

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω -rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo.** Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Sor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

- 113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
- 114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
- 116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
- 121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
- 122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAЕ UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh.** Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas.** Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi.** Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich.** Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka.** Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinmaa.** Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto.** Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.