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. In contrast to its advantages of convenience and speed, social media can spread any kind of information since no regulatory authority checks the news. 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 [1] 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.

When presented as news, false or misleading information is called “fake news” (Pennycook & Rand, 2021). In the social media context, fake news is false and misleading information that is dispersed and seeks to deceive people. (Lazer, Baum, Benkler, et al., 2018). In literature, false and misleading news is also referred to as disinformation, misinformation, or fake news. We distinguish disinformation and fake news from misinformation since misinformation does not seek to deceive people. Rather, it is incorrect information, however unbeknownst to the user who shared it. We will use the term “fake news” instead of “disinformation” throughout this thesis.

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

[1] https://www.statista.com/statistics/718019/social-media-news-source/

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). Moreover, 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 a 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 SOTA 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. For instance, a study in 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. Consequently, 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.

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 talk about the challenges when detecting fake news, both from human and computer perspectives. We make an excursion into human psychology, delivering insights into why humans fall for fake news. Then we talk about the difficulties when creating a fake news detection model.

2.1.1 Overview

The term fake news has existed for over a century (Lazer, Baum, Benkler, et al., 2018). Throughout history, various forms of widespread fake n ews have been recorded. For instance, 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 customers and almost no backlash after the newspaper admitted that the articles mentioned earlier were a hoax [1,2]. It rose to prominence with the 2016 U.S. Presidential Election (Beckwith, 2021). It is clear that fake news has an effect not only on individuals but also on countries. Nowadays, fake news is still at large because it is challenging to detect for humans and computers (Newman, Fletcher, Robertson, et al., 2022).

2.1.2 Challenges

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.

2.3 Explainable AI for FND

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

[2] https://www.history.com/this-day-in-history/the-great-moon-hoax

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

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

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

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

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.

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

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

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*.

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*.

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

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

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

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*.

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