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

Acknowledgments **iii**

Abstract **iv**

1 Introduction **1**

2 Background and Related Work **3**
2.1 Fake News Detection 3
2.1.1 Fake News 4
2.1.2 Foundations of Fake News 6
2.1.3 Evolution of Fake News Detection 10
2.2 Explainable Artificial Intelligence 12
2.2.1 Importance of Explainable Artificial Intelligence 12
2.2.2 A Good Explanation 12
2.2.3 Overview of Explainable Artificial Intelligence 12

3 Fake News Detection Models **13**
3.1 Content Based Models 13
3.1.1 Definitons 13
3.1.2 Dataset 13
3.1.3 Tokenizer 13
3.1.4 Model 13
3.1.5 Explainability and Explanation 13
3.2 Social Context Based Models 13
3.2.1 Geometric Deep Learning 14
3.2.2 Dataset 14
3.2.3 Models 14

4 Explainability of Fake News Detection Models **15**
4.1 Explanation Techiques 15
4.1.1 SHAP, DeepSHAP 15
4.1.2 GNNExplainer 15
4.1.3 Explainability vs. Explanation 15

4.2 Content Based Fake News Detection Models 15
4.2.1 Explaining Content Based Fake News Detection Models 15
4.2.2 Introducing Unseen Data 15
4.2.3 Results 15
4.3 Content and Social Feature Based Fake News Detection Models 15
4.3.1 Explaining Content and Social Feature Based Fake News Detection Models 15
4.3.2 Introducing Unseen Data 15
4.3.3 Results 15

5 Conclusion **16**

List of Figures **17**

List of Tables **18**

Bibliography **19**

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 et al., 2021; Nakamura et al., 2020; Santia & Williams, 2018; Shu et al., 2017; Tacchini et al., 2017; Wang, 2017), and some studies leveraged the power of *Machine Learning* (ML) to automatically detect fake news (Bian et al., 2020; Han et al., 2020; Monti et al., 2019; X. Zhou et al., 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 et al., 2018), representing the propagation networks as graphs (Dou et al., 2021), and using *Graph Neural Networks* (GNN) to produce *State Of The Art* (SOTA) models (Monti 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 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 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 2.1.3, we first talk about the evolution of detection algorithms, then we classify FND algorithms with respect to their input data type and

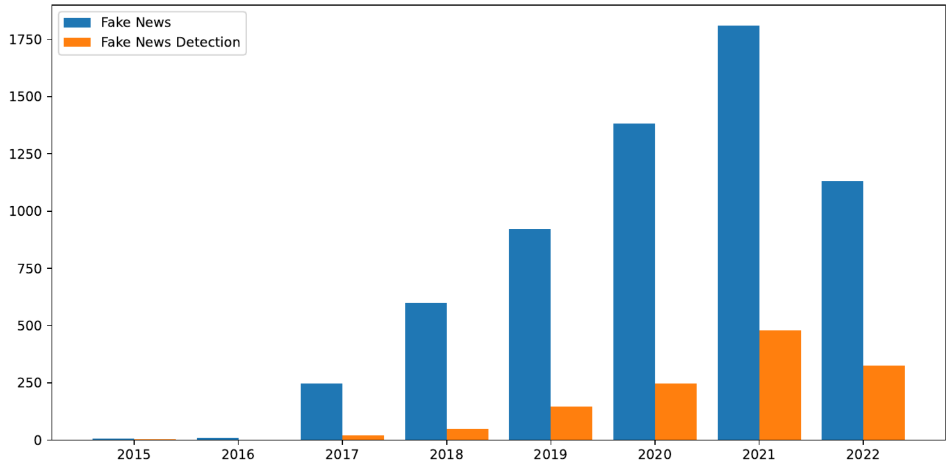

Figure 2.1: Total number of publications that include (1) *Fake News* (blue) and (2) *Fake News Detection* (orange) publications by year. Source: Scopus; Search Arguments: (1) TITLE-ABS-KEY("fake news*") PUBYEAR AFT 2014 (2) TITLE-ABS-KEY("fake news detection")

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 et al., 2017). This definition is widely used in other studies (Conroy et al., 2015; Mustafaraj & Metaxas, 2017; Shu et al., 2017). Furthermore, recent social sciences studies (Lazer 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 et al., 2013; Jin et al., 2016; V. Rubin et al., 2016). Further definitions include hoaxes, satires, and obvious fabrications (V. L. Rubin et al., 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 et al., 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).
Digital News Report 2022 (N. Newman et al., 2022) reports in its key findings that trust

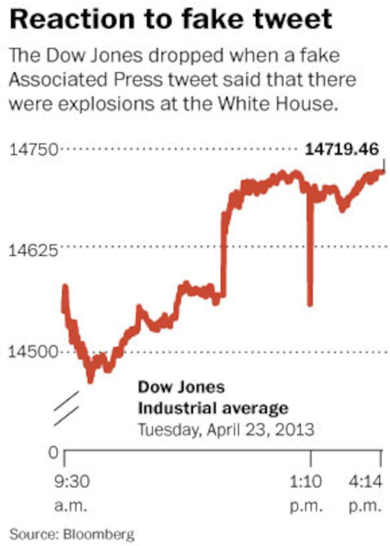

Figure 2.2: The market's reaction to the fake tweet. The sharp decline caused by a single tweet. Image obtained from (ElBoghdady, 2013)

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* and *confirmation bias*, examine why humans fall for fake news. The first refers to a

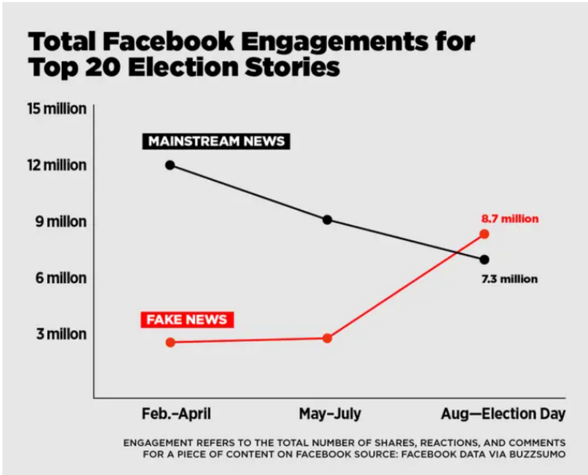

Figure 2.3: The rising engagement for fake news stories observed after May-July, just before Presidential Elections. Image obtained from (Silverman, 2016)

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 et al., 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 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 et al., 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 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 et al., 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.

Data-Oriented Foundations. We define features for news content and social context to represent the news pieces in a structured manner. First, we introduce attributes for news content (Shu et al., 2017):

- *Source*: Publisher of the news piece.
- *Headline*: Short title text that aims to catch the readers' attention and describes the article's main topic.
- *Body Text*: The main text piece that details the news story.
- *Image/Video*: Part of the body content supplies visual input to articulate the story.

Using these attributes, we extract two types of features for news content:

Linguistic-based features: The news content is heavily based on textual content. Thus, the first feature that belongs to this class is lexical features which make use of character and word level frequency information which can be obtained by the utilization of *term frequency-inverse term frequency* (TF-IDF) (Jones, 1972; Luhn, 1957). The second feature is based on syntactic features which include sentence-level features that can be obtained via n-grams and bag-of-words (BoW) and punctuation and parts-of-speech (POS) (Daelemans, 2010) tagging. We can extend these features to domain-specific ones, such as external links and the number of graphs (Potthast et al., 2017).

Visual-based features: Particularly for fake news, the visual content is a strong tool for establishing belief (Dan et al., 2021). Hence, the features that reside in images and videos become significant. Fake images and videos which brings the fake story together are commonly used(e.g. Harding, 2012; Sawer, 2020). To counteract the effects of misleading visual input, recent studies (Qi et al., 2019) examined visual and statistical information for fake news detection. Visual features consist of clarity score, similarity distribution histogram, diversity score, and clustering score. Statistical features are listed as count, image ratio, multi-image ratio etc. (Shu et al., 2017).

Now, we define features for social context, which has recently drawn much attention from the research community (Shu et al., 2020; Shu et al., 2019). Overall, there are three aspects of social context data that we will concern, namely, user-based, post-based, and network-based features.

User-based: As mentioned in the Technical Foundations part of this section, fake news has various ways of disseminating, such as via echo chambers or bots. Therefore, analyzing user-based information can prove useful. We distinguish user-based features at the group and individual levels (Shu et al., 2017). Individual levels are extracted to deduce the credibility of each user by utilizing, for example, the number of followers and followees, the number of tweets authored by a user, etc (Castillo et al., 2011). On the other hand, group-level user-based features are the general characteristics of groups of users related to the news (Yang et al., 2012). Parallel to the echo chambers idea, the assumption is the consumers of real and fake news tend to form different groups, which may lead to unique characteristics. Typical group-level features stem from individual-level features by obtaining the share of verified users, and the average number of followers and followees (Ma et al., 2015).

Post-based:

Network-based:

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 Evolution of Fake News Detection

Fake news detection is as old as fake news itself. Before social media became a hub for news consumers, fact-checkers, i.e., fake news detectors, were only journalists and literate people. Following the source shift of the news from printed paper to online, then social media, detection of fabricated news have become costly, cumbersome, and not as rewarding due to the endless stream of information and decreasing trust in journalism. Automatic detection for news thus became a necessity in our world (Chen et al., 2015).
The initial approaches for automated fake news detection focus on news context and stem from deception detection in language. The first study of detecting deception in language (Undeutsch, 1954) hypothesized that the truthfulness of the statement is more important than the integrity of the reporting person, and there exist definable and descriptive criteria that form a crucial mechanism for the determination of the

truthfulness of statements. Even though this study is from experimental psychology, it stresses the feasibility of defining a set of rules that determine the truthfulness of a statement.

An early study from criminology, Scientific Content Analysis (SCAN) (Sapir, 1987), analyzes freely written statements. In this process, SCAN claims to detect potential instances of deception in the text but cannot label a statement as a lie or truth. The next study for SCAN (Smith, 2001) is the first known study that correlates linguistic features with deceptive behavior using high-stakes data. Similar to SCAN, the consequent studies (Adams, 2002; M. L. Newman et al., 2003) that link linguistic features to deception classify the owner of the statement as truth-teller or liar according to the frequency of deception indicators in the statement.

Although for automated deception detection, defining a methodology is more challenging (DePaulo et al., 1997), early studies have shown that this task is achievable. A detailed study (L. Zhou et al., 2004) makes a structured approach using linguistic-based cues and draws attention to further studies for automating deception detection. In this study, the authors extend linguistic-based cues with complexity, expressivity, informality, and content diversity. Instead of using humans as cue identifiers, authors use *Natural Language Processing* (NLP) techniques, namely an NLP tool called iSkim (L. Zhou et al., 2002), to extract cues automatically. Another study (Burgoon et al., 2003) also focuses on linguistic cue analysis. With a small dataset and employing the C4.5 (Salzberg, 1994) algorithm, the authors reach 60.72% accuracy using 15-fold cross-validation.

Similarly, in (Bachenko et al., 2008), the authors developed a system for automatically identifying 275 truthful or deceitful statements with the use of verbal cues using Classification and Regression Tree (CART) (Breiman et al., 1984). Additionally, the studies (Hancock et al., 2007; V. L. Rubin, 2010) make use of a relatively small dataset and analyze linguistic-based cues. Rubin's series of studies (V. Rubin et al., 2015; V. L. Rubin, 2010; V. L. Rubin & Lukoianova, 2015; V. L. Rubin & Vashchilko, 2012) makes use of Rhetorical Structure Theory (RST) and Vector Space Modeling (VSM). The first captures the coherence of a story using functional relations among meaningful text units and delivers a hierarchical structure for each news story (Mann & Thompson, 1988). The second is the way to represent rhetorical relations in high-dimensional space. The authors used logistic regression as their classifier and reached 63% accuracy. (V. Rubin et al., 2015).

Furthermore, a study from Afroz and colleagues (Afroz et al., 2012) investigates stylistic deception and uses lexical, syntactic, and content-specific features. Lexical features include both character- and word-based features. Syntactic features represent sentence-level style and include frequency of function words from LIWC (Pennebaker et al., 2007), punctuation and *parts-of-speech* (POS) tagging in which a text is assigned its morphosyntactic category (Daelemans, 2010). Finally, content-specific features are keywords for a specific topic. For classification, the authors then leveraged Support Vector Machines (SVM) (Hearst et al., 1998).

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

2.1 Fake News and Fake News Detection Publications by Year 4
2.2 Market Reaction to Fake Tweet 6
2.3 Total Facebook Engagements for Top 20 Election Stories 7

List of Tables

Bibliography

Adams, S. H. (2002). Communication under stress: Indicators of veracity and deception in written narratives.

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. *2012 IEEE Symposium on Security and Privacy*, 461–475. https://doi.org/10.1109/SP.2012.34

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.

Bachenko, J., Fitzpatrick, E., & Schonwetter, M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 41–48.

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

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis.

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

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. In H. Chen, R. Miranda, D. D. Zeng, C. Demchak, J. Schroeder, & T. Madhusudan (Eds.), *Intelligence and security informatics* (pp. 91–101). Springer Berlin Heidelberg.

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

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675–684. https://doi.org/10.1145/1963405.1963500

Chen, Y., Conroy, N. K., & Rubin, V. L. (2015). News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/https://doi.org/10.1002/pra2.2015.145052010081

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

Daelemans, W. (2010). Pos tagging. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 776–779). Springer US. https://doi.org/10.1007/978-0-387-30164-8_643

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual mis- and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 641–664. https://doi.org/10.1177/10776990211035395

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception [PMID: 15661668]. *Personality and Social Psychology Review*, *1*(4), 346–357. https://doi.org/10.1207/s15327957pspr0104_5

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

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1–23. https://doi.org/10.1080/01638530701739181

Harding, L. (2012). Putin seen behind bars in spoof video.

Hearst, M., Dumais, S., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

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

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

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

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309–317. https://doi.org/10.1147/rd.14.0309

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754. https://doi.org/10.1145/2806416.2806607

Mann, W. C., & Thompson, S. A. (1988). *Text - Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281. https://doi.org/doi:10.1515/text.1.1988.8.3.243

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, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles [PMID: 15272998]. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

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.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count (liwc2007).

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

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *CoRR*, *abs/1702.05638*.

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. *CoRR*, *abs/1908.04472*.

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. (2015). Towards news verification: Deception detection methods for news discourse. https://doi.org/10.13140/2.1.4822.8166

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. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the 73rd ASIST Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*.

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

Rubin, V. L., & Lukoianova, T. (2015). Truth and deception at the rhetorical structure level. *J. Assoc. Inf. Sci. Technol.*, *66*(5), 905–917. https://doi.org/10.1002/asi.23216

Rubin, V. L., & Vashchilko, T. (2012). Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 97–106.

Salzberg, S. L. (1994). C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, *16*(3), 235–240. https://doi.org/10.1007/BF00993309

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

Sapir, A. (1987). *Scientific content analysis (SCAN)*. Laboratory of Scientific Interrogation.

Sawer, P. (2020). 'deepfake' queen's speech: Channel 4 criticised for 'disrespectful' christmas message.

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., Mahudeswaran, D., Wang, S., & 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.

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

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320. https://doi.org/10.1145/3289600.3290994

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

Smith, N. (2001). *Reading between the lines: An evauluation of the scientific content analysis techniques (scan)*. Home Office.

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.

Undeutsch, U. (1954). *Die entwicklung der gerichtspsychologischen gutachtertaetigkeit*. Hogrefe.

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.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. https://doi.org/10.1145/2350190.2350203

Zhou, L., Booker, Q., & Zhang, D. (2002). Rod - toward rapid ontology development for underdeveloped domains. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, 957–965. https://doi.org/10.1109/HICSS.2002.994046

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, *13*(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

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