DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis

Explainability of Fake News Detection Models for Social Media

Batuhan Erdogdu

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis

Explainability of Fake News Detection Models for Social Media

Erklärbarkeit von Modellen zur Fake-News-Erkennung in Sozialen Medien

Author:	Batuhan Erdogdu
Supervisor:	Prof. Dr. Georg Groh
Advisor:	M.Sc. Carolin Schuster
Submission Date:	Submission date

I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, Submission date Batuhan Erdogdu

Acknowledgments

Abstract

Contents

1 Introduction

With the rapid development of communication technologies, social media has become one of the most frequently used news sources since it is easier, faster, and offers interaction with people. For example, a study from Pew Research Center (Walker & Matsa, 2021) reports that in 2021, 48% of U.S. adults get their news from social media "often" or “sometimes”. Furthermore, global data from 2022 (Watson, 2022) shows that over 70% of adults from Kenya, Malaysia, Phillippines, Bulgaria, and Greece use social media as one of their news sources, while this share is lower than 40% for the adults in the United Kingdom, The Netherlands, Germany, and Japan. These examples show that a considerable percentage of the population uses social media as a news source.
In contrast to its convenience, interactivity, and speed, social media can spread any kind of information since no regulatory authority checks the posts. As a result, a flood of false and misleading information is observed on social media (Allcott & Gentzkow, 2017).
The research community introduced numerous approaches to counteract the uncontrolled dissemination of fake news. For instance, some studies focused on building datasets (Dou, Shu, Xia, et al., 2021; Nakamura, Levy, & Wang, 2020; Santia & Williams, 2018; Shu, Sliva, Wang, et al., 2017; Tacchini, Ballarin, Della Vedova, et al., 2017; Wang, 2017), and some studies leveraged the power of *Machine Learning* (ML) to automatically detect fake news (Bian, Xiao, Xu, et al., 2020; Han, Karunasekera, & Leckie, 2020; Monti, Frasca, Eynard, et al., 2019; Zhou, Wu, & Zafarani, 2020) by learning features from the data. Due to the number of posts and the limitation of staff to check the posts, ML-based techniques can reduce manual labor when used with human supervision to counter the spreading of fake news. However, ML-based techniques with high complexity, such as *Deep Neural Networks* (DNN), are harder to understand and interpret since they act like black-boxes (Castelvecchi, 2016).
The integration of ML-based methods into human society impacts more people every day. While incredibly helpful in some aspects, ML-based techniques do not offer a reason for a particular prediction. Furthermore, we can not simply accept classification accuracy as a metric to evaluate real-world problems (Doshi-Velez & Kim, 2017). Integrating ML-based methods into human society makes interpretability a requirement to increase social acceptance (Molnar, 2022).
Consequently, a new research field called *eXplainable Artificial Intelligence* (XAI) surfaced

to fill this missing link between humans and *Artificial Intelligence* (AI). XAI proposes creating a set of ML techniques that deliver more explainable models while preserving learning performance, and help humans to understand, properly trust, and effectively handle the emerging generation of artificially intelligent partners (Gunning & Aha, 2019). While incorporating XAI increases social acceptance, it also aims to create more privacy-aware (Edwards & Veale, 2017), fairer, and trustworthy systems (Lipton, 2016). Like all ML techniques, *Fake News Detection* (FND) models need interpretability, particularly when implementing countermeasures for fake news. However, the interpretability of a model is not often considered despite the large amount of research produced in the last decade. Incorporating social context (Shu, Mahudeswaran, Wang, et al., 2018), representing the propagation networks as graphs (Dou, Shu, Xia, et al., 2021), and using *Graph Neural Networks* (GNN) to produce *State Of The Art* (SOTA) models (Monti, Frasca, Eynard, et al., 2019) have increased the complexity, but also the performance of FND models. For instance, using social context data alone has proved to be more effective than textual data alone in recent studies (Dou, Shu, Xia, et al., 2021). However, it is not clear which social features impact the decision process of these models.
This thesis focuses on the explainability of FND models using tools from the XAI suite. Specifically, we focus on content-based models and social context-based models to elaborate on their interpretability. Thus, we define three research objectives:

RO1 Determine the interpretation tools for explaining FND models.

RO2 Show that interpretations of FND models play an essential role in understanding the shortcomings of the FND models.

RO3 Determine which features impact the outcome the most.

From here on, talk about the structure of the thesis.

2 Background and Related Work

We explain two research fields that create the bedrock of this thesis, namely, fake news detection and explainable artificial intelligence. Both areas provide the foundation of tools that were used in this work. The first provides the mechanisms and approaches to detect fake news, and the second offers a suite of techniques to interpret these mechanisms and approaches.
Initially, in 2.1, we discuss societal challenges, the characteristics, and the history of fake news. Then we talk about the detection methods that were developed over the years. After showing the challenges of creating FND models, we conclude the first section with SOTA FND models.
After fake news detection, in 2.2, we first examine when XAI is necessary and its importance. Then, we define the suite of explainable artificial intelligence and the goals of XAI, and finally, we determine the suite that aims to satisfy these goals.

2.1 Fake News Detection

In the past decade, social media has become a place where anyone can share information. Although fast, free, and easy to access, obtaining real news from social media can be difficult, and one should do so at their own risk and always check the facts (Allcott & Gentzkow, 2017; Lazer, Baum, Benkler, et al., 2018). But the news stream never ends; thus, the need to verify the credibility of news using automated systems arises. To address this necessity, the number of studies involving *Fake News* or *Fake News Detection* has dramatically increased in the last decade (Fig. 2.1).
In 2.1.1, we briefly present the history of fake news and look at studies that display the impact of fake news on society. In this section, we also define the terms fake news, disinformation, and misinformation.
In 2.1.2, we make an excursion into social sciences and human psychology, delivering insights into why humans fall for or tend to believe fake news. Furthermore, we draw some insights from the socio-technical foundations of fake news.
We then list the available datasets used in FND and talk about their advantages and disadvantages in 2.1.3. Finally, in 2.1.4, we first talk about the evolution of detection algorithms, then we classify FND algorithms with respect to their input data type and

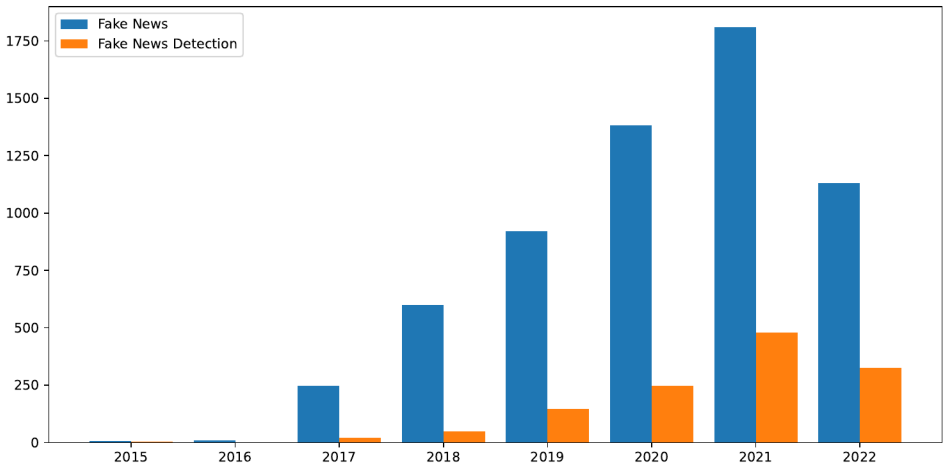

Figure 2.1: Total number of publications that include (1) *Fake News* (blue) and (2) *Fake News Detection* (orange) publications by year. Source: Scopus; Search Arguments: (1) TITLE-ABS-KEY("fake news*") PUBYEAR AFT 2014 (2) TITLE-ABS-KEY("fake news detection")

what they focus on that data.

2.1.1 Fake News

Throughout history, various forms of widespread fake news have been recorded. For instance, in the thirteenth century BC, Rameses the Great decorated his temples with paintings that tell stories of victory in the Battle of Kadesh. However, the treaty between the two sides reveals that the outcome of the battle was a stalemate (Weir, 2009). Just after the printing press was invented in 1439, the circulation of fake news began. One of history's most famous examples of fake news is the "Great Moon Hoax" (Foster, 2016). In 1835, The Sun newspaper of New York published articles about a real-life astronomer and a made-up colleague who had observed life on the moon. It turns out that these fictionalized articles brought them new customersand almost no backlash after the newspaper admitted that the articles mentioned earlier were a hoax[1].
In order to highlight the difference, using the definitions from (Pennycook & Rand, 2021), we formally define the terms disinformation and misinformation as follows,

Definition 2.1.1 (*Disinformation*)**.** Information that is false or inaccurate and was created with a deliberate intention to mislead people.

[1]https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/

Definition 2.1.2 (*Misinformation*)**.** Information that is false, inaccurate, or misleading. Unlike disinformation, misinformation does not necessarily need to be created deliberately to mislead.

There is no fixed definition for fake news. Thus, we elaborate on the definitions of fake news. A limited definition is news articles that are intentionally or verifiably false (Allcott & Gentzkow, 2017). This definition stresses authenticity and intent. The inclusion of false information that can be confirmed refers to authenticity. On the other hand, intent refers to the deceitful intention to delude news consumers (Shu, Sliva, Wang, et al., 2017). This definition is widely used in other studies (Conroy, Rubin, & Chen, 2015; Mustafaraj & Metaxas, 2017; Shu, Sliva, Wang, et al., 2017). Furthermore, recent social sciences studies (Lazer, Baum, Benkler, et al., 2018; Pennycook & Rand, 2021) define fake news as fabricated information that mimics news media content in form but not in organizational process or intent. Similarly, this definition covers authenticity and intent; additionally, it includes the organizational process. More general definitions for fake news consider satire news as fake news due to the inclusion of false information even though satire news aim to entertain and inherently reveals its deception to the consumer (Balmas, 2014; Brewer, Young, & Morreale, 2013; Jin, Cao, Zhang, & Luo, 2016; V. Rubin, Conroy, Chen, & Cornwell, 2016). Further definitions include hoaxes, satires, and obvious fabrications (V. L. Rubin, Chen, & Conroy, 2015) In this thesis, we are not interested in the organizational process and do not consider conspiracy theories (Sunstein & Vermeule, 2009), superstitions (Lindeman & Aarnio, 2007), rumors (Berinsky, 2017), misinformation, satire, or hoaxes. Therefore, we use the limited definition from (Allcott & Gentzkow, 2017) and formally introduce it as follows:

Definition 2.1.3 (*Fake News*)**.** News articles that are intentionally or verifiably false.

Fake news can lead to disastrous situations, such as crashes in stock markets, resulting in millions of dollars. For example, Dow Jones industrial average went down like a bullet (see Fig. 2.2) after a tweet about an explosion injuring President Obama went out due to a hack (ElBoghdady, 2013).
The detrimental impacts of fake news further extend to societal issues. When fake news rose to prominence with the 2016 U.S. Presidential Election (Beckwith, 2021), a man, convinced by what he read on social media about a pizzeria trafficking humans, went on a shooting spree in that pizzeria. Later named Pizzagate (Fisher, Cox, & Hermann, 2016), this incident illustrates the deadly impact of fake news. In fact, fake news can even affect presidential elections (Allcott & Gentzkow, 2017; Read, 2016).
Recent history exhibits that some fake news spreads like wildfires through social media. Evidence shows that the most popular fake news stories were more widely shared than the most popular mainstream news stories (Silverman, 2016).

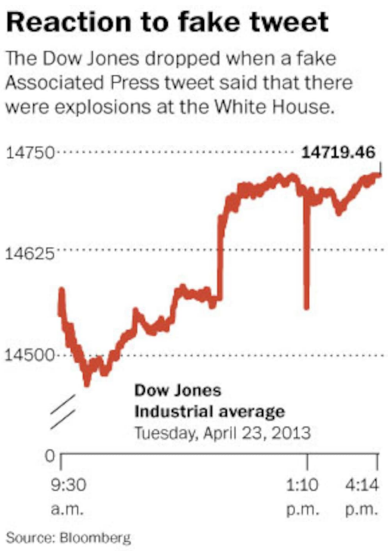

Figure 2.2: The market's reaction to the fake tweet. The sharp decline caused by a single tweet. Image obtained from (ElBoghdady, 2013)

Digital News Report 2022 (Newman, Fletcher, Robertson, et al., 2022) reports in its key findings that trust in the news is 42% globally, the highest (69%) in Finland, and the lowest (26%) in the U.S.A. Additionally, the same study shows that in early 2022, in the week of the survey, between 45% and 55% of the surveyed social media consumers worldwide witnessed false or misleading information about COVID-19. The same study also reports the appearance of fake news in politics was between 34% and 51%, and between 9% and 48% for fake news about celebrities, global warming, and immigration (Watson, 2022).

2.1.2 Foundations of Fake News

The environment for fake news has been the traditional news media for a long time. First started with newsprint, then continued with radio and television, and now with social media and the web, the dissemination of fake news reached its peak. Next, we discuss the psychological and social foundations of fake news to stress the importance of human psychology, especially when accepting fake news as genuine and sharing it with others. Then we focus on the technical foundations where we discuss how social media and technologiy have accelerated the diffusion of fake news.

Psychological Foundations. Understanding the difference between real and fake news is not an easy task for a human. Two psychological theories, namely, *naive realism*

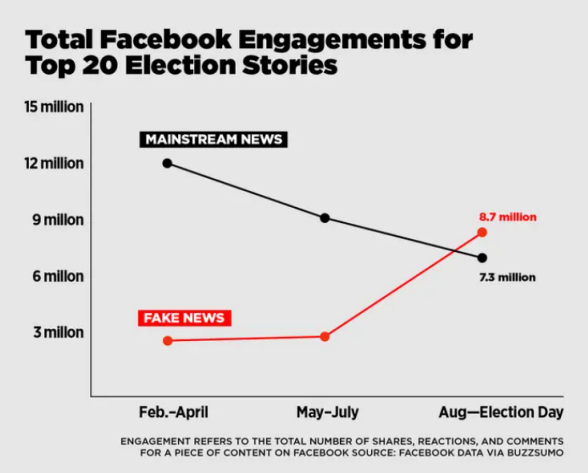

Figure 2.3: The rising engagement for fake news stories observed after May-July, just before Presidential Elections. Image obtained from (Silverman, 2016)

and *confirmation bias*, examine why humans fall for fake news. The first refers to a person's disposition to believe that their point of view is the mere accurate one, while people who believe otherwise are uninformed or biased (Reed, Turiel, & Brown, 2013). The second, often called selective exposure, is the proclivity to prefer information that confirms existing views (Nickerson, 1998).
Another reason for human fallacy in fake news is that once a misperception is formed, it becomes difficult to correct. In fact, it turns out that correcting people leads them to believe false information more, especially when given factual information that refutes their beliefs (Nyhan & Reifler, 2010).
Social Foundations. The prospect theory explains the human decision-making process as a mechanism based on maximizing relative gains and minimizing losses with respect to the current state (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). This inherent inclination to get the highest reward also applies to social cases in which a person will seek social networks that provide them with social acceptance. Consequently, people with different views tend to form separate groups, which makes them feel safer, leading to the consumption and dissemination of information that agrees with their opinions. These behaviors are explained by social identity theory (Ashforth & Mael, 1989) and normative social influence (Asch & Guetzkow, 1951). Two psychological factors play a crucial role here (Paul & Matthews, 2016). The first, social credibility, is explained by a person's tendency to recognize a source as credible when that source is deemed credible by other people. The second, called the frequency heuristic, is the acceptance of a news piece by repetitively being exposed to it. Collectively, these

psychological phenomena are closely related to the well-known filter bubble (Pariser, 2011), also called echo chamber, which is the formation of homogenous bubbles in which the users are people of similar ideologies and share similar ideas. Being isolated from different views, these users usually are inclined to have highly polarized opinions (Sunstein, 2001). As a result, the main reason for misinformation dispersal turned out to be the echo chambers (Vicario, Bessi, Zollo, et al., 2016).

Technical Foundations. Social media's easy-to-use and connected nature give rise to more people selecting or even creating their own news source. Naturally, this gives way to more junk information echoing in a group of people on social media. As algorithms evolve to understand user preferences, social media platforms recommend similar people or groups to those in echo chambers. A recent study (Cinus, Minici, Monti, & Bonchi, 2022) shows that these recommenders can strengthen these echo chambers. They discuss that some of these recommenders contribute to the polarization on social media. In other words, people can convince themselves that any fake news is real by staying in their echo chambers. One main reason that some fake news spreads so rapidly on social media is the existence of malicious accounts. The account user can be an actual human or a social bot since creating accounts on social media is no cost and almost no effort. While many social bots provide valuable services, some were designed to harm, mislead, exploit, and manipulate social media discourse. Formally, a social bot is a social media account governed by an algorithm to fabricate content and interact with other users (Ferrara, Varol, Davis, et al., 2016). A more recent study from the same author shows that malicious social bots were heavily used in the 2016 U.S. Presidential Elections (Bessi & Ferrara, 2016). On the other hand, malicious accounts that are not bots, such as online trolls who aim to trigger negative emotions and humans that provoke people on social media to get an emotional response, contribute to the proliferation of fake news (Cheng, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017).

Building upon three foundations, we draw some results for fake news to be considered when building a fake news detection model:

1. *Invasive*: Fake news can appear on anyone's feed if it spreads for a sufficient amount of time.
2. *Hard to discern*: Fake news is fabricated in such a way that it resembles the authenticity of a real news source. This indistinguishability leads to issues when working with news-content-based FND models.
3. *The source is crucial*: The credibility of a news source is essential. We can use news from credible sources to teach the model to distinguish genuine from fabricated.
4. *Fake news has hot spots*: The echo chambers are invaluable examples when trying

to understand the behaviors of fake news. We can leverage this attribute and use social models, such as graphs, to successfully detect fake news.

5. *Early detection is essential*: As discussed in psychological foundations, the volume of exposure to a piece of fake news can significantly affect one's opinions, thus leading to more misinformed individuals.

Next, we discuss general detection methods and how they have evolved. Then, we focus on fake news detection and widely used datasets offered by the research community.

2.1.3 Datasets

2.1.4 Evolution and Current State of Fake News Detection Models

2.2 Explainable Artificial Intelligence

2.2.1 Importance of Explainable Artificial Intelligence

2.2.2 A Good Explanation

2.2.3 Overview of Explainable Artificial Intelligence

only include what you use.

3 Fake News Detection Models

3.1 Content Based Models

3.1.1 Definitons

Talk about text based models, tf-idf, bag of words(BoW), how BERT is used in these tasks, (in the end) just assert that only text based models are not sufficient.

3.1.2 Dataset

Used the Kaggle competition dataset. -> Talk about the general analysis of the dataset. (How many instances, real/fake instances,)

3.1.3 Tokenizer

Used DistilRoBERTa tokenizer. (check the tokenizer of the model and talk about it)

3.1.4 Model

Used the model in transformers repository. The model from GonzaloA was used since it also provided its dataset and their train/val/test splits.

3.1.5 Explainability and Explanation

The model seems to have memorized some basic patterns and rely on that. Talk about the properties of explanation techniques. (Localization,) Define explainability. Define explanation. Input perturbation Explain a novel news (use test data)

3.2 Social Context Based Models

Talk about models that incorporate social context, spatiotemporal information and other context with text data. Can be any kind of model.

3.2.1 Geometric Deep Learning

Talk about Graph Neural Networks

3.2.2 Dataset

FakeNewsNet, UPFD, explain the dataset, no of edges/nodes. Which models use this dataset,

3.2.3 Models

SAGE GNN UPFD GCNFN

4 Explainability of Fake News Detection Models

4.1 Explanation Techiques

4.1.1 SHAP, DeepSHAP

4.1.2 GNNExplainer

4.1.3 Explainability vs. Explanation

4.2 Content Based Fake News Detection Models

4.2.1 Explaining Content Based Fake News Detection Models

4.2.2 Introducing Unseen Data

4.2.3 Results

4.3 Content and Social Feature Based Fake News Detection Models

4.3.1 Explaining Content and Social Feature Based Fake News Detection Models

4.3.2 Introducing Unseen Data

4.3.3 Results

5 Conclusion

List of Figures

List of Tables

Bibliography

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–36. https://doi.org/10.1257/jep.31.2.211

Asch, S. E., & Guetzkow, H. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, 222–236.

Ashforth, B. E., & Mael, F. (1989). Social identity theory and the organization. *The Academy of Management Review*, *14*(1), 20–39.

Balmas, M. (2014). When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, *41*(3), 430–454. https://doi.org/10.1177/0093650212453600

Beckwith, D. C. (2021). United states presidential election of 2016. In *Encyclopedia britannica*.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*(2), 241–262.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, *21*(11).

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. https://doi.org/10.48550/ARXIV.2001.06362

Brewer, P. R., Young, D. G., & Morreale, M. (2013). The Impact of Real News about "Fake News": Intertextual Processes and Political Satire. *International Journal of Public Opinion Research*, *25*(3), 323–343. https://doi.org/10.1093/ijpor/edt015

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature*, *538*, 20–23. https://doi.org/10.1038/538020a

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Cinus, F., Minici, M., Monti, C., & Bonchi, F. (2022). The effect of people recommenders on echo chambers and polarization. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 90–101.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/https://doi.org/10.1002/pra2.2015.145052010082

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/ARXIV.1702.08608

Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). User preference-aware fake news detection. *CoRR*, *abs/2104.12259*.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. *Duke law and technology review*, *16*, 18–84.

ElBoghdady, D. (2013). Market quavers after fake ap tweet says obama was hurt in white house explosions.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Commun. ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Fisher, M., Cox, J. W., & Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in d.c.

Foster, V. S. (2016). The great moon hoax. In *Modern mysteries of the moon: What we still don't know about our lunar companion* (pp. 11–44). Springer International Publishing. https://doi.org/10.1007/978-3-319-22120-5_2

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Han, Y., Karunasekera, S., & Leckie, C. (2020). Graph neural networks with continual learning for fake news detection from social media. https://doi.org/10.48550/ARXIV.2007.03316

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1). https://doi.org/10.1609/aaai.v30i1.10382

Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica*, *47*, 263–291.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lindeman, M., & Aarnio, K. (2007). Superstitious, magical, and paranormal beliefs: An integrative model. *Journal of Research in Personality*, *41*(4), 731–744. https://doi.org/https://doi.org/10.1016/j.jrp.2006.06.009

Lipton, Z. C. (2016). The mythos of model interpretability. https://doi.org/10.48550/ARXIV.1606.03490

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *CoRR*, *abs/1902.06673*.

Mustafaraj, E., & Metaxas, P. T. (2017). The fake news spreading plague: Was it preventable? *CoRR*, *abs/1703.06988*.

Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6149–6157.

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). Reuters institute digital news report 2022. *Digital News Report 2022*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.

Paul, C., & Matthews, M. (2016). The russian "firehose of falsehood" propaganda model: Why it might work and options to counter it.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/https://doi.org/10.1016/j.tics.2021.02.007

Read, M. (2016). Donald trump won because of facebook.

Reed, E. S., Turiel, E., & Brown, T. (2013). Naive realism in everyday life: Implications for social conflict and misunderstanding.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 7–17. https://doi.org/10.18653/v1/W16-0802

Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/https://doi.org/10.1002/pra2.2015.145052010083

Santia, G., & Williams, J. (2018). *Buzzface: A news veracity dataset with facebook user commentary and egos*.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and spatialtemporal information

for studying fake news on social media. https://doi.org/10.48550/ARXIV.1809.01286

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Silverman, C. (2016). This analysis shows how viral fake election news stories outperformed real news on facebook.

Sunstein, C. R. (2001). *Echo chambers: Bush v. gore, impeachment, and beyond*. Princeton University Press.

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures*. *Journal of Political Philosophy*, *17*(2), 202–227. https://doi.org/https://doi.org/10.1111/j.1467-9760.2008.00325.x

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. https://doi.org/10.48550/ARXIV.1704.07506

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Vicario, M. D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Walker, M., & Matsa, K. E. (2021). News consumption across social media in 2021.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *CoRR*, *abs/1705.00648*.

Watson, A. (2022). Usage of social media as a news source worldwide 2022.

Weir, W. (2009). In *History's greatest lies: The startling truths behind world events our history books got wrong* (pp. 28–41). Fair Winds Press.

Zhou, X., Wu, J., & Zafarani, R. (2020). Safe: Similarity-aware multi-modal fake news detection. https://doi.org/10.48550/ARXIV.2003.04981