Saarland University

Faculty of Mathematics and Computer Science
Department of Computer Science
Master Thesis

Characterisation of Users' Behaviours towards Fake News through the Analysis of their Networks

submitted by
Mahmoud Fawzi
Saarbrücken
November 2022

Supervisor & Reviewer 1:
Prof. Dr. Vera Demberg
Computer Science and Computer Linguistics
Saarland Informatics Campus
Saarbrücken, Germany

Advisor & Reviewer 2:
Dr. Walid Magdy
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Acknowledgements

The completion of this work wouldn't have been possible without the following great people and entities:

- ***Dr. Walid Magdy***: I am very grateful to my advisor who trusted that I would be able to work on a project with such a complicated and interdisciplinary nature while it was expected to be undertaken by a postdoctoral researcher. His tips, ideas, and guidance were the main ingredients of the success of this work. This collaboration was the most enriching research experience I have ever had in my entire career so far. It was also a blessing on a personal level.

- ***Prof. Dr. Vera Demberg***: It is very hard to describe how much of a nice person Professor Vera is. I can't thank her enough for her belief in my thesis proposal and my ability to accomplish it, her smooth communication, her availability whenever needed, and her valuable comments and hints on my analysis. I honestly hope that many of the other professors I encountered learn to have the same encouraging spirit while dealing with young researchers.

- **FakeDet Team**: This team of brilliant researchers tackled the misinformation problem in many forms and gave me many relevant tips which allowed me to have very diverse knowledge about the topic. Special thanks to *Amr Keleg* and *Ibrahim Abufarha* from The University of Edinburgh and *Prof. Dr. Tamer Elsayed*, *Zein Sheikh Ali*, *Maram Hasanain*, and *Fatima Haouari* from Qatar University.

- **Al Jazeera Media Network**: I am proud to have worked on a project sponsored by Al Jazeera. It is such a pleasure to see an entity from an Arabic Islamic country investing in an interdisciplinary advanced research field like Computational Social Sciences. I wish them all the best with their organisational activities during the 2022 World Cup in Qatar.

Abstract

The detection and analysis of fake news and its origins has become a main task associated with the overall objective of social media regulation in recent years. The majority of work was towards detecting misinformation with some focus on analysing the flow of fake news over social networks. However, there is less attention on understanding the characteristics of social media users who consume these fake news. In this work, we investigate the possibility of predicting users' reactions towards fake news and defining some network characteristics for each users group. We utilised a set of fact-checking websites in the Arab world that report social media posts spreading fake news and the interactions with them. We defined three sets of users: 1) Spreaders, who spread fake news, 2) Checkers, who constantly share fact-checked news, and 3) Refuters, who respond to fake-news posts declaring their inaccuracy. We build a classifier that uses users' network graph to predict their reactions with an accuracy exceeding 93%. We applied further analysis for the most effective features of each users group and noticed that spreaders interact with more accounts that use their mother tongue and more accounts that get suspended while checkers and refuters interact with more foreign accounts and news-reporting entities.

Contents

1 Introduction **1**

1.1 History of Fake News 1

1.2 Fake News and Social Media 4

1.3 Twitter's Content Moderation 6

1.4 Thesis Contribution 7

2 Related Work **9**

2.1 Misinformation 9

2.2 Accounts' Behaviour 12

2.3 Accounts and Misinformation 13

3 Background **18**

3.1 Support Vector Machines 18

3.2 Multi-label Binarization 21

4 Methodology **23**

4.1 Users class definitions 23

4.2 Classification of User Accounts 25

5 Experimental Setup **27**

5.1 Ethical and Privacy Considerations 27

5.2 Data Collection 28

5.3 Data Statistics 30
5.4 Modelling Pipeline 30

6 Results & Discussion **33**
6.1 Classification Results 33
6.2 Analysis 33
6.3 Limitations 39

7 Conclusion **41**
7.1 Summary of Contributions 41
7.2 Future Work 42

List of Figures **43**

List of Tables **45**

Bibliography **46**

Chapter 1
Introduction

Rumours and false stories have been around as long as humans have lived in groups where power matters (Burkhardt, 2017). Nevertheless, over the past decade, fake news has become a famous term that has been defined, labelled, and classified in various ways (Safieddine, 2020). In order to understand the potential reasons behind this development, we need to quickly walk through the history of fake news.

1.1 History of Fake News

Burkhardt (2017) classifies the timeline of fake news into four eras as shown in figure 1.1:

1. **Pre-Printing Press Era**
   Primitive forms of writing inscribed on stone, clay, or papyrus appeared thousands of years ago. They are characterised by being usually limited to significant people like emperors, priests, or military leaders. The fact that such minorities controlled information gave them power over others and contributed to the existence of

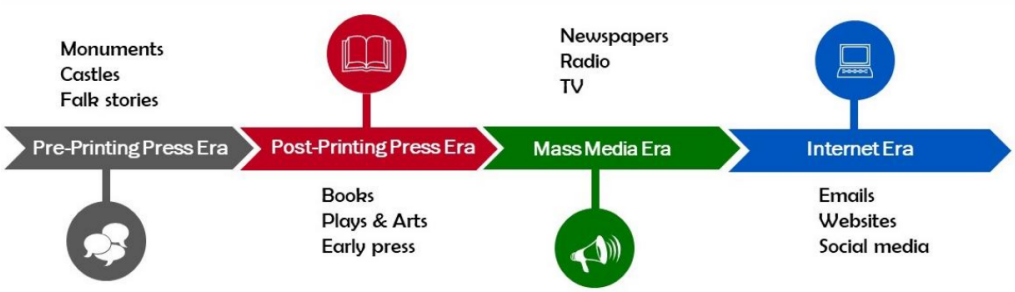


Figure 1.1: The Timeline of Fake News as described by Burkhardt (2017) and visualised by Safieddine (2020).

social hierarchies we know today. There was no means to verify the claims introduced in such forms. In the sixth century AD, the principal historian of Byzantium used fake news to smear the emperor Justinian and his wife, despite supporting Justinian during his lifetime. There could be no retaliation, questioning or investigations about these fake news due to the fact that the new emperor didn't favour Justinian.

2. **Post–Printing Press Era**
   The main pillars of this era are the invention of printing and the coincident spread of literacy. In the beginning literate people had similar power to significant people from the previous era over the illiterate. In 1522, the Italian author Pietro Aretino wrote wicked poems and plays to blackmail former friends and patrons and ask them for money. Canards were a series of fake news that contaminated France during the 17th century. They reported myths about some members of the royal family like Marie Antoinette which contributed to a harsh treatment to them after the French Revolution. As literacy rates increased, this basic misleading power started to diminish. Disinformation required more advanced skills like authoritative and convincing writing. Manipulation through printed information in general required funding and this is how talented writers started to be paid to write for the benefit of their employer. In 1844, the American writer Edgar Allan Poe wrote an article about

a balloonist who had crossed the Atlantic in a hot air balloon in only three days. The article was carefully written to cover scientific details plausibly and many people believed the story until they failed to find the balloon or the balloonist. The story was retracted four days after publication. In 1710, Jonathan Swift wrote *"Falsehood flies, and truth comes limping after it, so that when men come to be undeceived, it is too late; the jest is over, and the tale has had its effect"*. This statement is remarkably similar to the recent analyses published about the effect of disinformation (Siar, 2021; Reglitz, 2022). They usually discuss a deeper nature of the negative impact of fake news that's not related to the belief in them but to the threat they impose on democracy, integrity of elections, and public health. Vaccari and Chadwick (2020) link disinformation with uncertainty and distrust by exploring the case of Deepfakes (Westerlund, 2019). Even though some people might not be fooled by them, Deepfakes contribute to generalised indeterminacy and cynicism.

3. **Mass Media Era**
   What characterises this era is the wider reach associated with the conventional media that includes the radio and the television. As a result, fake news caused public panics. In 1926, BBC Radio reported that London was being attacked by communists with an early disclaimer that the broadcast was a spoof and not an actual news broadcast. This caused public panic until the story could be explained. The same problem occurred when the science fiction book *War of the Worlds* was broadcast in 1938 in the United States.

4. **Internet Era**
   This era introduces access to information that's almost instant. In the early years of web, disinformation started through some hoax websites like *DHMO.org* which was claiming that *Dihydrogen Monoxide* is present everywhere and that it causes cancer, acid rain, and global warming. One realises that this is a joke after noticing that Dihydrogen Monoxide is actually water! The emergence of advertisers as sponsors for most websites accelerated the rates of content creation

in general and interesting content creation in particular. People are interested in gossip, rumour, and scandals which are more likely to include disinformation. Websites are being created to capitalise on this nonintellectual side of the human nature.

We conclude from this brief overview that disinformation isn't a rising phenomenon on its own and existed since human have lived. However, the cultural evolution and the technological advances alter its nature, intensity, and impact continuously. Disinformation waves may follow similar patterns and share similar characteristics; however, almost every disinformation wave has a unique form and hence a unique impact and information propagation paradigm. This motivates continuous development of disinformation analysis and detection techniques.

1.2 Fake News and Social Media

We focus on the most recent disinformation waves as they were closely linked with social media. Social media facilitated the spread of fake news (Martens et al., 2018). This is because algorithm-driven news distribution platforms separate the role of content editing from the distribution process, which mainly focuses on maximising traffic and advertising revenue. In addition, social media boosts what is called horizontal propaganda. Vlăduţescu (2014) explains that the classical propaganda is vertical such that it is done by a leader or an authority acting by influence from the height of their prestige on a passive crowd placed in a position of inferiority. While the horizontal propaganda develops on the scientific base of the theory of dynamics of groups (Ellul, 1962). It is characterised by dimming the contrast between the propagandist and the audience such that every individual can be a source and a target at the same time. Last but not least Polage (2012) demonstrated the crucial effect of information repetition in the context of fake news. Earlier research had shown that repeated information is more likely to be rated as true than the information that has not been heard before. The study shows that this general theory holds

true for fake news as well. Moreover, it shows that repeating false claims creates a false memory about the sources of such claims. In other words, repeating fake news confuses the users about whom they have heard them from before. This effect is augmented further with horizontal propaganda as sources and targets are generally not distinct groups.

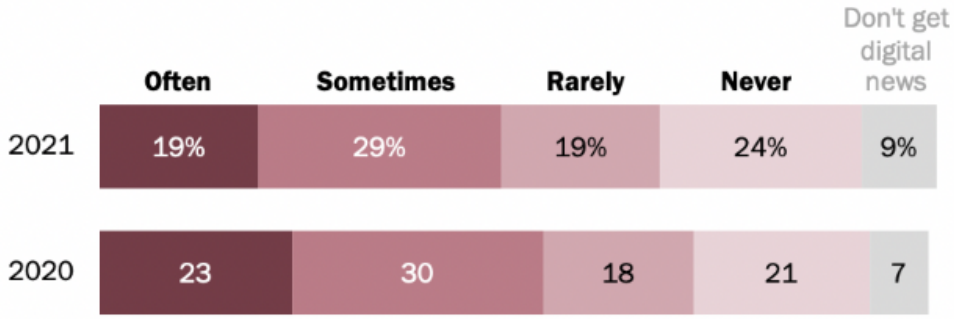


Figure 1.2: Americans' News Consumption on Social Media in 2020 and 2021 as reported by (Walker and Matsa, 2021).

The impact of these properties of social media appeared in the yearly survey conducted by the Pew Research Center Walker and Matsa (2021) to measure news consumption across social media among Americans. In the recent 2021's report, it shows a 5% decrease in the percentage of Americans who get news on social media at least sometimes compared to 2020 (Figure 1.2).

This decline is even present within the platforms where the majority of users used to regularly get news from namely Twitter and Facebook (Figure 1.3). This makes content regulation necessary not only for ethical and legal reasons but also for maximising the usage and consequently profit. A recent study attempts to quantitatively measure the moral impact of fake news (Olan et al., 2022). They develop a novel conceptual framework for this purpose and find out that societies are split on differentiating true news from fake ones as a result of disinformation. They even observe splits in core societal values and advise social media platforms to continuously upgrade their fact-checking technologies to tackle new tricks and strategies used in cascading fake news in the society.

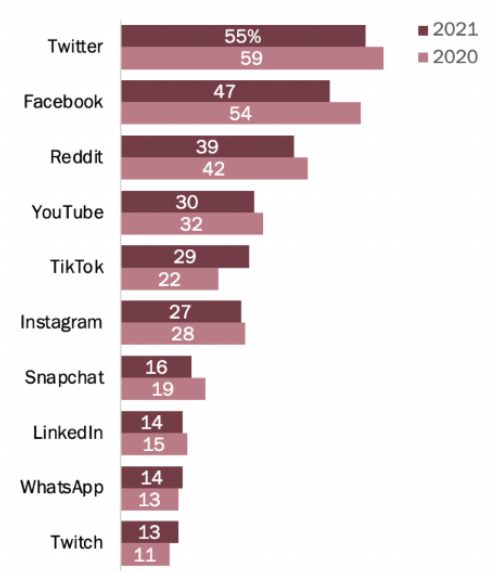


Figure 1.3: Percentage of each social media platform's users who regularly consume news there as reported by (Walker and Matsa, 2021).

1.3 Twitter's Content Moderation

Twitter's soft moderation strategies introduced amid the infodemic associated with COVID-19 are a successful example of the platform's efforts in this regard (Roth and Pickles, 2020). The first strategy is to display a warning cover before the tweet is displayed and the second is to display a warning tag below the Tweet (Figure 1.4). Sharevski et al. (2021) show that although the warning tag is not effective, the warning cover is. Bhuiyan et al. (2021) propose a more interactive technique namely NudgeCred that achieves better awareness through nudges (Figure 1.5). Twitter also publishes periodically sets of accounts which appeared to be state-sponsored to manipulate public opinion through misinformation, however, it does not label these accounts directly as fake news spreaders nor label a specific set of their tweets as fake news (Twitter, 2012). Zannettou et al. (2019) compare the behaviour of a set of these accounts to a set of random Twitter users and find interesting differences in the content they share, the evolution of accounts and the general use of Twitter.

Figure 1.4: An example of Twitter's warning cover and warning tag visualised by (Sharevski et al., 2021).

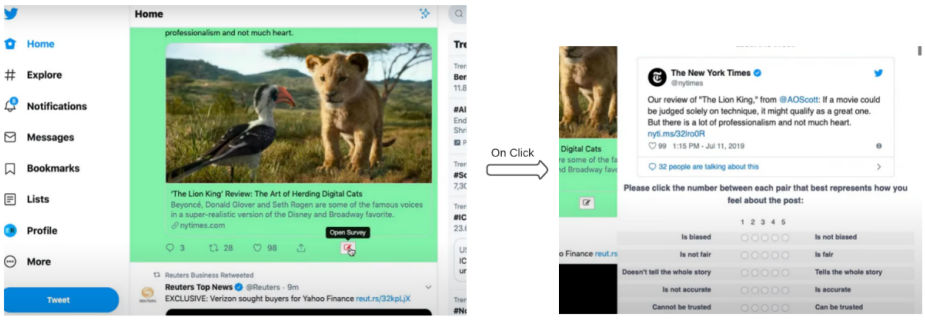


Figure 1.5: NudgeCred five-item credibility questionnaire.

1.4 Thesis Contribution

In recent years, there have been few initiatives to understand how misinformation might spread on social media (Cha et al., 2021; Kumar et al., 2016) and how to prevent it, usually through techniques that detect fake news (Shu et al., 2020; Resende et al., 2019) and block them or at least warn users. Nevertheless, there is less amount of work to understand the characteristics of users who consume these pieces of fake news on social media and understand their reactions towards a piece of fake news, where some users would believe and spread, while others check the news and might refute it. In our study here, we try to fill this gap and provide a study on the social media users reactions to misinformation on social media. This is seen highly important for the computational social science community to enable the design of further techniques that can stop the

spread of misinformation on social media. In particular, we investigate the following two research questions:

- **RQ1**: Is it possible to classify a user's general behaviour towards fake news either being a spreader or a checker/refuter?
- **RQ2**: What are the main differences between the networks of users who spread fake news in comparison with those who check or refute it?

By answering these two questions, we can have an insight about the characteristics of users who are more vulnerable to fake news compared to those who are more vigilant. This can be an important step towards understanding why some people fall for fake news more than others. For our study, we rely on a set of tweets labelled manually as fake news by multiple Arabic fact checking platforms as well as the set of tweets tweeted by the Twitter accounts of these fact-checking platforms in which they clarify the truth behind claims and rumours. We collected those tweets and fetched all the accounts which interacted with them either by liking, retweeting, or replying. We define three types of users: 1) Spreaders: who tweet/retweet fake news; 2) Checkers: who consistently retweet/like fact-checking tweets from the fact-checking accounts; and 3) Refuters: who reply on the fake-news tweets refuting the information mentioned.

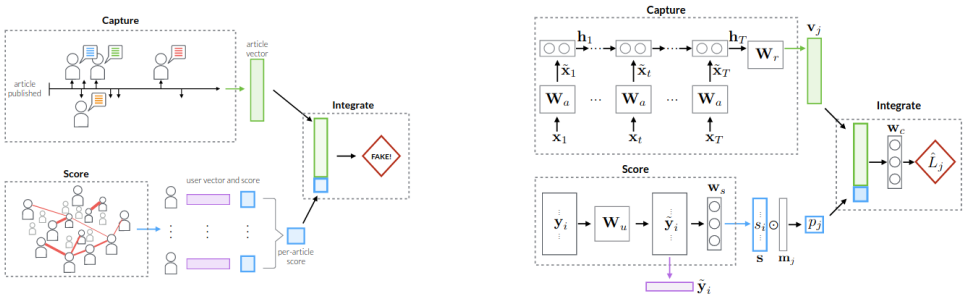
We built two classifiers that use network information of the accounts to classify them into spreaders against checkers and spreaders against refuters with accuracy reaching 93.4% and 93.7% respectively. Finally, we analysed the most distinctive features for each class and noticed that both refuters and checkers are identified by more foreign accounts, while spreaders are identified by accounts that have higher potential to be suspended. The findings of our study contribute to the efforts of limiting fake news from being spread, but this time from the user's side. We show that the coexistence of some characteristics makes some users more likely to spread fake news, which should motivate for designing more robust techniques for limiting the spread of fake news.

Chapter 2
Related Work

In this section, we introduce the most remarkable studies on the topic of fake news on social media. First, we explore the recent literature involving misinformation analysis and detection in general. Second, we explore the works utilising the behaviour of social media accounts belonging to specific groups. Finally, we review the approaches that analyse the behaviour of social media accounts in the context of misinformation.

2.1 Misinformation

Ruchansky et al. (2017) is one of the most justifiable architectures used for fake news detection. It was the first model to capture the three common features of fake news namely its text, response, and source. The model named *CSI* consists of three main parts, *Capture* for extracting a temporal representation of news articles to account for the response and text features, *Score* for representing and scoring the behaviour of users to account for the source feature, and *Integrate* for integrating the scores of *Capture* and *Score* as shown in figure 2.1. This model was used by Sharma et al. (2021) to process misinformation data linked with the 2020 US presidential



(a) Intuition behind *CSI*. Here, *Capture* receives the temporal series of engagements, and *Score* is fed an implicit user graph constructed from the engagements over all articles in the data.

(b) The *CSI* model specification. The *Capture* module depicts the LSTM for a single article a_j , The *Score* module operates over all users and its output is filtered with relevance to a_j.

Figure 2.1: An illustration of the CSI model.

election. The formulation of input data in this application was itself a valuable contribution. Cascades are constructed such that each cascade corresponds to a time-ordered sequence of engagements. Engagements in this case were retweets, quotes, and replies for Twitter tweets.

Benkler et al. (2020) is an extensive study that investigates the same effective disinformation campaign. It shows that this wave relied on elite institutional focus that would always share what the president says because simply it is news! Although the characteristics of spreader accounts are not analysed thoroughly, a topology describing them is identified (Figure 2.2). Interestingly, the study concludes that with this particular topology, current misinformation detection and regulation techniques used by popular platforms like Facebook and Twitter are not protective enough calling for different approaches. Even though Mosleh and Rand (2021) took a decent step to measure the exposure to misinformation through public figures which they call "Elite Misinformation", the topology of spreader accounts aforementioned is not totally covered since it is not limited to public figures.

Spezzano et al. (2021) point out the importance of news properties other than its text by showing that better accuracy in detecting fake news is

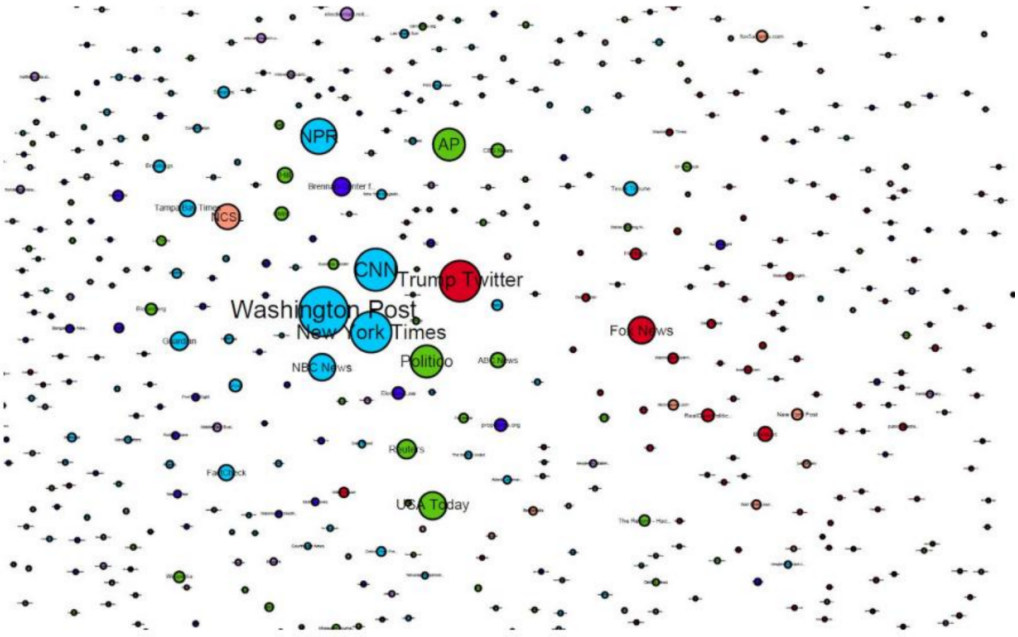


Figure 2.2: Media outlets sized and located by interlinking among their web stories about a given topic (Mail-in Voting Fraud in this case). Trump on Twitter is central to the overall dynamic. Fox News is central on the right. Notable is the significant role of Reuters, The New York Times, and CNN as leading media providers.

achieved by focusing on some elements of meta-data as well as the source of the news instead of its text. Another key result of this study is that automated detectors performed better than the human sample they tested on the same articles which motivates further improvement for these detectors especially after Pennycook et al. (2021) had demonstrated that humans value accuracy more than partisanship. With this mature responsible behaviour, reminding users about the accuracy of potential shareable threads is sufficient to limit misinformation to a large extent. Recuero et al. (2020) analyses misinformation data linked with Brazil's 2018 presidential election and claims that hyperpartisanship boosts misinformation which has a better chance of propagating in polarised circles due to the filter bubbles that arise under such conditions. This supports the idea of the potential existence of characteristic patterns among fake news spreaders that should not be necessary related to fake news. In many cases partisan-

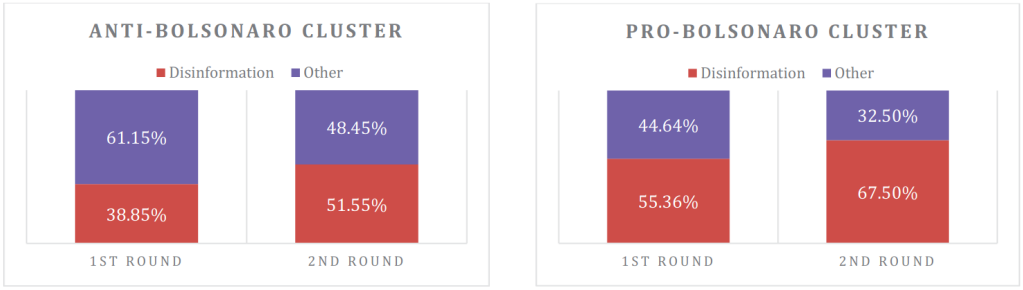


Figure 2.3: Disinformation content increasing among polarised crowds during the second round of the Brazilian elections compared to the first round.

ship can be detected and measured much more easily than the truth or falsehood of content and hence the correlation between the two aspects makes utilising partisanship to detect misinformation very possible.

2.2 Accounts' Behaviour

Users' engagements have been used to predict many targets including themselves. Hu et al. (2021) constructed statistical models to examine the predictive power of multiple factors in predicting the presence and the degree of users' engagements such as posting, retweeting or replying to tweets about a set of 643 real-world events. They come to discover that the most powerful predictors are prior users' engagements in addition to social network structures, topical interests and geolocation. Islam et al. (2014) use the network structure in addition to the engagements as a basis for a recommender system that recommends whom to follow. A stated limitation of this work is that it assumes that users would always be interested in following people who resemble those whom they already follow and interact with based on some matching criteria, however, people might introduce some browsing behaviour that shows their need for something different. Events can also be recommended based on interactions, Magnuson et al. (2015) did that through capturing geotagging information associated with Twitter's traffic to offer geographic recommendations,

however, this approach faces the challenge of the sparsity of geotagged threads in comparison with untagged ones. Jurgens et al. (2021) analyses and compares different approaches (Backstrom et al., 2010; Kong et al., 2014; McGee et al., 2013) that can help with this challenge through predicting geotagging information using network information turning the problem to a semi-supervised one. What matters for us in this case is the fact that network information is again useful in predicting geolocation. Aldayel and Magdy (2019) shows that network information can be used to predict a user's stance about a given topic even if the user's has not posted, shared, or interacted with anything related to this topic. The homophily of social networks makes such prediction possible and accurate to a large extent. This study is the only one we are aware of that performs the same analysis of network information that we perform and does not only focus on the accuracy of classification but also on the characteristics of the decisive features. Finally, target advertisement is one of the most important applications in this context since it is the main source of income for social media platforms. Accurate target advertisement using user's behavioural data covered in many works such as (Bhatia and Hasija, 2016) reflect clearly the value of such data.

2.3 Accounts and Misinformation

A recent example of a critical misinformation wave is the one linked with the outbreak of COVID-19. Hadlington et al. (2022) performed a qualitative study in an attempt to understand users' interactions with fake news and the motives behind them. An interesting finding was that people's interactions with information on social media are influenced by the aim of staying social. This suggests that peoples' social positions described more formally as their positions in a social network graph influence how they interact with information. Caldarelli et al. (2021) followed this hypothesis and used the mutual retweets counts among a given set of users to identify how close their opinions are and to analyse their tendency to spread fake news about COVID-19 (Figure 2.4). This work has two inter-

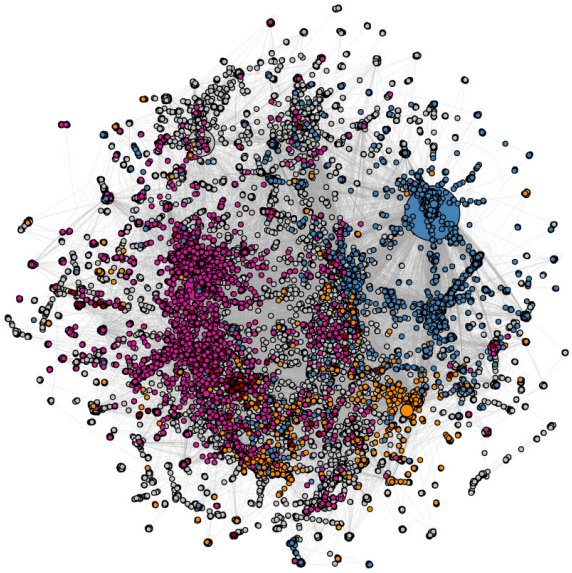


Figure 2.4: The projection of the retweet activity network: the communities have been highlighted according to the political discursive groups they take part to. All nodes not belonging to political discursive communities are in grey. Nodes' dimensions are proportional to their out degree.

esting properties in common with ours. Firstly, it performs the analysis using tweets written in an understudied language namely Italian. In our case we use tweets written in Arabic. Secondly and more importantly, its methodology is language agnostic and can be applied on sets of tweets in any combination of languages which is also the case for our methodology. The methodology however lacks a solid labelling technique for fake news and is content with assuming that information from infamous sources is more likely to include misinformation.

Weinzierl et al. (2021) utilise another interaction type which is replying (Table 2.1) to analyse whether users adopt or reject a given piece of fake news about COVID-19. This is done using an architecture composed of BERT in addition to Graph Attention Networks (GATs) that model lexical, emotional, and semantic features (Figure 2.5). Yang et al. (2020) find that although the ratio of accounts suspected to be bots sharing low-credibility information is higher than the ratio of such accounts sharing other content, the majority of of this content is still generated or shared by likely humans.

Misinformation Target: *Shaking hands cannot infect anyone since it is Sunnah.*
STANCE: **Agree** *Tweet:* NEW CASE: Illinois Health Officials say despite this new case, people do not need to alter their daily lives, but should work to stop the spread of germs.
STANCE: **Disagree** *Tweet:* Still many Indians follow this dharma intentionally and also unintentionally as well. Now WHO says, to maintain 3 feet distance from one individual to another to prevent from attack of Corona virus. Never touch anyone. Wash ur hands everytime..
Misinformation Target: *The coronavirus outbreak is a cover-up for a 5G-related illness.*
STANCE: **Agree** *Tweet:* @DineshDSouza Who ever said the Coronavirus is a hoax is correct. It's 5G radiation disease and its only going to get worse!!
STANCE: **Disagree** *Tweet:* I just read a tweet where someone claimed Coronavirus was actually a result of 5g exposure. These idiots walk among us.

Table 2.1: Examples of COVID-19 misinformation and tweets adopting or rejecting it from the COVIDLIES dataset.

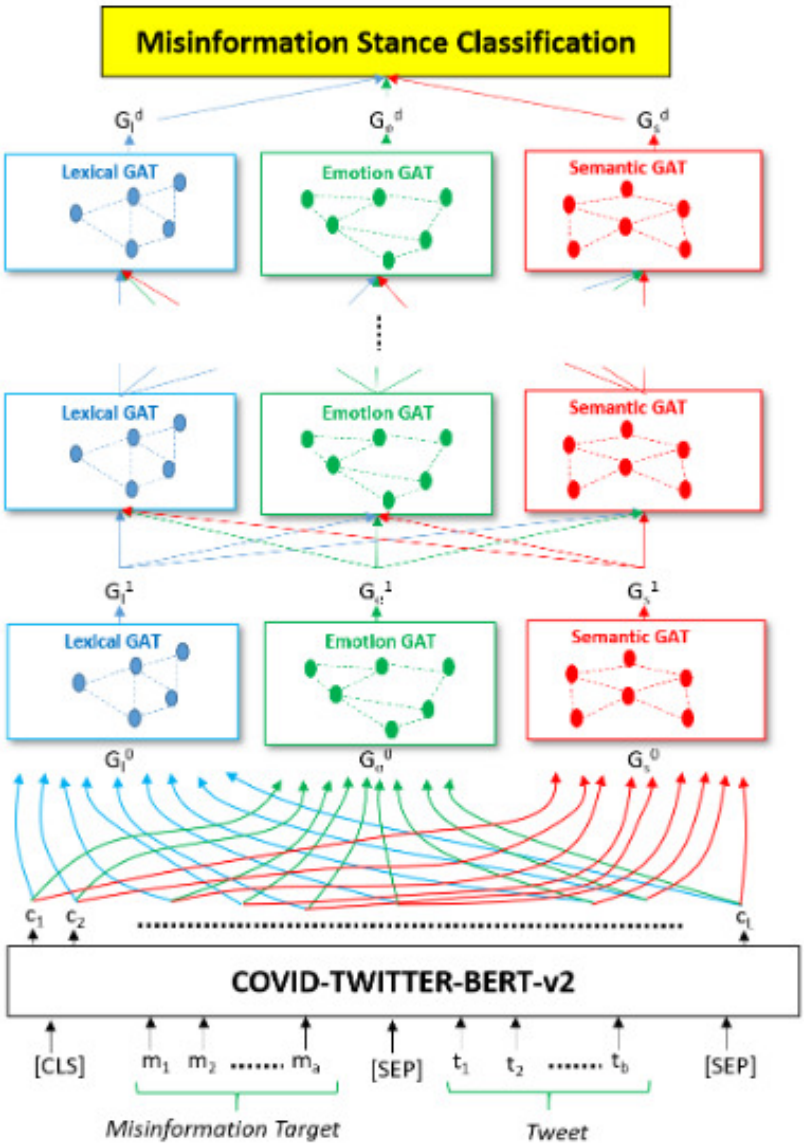


Figure 2.5: Neural architecture of the Lexical, Emotion, and Semantic Graph Attention Network for Stance Identification(LES-GAT-StanceId) system.

Some works such as (Vogel and Meghana, 2020) had a very similar problem definition to ours. They try to design a model which detects fake news spreaders without being language-dependent. However, they achieve a significantly worse performance for Spanish than English suggesting partial dependence on language. One other major concern in this work are the class labels. The proposed model objective is to distinguish fake news spreaders from people who share credible information. The 1st class is defined such that it includes anybody who shared fake news in the past, while the other class is defined to include anybody who never did. This assumption doesn't only mean that sharing a single instance of fake news makes a user a fake news spreader but also more importantly it relies on having a predefined set of all content defined as fake news which is not the case for real world scenarios. Leonardi and Rizzo (2021) change the name of the second class to be checkers, however, they still define it as people who share true news and support them. They don't provide a formal description of the checking behaviour. Rath et al. (2020) had the same classes, however their prediction technique is based on Graph Neural Networks (GNNs) (Scarselli et al., 2005) which is a convenient choice for this problem and expected to capture strong predictive power from the network information. This study however lacks the analysis of the accounts associated with the most influential features which is understandable as the explainability of GNNs in general is not trivial. Most recently, Mu et al. (2022) have tackled the problem of classifying spreaders against refuters using the features of users' language. The robust definition of the refuters class in this work which makes use of the reporting scheme of Sina Weibo social media platform (Corporation, 2009) makes the classification task much more meaningful. Refuters or "Active Citizens" as the authors denote them are users who had reported threads as fake news which were later manually labelled as such by the site administration. They also performed a detailed analysis on the language features of each of the classes and gave a hint that a similar analysis could be performed on network features.

Chapter 3
Background

In this section we explore in depth some technical aspects related to the design choices of our classifiers. First, we explain why the Support Vector Machine is the relevant choice to our type of analysis. Then, we discuss our feature representation and how it improves our classification performance and serves our feature analysis.

3.1 Support Vector Machines

Support Vector Machine (SVM) is a supervised learning model that seeks finding a hyperplane that classifies the data points for a classification task or fits them for a regression task. The hyperplane's position and orientation are influenced by the data points lying on the borders of the decision boundary which are called support vectors. The optimisation of such a hyperplane shall yield a maximum margin for a classification task which is a pleasant property for generalisation and avoiding overfitting.

We use SVMs with a linear kernel for our experiments. Discarding deep models is motivated by our focus on the analysis of input features which

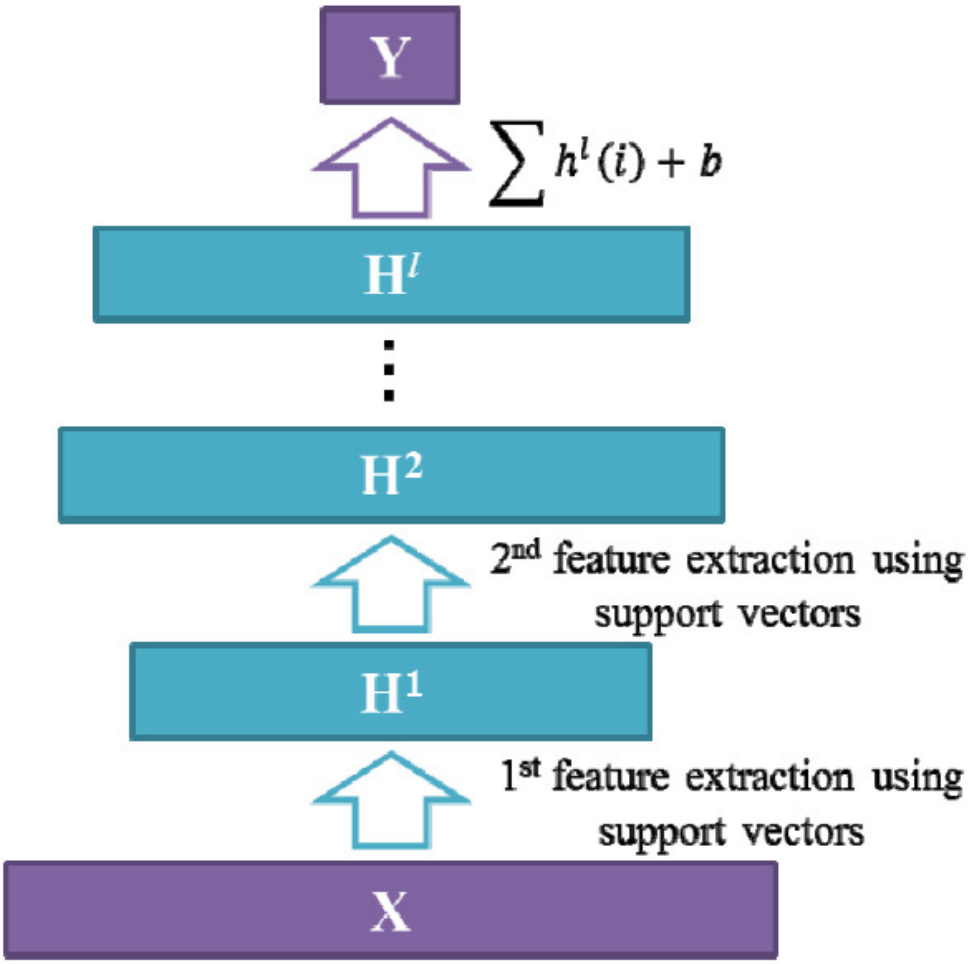


Figure 3.1: SVMs used in neural networks by Kim et al. (2013) due to their feature extraction ability.

have direct relation with the output label for linear models unlike deep models. The existing explainable AI techniques like LIME (Ribeiro et al., 2016) could theoretically enable performing our analysis with deep models. Nevertheless, they are proved not to be robust enough for high dimensional problems (Alvarez-Melis and Jaakkola, 2018) as their explanations are very unstable with respect to slight variations of input. Moreover, GNNs, which have become very popular among deep models to fit social data best (Fan et al., 2019), have technical sources of bias by design like distance-related (Ma et al., 2021) and degree-related biases (Tang et al., 2020) which are not present in linear models by definition.

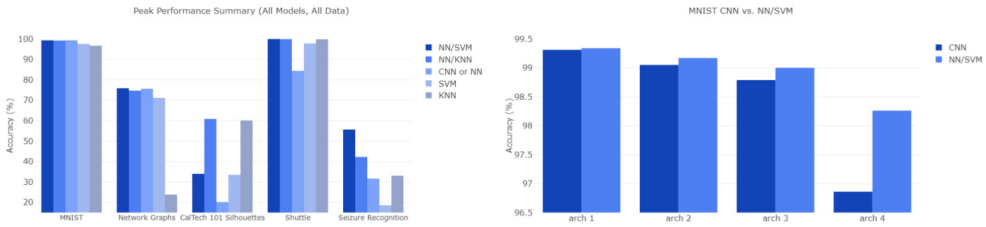


Figure 3.2: Figure 2: Left: Summary of neural networks and SVMs performance across 5 datasets (highest accuracy across all architectures), Right: A close up on the performance on MNIST for 4 CNN architectures against SVMs

Among linear models we select SVMs for a number of reasons:

1. SVMs have inherent ability to select data points that are most important for classification with good generalisation capabilities (Kim et al., 2013) which means that they can effectively discriminate features (Figure 3.1). This is very relevant for our analysis task as we are interested to know the most significant features that identify each set of users who perform some prescribed behaviour.

2. SVM is a better classifier when used as the final layer of a Neural Network than the conventional fully-connected layer (Notley and Magdon-Ismail, 2018). This experiment is the origin of the claim that Neural Networks are actually better feature extractors than

classifiers. With the same feature extraction backbone of neural network layers, SVM outperformed all other neural final layers on 5 different datasets (Figure 3.2).

3. Among linear models we select SVMs because they proved to perform best on social data (AlDayel and Magdy, 2020).

3.2 Multi-label Binarization

In order to better understand Multi-label binarization, we will quickly visit its simpler form namely one hot encoding. This way of representing data is recommended for enumerating categorical variables whose values aren't related to each other numerically. For example, if a variable represents *Fruits*, assigning a numeric value for each fruit can be done arbitrarily since an apple doesn't have to be more than an orange or much less than a banana. However, such an arbitrary choice causes machine learning models to behave differently as they interpret numerical values as comparable magnitudes and not just codes for unrelated things. One hot encoding tackles this issue by expanding this variable into a set of variables each representing a unique value for the expanded variable. Each data record will have the value *1* for the variable representing the unique value this record had for the expanded variable and *0* for the rest of the variables in the set (Figure 3.3).

In many cases, categorical variables can be multi-value attributes. If we imagine that *Fruits* represents the list of fruits offered by a store, it can hold the values *Apple* and *Banana* for the same data record. Multi-label binarization models this scenario by putting the value *1* for all variables representing the fruits offered by the store and the value *0* for fruits that are not.

The variables that we have with this nature within our problem represent sets of accounts. This includes the set of accounts some user follow, the set of accounts whom the user shares their content, and the set of accounts whom the user likes their content. Although all these variables repre-

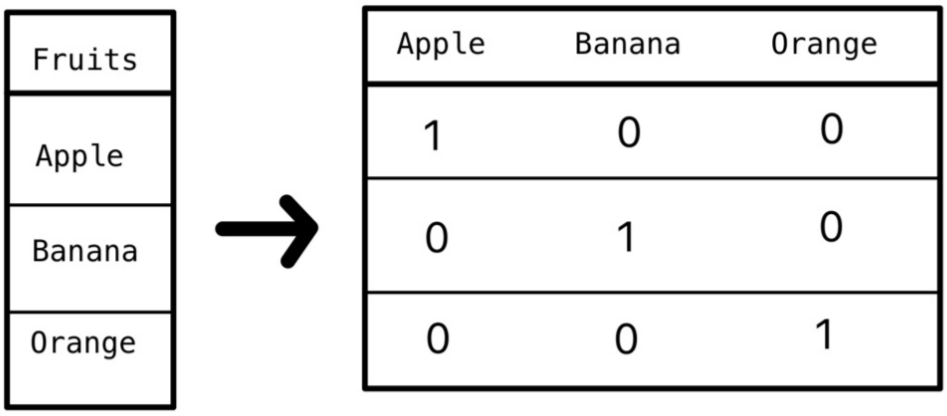


Figure 3.3: An example of one hot encoding over the variable *Fruits*.

sent social media accounts, Multi-label binarization should be applied over each of them separately because each set of variables shall model a different behaviour. We are interested in knowing the most decisive behaviours in characterising fake news consumers and not just the most decisive accounts. In addition, this separation allows the SVM to model different types of features. Following an account is a different feature from retweeting content for the same account. This flexibility yields higher classification accuracy.

Chapter 4
Methodology

In this section, we explain our methodology for the analysis process of users. Initially, we define the types of user accounts according to their reactions to fake news. Secondly, we discuss the features of those accounts that we are interested to analyse, and that are used to build a classifier to distinguish those different account groups. Finally, we describe our analysis approach.

4.1 Users class definitions

In this work we define three classes of users according to their behaviour in the context of fake news spread:

- **Spreaders**: The accounts that frequently tweet or retweet information labelled as fake news.
- **Checkers**: The accounts that frequently like and/or retweet information shared by a collection of fact checking accounts.
- **Refuters**: The accounts that reply on tweets sharing information

labelled as fake news to refute them.

For all the classes of accounts above, our label is based on a noticed behavior. However, this does not imply any label to the intention of the account holder. For example, spreaders who tweet the fake news themselves do not have to be the source of fake information, but they simply can be victims of a piece of fake news they believed and decided to tweet about. Similarly, checkers and refuters aren't necessarily immune against fake news, but they are just the accounts that we recorded news checking/refuting behavior for them.

Our main task in this study is to understand the main differences, if any, in their social networks. Particularly, the characteristics of fake-news spreaders on one side versus the checkers or refuters on the other side.

We split our task into two binary classification subtasks where accounts are either classified into spreaders and checkers or classified into spreaders and refuters. This is justified by the intuition that checkers and refuters are very likely to overlap with each others and cannot generally be regarded as mutually exclusive behaviors. Apparently, formalizing the aforementioned class definitions requires setting a threshold of interactions for each class that qualify an account to belong there. We believe these thresholds should be tuned based on the distribution of the dataset; thus, we will mention the exact numbers we used in the experimental setup section. The tuning steps that we applied and that we believe to be applicable on any similar dataset are as follows:

1. Given a set of accounts that have interacted with threads in the prescribed behavior of a class at least once, compute the total number of interactions made by each account.

2. For all the possible numbers of interactions per account ($n : 1 \rightarrow N$) where N is the number of interactions of the most interacting account, compute the total numbers of accounts that made n interactions.

3. Check which of the two classes has a superior interaction distribution

i.e. has more ns where for all of them, it has more interactions than the second class.

4. Pick the highest possible n for the superior class that will qualify sufficient accounts to train your classifier.

5. Pick n for the second class such that it qualifies roughly the same number of accounts as the superior class to have a balanced dataset.

4.2 Classification of User Accounts

For the purpose of analyzing the differences among those user accounts, we build a couple of classifiers to automatically detect if a user is a spreader vs checker or a spreader vs refuter. Our main objective is to see if there are any significant signals in those users' networks that can be predictive to their general behavior with fake-news. Our hypothesis is that users who fall for fake news to the degree they tweet/retweet about them might have different social network positioning than those who are more careful with fake news to the degree that they consistently tweet about news correction content or respond to fake news tweets refuting them. Since we do not have any preliminary assumptions about those differences in the networks, we decided to use all their network information and then apply analysis to the most significant features that have the most predictive power to understand those differences.

Following the work by Aldayel and Magdy (2019) which showed the effectiveness of different types of network features in predicting users' stance, we define the following types of network features:

- **Preference Network**: This feature set models the users preferences based on the tweets they like. Instead of the mentioned accounts and the linked website domains in (Aldayel and Magdy, 2019), we model this network using the authors of such tweets because we believe they are more representative of a general pattern to be captured about the liking user, especially in our task.

- **Connection Network**: This feature set models the long-term ties between the users. Instead of including both the accounts that the users follow and those who follow them, we drop the accounts that follow the users because in most cases the users have no control over them so they are not representative of their choices and we focus only on the network of users followed by our accounts.

- **Interaction Network**: This feature set models the networks the users interact with in their posts by mentioning or retweeting. We decide to split this into two feature sets where the first includes the retweets only and the second includes the other types of mentions (plain mentions or replies). We hypothesize that the retweets feature set indicate a different signal that replies/mentions since it usually implies agreement on the content, while replies can be agreement or objection. Note, quoted retweets are not counted in our setup as retweets, since they can indicate objection to content.

All the above network features are language-agnostic and do not require custom handling based on the language of the dataset, which allows applying our methodology on multilingual datasets. In addition, these features do not include sensitive attributes about the target user such as gender or race. We are aware that non-sensitive attributes might correlate with sensitive attributes in some datasets (Zhao et al., 2022) introducing bias to the classifier. However, this bias is still much less than using sensitive attributes directly especially because we are not using deep models as explained earlier.

Our approach is to experiment with the different set of features to identify the set that achieves the best classification performance, which would indicate the presence of distinctive features for each user group that allow this classification. Afterwards, we apply analysis on distinctive features to understand the main difference among those users who spread fake news vs those who check/refute them.

Chapter 5
Experimental Setup

5.1 Ethical and Privacy Considerations

Detecting and combating misinformation is generally an ethical objective but it should not be done in a privacy-violating fashion. Thus, in our study, we base our study on public tweets only. Furthermore, Twitter as a platform does not enforce users to share any of their demographic information upon registration. Therefore, Twitter accounts are not necessarily linked to the physical identity of users. We again emphasise few points about our study: 1) the labels we give to user accounts in this study are based on a noticed online behaviour not a label to their intention. 2) The purpose of the study is NOT to classify users as fake news spreaders but mainly to understand the differences between users who fall for fake news vs those who are more vigilant. 3) We do not share any information about the users in our dataset that to identifying any of the individuals in our study. Finally, we have obtained an ethical approval to conduct this study from our host institute.

5.2 Data Collection

Data collection follows the following steps: 1) collection of tweets including fake news and others that include fact-checking information; 2) identifying the most engaging users with those tweets then collecting their network information. For the whole data collection process, the Twitter API V2 (Twitter, 2020) was used.

Fake Tweets: We use the *Arafacts* dataset (Sheikh Ali et al., 2021) as our source for fake news claims. It is a set of claims (news stories) that have been verified by five Arabic fact-checking websites for being either accurate or fake. We contacted the authors and they provided us with a more recent version than the one in their publication which includes a total of 8,957 claims from April 2016 till December 2021. The recent version also has dedicated fields for Twitter information. There is a field that categorises claims into five categories based on their fact checking results: *True*, *False*, *Partly-false*, *Sarcasm*, *Unverifiable*. We focus in our analysis on claims that are associated with tweets only and that are labelled as completely *False*, where their count is 1,252 (14% of the total data) claims (i.e. fake news). We tried to utilize other false claims that include external links instead of tweets to retrieve tweets that share this link following the data collection approach of Mosleh et al. (2022) but we found that these tweets are too few for Arafacts. A claim can be associated with multiple tweets spreading the same rumour. We have a total of 3,300 tweets associated with false claims. For such tweets, we collect their authors, retweeters, as well as the replies. While authors and retweeters will be the source for our spreaders class, for the replies, we collect the repliers who included phrases that indicate refuting the claim of the news. A full list of these terms and their English translation using Google Translate (Google, 2006) can be seen in table 5.2. These repliers will be the source for our refuters class as described later. Although earlier studies show that refutes don't necessarily decrease misinformation (Mosleh et al., 2021) and that including logical reasoning in the refute is the way that pushes the fake news spreader to respond positively (Martel et al., 2020), our objective focuses

on the refuting behaviour and not its consequences.

Verification Tweets: We obtained the corresponding Twitter accounts of the five fact checking websites that are the source of the annotated the data in Arafacts dataset. These fact-checking accounts are `AFP Fact Mena`, `Misbar`, `Maharat News`, `Fatabyyano`, and `Taakkad`. We then collected their recent timelines of tweets, which include a total of 11,158 fact-checking tweets that are sharing a fact about a trending rumour in the Arab world. Similarly, we collected all the accounts that engaged with these fact-checking tweets, including the likers and retweeters to be the source for our checkers class.

Network Information: Interestingly but not surprisingly, the superior class as defined in our methodology appears to be the spreaders class for both binary classification tasks. This means that rumours get more interactions than corrections and that refuting replies are less than the shares for fake news. We pick a threshold n=5 and this results in 1829 accounts spreading fake news more than five times. For n=5 for the checkers class, we get 1681 accounts doing more than five interactions. To balance the dataset, we add 148 accounts from those who did exactly five interactions to have 1829 checkers. To make sure that our refuting definition indicate actual refutes, we manually label the replies that included our refuting terms. We found that 77% of them were actual refutes whereas the rest didn't discuss the falsehood of the main thread despite including refuting terms. For n=2, we find 337 accounts the refuters class and we add 1429 which did exactly one interaction to have 1829 refuters.

For all these sets of accounts we collect the following:

- The Twitter accounts they follow
- Their last 3,200 Tweets since this is the number of tweets allowed by Twitter
- The authors of the last 3,200 Tweets they liked

Class	Accounts	Tweets	Sources
Spreaders	1829	913	772
Checkers	1829	8092	5
Refuters	1829	326	301

Table 5.1: The number of collected accounts per class, the unique tweets they interacted with, and the unique sources tweeting these tweets

- The meta-data associated with the tweets that show if they were retweeting the content of another account or replying to another account

5.3 Data Statistics

We investigated the diversity of these accounts in terms of the number of tweets they interact with and the number of sources which tweeted these tweets (Table 5.1). It was confirmed that these accounts did their prescribed behaviour on hundreds of tweets that have hundreds of unique authors.

5.4 Modelling Pipeline

We multi-hot-encode the given set of accounts such that the feature vector's length is equal to the count of all accounts that appeared in the whole dataset with values equal to 1 for accounts that exist for a particular data record and 0 otherwise. For example if we consider the connection network of some user u in a dataset U then its feature vector length will be equal to the count of all followed accounts in this given dataset with 1's in the positions of accounts followed by u and 0's elsewhere. We optimise this long list of features by dropping accounts that appear only once in the whole dataset since they do not introduce any predictive power and

increase the length of the feature vector and the memory consumption unnecessarily. For the connection feature set in case of classifying spreaders against checkers, we *drop* the five fact-checking Twitter accounts that were used to construct the Checkers class from the feature vectors to prevent the model from learning to classify the accounts that follow these accounts regardless of their misinformation behaviour. We train and test SVMs using Scikit-learn library (Pedregosa et al., 2011) with balanced records of the two classes to be predicted and with a 90-10% train-test split. We also repeat our experiments using SVM Light (Joachims, 1999) for validation [1].

[1]Code for data collection and model deployment is available at `git.ecdf.ed.ac.uk/mfawzi/analyzing_engagement_with_fact_checks` and `gitlab.com/bigirqu/fakedet` respectively. Contact `mhmud.fwzi@gmail.com` to get access.

Arabic Word	Translation	Arabic Word	Translation
كذب	False	خرافة	Myth
كاذب	False	تلفيق	Fabrication
زيف	False	ملفق	Fabricated
زائف	False	كذوب	Liar
إشاعة	Rumour	هراء	Bullshit
غير صحيح	Not True	غش	Cheat
مو صحيح	Not True	خدع	Deceive
غير حقيقي	Not True	خادع	Deceptive
خطأ	False	دلس	Delusion
خاطئ	False	تدليس	Fraud
مخطئ	Wrong	خدع	Deceive
خاطىء	Wrong	اختلاق	Fabrication
تزوير	Forgery	مختلق	Feigned
تضليل	Fabrication	إدعاء	Claim
زور	Falsify	مستحيل	No way
مضلل	Misleading	أوهام	Illusions
فبرك	Make up	افتعال	Fabricate
شائعات	Rumours	مفتعل	Artificial
شائعة	Rumour	غير دقيق	Inaccurate
تزييف	Falsification	مو دقيق	Not Accurate
ليس صحيح	Not right	لم يحدث	Did not happen
تركيب	Composition	لا دليل	No Proof
فوتوشوب	Photoshop	فوتشوب	Photoshop

Table 5.2: The list of words that indicate refuting a claim and their English translation via Google (2006).

Chapter 6
Results & Discussion

6.1 Classification Results

As shown in table 6.1 and table 6.2, network features in general are able to classify spreaders against checkers with accuracy ranging from 83.9% up to 93.4% and against refuters with accuracy ranging from 85.8% to 93.7%. As shown, the performance of the different types of networks is similar for both classification tasks. Retweets network is the most effective network for predicting users' behaviour with fake-news, while mentions network is the least effective. Combining the preference, connection and retweets network together improves the performance further for the two tasks reaching an accuracy and F-scores over 93%

6.2 Analysis

We consider the distribution of the top significant features in the best performing models that combine the networks as features for classification

Network	Accuracy	Precision	Recall	F1-Score
Preference	87%	87.9%	85.4%	86.6%
Connection	88.6%	90.1%	86.5%	88.3%
Interaction	91.4%	92%	90.4%	91.2%
Retweets	92%	93.6%	89.9%	91.7%
Mentions	83.9%	85.3%	91.5%	83.3%
Preference + Connection + Retweets	93.4%	94.3%	92.1%	93.2%

Table 6.1: Fake news spreaders versus checkers performance using different sets of features to train an SVM for binary classification.

Network	Accuracy	Precision	Recall	F1-Score
Preference	85.8%	85.6%	85.6%	85.6%
Connection	84.7%	87.8%	80%	83.7%
Interaction	91.5%	91.6%	91.1%	91.4%
Retweets	92.3%	94.1%	90%	92%
Mentions	83.8%	83.1%	84.4%	83.7%
Preference + Connection + Retweets	93.7%	92.9%	94.4%	93.7%

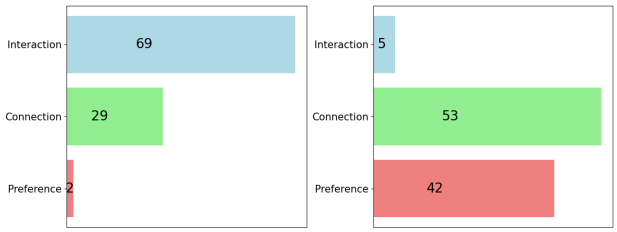
Table 6.2: Fake news spreaders versus refuters performance using different sets of features to train an SVM for binary classification.

(Preference + Connection + Retweets). We check the features coefficients of the SVM models and extract the most significant 100 features with the highest coefficient for each class for both tasks.

As shown in figure 6.1 and figure 6.2 for the spreaders and refuters classes, most features are from the retweeters network while for the checkers class, the connection network dominates. Likes network barely contributes to the spreaders class but it contributes more to the checkers and refuters classes. In addition, after a month of data collection, we noticed that seven of those accounts in the top 100 features were suspended accounts among those accounts for the spreaders, while none were suspended for the checkers class, and only one for the refuters class. This is an indication that those accounts that spreaders interact with on Twitter might be violating Twitter terms somehow and thus got suspended.

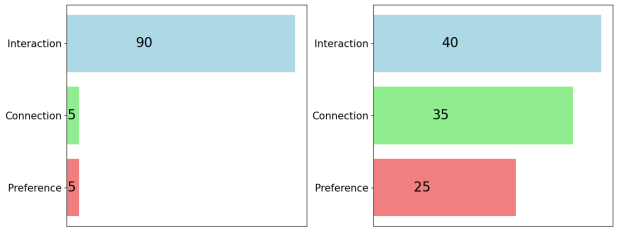
Finally, for the top 100 feature-accounts, we discard personal accounts and classify the non-personal accounts among the top features into 1) Professionals/Institutions reporting news in Arabic, 2) Other Arabic Institutions/Public Figures, 3) Professionals/Institutions reporting news in other languages than Arabic, and 4) Other non-Arabic Institutions/Public Figures (Table 6.3). We observe that the third category highly identifies the checkers class (56% of the features) and that the non-Arabic accounts identifying the spreaders class are usually not related to news reporting. It was very interesting to find that `Berkeley Advanced Media Institute`, which offers training about professional journalism and story telling, identifies the checkers class. We recognise `David Emery`, a famous editor who calls himself `debunker` among the features of the likers network that most identify the checkers class. For the second classification task, 81% of the accounts identifying the spreaders class share content in Arabic compared to 55.6% only for the refuters (Table 6.4).

In this study, our aim is to understand the main differences in network characteristics of accounts who help in spreading misinformation on social media compared to those who are more careful with fake news. We apply our analysis on a set of users who shared/refuted fake news in the Arab world.



(a) Spreaders (b) Checkers

Figure 6.1: The distribution of the top 100 significant features in the combination classifier that identify (a) The Spreaders Class (b) The Checkers Class.



(a) Spreaders (b) Refuters

Figure 6.2: The distribution of the top 100 significant features in the combination classifier that identify (a) The Spreaders Class (b) The Refuters Class.

Spreaders			Checkers		
fahddeepaji1	halgawi	US_SOCEUR	binamanhaji	pierrerabbat	PoliWork
solarimpulse	TurkiHAlhamad1	AJABreaking	theawayfans	gulfamber	USATODAY
KingSalman	AJASports	SaudiIntlGolf	OmanNewsAgency	hedayh_ar	maischberger
goodreads	amfozan	ColinYeo1	AMI_Berkeley	praddenkeefe	HyateSukar
TPUSA	saracihan61	Saudi_Vi2030	Guderian_Xaba	philmcnulty	danwootton
ajyad_3	AliSiamPress	NewsHour	HeidiReports	taghreedrisheq	AsemAlnabeh
aljouiabdullah	Dr_Alkadi	LottevBeek	muttons	joeroganhq	Ataya1_2020
UDH_Friends	AmnestySyrien	LaurentMeillan	apantazi	BrighamWomens	LionelMedia
AlsheikhSultan	kasta_app		aabackhaus		

Table 6.3: Non-personal accounts identifying spreaders and checkers classified into (1) Professionals/Institutions reporting news in Arabic, (2) Other Arabic Institutions/Public Figures, (3) Professionals/Institutions reporting news in other languages than Arabic, and (4) Other non-Arabic Institutions/Public Figures.

Spreaders			Refuters	
fahdalruqi	halgawi	AJASports	nuhaaldossary	danielegiazzi
ExtraMadia	TurkiQashlan	Ben_Sulayem	holmakhdar	waffootball
SEkramyofficial	AlAnba_News_KW	AJAPodcasts	AlAraby_Gulf	logomakerCA
RBouzoAH	alkhabar_ye	Turkey_Affairs	CGCKuwait	MadawiDr
arab_media2	abdullahagar2	TheRedPlatform	humor_economico	
BaderWeatherMap	ThePodShow	Islamic_Wisdom_		
ArabHoroscope	Ta3meed	KSAembassyFI		

Table 6.4: Non-personal accounts identifying spreaders and refuters classified into (1) Professionals/Institutions reporting news in Arabic, (2) Other Arabic Institutions/Public Figures, (3) Professionals/Institutions reporting news in other languages than Arabic, and (4) Other non-Arabic Institutions/Public Figures.

We have two specific research questions for our study. Our RQ1 was whether or not it is possible to classify general behaviours of users towards fake news into spreading or checking/refuting. The answer to this question was noticed to be "Yes", it is surprisingly easy to classify those users who spread fake news from those who check/refute fake news. Using only network features of those users, we have detect spreaders from checker or refuters with an accuracy over 93%. This performance exceeded our initial expectation. The users were collected from interactions with a diverse set of news, and thus it was expected to find sparse data and consequently find it hard to capture common signals that are shared among those users to be able to classify them with this high performance. However, it was shown that a simple linear SVM classifier can easily distinguish between those users. These findings highly motivated our RQ2, which is what are those common signals among spreaders that make them distinctive from other user groups?

Our analysis to the most predictive network features from each group showed obvious differences in terms of accounts types of accounts each user group follows. Those who check news seem to follow more authentic accounts, including international sources, while those who spread fake news are more local and less exposed to authentic sources. Spreaders also follow accounts that have higher chance of getting suspended.

Our findings highlight the importance of user's network choices on social media and how this can affect their behaviour towards fake news. Our work contributes to the previous work in this area (Mu et al., 2022; Hadlington et al., 2022), and have implications especially for social medial platform designers who are interested in reducing the spread of fake news and increasing the awareness of unauthentic sources for getting information.

6.3 Limitations

- Although our classifiers are language-agnostic, we applied our analysis on an Arabic dataset only. We believe that more interesting findings could be found when applying the same pipeline on other datasets with the same nature in other languages and that lifting Twitter API's limits on data retrieval shall allow researchers to utilise network features to better detect misinformation and fake news spreaders. It will also allow further understanding of the nature of these networks which is a good initial step towards social and psychological theories in this direction.

- The study is limited to Twitter only, class definitions should be adjusted to adapt other types of interactions on other platforms.

- Only around 25% of the most significant features are non-personal accounts. This ratio is even much less for the refuters class. This means that personal accounts have higher influence when it comes to the propagation of fake news. Such a finding matches Halpern et al. (2019)'s theoretical model about the influence of personal factors on the belief of misinformation and confirms our comment about (Mosleh and Rand, 2021) that the topology that causes elite misinformation (Benkler et al., 2020) isn't restricted to public figures. Indeed a deeper understanding of the nature of these accounts could be achieved by modelling the interactions among personal accounts in the same way cascades are constructed in Sharma et al. (2021) but this would encounter complicated challenges related to privacy and explainability.

- While labelling refutes we recognise a common behaviour that we did not consider where many users confirm that fake news are correct and fight for them. Our initial observation tells that this usually happens when the fake news is in favour of the stance of the user; however, this correlation requires a quantitative study to be confirmed.

- Classifying spreaders against refuters or checkers isn't a real problem as users can belong to both classes regarding different topics or even the same topic. Moreover, for users who might belong to only one of the two classes, they aren't expected to form a balanced dataset in the real world.
  We remind about the fact that our objective was mainly to define the characteristics associated with such behaviours but not to create a hard line of classifications among users. We also propose cascading our two models to deploy them for fake news spreaders detection only. This means that the models' decisions won't include labelling an account as a refuter or a checker. They will only include a confidence score about the likelihood than an account is a spreader.

Chapter 7
Conclusion

7.1 Summary of Contributions

In this work, we successfully classify accounts that interact with fake news on social media into spreaders and checkers/refuters. We observe many differences between the networks of different classes. Networks of fake news spreaders tend to include suspended accounts, more accounts using the same language of the spreader, and less accounts specialised in news reporting. Networks of fact checkers tend to include more foreign accounts specialised in news reporting. Most significant features that define the misinformation behaviour are personal accounts and not famous ones. We hope that these findings motivate further work that utilises these effective features to improve misinformation detection techniques and to construct potential social and psychological theories by analysing the features more qualitatively while keeping in mind the ethical considerations of such analysis.

7.2 Future Work

There are many potential extensions to our work that can help social media platforms to combat misinformation:

- Building a GNN classifier that utilises our extracted network features might yield an even higher classification accuracy. This will be helpful for our proposed cascaded pipeline where we pool the decisions of the spreader vs checker classifier and the spreader vs refuter classifier to identify the likelihood of an account being a spreader.
- Pooling the decisions of two classifiers can be done in multiple ways. Averaging can be a naive but an effective way. Creating a pooling layer and fine-tuning it can also be effective.
- Examining how far the personal accounts we found within the most significant features confirm Halpern et al. (2019)'s model is an interesting research question.

The work in this thesis has been published later in CSCW 2024 as a main conference paper entitled *"Pinocchio had a Nose, You have a Network!": On Characterizing Fake News Spreaders on Arabic Social Media* (Fawzi and Magdy, 2024). Nevertheless, the content of both is different as the thesis provides more details about the background and the setup while the paper includes additional computations and analyses that were done in response to the review process.

List of Figures

1.1 The Timeline of Fake News as described by Burkhardt (2017) and visualised by Safieddine (2020). 2

1.2 Americans' News Consumption on Social Media in 2020 and 2021 as reported by (Walker and Matsa, 2021). 5

1.3 Percentage of each social media platform's users who regularly consume news there as reported by (Walker and Matsa, 2021). 6

1.4 An example of Twitter's warning cover and warning tag visualised by (Sharevski et al., 2021). 7

1.5 NudgeCred five-item credibility questionnaire. 7

2.1 An illustration of the CSI model. 10

2.2 Media outlets sized and located by interlinking among their web stories about a given topic (Mail-in Voting Fraud in this case). Trump on Twitter is central to the overall dynamic. Fox News is central on the right. Notable is the significant role of Reuters, The New York Times, and CNN as leading media providers. 11

2.3 Disinformation content increasing among polarised crowds during the second round of the Brazilian elections compared to the first round. 12

2.4 The projection of the retweet activity network: the communities have been highlighted according to the political discursive groups they take part to. All nodes not belonging to political discursive communities are in grey. Nodes' dimensions are proportional to their out degree. 14

2.5 Neural architecture of the Lexical, Emotion, and Semantic Graph Attention Network for Stance Identification(LES-GAT-StanceId) system. 16

3.1 SVMs used in neural networks by Kim et al. (2013) due to their feature extraction ability. 19

3.2 Figure 2: Left: Summary of neural networks and SVMs performance across 5 datasets (highest accuracy across all architectures), Right: A close up on the performance on MNIST for 4 CNN architectures against SVMs 20

3.3 An example of one hot encoding over the variable *Fruits*. . 22

6.1 The distribution of the top 100 significant features in the combination classifier that identify (a) The Spreaders Class (b) The Checkers Class. 36

6.2 The distribution of the top 100 significant features in the combination classifier that identify (a) The Spreaders Class (b) The Refuters Class. 36

List of Tables

2.1 Examples of COVID-19 misinformation and tweets adopting or rejecting it from the COVIDLIES dataset. 15

5.1 The number of collected accounts per class, the unique tweets they interacted with, and the unique sources tweeting these tweets 30

5.2 The list of words that indicate refuting a claim and their English translation via Google (2006). 32

6.1 Fake news spreaders versus checkers performance using different sets of features to train an SVM for binary classification. 34

6.2 Fake news spreaders versus refuters performance using different sets of features to train an SVM for binary classification. 34

6.3 Non-personal accounts identifying spreaders and checkers classified into (1) Professionals/Institutions reporting news in Arabic, (2) Other Arabic Institutions/Public Figures, (3) Professionals/Institutions reporting news in other languages than Arabic, and (4) Other non-Arabic Institutions/Public Figures. 37

6.4 Non-personal accounts identifying spreaders and refuters classified into (1) Professionals/Institutions reporting news in Arabic, (2) Other Arabic Institutions/Public Figures, (3) Professionals/Institutions reporting news in other languages than Arabic, and (4) Other non-Arabic Institutions/Public Figures. 37

Bibliography

[1] Aldayel, A. and Magdy, W. (2019). Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–20.

[2] AlDayel, A. and Magdy, W. (2020). Stance detection on social media: State of the art and trends.

[3] Alvarez-Melis, D. and Jaakkola, T. (2018). On the robustness of interpretability methods.

[4] Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. pages 61–70.

[5] Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., Kaiser, J., and Schmitt, C. (2020). Mail-in voter fraud: Anatomy of a disinformation campaign. *SSRN Electronic Journal*.

[6] Bhatia, V. and Hasija, V. (2016). Targeted advertising using behavioural data and social data mining. pages 937–942.

[7] Bhuiyan, M., Horning, M., Lee, S., and Mitra, T. (2021). Nudgecred: Supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–30.

[8] Burkhardt, J. M. (2017). *Combating fake news in the digital age*. Chicago, IL : ALA TechSource ©2017.

[9] Caldarelli, G., De Nicola, R., Petrocchi, M., Pratelli, M., and Saracco, F. (2021). Flow of online misinformation during the peak of the covid-19 pandemic in italy. *EPJ Data Science*, 10:34.

[10] Cha, M., Cha, C., Singh, K., Lima, G., Ahn, Y.-Y., Kulshrestha, J., and Varol, O. (2021). Prevalence of misinformation and factchecks on the covid-19 pandemic in 35 countries : Observational infodemiology study. *JMIR Human Factors*.

[11] Corporation, S. (2009). Sina weibo.

[12] Ellul, J. (1962). *Propagandes*. Librairie Armand Colin.

[13] Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., and Yin, D. (2019). Graph neural networks for social recommendation.

[14] Fawzi, M. and Magdy, W. (2024). "pinocchio had a nose, you have a network!": On characterizing fake news spreaders on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–20.

[15] Google (2006). Google translate.

[16] Hadlington, L., Harkin, L., Kuss, D., Newman, K., and Ryding, F. C. (2022). Perceptions of fake news, misinformation, and disinformation amid the covid-19 pandemic: A qualitative exploration. *Psychology of Popular Media*.

[17] Halpern, D., Valenzuela, S., Katz, J., and Orrego Miranda, J. (2019). *From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News*, pages 217–232.

[18] Hu, Y., Farnham, S., and Talamadupula, K. (2021). Predicting user engagement on twitter with real-world events. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):168–177.

[19] Islam, M., Ding, C., and Chi, C.-H. (2014). Personalized recommender system on whom to follow in twitter. In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, pages 326–333.

[20] Joachims, T. (1999). Making large scale svm learning practical. *Advances in Kernel Methods: Upport Vector Machines*.

[21] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y., and Ruths, D. (2021). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):188–197.

[22] Kim, S., Kavuri, S., and Lee, M. (2013). Deep network with support vector machines. In Lee, M., Hirose, A., Hou, Z.-G., and Kil, R. M., editors, *Neural Information Processing*, pages 458–465, Berlin, Heidelberg. Springer Berlin Heidelberg.

[23] Kong, L., Liu, Z., and Huang, Y. (2014). Spot: Locating social media users based on social network context. *Proc. VLDB Endow.*, 7(13):1681–1684.

[24] Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes.

[25] Leonardi, S. and Rizzo, G. (2021). Automated classification of fake news spreaders to break the misinformation chain. *Information*, 12:248.

[26] Ma, J., Deng, J., and Mei, Q. (2021). Subgroup generalization and fairness of graph neural networks.

[27] Magnuson, A., Dialani, V., and Mallela, D. (2015). Event recommendation using twitter activity. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, page 331–332, New York, NY, USA. Association for Computing Machinery.

[28] Martel, C., Mosleh, M., and Rand, D. (2020). You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online.

[29] Martens, B., Aguiar, L., Gomez, E., and Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. *SSRN Electronic Journal*.

[30] McGee, J., Caverlee, J., and Cheng, Z. (2013). Location prediction in social media based on tie strength. pages 459–468.

[31] Mosleh, M., Martel, C., and Eckles, D. (2021). Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. pages 1–13.

[32] Mosleh, M., Martel, C., Eckles, D., and Rand, D. (2022). Promoting engagement with social fact-checks online.

[33] Mosleh, M. and Rand, D. (2021). Falsehood in, falsehood out: A tool for measuring exposure to elite misinformation on twitter.

[34] Mu, Y., Niu, P., and Aletras, N. (2022). Identifying and characterizing active citizens who refute misinformation in social media.

[35] Notley, S. and Magdon-Ismail, M. (2018). Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and k-nearest neighbor classifiers.

[36] Olan, F., Jayawickrama, U., Arakpogun, E., Suklan, J., and Liu, S. (2022). Fake news on social media: the impact on society. *Information Systems Frontiers*.

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[38] Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., and Eckles, D. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592:1–6.

[39] Polage, D. (2012). Making up history: False memories of fake news stories. *Europe's Journal of Psychology*, 8:245–250.

[40] Rath, B., Salecha, A., and Srivastava, J. (2020). *Detecting Fake News Spreaders in Social Networks Using Inductive Representation Learning*, page 182–189. IEEE Press.

[41] Recuero, R., Soares, F. B., and Gruzd, A. (2020). Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):569–578.

[42] Reglitz, M. (2022). Fake news and democracy. *Journal of Ethics and Social Philosophy*, 22:162–187.

[43] Resende, G., Melo, P., Reis, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). Analyzing textual (mis)information shared in whatsapp groups.

[44] Ribeiro, M., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. pages 97–101.

[45] Roth, Y. and Pickles, N. (2020). Updating our approach to misleading information.

[46] Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI: A hybrid deep model for fake news. *CoRR*, abs/1703.06959.

[47] Safieddine, F. (2020). *Chapter 1: History of Fake News*.

[48] Scarselli, F., Yong, S. L., Gori, M., Hagenbuchner, M., Tsoi, A. C., and Maggini, M. (2005). Graph neural networks for ranking web pages. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 666–672.

[49] Sharevski, F., Alsaadi, R., Jachim, P., and Pieroni, E. (2021). Misinformation warning labels: Twitter's soft moderation effects on covid-19 vaccine belief echoes.

[50] Sharma, K., Ferrara, E., and Liu, Y. (2021). Characterizing online engagement with disinformation and conspiracies in the 2020 u.s. presidential election.

[51] Sheikh Ali, Z., Mansour, W., Elsayed, T., and Al-Ali, A. (2021). Arafacts: The first large arabic dataset of naturally-occurring professionally-verified claims.

[52] Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):626–637.

[53] Siar, S. (2021). Fake news, its dangers, and how we can fight it.

[54] Spezzano, F., Shrestha, A., Fails, J., and Stone, B. (2021). That's fake news! reliability of news when provided title, image, source bias & full article. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–19.

[55] Tang, X., Yao, H., Sun, Y., Wang, Y., Tang, J., Aggarwal, C., Mitra, P., and Wang, S. (2020). Investigating and mitigating degree-related biases in graph convoltuional networks. pages 1435–1444.

[56] Twitter (2012). Twitter transparency center.

[57] Twitter (2020). Twitter api v2.

[58] Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6:205630512090340.

[59] Vlăduţescu, (2014). Communicational types of propaganda. *International Letters of Social and Humanistic Sciences*, 33:41–49.

[60] Vogel, I. and Meghana, M. (2020). Detecting fake news spreaders on twitter from a multilingual perspective. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 599–606.

[61] Walker, M. and Matsa, K. E. (2021). News consumption across social media in 2021. *The Pew Research Center*.

[62] Weinzierl, M., Hopfer, S., and Harabagiu, S. M. (2021). Misinformation adoption or rejection in the era of covid-19.

[63] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9:39–52.

[64] Yang, K.-C., Torres-Lugo, C., and Menczer, F. (2020). Prevalence of low-credibility information on twitter during the covid-19 outbreak.

[65] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. pages 218–226.

[66] Zhao, T., Dai, E., Shu, K., and Wang, S. (2022). Towards fair classifiers without sensitive attributes: Exploring biases in related features. pages 1433–1442.