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. For example, a study from Pew Research Center[1] reports that in 2021, 48% of U.S. adults get their news from social media "often" or "sometimes". As another example, global data from 2022[2] 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 statistics indicate that social media is a crucial news source. Still, one should follow the news on social media carefully since there is no regulatory authority to check the news. 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 in early 2022, in the week of the survey, between 45% and 55% of surveyed social media consumers worldwide have witnessed false or misleading information about COVID-19.

The term fake news has existed for over a century (Lazer, Baum, Benkler, et al., 2018). It rose to prominence with the 2016 U.S. Presidential Election (Beckwith, 2021). With its prevalence, the research community has focused on the automatic detection of fake news (Monti, Frasca, Eynard, et al., 2019; Reis, Correia, Murai, et al., 2019; Shu, Mahudeswaran, Wang, et al., 2018; Shu, Sliva, Wang, et al., 2017; Wang, 2017).

1.0.1 Subsection

See Table 1.1, Figure 1.1, Figure 1.2, Figure 1.3.

[1] https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/

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

Table 1.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

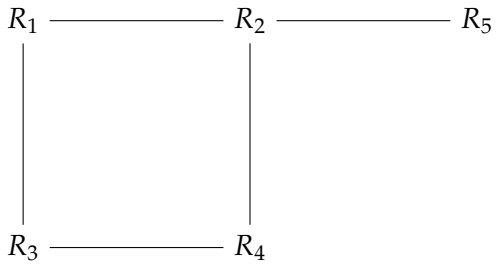

Figure 1.1: An example for a simple drawing.

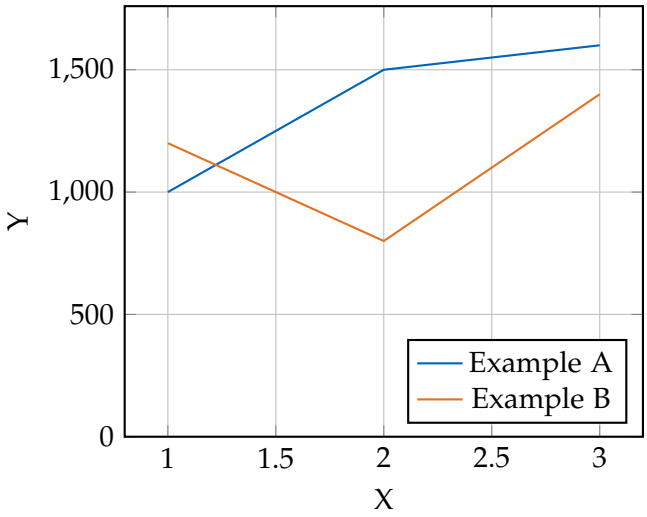

Figure 1.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 1.3: An example for a source code listing.

2 Background and Related Work

2.1 Fake News Detection

2.1.1 Overview

2.1.2 Datasets

2.1.3 Challenges

2.1.4 Evolution of Fake News Detection Models

2.1.5 Current Approaches

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

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 Content and Social Feature 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 Definitions

-> Talk about Graph Neural Networks

3.2.2 Dataset

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

3.2.3 Model

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*.

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

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*.

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

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, *34*(2), 76–81. https://doi.org/10.1109/MIS.2019.2899143

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

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