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 4
2.1 Fake News Detection . . . 4
2.1.1 Fake News . . . 5
2.1.2 Foundations of Fake News . . . 7
2.1.3 Evolution of Fake News Detection . . . 12
2.2 Explainable Artificial Intelligence . . . 20
2.2.1 Foundations of Explainable Artificial Intelligence . . . 21
2.2.2 What Makes A Good Explanation . . . 24
2.2.3 Overview of Techniques in Explainable Artificial Intelligence . . . 28

3 Fake News Detection Models 31
3.1 News Content Models . . . 31
3.1.1 Notation and Definitions . . . 31
3.1.2 Transformer Architecture . . . 36
3.1.3 Model, Dataset, Tokenizer Analysis . . . 40
3.2 Social Context Models . . . 44
3.2.1 Overview of Graphs . . . 44
3.2.2 Graph Neural Networks . . . 46
3.2.3 Dataset . . . 46
3.2.4 Models . . . 46
3.3 Early Fake News Detection and Model Aging . . . 46

4 Explainability of Fake News Detection Models 47
4.1 Explainability of News Content Models . . . 47
4.1.1 SHAP, DeepSHAP . . . 47
4.1.2 SHAP in Action . . . 47
4.1.3 Introducing Unseen Data . . . 47

4.1.4 Results 47
4.2 Explainability of Social Context Models 47
4.2.1 GNNExplainer 47
4.2.2 GNNExplainer in Action 47
4.2.3 Introducing Unseen Data 47
4.2.4 Results 47

5 Conclusion **48**

List of Figures **49**

List of Tables **50**

Bibliography **51**

1 Introduction

With the rapid development of communication technologies, social media has become one of the most frequently used news sources. 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 got 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* (DNNs), 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 (Z. C. 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* (GNNs) 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.

In the next section, we elaborate on fake news, FND methods and xAI. We give foundations of fake news and define its characteristics. Then we categorize FND models and give important examples from literature. After examining fake news and detection methods for it, we focus on characterization of xAI, give definitions that will be used throughout this thesis.
In the third section, we examine FND models that were used in this thesis, and characterize them as defined in 2.1.3. Furthermore, we deliver information about the model architecture, technologies and datasets used, and a detailed mathematical background on DNNs and GNNs. We also draw attention to some crucial matters such as model aging.
In the fourth section, we focus on xAI techniques that are used in this thesis. We give a comprehensive explanations and show their importance when dealing with complex models. We illustrate results from our experiments, draw attention to shortcomings of models, and show a model is fair or not. We also discuss the plausability of the

produced explanantions in this section.
In the last section, we talk about our overall findings. We show that research objectives are satisfied. Additionally, we discuss the limitations that we have encountered, and possible future works.

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 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 strategies.
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). Nevertheless, 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 social, technical, and data-oriented foundations of fake news.
We then list the available datasets used in FND and deliberate their advantages and disadvantages in 2.1.3. Finally, in 2.1.3, we summarize the evolution of detection

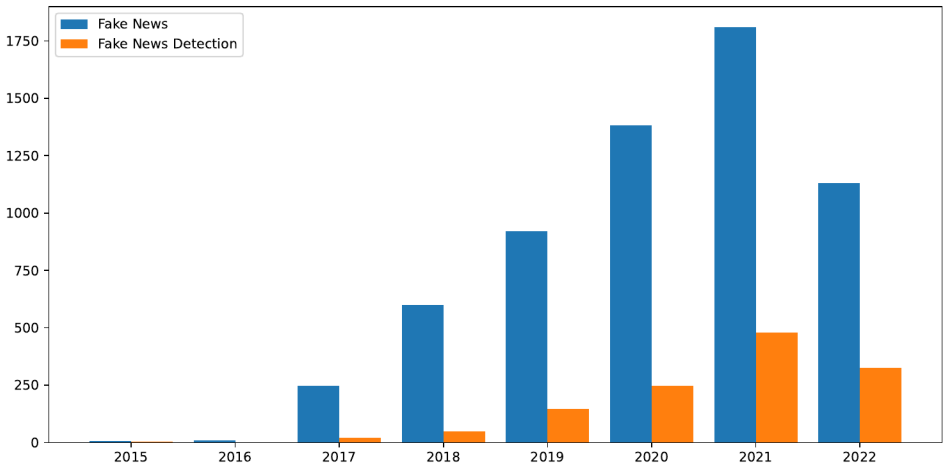


Figure 2.1: Total number of publications that include (1) *Fake News* (blue) and (2) *Fake News Detection* (orange) publications by year. Source: Scopus; Search Arguments: (1) TITLE-ABS-KEY("fake news*") PUBYEAR AFT 2014 (2) TITLE-ABS-KEY("fake news detection")

algorithms, then we classify FND algorithms with respect to their input data type and 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 battle's outcome 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 customers and 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 introduce 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."* (Pennycook & Rand, 2021)

[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."* (Pennycook & Rand, 2021)

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.

Definition 2.1.3 (*Fake News*)**.** *"News articles that are intentionally or verifiably false."* (Allcott & Gentzkow, 2017)

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).

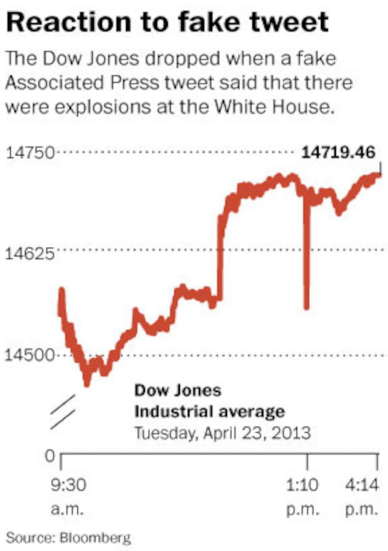


Figure 2.2: The market's reaction to the fake tweet. The sharp decline caused by a single tweet. Image obtained from (ElBoghdady, 2013)

Digital News Report 2022 (N. Newman 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* and *confirmation bias*, examine why humans fall for fake news. The first refers to a person's

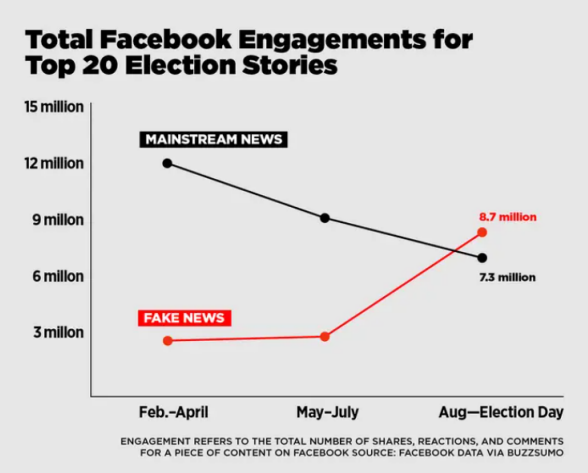


Figure 2.3: The rising engagement for fake news stories observed after May-July, just before Presidential Elections. Image obtained from (Silverman, 2016)

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) or bag-of-words (BoW). The second feature is based on syntactic features which include sentence-level features that can be obtained via *n-grams* 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, we will concern three aspects of social context data: user-based, post-based, and network-based features.

User-based: As mentioned in the Technical Foundations part of this subsection, fake news has various ways of disseminating, such as via echo chambers, malicious accounts, 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 social identity theory and normative social influence 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: Analysis of reactions by users can prove helpful when determining whether a news piece is real or not. For example, if a news piece is getting doubtful comments, this can help determine the news piece's credibility. As such, post-based features are based on inferring the integrity of a news piece from three levels. Namely, post-level, group-level, and temporal-level (Shu et al., 2017). Post-level features can be embedding values for each post or take forms as mentioned in linguistic-based features, e.g., n-grams, BoW, etc. For post-level features, we can also consider general approaches such as topic extraction (e.g., using latent Dirichlet allocation (LDA) (Blei et al., 2003)), stance extraction, which provides information about users' opinions (e.g., supports, opposes (Jin et al., 2016)), and finally credibility extraction, which deals with estimating the degree of trust for each post (Castillo et al., 2011). Group-level post-based features collect feature values for all relevant posts and apply an operation to extract pooled information. When determining the credibility of news, group-level features proved to be helpful (Jin et al., 2016). Temporal-level features deal with changes in post-level features over time. Typically, unsupervised learning methods such as Recurrent Neural Networks (RNN) are employed to capture the changes over time (Ma et al., 2016).

Network-based: As discussed in the Technical Foundations part, fake news is likely to give rise to echo chambers, which leads to the idea of a network-based approach. When represented as networks, the propagation behavior of fake news can be analyzed further, and patterns can be discovered (Shu et al., 2017). In literature, various types of networks exist, the most common ones are stance networks, occurrence networks, and friendship networks. Stance networks are constructed

upon stance detections which is a part of sentiment analysis and deal with determining a user's viewpoint using text and social data (Du et al., 2017). Using all users' stances, a network is built in which the nodes are the tweets relevant to the news piece and the edges represent the similarity of stances between nodes (Jin et al., 2016; Tacchini et al., 2017). On the other hand, occurrence networks leverage the frequency of mentions or replies about the same news piece (Kwon et al., 2013). Friendship networks are based on the follower/followee relationship of users who share posts connected to the news piece. Derived from friendship networks, in the form of one of the datasets we use in our experiments (Dou et al., 2021), diffusion networks are designed to track the course of the dissemination of news (Kwon et al., 2013). Briefly, a diffusion network consists of nodes representing users and diffusion paths representing the relationship and interaction between users. In detail, a diffusion path between two users u_i and u_j exists if and only if u_j follows u_i, and u_j shares a post about a news piece that u_i has already shared a post about (Shu et al., 2017). It has been shown that characterizing these networks is possible (Kwon et al., 2013). Approaches for these networks have gained traction recently, especially with some SOTA GNNs, e.g., (Monti et al., 2019).

To conclude this subsection, we have covered psychological, social, technical, and data-oriented foundations in this section. We established that, from different aspects, there are various reasons for the dissemination of fake news. Accordingly, we consider these reasons when building FND systems. In the next section, we discuss FND approaches and how they have evolved. Moreover, we characterize FND models and talk about each type of approach.

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).
Similar to what we did in the Data-Oriented Foundations part of the previous subsection, we classify fake news detection models as *News Content Models* and *Social Context Models* (see 2.4) and start with News Content Models by following the classification principles in (Shu et al., 2017).

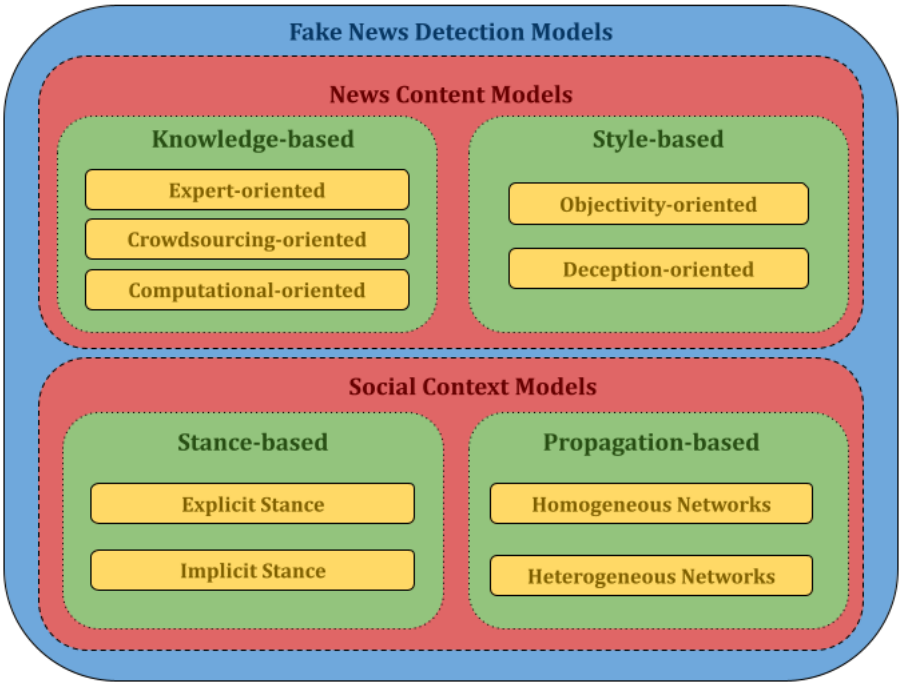


Figure 2.4: Characterization of Fake News Detection Models, Figure inspired by Figure 1 in Shu et al., 2017.

News Content Models. Based on news content and fact-checking methodologies, these models are the starting point of fake news detection. News content models are classified as Knowledge-based and Style-based. We first introduce style-based models as they are the initial approaches for FND.

Style-based: Previous research in psychology has mainly focused on style-based approaches to detect *manipulators* in the text. Particularly deception detection techniques were popular and commonly developed in early works in criminology and psychology. We describe two different ways to approach style-based news content models, namely, *Deception-oriented* and *Objectivity-oriented* (Shu et al., 2017).

- *Deception-oriented*: The initial approaches for automated fake news detection focus on news context and stem from deception detection in language. The first study that focuses deception detection 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 subsequent 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 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 utilized logistic regression as their classifier and reached 63% accuracy.
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 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). More comprehensive and modern approaches such as (Wang, 2017) also leveraged the power of *Convolutional Neural Networks* (CNNs) to determine the veracity of news.

- *Objectivity-oriented*: Objectivity-oriented news content models aim to detect indicators of the lessening of objectivity in news content (Shu et al., 2017). These indicators are observed in the news from misleading sources, such as hyperpartisan sources which display highly polarized opinions in favor of or against a particular political party. Consequently, this polarized behavior motivates the fabrication of news that supports the sources' political views or undermines the opposing political party. *Hyperpartisan news* are a subtle form of fake news and defined as misleading coverage of events that did actually occur with a strong partisan bias (Pennycook & Rand, 2019). Since the spread of hyperpartisan news can be detrimental, many approaches to detect hyperpartisanship in news articles have been developed. For instance, in (Potthast et al., 2017), the authors take a stylometric methodology to

detect hyperpartisan news. In this study, the authors employ 10 readability scores, and dictionary features where each feature represent the frequency of words from a carefully crafted dictionary in a given document with the help of General Inquirer Dictionaries (Stone et al., 1966). A competition for detecting hyperpartisan news (Kiesel et al., 2019) hosted several teams with a variety of ideas which include the utilization of n-grams, word embeddings, stylometry, sentiment analysis etc. The most popular method was the usage of embeddings, particularly the models that leveraged BERT (Devlin et al., 2018).

Also used for dissemination of hyperpartisan news (Kiesel et al., 2019), another form of fake news that is evaluated under this focus is *Yellow-journalism*, which utilizes clickbaits such as catchy headlines, images etc. that invokes strong emotions, and it aims to generate revenue (Agrawal, 2016; Palau-Sampio, 2016). Studies that aim to detect clickbaits mainly focus on headlines. For example, in (Rony et al., 2017), the authors construct a DNN in which they use distributed subword embeddings (Bojanowski et al., 2016; Joulin et al., 2016) as features with an extension of skip-gram model (Mikolov, Sutskever, et al., 2013).

Knowledge-based: Being the most direct way of detecting fake news, these approaches make use of external fact-checkers to verify the claims in news content (Shu et al., 2017). Fact-checkers are either sophisticated algorithms, domain experts or crowdsourced to assess the truthfulness of a claim in a specific context (Vlachos & Riedel, 2014). With growing attention on fake news detection, automated fact-checking has drawn much attention, and considerable efforts have been made in this area (Barrón-Cedeño et al., 2020; Thorne & Vlachos, 2018). We categorize knowledge-based news content models as *Expert-oriented*, *Crowdsourcing-oriented*, and *Computational-oriented*.

- *Expert-oriented*: These approaches are essentially dependent on human domain experts who investigate the integrity of a news piece collecting relevant information and documents to come up with a decision about the truthfulness of a claim [1]. Platforms like Politifact [2] and EUfactcheck [3] are examples for expert-oriented fact-checking for all news from a variety of sources. These platforms label news in a range such that the label reflects the veracity of the news. A different approach for labeling is exercised by Snopes [4], which

[1] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/
[2] https://www.politifact.com/
[3] https://eufactcheck.eu/
[4] https://www.snopes.com/

extends the same logic of Politifact by including different aspects of fact-checking such as Scam, Miscaptioned, Outdated, etc [1]. Recently replaced by an irrelevant magazine website, another instance was Gossipcop [2], which dealt with celebrity fact-checking and contributed to the creation of fake news dataset (Shu et al., 2018). Even though expert-based fact-checking is reliable, with the increasing magnitude of news stream and speed of spread, it is not scalable to fact-check every news piece by hand; thus, manual validation alone becomes insufficient (Guo et al., 2022).

- *Crowdsourcing-oriented*: Powered by the wisdom of crowds (Galton, 1907), crowdsourcing-oriented fact-checking is a collection of annotations that are afterward aggregated to obtain an overall result indicating the veracity of the news. Unlike professional fact-checkers, who are in short supply, this approach is scalable given that the crowd contains enough literate people (Allen et al., 2021). For instance, Twitter launched a program called Birdwatch [3], in which the users are able to leave notes for tweets that they think contain misinformation. Furthermore, this tool allows users to rate each other's notes, leading to the diversity of perspectives [4]. Another example is from Facebook [5], which uses a third party of crowdsourced fact-checkers called International Fact-Checking Network [6] (IFCN).
- *Computational-oriented*: Heavily dependent on external sources, computational-oriented models are scalable, automated systems that are designed to predict whether a claim is truthful or not. The studies that focus on this type of approach mainly try to solve two issues: (i) identifying check-worthy claims and (ii) estimating the integrity of claims (Shu et al., 2017). The first issue requires the extraction of factual claims from news content or other related textual content. For example, in (Hassan et al., 2015), the authors collect presidential debate transcripts, then label them into three classes with the help of crowdsourcing. Using annotated data and supervised learning techniques, the authors uncover some interesting patterns in these transcripts. Another study that covers both issues uses Wikipedia information to generate factual claims and then check whether a given claim is truthful or not (Thorne et al., 2018). The second issue, compared to the first one, requires the utilization of structured external sources. *Open web* and structured *knowledge graphs* are the

[1] https://www.snopes.com/fact-check-ratings/
[2] https://web.archive.org/web/20190807002653/https://www.gossipcop.com/about/
[3] https://twitter.github.io/birdwatch/overview/
[4] https://twitter.github.io/birdwatch/diversity-of-perspectives/
[5] https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works
[6] https://www.poynter.org/ifcn/

two most prominent tools when tackling this issue. Open web tools analyze features like mutual information statistics (Etzioni et al., 2005), frequency, and web-based statistics (Magdy & Wanas, 2010). On the other hand, knowledge graphs are interconnected. One noteworthy example is ontologies such as DBPedia (Auer et al., 2007), using which one can define semantic relations and rules in order to infer whether a claim is correct (Braşoveanu & Andonie, 2019).

Social Context Models. The interconnected design of social media can be leveraged by extracting user-based, post-based, and network-based features and utilizing these features to supplement news content models. Social context models exploit related user engagements for a news piece by capturing this external information from multiple angles. Two types of social context models are prominent: *Stance-based* and *Propagation-based* (Shu et al., 2017).

- *Stance-based*: Given a news piece, these approaches estimate the user's stance toward a specific news topic. More formally, stance detection in social media deals with users' viewpoints toward particular topics by means of various aspects related to users' posts and characteristic traits (ALDayel & Magdy, 2021). The user's stance information can be extracted either implicitly or explicitly. Implicit stance can be automatically obtained from social media posts with the help of NLP tools such as sentiment analysis (Mohammad et al., 2016). Explicit stances are rather easier to obtain since they are direct expressions of opinions or emotions. For example, "like" on Twitter or Facebook, "upvote" and "downvote" ratings on Reddit, and "like" and "dislike" on Youtube are explicit stances of users. A study that utilizes explicit stances called Some Like it Hoax, uses logistic regression and harmonic boolean label crowdsourcing for classification on a dataset they curated from Facebook. In the stance classification process, they consider the likes and the issuer of likes for each post. They state that logistic regression comes short in this task since it cannot learn anything about posts without likes (Tacchini et al., 2017). An early example of implicit stance detection leverages the dialogic relations between authors by constructing graphs that represent the interaction between authors. They show that this information can improve the performance of stance-detection models (Walker et al., 2012). A more detailed study (Addawood et al., 2017) investigates stance classification considering lexical, syntactic, twitter-specific and argumentation feature types. Although some twitter-specific features can be considered as explicit stances, such as if the tweet is a retweet, if a tweet contains the title to an article, if a tweet contains a hashtag, etc., those features are later aggregated before it is fed to the classifier. The authors reach the highest F1 score using lexical and argumentation features. In literature, there

are also implicit stance-based approaches that aim to detect the veracity of a news piece by exploring the relationship between a headline and the article (Ghanem et al., 2018; Hanselowski et al., 2018).
Another variation of stance-based detection is rumor detection. One example of a rumor detection model is a Bayes classifier that utilizes content-based, network-based, and twitter-specific meme features through *Information Retrieval* (IR) techniques (Qazvinian et al., 2011). In this study, the authors propose a general framework that leverages statistical models and maximizes a linear function of log-likelihood ratios to retrieve rumorous tweets. They show that the features they used contribute to their model's overall performance.

- *Propagation-based*: Inspired by the assumption that the veracity of a news event is highly correlated with the credibilities of related social media posts, propagation-based models employ the interrelations of related social media posts to classify a news piece as truthful or not (Shu et al., 2017). These models can be based on either *homogeneous networks* or *heterogeneous networks*. Homogeneous networks are built upon a single type of entity, such as a post or event (Jin et al., 2016). A study by Jin et al., 2016 created homogeneous credibility networks for each topic which is extracted using an unbalanced version of the Joint Topic Viewpoint Model (Trabelsi & Zaiane, 2014). These credibility networks consist of nodes as tweets and edges as links, defined by either supporting or opposing. On the other hand, heterogeneous networks can contain multiple types of entities such as events, sub-events, posts, comments, etc. For example, in (Jin et al., 2014), the authors build a hierarchical propagation graph that contains events, sub-events, and messages from parent to child, respectively. Using an iterative method, they provide a globally optimal solution for the graph optimization problem in the study.
  Furthermore, an interesting study from a decade ago based its credibility estimation algorithm on PageRank and similarity scores. Their propagation network consists of graphs in which possible nodes are events, tweets that were posted about that event, and users who posted those tweets. A more recent study, which we also use in this thesis, is *User Preference-Aware Fake News Dataset* (UPFD) (Dou et al., 2021). This dataset houses two different datasets, one from Politifact and one from Gossipcop. Its root nodes are news pieces, the child of the root node are the users who retweeted the news piece, and the children of the child node are the users who are assumed to have retweeted the news piece after its parent in terms of time. The authors use news content and social engagement information to construct the graph. The best-performing model is based on news and social context. It uses GraphSAGE (Hamilton et al., 2017) as graph encoder and

BERT (Devlin et al., 2018) as the text encoder and reaches 84.62% and 97.23% accuracy on Politifact and Gossipcop, respectively.

We examined two types of FND models, namely, news content and social context models. For each type, we further categorized then defined each type of model, and we gave examples for each. It is crucial to note that approaches are not necessarily purely news content or social context-based; they can be based on both news content and social context. For instance, like the example we gave in propagation-based social context models, GraphSAGE, there are models such as GNN-CL (Han et al., 2020) or GCNFN (Monti et al., 2019), which are baseline models for UPFD (Dou et al., 2021) and will be discussed in detail in the next section.
To summarize this section, we have introduced the history and definitions of fake news in subsection 2.1.1. Then, we investigated the foundations of fake news and gave motivations for developing automated FND systems in subsection 2.1.2. Following that, we examined the evolution and characterized FND models in section 2.1.3. We have included at least two examples for each type of model and briefly summarized their approaches. We also briefly examined one of the datasets (Dou et al., 2021) and models (Hamilton et al., 2017) used in this thesis; however, in-depth information will be provided in the next chapter.
In 2.2, we elaborate on the techniques available in explainable artificial intelligence. We discuss the qualities of a reasonable explanation, and we highlight the importance of the interpretability of a model. We give essential definitions that will be used throughout this thesis.

2.2 Explainable Artificial Intelligence

Understanding and interpreting a model's prediction is very important nowadays since this understanding allows to validate the reasoning of the model and extract rich information for a human expert, and can lead to increased trust in the model (Ribeiro et al., 2016). Furthermore, explanation of a model can help to improve the model (Lundberg & Lee, 2017) and alleviate concerns raised by Ethical AI (Angwin et al., 2016; Goodman & Flaxman, 2017). In this section, we introduce the background for XAI techniques that were used in this thesis. First, in 2.2.1 we characterize XAI by following works from Z. C. Lipton, 2016 and Barredo Arrieta et al., 2020 and give definitions to clarify the taxonomy. Then, in 2.2.2, we discuss the properties of good explanations, the goals of XAI and the evaluation techniques for explanation ethods. Finally, in 2.2.3 we briefly lay out the most frequently mentioned explanation methods in the literature, along with the ones we use in this thesis. We summarize each of them and cover explanation techniques offered to any kind of neural network.

2.2.1 Foundations of Explainable Artificial Intelligence

Initial AI methods were not sophisticated enough to require additional explanation schemes. In the last years, expanding applications of DNNs have led to the adoption of these opaque systems even more. Although empirically successful thanks to enormous parameter spaces and numerous layers, DNNs are complex *black-box* models in terms of interpretability (Castelvecchi, 2016).
In the XAI context, *black-box* or *opaque* models are considered to be the opposite of *transparent* because they require a further search to understand their inner workings (Z. C. Lipton, 2016). Accordingly, humans hesitate to use systems that are not directly interpretable and reliable (J. Zhu et al., 2018), making *interpretability* essential. Moreover, from a legal perspective, the notion *right to explanation* brings more attention to interpretabilty (Z. C. Lipton, 2016). Particularly in situations such as when:

- The prediction of AI directly affects human life, e.g., fully autonomous cars in traffic, medical AI assistants etc.
- The reasons behind an AI system's decision can not be clearly determined.

With the additional demand from the Ethical AI field (Goodman & Flaxman, 2017), the research community has put in a great amount of effort to gap the bridge between a black-box model and its interpretability. However, the lack of consensus on taxonomy has led to synonymous usages of interpretability. The early definitions for interpretability were too broad, describing it as essentially an additional design driver when building a model (Kim, Rudin, et al., 2015) or a requirement for *trust* (Kim, 2015). But can trust be defined in an objective way? Is the accuracy or F1 score of the model is enough to trust a model? To answer the first question, Z. C. Lipton, 2016 argues that trust is subjective and it is not technically defined. To answer the second question, taking only the performance metrics as a baseline for trust in the model is shown to be an incorrect approach, particularly studies that analyze models with *adversarial examples* (Liang et al., 2021; Yuan et al., 2017). Moreover, Doshi-Velez and Kim, 2017 argues that the need for interpretability comes from the *incompleteness* of the problem formalization.
Instead of trying to find a technical definition for interpretability, we can categorize existing systems in terms of their transparency. Z. C. Lipton, 2016 states two properties for interpretable models: *transparency* and *post-hoc interpretability*. The definition for the first and its related terms are given as in the following,

Definition 2.2.1 (*Understandability*)**.** "*Denotes the characteristic of a model to make a human understand its function - how the model works - without any need for explaining its internal structure or the algorithmic means by which the model processes data internally*" (Barredo Arrieta et al., 2020; Montavon et al., 2018).

Definition 2.2.2 (*Transparency*)**.** "*A model is considered to be transparent if by itself it is understandable*" (Barredo Arrieta et al., 2020).

To elaborate further, we discuss degrees of transparent models as not all models provide the same extent of understandability (Barredo Arrieta et al., 2020). Both in Z. C. Lipton, 2016 and Barredo Arrieta et al., 2020 the categorization is made as: *simulability*, *decomposability*, and *algorithmic transparency*. We discuss each of them briefly.

- *Simulability*: Denotes the model's characteristic to be simulated or thought only by a human. Thus, the complexity of a model plays an important role here. Models that can be presented to a human in terms of text and visualizations are considered interpretable (Ribeiro et al., 2016), and in this case, models this elementary fall into simulatable models category (Barredo Arrieta et al., 2020; Tibshirani, 1996).

- *Decomposability*: Represents the model's characteristic to explain each part of the model. Basically, when all components of a model are simulable, then that model is decomposable (Z. C. Lipton, 2016) given that the inputs are already interpretable (Barredo Arrieta et al., 2020).

- *Algorithmic Transparency*: Deals with the user's comprehension of the input's journey from entering the model to becoming a prediction (Barredo Arrieta et al., 2020; Z. C. Lipton, 2016). For example, linear models can be considered algorihtmically transparent since the user can understand how the model can act in a given situation (James et al., 2013).

The second property of interpretable models, *post-hoc explainability*, aims to improve the interpretability of not readily interpretable models. It does so by means of *text explanations*, *visual explanations*, *local explanations*, *explanations by example*, *explanations by simplification*, and *feature relevance explanations* techniques (Barredo Arrieta et al., 2020; Z. C. Lipton, 2016). In a more general sense, post-hoc explainability methods can be grouped into three categories in terms of the knowledge of the target model, granularity of focus, and form: *model-specific or model-agnostic*, *local or global* and *form*. The first category refers to the explainability method's assumption on the model's structure. *Model-specific* techniques can be utilized with a limited set of models since these techniques make an assumption on the model to be explained. On the other hand, *model-agnostic* techniques are designed in such a way that does not require knowledge about model's inner workings (Barredo Arrieta et al., 2020; Guidotti et al., 2018). The second category denotes the explanation's domain. *Local* explanations reason about a particular prediction of a model at feature level (Doshi-Velez & Kim, 2017) (e.g. compute a saliency map by taking the gradient of the output with respect to a given

input vector (Z. C. Lipton, 2016)), whereas global explanations aim to outline the model's general behavior on the dataset (Barredo Arrieta et al., 2020; Doshi-Velez & Kim, 2017; Guidotti et al., 2018). Global explanations are usually presented in the structure of a series of rules (Lakkaraju et al., 2016). The third and last category, form of the explanation is the manner that is conveyed to the user. We will discuss specific forms of explanation in detail after we give a definition of *explainability*, since from now on we talk about the explainability of a model rather than its interpretability.

Definition 2.2.3 (Explainability)**.** "*Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.*" (Barredo Arrieta et al., 2020; Guidotti et al., 2018)

- *Text explanations* are techniques that learn to produce textual expressions that assist user to understand the outcomes of the model (Bennetot et al., 2019).
- *Visual explanations* are techniques that supplement a model's explainability by visualizing the model's behavior. Due to the mismatch between high-dimensional nature of complex ML systems and the capacity of human reasoning (Burrell, 2016), visual explanations often employ dimensionality reduction practices (Barredo Arrieta et al., 2020).

From the perspective of explainability, one would intuitively prefer transparency since transparent models can be explained with ease. However, some works argue that as transparency of a model increases their performance usually tends to decrease (Dosilovic et al., 2018), although other works argue that this is not necessarily true, particularly in cases where the data is well structured and the quailty and value of available features is outstanding (Rudin, 2018). Furthermore, considering FND models, the need for big and complex models can not be avoided since news pieces tend to be long texts, or social networks are represented as graphs, thus forcing SOTA FND models to utilize complex approaches that decreases transparecy such as word embeddings, data fusion, graph data structure (Dou et al., 2021).
On the other hand, although global explanations can be helpful to domain experts by providing information about what model has learnt on a global level, it might be difficult to obtain (Z. C. Lipton, 2016). Instead, local explanation methods are more easier to obtain and more practical for real-world applications. For example, if a user requests an explanation for a prediction, local explanations can provide it, which also complies with "right to explanation" (Goodman & Flaxman, 2017).
Depending on the model adopted for FND, we might be required to use model-specific or model-agnostic approaches. For instance, when dealing with a GNN,

model-agnostic approaches do not provide easily interpretable explanations, thus requiring a model-specific explanation method. On the other hand, when dealing with a DNN model-agnostic post-hoc approaches are usually the choice (Barredo Arrieta et al., 2020). Therefore, for FND models used in this thesis, we are required to adopt both model-agnostic and model-specific post-hoc approaches. Below we list types of post-hoc approach as listen in (Barredo Arrieta et al., 2020).

- *Explanations by example*, a method suggested by Caruana et al., 1999, focus on obtaining representative information from a model by providing explanations for an example that sufficiently illustrates the inner workings of the model (Baehrens et al., 2010; Barredo Arrieta et al., 2020).
- *Explanations by simplification* techniques construct a new and simplified system to provide explanations for a model. These simplified systems keep the performance of the original model while displaying less complexity (Barredo Arrieta et al., 2020).
- *Feature relevance explanations* compute feature relevance scores for a model's variables in order to determine the effect of a feature has upon a model (Barredo Arrieta et al., 2020).

We discussed what kind of explanation methods we can adopt and how these methods can shape the design of a model and the forms of explanations that aims to convey information about the model's behavior. It is possible to see a combination of the previously mentioned explanation forms. In order to present the user with an comprehensible explanation, we characterize a good explanation and define its important properties in the next subsection.

2.2.2 What Makes A Good Explanation

In literature, the requirements for a *good explanation* are rigorously researched. However, there is not a clear definition of how an explanation should look like or convey to the user. It is challenging to objectively define what makes a good explanation (Barredo Arrieta et al., 2020). In order to tackle the issue of subjectivity of explanations, XAI draws wisdom from social and cognitive sciences. A comprehensive study on social sciences and XAI by Miller, 2017, analyzes explanations in terms of the content, the explainer and the explainee. The author argues that the research in AI is lacking knowledge about the properties and structure of an explanation. The major findings on how a good explanation should be are outlined below.

- Explanations are *contrastive* (P. Lipton, 1990; Miller et al., 2017). When presented with counterfactual explanations, users can understand the decision made by

the model easier (Byrne, 2019; Escalante et al., 2018; Lopez-Paz et al., 2016). For example, rather than asking why event A occurred, we ask why event A occurred instead of an event B (Barredo Arrieta et al., 2020; Miller, 2017).

- Explanations are *selective*. Presenting all the causes for an event to a human is pointless, since humans are inclined to select a couple main causes out of numerous, sometimes countless, causes as shown in (Herlocker et al., 2000). Accordingly, Miller, 2017 argues that this selection process is shaped by specific cognitive biases.

- Explanations are *social*. They are conveyed from explainer to explainee via a social interaction. Hence, explanations are transferred through the frame of explainer's beliefs about the explainee's beliefs (Miller, 2017).

- Probabilities probably don't matter. Even though probabilities matter when creating the explanations, the usage of these statistical relations in explanations is not as effective as that of causes. If the underlying causal explanation is not included, then the utilization of statistical generalisations is not sufficient (Miller, 2017).

It should be noted that the characteristics of a good explanation is not limited to the ones mentioned above. These are the most prominent characteristics of numerous which is discussed in (Miller, 2017) detail. An important aspect here is that the explanations are provided to an *explainee* who is the *audience* in (Barredo Arrieta et al., 2020), which refers to the person receving the explanation. It is further noted in (Barredo Arrieta et al., 2020) that explanations are dependent on the audience, i.e., an explanation meant for an end-user will not be enough for a domain expert or an explanation for a domain expert might be too complicated for an end-user. Also, we will refer to the expainee as *audience* from now on.
Main target audience is the main driver when considering the needed outcomes for an explanation. Barredo Arrieta et al., 2020 summarizes the pursued goals when trying to attain explainability. These are listed below.

- *Trustworthiness* deals with the assurance of a model's intended behavior when the model is presented with real-world scenarios (Antunes et al., 2008; Z. C. Lipton, 2016). Some studies highlight the importance of *trustworthiness* as a requirement for explainability (Kim, Glassman, et al., 2015; Ribeiro et al., 2016). The main target audience for this goal is domain experts, and users affected by the model decisions (Barredo Arrieta et al., 2020).

- *Confidence* refers to the robustness and stability of a model (Barredo Arrieta et al., 2020). Yu, 2013 argues that stability is a prerequisite when obtaining explanations

from a model. Moreover, (Barredo Arrieta et al., 2020) argues that trustworthy explanations should not be obtained from unstable models. The audience relevant for this goal are domain expers, developers, managers, and regulatory entities.

- *Fairness* refers to a model's potential to ensure a fair prediction for a user affected by the model's prediction on the basis of characteristics such as age, race, gender, etc (Barredo Arrieta et al., 2020; Oneto & Chiappa, 2020). The audience for this goal consists of users affected by model decisions and regulatory entities.
- *Transferability* refers to the model's capability to perform on unseen data. It is a desired goal to have not just for explainability but also for obtaining a good performance from the model (Kuhn & Johnson, 2013). The audience for this goal is domain experts and data scientists.
- *Causality* denotes the causal relationships between variables of a model (Pearl, 2009). It aims to provide causal information among the data variables. Its main audience is domain experts, managers, and regulatory entities (Barredo Arrieta et al., 2020).
- *Informativeness* is a goal meant for all users and it deals with the information provided by the model. In order to fill the gap between user's decision and the prediction of a model, a massive amount of information about the problem at hand needs to be conveyed to the end user (Barredo Arrieta et al., 2020).
- *Accessibility* refers to the possibility of end users getting more involved in a model's development or improvement (Miller et al., 2017). The audience for this goal includes product owners, managers, and user affected by model decisions.
- *Interactivity* denotes a model's capability to interact with the user (Kim, 2015). The main audience consists of domain experts and users affected model decisions.
- *Privacy awareness* is a goal not frequently seen in the literature. It deals with the learnings of a model's internal representation such that these learnings might pose a privacy breach. From the opposite perspective, it is a differential privacy breach when an unauthorized third party can explain the inner workings if a trained model (Barredo Arrieta et al., 2020).

On the other hand, for a given explanation and its preferred characteristics, how do we objectively evaluate an explanation? For example, considering a model, the evaluation metrics obtained from the test set reflect the model's overall performance on unseen data and allow to compare different models that use the same dataset (Olson et al., 2017). For example, metrics like accuracy, F1 score, recall, and precision are often used in the

evaluation of models. Given that there are numerous metrics, it should be noted that different domains and models may require different evaluation metrics (Handelman et al., 2019; Hossin & Sulaiman, 2015; McNee et al., 2006). For a comprehensive study on the evaluation metrics of ML models, please refer to (Japkowicz & Shah, 2011). Similar to models, explanations require evaluation methods that can quantify their performance. So far, we have seen that explanations might have different audiences, they can take several forms, and they have desired properties. Therefore, like models, there should be a set of explanation evaluation methods which focus on different categories of explanation approaches. In fact, a rigorous study by Doshi-Velez and Kim, 2017 lays out the categorization of explanation evalution approaches. The authors split the evaluation methodologies into three:

1. *Application-grounded evaluations* involve conducting experiments on real humans who are domain experts interacting with explanations in a real-world application. This kind of evaluation directly tests the objective of the system, thus, attaining high performance with respect to application-grounded evaluation suggests good evidence of the explanation's success. The fact that we need humans interacting with a real-world application in an environment which can be observed for experimentation makes this type of evaluation more specific and thus the most costly of all three types of evaluation (Doshi-Velez & Kim, 2017).

2. *Human-grounded evaluations* are constructed by simpler experiments conducted on real humans who are not necessarily domain experts. Although this type of evaluation is less specific compared to the application-grounded evaluations, it offers more flexibilty and less costly. It is a good choice when the task is to test the quality of an explanation in a general sense (Doshi-Velez & Kim, 2017). For example, a recent study (Mohseni et al., 2018) used human attention maps that overlay on images as explanations and asked users to rate the decision made by the model. The study further argues that the evaluation on these attention maps can be utilized to understand the *trustworthiness* of a model.

3. *Functional-grounded evaluations* do not include real humans, instead this kind of evaluations use a formal definition of interpretability as a proxy to assess the explanation's quailty. The lack of human dependency makes them favorable due to the low cost. Typically, these evaluations are preferred when conducting experiments with humans in the loop might be unethical. The challenge with these evaluations is to select the right proxy models. Accordingly, when possible, it is considered good practice to first obtain proxies that were verified, for instance, by human-grounded evaluations (Doshi-Velez & Kim, 2017).

From high cost to low cost, and more specific to more broad, one can opt for an evaluation technique to obtain a performance indicator of an explanation. As discussed above, each approach require a completely different setting, which brings their shortcomings with it. For instance, depending of the availability of resources such as time, finances, expertise of the user or sufficiency of human subjects one might have to opt for a different evaluation technique.
Having highlighted important characteristics of a good explanation, we now move forward to the frequently mentioned techniques used in XAI. We mostly focus on post-hoc local explanation techniques and outline their contribution to this thesis.

2.2.3 Overview of Techniques in Explainable Artificial Intelligence

As discussed in the last section, when contructing an explanation method, one has to consider the audience, opt between model-specific or model-agnostic, local or global explanations, and also, utilize various forms of explanation. In literature, there exist various combinations of previously mentioned options. For transparent models, no further explanation method needed, one can obtain relevant information in the forms of weights or attention scores, given that the features are simple enough (Barredo Arrieta et al., 2020). In particular, we talk about explanation methods that were frequently mentioned in studies and relevant for this thesis.
First, we discuss the initial methodologies aimed to gain insight from a black-box model. The one of the initial approaches was to ask the question: *What happens if we remove this part of the input? Sensitivity Analysis* (SA) deals with analytical assessment of the effect of an omitted input variable on the uncertainty of a model (Novak et al., 2018). SA can be done on two levels, local and global. *Local Sensivity Analysis* (LSA) assesses the impact of the changes in the input on the output whereas *Global Sensitivity Analysis* (GSA) examines the effect of each variable (feature) with respect to the variations of all parameters (Rao et al., 2019). In literature, there are a variety of approaches for both GSA and LSA. For instance, Novak et al., 2018 constructs a GSA that employs the partial derivative of each parameter in the back-propagation algorithm to explore the change rule, which admits the *Input-Perturbation-Sensitivity* (IPS) that allows to obtain global sensitivity. An interesting example of a GSA and LSA fusion approach, Kowalski and Kusy, 2018, utilizes LSA to reduce the number of input features and GSA to reduce the number of patterns learned by a model.
Another approach was to calculate relevance scores for each feature using saliency maps. The usage of saliency maps first appeared in CNNs for images (Simonyan et al., 2014), then extended to NLP in Bansal et al., 2016; Denil et al., 2014. Typically, salience maps compute a gradient to get a relevance score to an input feature. In other words, they convey information about the model's sensitivity with respect to the input.

A popular method used in XAI is *Layer-wise Relevance Rropagation* (LRP) which was first introduced for *Fully Connected Networks* (FCNs) and CNNs in Lapuschkin et al., 2015, then extended to *Recurrent Neural Networks* (RNNs) in Arras et al., 2017. LRP assumes that a model can be *decomposed* into several layers which can contain feature relevant information. From the last layer to the input layer, LRP computes a relevance score for each dimension of the vector at a layer, and as LRP moves backwards in the layers, the sum of relevance scores do not change, staying always equal to the prediction probability (Lapuschkin et al., 2015).
Similar to LRP, a study for explaining DNNs offers another solution named *Deep Learning Important FeaTures* (DeepLIFT) (Shrikumar et al., 2017). This approach addresses two shortcomings of LRP, namely, the failure to model saturation caused by activation functions, and the possibility of getting misleading importance scores due to discontinuous gradients. Combining techniques from LRP and integrated gradients (Sundararajan et al., 2016), DeepLIFT computes importance scores based on the *difference-from-reference* approach which allows propagation of information even if the gradient is zero. Difference-from-reference is a method which involves determining a reference then getting the difference between the reference and the output. This method is also later adopted by to create DeepSHAP (Lundberg & Lee, 2017).
In contrast to model-specific approaches like LRP and DeepLIFT, *Locally Interpretable Model-agnostic Explanations* (LIME), as the name suggests, is a model-agnostic method. LIME interprets the predictions of any black-box model by approximating the model around a prediction. This approximation allows to obtain a locally faithful and interpretable version of the model (Ribeiro et al., 2016).
So far, there is no study that unifies all the works to create one explainability framework. To address this lack of unification, Lundberg and Lee, 2017 offers *SHapley Additive exPlanation* (SHAP) framework, in which the authors utilize recent studies from game theory based on (Shapley, 2016). These studies are *Shapley regression values* (Lipovetsky & Conklin, 2001), *Shapley sampling values* (Štrumbelj & Kononenko, 2014)*Quantitative input influence* (Datta et al., 2016), and recent approaches like LIME, DeepLIFT are utilized to create a model-agnostic and model-specific explainers. SHAP values measure the feature importance and obtained via Shapley values of conditional expectation function of a model (Lundberg & Lee, 2017). Model-agnostic SHAP values are computed using Shapley sampling values method which uses an approximation of a permutation adaption of of classic Shapley value estimation. For example, *KernelSHAP* employs LIME with linear explanations and Shapley values to find a weighting kernel that enables regression based estimation of SHAP values. On the other hand, the authors propose *LinearSHAP* which can approximate Shapley values using weights for linear models, and *DeepSHAP* which connects DeepLIFT with Shapley values, *Low-order SHAP*, and *Max SHAP* for model-specific explainers. We will discuss SHAP values further in

Chapter 4.
In literature, there is a lack of explanation methods for GNNs. GNNs require graphs as input and an output for either graph or node depending on the focus of the task (Zhang et al., 2018). Graphs are capable of representing rich relational information between nodes and the node feature information (Zhang et al., 2018; J. Zhou et al., 2018). GNNs are powerful tools that are able to learn relational information between nodes as well as node features, making them a perfect candidate for analyzing social media networks (Zang et al., 2016). In our case, we want to understand how a GNN behaves when classifying fake and real news pieces and their propagation networks. A study by Ying et al., 2019 proposes a model-agnostic approach called *GNNExplainer* to explain predictions made by GNNs. *GNNExplainer* takes a trained GNN, input graph(s) and its prediction(s) ad it returns explanations in the form of subgraph(s) of input graph(s) along with the most influential node feaures for the prediction. These subgraphs are constructed by maximizing the mutual information between the subgraph and the input graph with respect to the prediction (Ying et al., 2019).
Bearing in mind the FND models and explanation methods discussed one can use LIME, DeepLIFT, or SHAP for news content models which are essentially DNNs with textual data as inputs. Especially for understanding which words or word groups are of the most importance, SHAP provides text plots and easily interpretable importance scores. Therefore, when assessing our choice of news content model, we employ SHAP framework, in particular, DeepSHAP. On the other hand, for GNNs the choice is straightforward as there is only one choice. Although, GNNExplainer can be helpful to identify the most important spreaders of a news piece which will be discussed in Chapter 4.
To conclude this chapter, it should be noted that due to the numerous studies in the literature, we did not cover all explanation methods, but an extensive study can be found in (Molnar, 2022). Moreover, we were not able to fully cover the psychological and social background of explanations as we did for fake news, however Miller, 2017 provides a rigorous research in that field. In the next chapter, we elaborate on FND models that were used in this thesis. We show how a neural network produces a prediction for a given input. We share our analysis on both textual and graph datasets. After talking about model parameters and hyperparameters, we evaluate our models and talk about the evaluation process. Also, we examine issues like early fake news detection and model aging, particularly for our SOTA FND models.

3 Fake News Detection Models

The automated detection of fake news on social media comes with its characteristic challenges. First, the fact that fake news are constructed to misguide its consumers makes them hard to distinguish by only using news content. Second, when we include social context into the model, the large-scale and noisy nature of social context data represents another issue (Shu et al., 2020). Moreover, from a broader perspective, fake news should be detected before it becomes widespread so that the amount of users affected can be minimized.
In this chapter, we examine how these challenges effect the model and dataset's design. We initially take a look at news content models in section 3.1. In first section, we lay out the definitions for the materials used. After we give a detailed analysis of the dataset, we talk about the tokenizer and model itself, discuss its performance on the dataset. In section 3.2 we investigate social context based and hybrid models. Similar to the first section, we give definitons for the used material, then talk about the dataset and model. In this section, we also examine issues such as early fake news detection and model aging.

3.1 News Content Models

The majority of approaches for FND models utilizes news content. Models that base their predictions only on news content focus on the patterns in the text, especially words or word groups that appear frequently in other instances of the same class. As discussed in Section 2.1, there exist a variety of approaches available for news content models, however, due to unavailable or outdated datasets, we were unable to work with most news content models.

3.1.1 Notation and Definitions

Here we introduce the notation utilized in this section. Note that these notations will appear in its context, which will provide concrete examples for each symbol defined in Table 3.1.
Using this notation we now define some relevant concepts. First, we talk about terms

$x^{raw} \in X^{raw}$	Input news article.
$y^{raw} \in Y^{raw}$	The label of news article.
T	Tokenizer function
ψ	Label mapping function
$x^{tok} \in X^{tok}$	Tokenized news article
$y \in Y$	Vectorized class value.
$\|x^{tok}\|$	The number of tokens in x^{tok}.
V	Vocabulary: A collection of tokens available to the tokenizer.
f	Classifier function, i.e., FND model.
y^*	Prediction of FND model.
$x \in X$	Numeric vector of x^{tok}
$\|x\|$	The length of input vector
l	Index of a layer
$l_{embedding}$	Embedding layer
$a_i^{(l)}$	The value of unit i in layer l
$w_{ij}^{(l)}$	Weight between units i in layer l and j in layer $l+1$
σ	Activation function

Table 3.1: Notation used in this section.

and definitions for *tokenization*, illustrate the mathematical insight in the tokenization process. First, to build upon a concrete foundation, let us consider a news article x^{raw} fed to the tokenizer.

Definition 3.1.1 (*Tokenizer*)**.** A tokenizer $T : X^{raw} \mapsto X^{tok}$ is a function that maps raw textual data to smaller units called tokens.

A token can be a word, character or a subword. Therefore, we define three types of tokenization techniques:

- *Word tokenization* splits the given text into individual words based on a delimeter such as whitespace, commma, etc. This approach creates a vocabulary (V) from the inputs it was trained on. All words do not appear in the vocabulary are replaced with unknown token ([UNK]), and this concept is called being *Out Of Vocabulary*(OOV). Depending on the task, the size of the vocabulary can grow quite large. The solution for exploding vocabulary sizes was introduced in subword tokenization. The commonly used examples for word tokenizers are Word2Vec (Mikolov, Chen, et al., 2013) and GloVe (Pennington et al., 2014).
- *Character tokenization* splits the text into single characters. Since the size of available characters is limited and known, the OOV problem is solved by encoding the

unknown word by means of its characters. Although looks like a good solution, the length of tokens can be massive for long texts.

- *Subword tokenization* splits the given text into subwords, also called *n-gram characters*. For instance, comparative words like harder is segmented into hard-er, or superlative words like hardest is segmented into hard-est. The most common method for subword tokenization is *Byte Pair Encoding* (BPE). BPE was introduced by Gage, 1994 but adapted to word segmentation by Sennrich et al., 2015. BPE iteratively merges the most frequently appearing character or character sequences. This approach allows for an efficient space usage thus smaller vocabularies (Sennrich et al., 2015).

We say that an input is *tokenized* after it is fed to the tokenizer. A tokenized news article x^{tok} is a vector of tokens in which the order of the words and characters in $x^{raw} \in X^{raw}$ are kept. Furthermore, to get a fixed length output, we pad the tokenized sequence x^{tok} with padding tokens ([PAD]) where the news article is not long enough. In case it is longer than the fixed length, then it is truncated.

$$T(x^{raw}) = x^{tok} = [x_1^{tok}, \ldots, x_n^{tok}], \text{ where } n = |x^{tok}|.$$

Furthermore, we denote the space of raw label $Y^{raw} = \{"fake", "real"\}$, with $y^{raw} \in Y^{raw}$. We use a label mapping function $\psi : Y^{raw} \mapsto Y$ that maps raw labels to classes, where $Y \in \mathbb{R}^2$ with,

$$\psi(y^{raw}) = y = \begin{cases} 0, & y^{raw} = "fake" \\ 1, & y^{raw} = "real" \end{cases}$$

In order to feed the input to the model, we need numeric data which can be obtained by numerous techniques. One widely used approach is BoW representation which produces features based on the number of occurrences of a word or token. An alternative BoW representation uses the presence/absence of word instead of frequencies. A more sophisticated approach is *Word2Vec*, which encodes words into numeric values by learning word associations. From the perspective of representation of a word, *Word2Vec* can capture different degrees of similarity between words which allows for preservation of semantic and syntactic relationships (Mikolov, Chen, et al., 2013). It is clear that the transformation of words into numeric vectors is a very crucial stage for FND since we need to maintain as much contextual information as possible. Yet, the SOTA is an even more sophisticated approach called *Transformer* which is utilized by many language models such as BERT. We will discuss Transformer architecture in 3.1.2. For ease of the notation, we refer to this stage as *Embedding Layer* and denote it with $l_{embedding}$.
The input transformation pipeline is illustrated in 3.1.

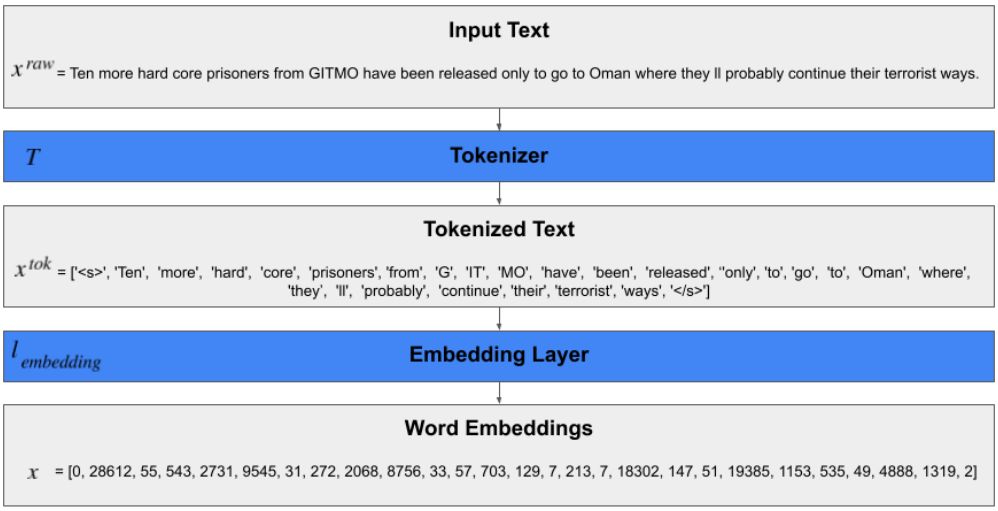


Figure 3.1: The preprocessing pipeline for a textual input.

Definition 3.1.2 (*FND Classifier*)**.** An FND classifier $f : X \mapsto Y$ is a function that outputs a predicted scores $f(x)_y$ for each class y for a given input x.

Definition 3.1.3 (*Prediction*)**.** A prediction y^* is the maximum of predicted scores $f(x)_y$ of an FND classifier.

$$y^* = argmax_{y \in Y} f(x)_y$$

Definition 3.1.4 (*FND Neural Network Classifier*)**.** A neural network classifier is a *classifier* f that comprises of layers l with $1 \leq l \leq L$, where L denotes the number of layers. Each layer has a set of units $a_i^{(l)}$ with i denoting the position of the unit in a layer l. We say that between two units $a_i^{(l)}$ belonging to layer l and $a_j^{(l+1)}$ belonging to layer $l+1$ have a weight value $w_{ij}^{(l)}$ that connects them. Along with a non-linear activation function σ, we can define the value of the j-th unit $a_j^{(l+1)}$ in terms of weights and unit values from previous layer for an FCN, with N is the number of units in layer l.

$$a_j^{(l+1)} = \sigma(\sum_{i=1}^{N} a_i^{(l)} w_{ij}^{(l)})$$

Neural networks iteratively optimize the weights between layers such that the pro-

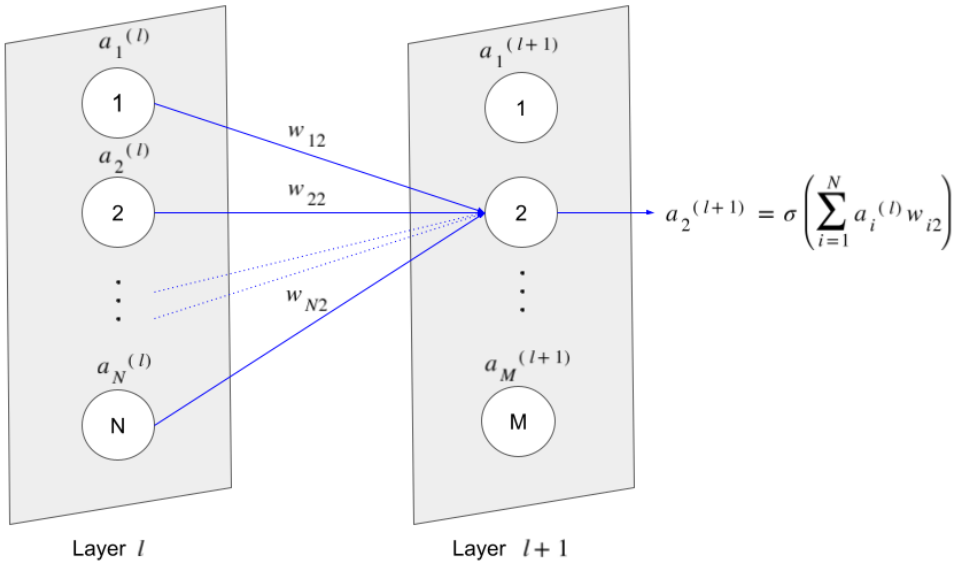


Figure 3.2: Units and layers of an FCN. For brevity, arrows and weights are only drawn for $a_2^{(l+1)}$

duced output is as close as to the expected output. This is done so using optimization methods such as *Gradient Descent* (GD) (Cauchy, 1847), *Stochastic Gradient Descent* (SGD) (Robbins & Monro, 1951), *Adaptive Moment Estimation* (Adam) (Kingma & Ba, 2014) to minimize the loss function. While there exist many optimization methods available for neural networks, we do not examine any of them.
For classification problems, we adopt a layer called *Softmax* (Bridle, 1989) outputs the predicted scores $f(x)_y$ for each class y by normalizing outputs $Z_y(x)$ from the previous layer.

$$f(x)_y = \frac{exp(Z_y(x))}{\sum_{\hat{y} \in Y} exp(Z_{\hat{y}(x)})}$$

The FCNs are the simplest architectures for neural networks. In an FCN, all units are connected which means there is a weight between each unit in layer l and $l+1$. While very simple to construct, usage of these architectures when modeling sequences is not preferred due to the number of trained parameters such as weights and biases. Instead, when modeling sequences like sentences and documents, a common approach is RNNs which allow previous outputs to be used while having hidden states. However, RNNs are not powerful enough to represent long-term dependencies (Bengio et al.,

1994) and suffer from vanishing/exploding gradients (Pascanu et al., 2012). Utilizing RNNs, an idea was proposed in which these shortcomings were addressed. Called *Long Short-Term Memory* (LSTM) (Hochreiter & Schmidhuber, 1997), the idea is to keep a cell state which is updated with previous cell's state. This cell state is conveyed to the consequent cells to form a chain that represents the document. More precisely, each cell corresponds to a token whose information will be shared with consequent tokens by the propagation of previously mentioned cell states. This approach is indeed very useful for long documents since news articles tend to be long and their sentences contextually relevant. LSTM is usually used in different variations based on the same idea. For instance, a consequent study has extended LSTMs with *peephole connections* in (Gers & Schmidhuber, 2000). LSTMs have been proven to deliver a good performance in NLP tasks such as *speech recognitions* (Xiong et al., 2016). LSTMs perform well however due to long training times and large memory requirements during training, they are being replaced with attention-based models.
One last thing to discuss is how the models are trained in terms of supervision. *Supervised* models are trained with label data. *Unsupervised* models work with unlabeled data and aims to find patterns in the data. A different setting is *semi-supervised* models. These models are often provided with small amounts of labeled data and large amounts of unlabeled data for training. There are two settings for semi-supervised learning. *Transductive* learning aims to predict unlabeled data whereas *inductive* learning samples unlabeled data from the same distribution to infer (Gammerman et al., 1998). Having introduced all necessary notation and definitions, in 3.1.2, we discuss the SOTA approach the Transformer models which are based on attention mechanism.

3.1.2 Transformer Architecture

Transductive learning has been successfully utilized along with encoder-decoder structure in many language tasks (Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017). Transduction is first proposed in Gammerman et al., 1998 to counteract with unlabeled data problem. In contrast to supervised learning, transductive learning does not require all data to be labeled, instead it utilizes the clustered behavior of data. Using the gaps between different clusters and a small set of labeled data, transductive learning assigns labels to unlabeled data. Accordingly, Transformer models are transductive models and use encoder-decoder structure to achieve that.
Encoder-decoder structure that takes into account the order of words was proposed in (Cho et al., 2014). This encoder-decoder structure consists of one RNN as encoder and one RNN as decoder. The encoder maps an input sequence to fixed-length vector, and the decoder maps this fixed-length vector to a target sequence. Transformer architecture adopts a similar approach which employs feed-forward and Multi-Head

Attention layers in both encoder and decoder which is illustrated in Fig. 3.3 with N=6 stacks of encoders and decoders.
In order to reduce sequential computation, CNNs have been adopted as building blocks that parallelly compute hidden representations for all input and output positions (Vaswani et al., 2017). Although aimed to reduce computation, the number of operations to convey information from one random input or output to another increases linealy in ByteNet (Kalchbrenner et al., 2016) and logarithmically in ConvS2S (Gehring et al., 2017). Contrary to CNNs, Transformers are able to fix the number of operations by averaging attention-weighed positions which decreases the effective resolution. However, this decrease in resolution is neutralized by the utilization of Multi-Head Attention (Vaswani et al., 2017). Initially suggested in the decoder of the model proposed in (Bahdanau et al., 2014), an attention mechanism works similar to human attention; it learns to put more importance on some words that convey the relevant information about the sentence. It does so by means of a context vector that depends on a sequence of *annotations*. An annotation h_i for a word (or embedding) x_i contains information about the complete input sentence but with a focus on the words that are closer to the word x_i. The context vector c_i for word x_i is obtained as a weighted sum of all these annotations h_i:

$$c_i = \sum_{j=1}^{|x|} \alpha_{ij} h_j.$$

The weight α_{ij} is obtained by applying softmax to associated energy e_{ij} which is an output of the alignment model a. The alignment model a is a feed forward neural network that jointly learns with the rest of the system. More precisely, we compute these values as follows:

$$\alpha_{ij} = \frac{e_{ij}}{\sum_{k=1}^{|x|} exp(e_{ik})}$$

where

$$e_{ij} = a(s_{i-1}, h_j)$$

with s_i representing the current and s_{i-1} the previous state of the model (Bahdanau et al., 2014). This is called *additive attention*.
For Transformer models, the authors define an attention function as *mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compability function of the query with the corresponding key.* (Vaswani et al., 2017). The Transformer model employs two different attention mechanisms, namely, *Scaled Dot-Product Attention* and *Multi-Head Attention*. By following the notation in (Vaswani et al., 2017), we denote that the input for the attention layers are matrices

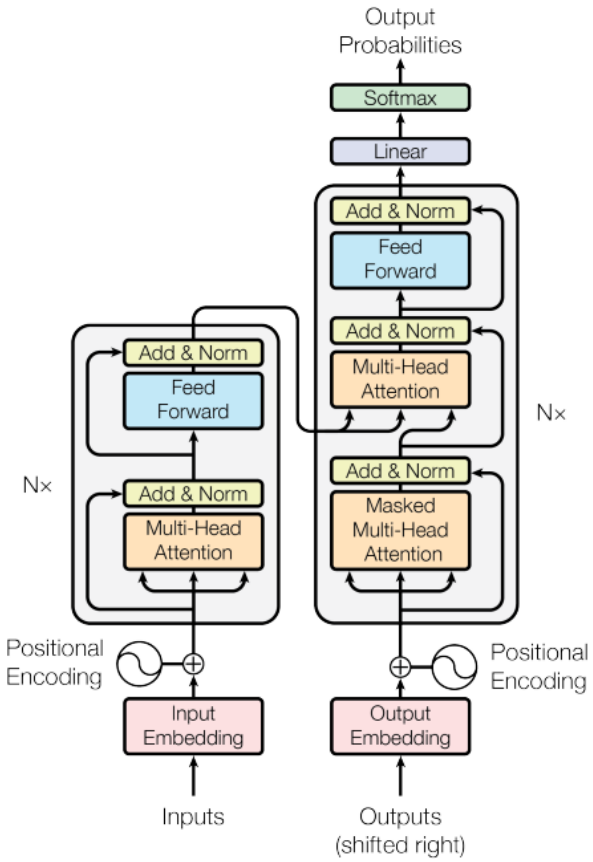


Figure 3.3: The Transformer Model Architecture (N=6). Figure obtained from (Vaswani et al., 2017)

called queries $Q \in \mathbb{R}^{d_{model} x d_k}$, keys $K \in \mathbb{R}^{d_{model} x d_k}$, and values $V \in \mathbb{R}^{d_{model} x d_v}$, with d_{model} being the model dimension. Scaled Dot-Product Attention computes the dot product of all queries $q_i \in Q$ with all keys K and scale the resulting wegihts with $\frac{1}{\sqrt{d_k}}$. After obtaining the softmax of the scaled weights, each weight is multiplied with the correspoding value to obtain attention values.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The second attention mechanism, Multi-Head Attention, uses multiple attentions each of which uses a different learned linear projection of Q, K, V. Output of each of these attentions are then concatenated to obtain the final result. More precisely, it is computed as follows,

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^O$$

where each $head_i$ is calculated as,

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

with projection for Q as $W_i^Q \in \mathbb{R}^{d_{model} x d_k}$, K as $W_i^K \in \mathbb{R}^{d_{model} x d_k}$, Q as $W_i^V \in \mathbb{R}^{d_{model} x d_v}$, and lastly, $W^O \in \mathbb{R}^{hd_v x d_{model}}$ (Vaswani et al., 2017).
We now discuss each part of the Transformer model briefly.

- *Encoder*: Consists of N=6 identical layers. Each of these layers have two sub-layers the first of which uses a Multi-Head Attention and layer normalization (Ba et al., 2016) along with a residual connection (He et al., 2015) around. The second sub-layer consists of a feed-forward layer and layer normalization as well as a residual connection around the feed-forward layer (Vaswani et al., 2017)
- *Decoder*: Same as the encoder, this part is also composed of N=6 layers. Additional to the previously discussed two sub-layers in encoder, decoder adopts a third sub-layer which computes the attention values over the output of encoder. As it was done for encoder, decoder also utilizes layer normalization at the end of each sub-layer as well as the residual connection (Vaswani et al., 2017).

Each of the feed forward networks in the sub-layers are position-wise, meanining that they are applied to each position separately and identically. These feed-forward networks use different parameters for each layer. The embeddings are obtained through transductive learning. Lastly, the positional encodings for input embeddings are calculated using sine and cosine functions of different frequencies (Vaswani et al., 2017).

We have summarized the Transformers architecture in order to lay foundations for the model we use, DistilRoBERTa (Liu et al., 2019; Sanh et al., 2019). Next, we initially introduce details of BERT, then RoBERTa, and lastly DistilRoBERTa which is our news content model.

3.1.3 Model, Dataset, Tokenizer Analysis

We employ a fine-tuned version of case-sensitive Transformer model DistilRoBERTa for our task of FND with news content. DistilRoBERTa is a distilled version of *A Robustly Optimized BERT Pretraining Approach* (RoBERTa) (Liu et al., 2019). It uses the same distillation procedure adopted for DistilBERT (Sanh et al., 2019) to *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2018). This distillation procedure is referred to as *Knowledge Distillation* and it compresses a model (Buciluâ et al., 2006) - the teacher - by means of training a smaller model - the student - to reproduce the same behaviour (Hinton et al., 2015). In our case, the teacher is RoBERTa and the student is DistilRoBERTa. First, in order to examine properties of RoBERTa, we discuss BERT in detail.
As the name suggests, the model architecture of BERT is a multi-layer bidirectional Transformer based on the Transformer model. BERT uses BookCorpus (Y. Zhu et al., 2015) and English Wikipedia as training dataset, with two training objectives, *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP). MLM procedure applies the following for each sentence sampled from a document in the cumulative dataset.

- Mask 15% of the tokens.
- In 80% of the cases, replace the masked tokens with [MASK].
- In 10% of the cases, replace the masked tokens with a different random vocabulary token.
- In the remaining 10% of the cases, the masked tokens are left unchanged.

NSP procedure is a binary classification loss that predicts whether two segments (sequences of tokens) are consecutive in the original text. *Positive* and *negative* examples are sampled with equal probability in this process. Positive examples are created with taking the consecutive sentences from the text corpus, whereas negative examples are generated by pairing segments from different documents (Liu et al., 2019).
RoBERTa is an optimized, longer pretrained with longer sequences version of a BERT implementation that uses only MLM as training objective. Contrary to BERT, RoBERTa keeps a dynamic masking pattern that changes in training. It is pretrained on reunion of five datasets (three more datasets than BERT) that size up to 160 gigabytes (GB):

BookCorpus (Y. Zhu et al., 2015), English Wikipedia ("English Wikipedia," 2022), CC-News (Nagel, 2016), OpenWebText (Radford et al., 2019), Stories (Trinh & Le, 2018).
RoBERTa tokenizes texts using BPE with a vocabulary size of 50,000 and maximum sequence length (maximum number of tokens) as 512. The beginning and end of each document (news article) is marked with <s> and </s> respectively. With MLM as training objective and Adam (2014) as the optimizer, the model reaches better results than BERT. Additionally, it should be noted that since these models are further trained for downstream tasks, thus we refer to training stage as pretraining to avoid any confusion.
DistilRoBERTa has the same general architecture as RoBERTa but the number of layers are reduced by a factor of two, the *token-type embeddings* and the pooler are removed. Then DistilRoBERTa is initialized with layers from the teacher. The distillation is done with very large batches (Sanh et al., 2019). Using RoBERTa as a teacher, the student DistilRoBERTa is pretrained on OpenWebTextCorpus (Gokaslan & Cohen, 2019) a reproduction of OpenWebText (Radford et al., 2019).
We employ a fine-tuned version of DistilRoBERTa from Huggingface[1]. It is trained on a dataset[2] curated from different sources. Although there exist better datasets and news content models, we opted for this particular model for two reasons. First, most SOTA news content models do not provide their code and dataset to reproduce results. Second, since Transformers are SOTA and this particular trained model provides us with not only the dataset but also with the train/test/validation splits which allows us to analyze its explanation. Now, we analyze the dataset, convey the distribution of labels and discuss potential peculiarities. Then we talk about the performance of the model, and reason about it.
Dataset. The news content dataset comprises of 40587 samples whose distribution of labels and train/validation/test splits are provided in Fig 3.4. The proportion of fake news is 46%, which is a fair distribution between real and fake news instances. The train/validation/test split is 60%/20%/20% which is common practice. We analyze 500 most frequently occurring tokens in the dataset using a WordCloud (Oesper et al., 2011) visualization for all samples of real and fake news separately in Fig 3.5. From the visualization, we can observe that samples from both datasets contain the words *new*, *state*, *President*, *Republican*. In fact, 500 most frequent tokens from fake news samples and real news samples share 63% of tokens. Furhtermore, we observe that one of the most frequent token is *Reuters* in real news instances. This might be give rise to a problem where the model learns the real news instances based on a couple tokens. This will be analyzed in detail in Section 4.1.

[1]https://huggingface.co/GonzaloA/distilroberta-base-finetuned-fakeNews
[2]https://huggingface.co/datasets/GonzaloA/fake_news

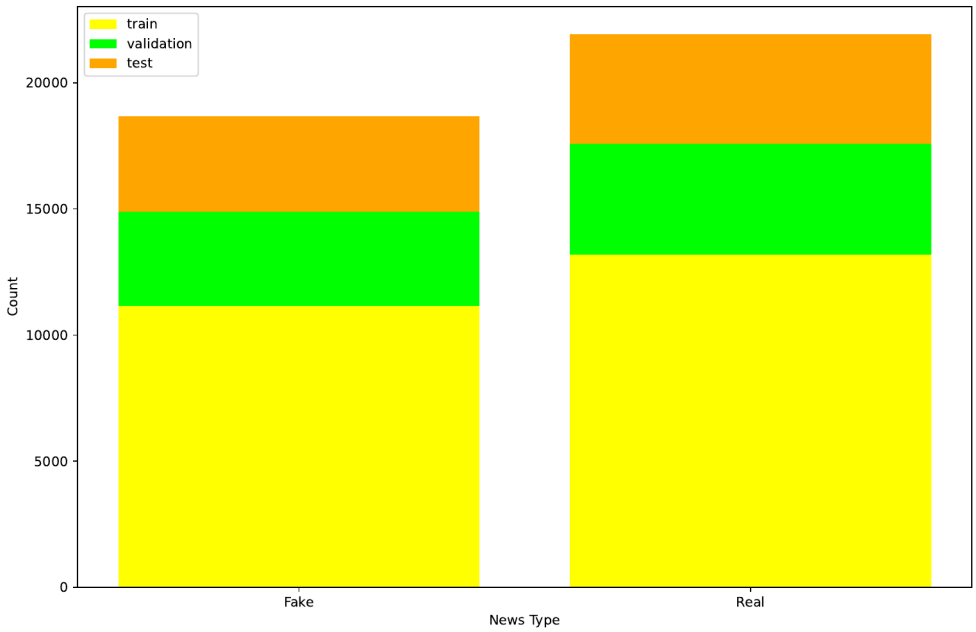
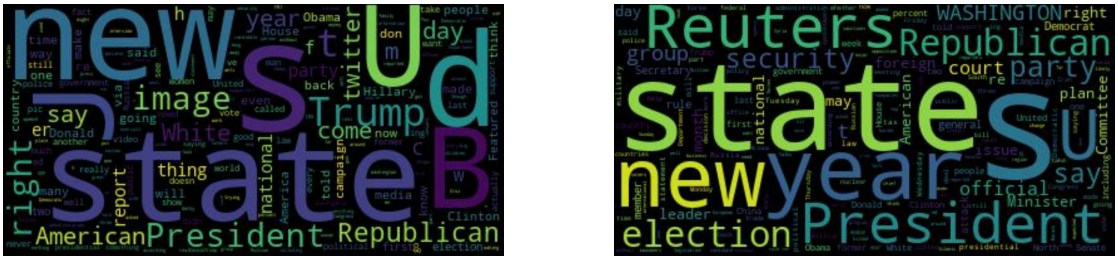


Figure 3.4: News content dataset distribution by label and train/validation/test split

Training. DistilRoBERTa has 6 layers, 12 attention heads and hidden size of 768 that totals up to 82M parameters (RoBERTa has 125M parameters). The vocabulary size is 50265 including special tokens. The model also uses a technique called *dropout* which is a regularization technique that drops some connections between layers based on a probability value (Srivastava et al., 2014). All model parameters are defined in 3.2.
Note that maximum position embeddings parameter also considers document start and end tokens when reporting its length. When feeding tokens x^{tok} to the model, the start (<s>) and end (</s>) tokens are added by the model pipeline, so this leaves us with maximum sequence length of 512, i.e., $|x| = 512$. Model is trained for 3 epochs with a batch size of 16, Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 08$) as the optimizer and a linear learning rate scheduler. The performance metrics for the model is provided in Table 3.3.
The model performs very well on the dataset. Only 4 samples out of 8117 are classified incorrectly in the test split of the dataset. We will closely work with those 4 samples when explaining our model and inspect the reasons behind this performance in Section 4.1.
We have laid out the foundations for the model we used, reported the characterstics, training details, and performance of the news content model in this section. Although seems powerful in this case, usually news content models are easily outdated as the



(a) Fake news

(b) Real news

Figure 3.5: WordCloud visualiations for most frequently occurring 500 tokens in samples of both classes

Attention dropout probability	0.1
Classifier dropout	None
Hidden layer activation function	gelu (Hendrycks & Gimpel, 2016)
Hidden size	768
Intermediate size	3072
Layer normalization factor (ϵ)	1e-05
Maximum position embeddings	514
Number of attention heads	12
Number of layers	6
Vocabulary size	50265

Table 3.2: Parameters of news content model

Accuracy	Precision	Recall	F1 score
98.85%	98.82%	99.03%	98.92%

Table 3.3: The performance metrics for news content model.

structure of news changes rapidly and fake news fabricators can replicate the properties that can be seen in real news content (Shu et al., 2020). Thus, we consider a different approach which primarily considers social context features for FND.

3.2 Social Context Models

As discussed in Section 2.1, social media's interconnected nature leads to fast dissemination of fake news. When a news piece is shared, it cascades through social media by means of friendship networks. These networks can be exploited to gather information about how fake news spread. Moreover, a user's historical information can prove effective when trying to analyze whether a news piece is real or not. This assumption stems from the psychological facts we discussed in Section 2.1. To recap, if a user is sharing fake news most of the time, i.e., the user's tweets are marked as fake, then it is likely that the next news piece they share would be fake as well. Usually, fake news are shared most within echo chambers, which gives rise to quick spread of fake news.
There exist many different approaches for social context models, however, it is a long and challenging task to create a dataset for social context models as the amount of data to be collected can grow dramatically. Thus, we selected a dataset that provides us with the social context information as well as the news content. Being a graph dataset, *User Preference-aware Fake Detection* (UPFD) (Dou et al., 2021) builds a propagation tree for the news. But in order to build a model that takes graphs as input we can no longer rely on standard deep learning approaches as the graph data has a different structure than news content data. It holds node information as well as edge information between nodes, allowing to store rich relational data (Dou et al., 2021).
In this section, we lay out foundations for the dataset and GNNs we used. Then we talk about the social context models along with their dataset UPFD that we used in this thesis. We also give some insights from (Shu et al., 2020) in which the authors of UPFD dataset analyze the data they later utilized to create the graph dataset UPFD.

3.2.1 Overview of Graphs

In this section, we introduce definitions to cover graphs and different types of GNNs. But we do not discuss each model type in detail due to GNNs' extensive background. Thus, we do not provide a notation for this section but we give mathematical definitions when required.
Graphs are an example of *non-euclidean* data, meaning that in contrast to *euclidean* data, they can represent complex relations and entities more accurate than one or

two dimensional representations. Non-euclidean data used in GDL can be grouped into grids, graphs, groups, geosedics, and gauges, which are also called the 5G's of GDL (Bronstein et al., 2021). We are interested in one G only, namely graphs. Since GNNs are a realization of GDL, we first briefly discuss the general framework.

Definition 3.2.1 (*Geometric Deep Learning (GDL)*)**.** A class of deep learning that aims to build models that can learn to predict on the non-euclidean domain.

GDL is an extensive field covering various techniques that can be appied to non-euclidean domain. Due to its complex nature, we do not provide a rigorous background on GDL. For a detailed background on GDL, we refer the reader to an extensive study on GDL (Bronstein et al., 2021). We only discuss parts related to GNNs that we utilized. First we define a graph.

Definition 3.2.2 (*Graph*)**.** A graph is a non-euclidean type of data that represents the relations between entities.

From the perspective of individual node connections, graphs can be categorized into two classes, *directed* and *undirected* graphs. Directed graphs have direction information in their edges, i.e., the information flows strictly from one node to another. On the other hand, undirected graphs do not have this limitation, the information flow is bidirectional. Since we are only interested in undirected graphs like our dataset, we give preliminaries for an undirected graph.
Following common notation on graphs, we define an undirected graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ as the set of nodes and $\mathcal{E}$ as the set of edges in a graph $\mathcal{G}$. We say that an edge $(\sqsubseteq_i, \sqsubseteq_j)$ exists between two nodes $\sqsubseteq_i$ and $\sqsubseteq_j$. Moreover, from the perspective of a graph's connectedness, there exist two types of graphs, *cyclic* or *acycylic*. Simply put, cyclic graphs have cycles in them, i.e., the graphs allows for at least one node $\sqsubseteq_i$ to have a series of different edges that creates a cycle back to the node $\sqsubseteq_i$. In contrast, acyclic graphs do not contain cycles. A concrete example for undirected acyclic graphs are trees. Also the structure of our dataset, trees have strictly one edge between two nodes.
When we look at graphs from node and edge type, we classify graphs as *homogeneous* and *heterogeneous* graphs. Homogeneous graphs have nodes and edges of the same type, whereas heterogeneous graphs have different types of nodes and edges. One example for homogeneous graphs are social networks in which the nodes are users and the edges represent the friendship between two users. If we modify this social network into a more detailed version in which we choose to represent another relationship between users, such as colleagueship, then we would have a heterogeneous graph, because there would be two types of edges in the graph.
Finally, if we consider graphs from a temporal aspect, we observe two kinds of graphs,

static and *dynamic*. Static graphs stay the same over time, their topology or features do not change. In contrast, dynamic graphs' features and topology vary over time, thus making time an important factor when working with dynamic graphs.

3.2.2 Graph Neural Networks

GNNs are neural networks that can take graphs as an input and produce predictions at three different levels: *node-level*, *edge-level*, and *graph-level*. Node-level tasks include node classification, node regression, node clustering. Node classification aims to categorize nodes into classes. Node regression deals with predicting continuous value for each node. Lastly, node clustering aims to group nodes into several clusters. Edge-level tasks consist of edge classification and link prediction. Edge classification tries to categorize an edge. Link prediction aims to find whether there is an edge between two nodes. Graph-level predictions are graph classification, graph regression, and graph matching. In all these settings, the model needs to learn representations of graph (J. Zhou et al., 2018).
In our experiments, we used two different models one of which uses a convolutional layer called GraphSAGE (Hamilton et al., 2017) and the other uses an convolutional attention layer called *Graph Attention* (GAT) (Veličković et al., 2017). Here we give background for both but we do not dive into details. We first give information about general framework.

GraphSAGE.

3.2.3 Dataset

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

3.2.4 Models

SAGE GNN UPFD GCNFN

3.3 Early Fake News Detection and Model Aging

4 Explainability of Fake News Detection Models

AI is getting more and more integrated into our lives, helping us with from the simplest tasks like playing music with voice command to more complex tasks like driving a car in open traffic.

4.1 Explainability of News Content Models

4.1.1 SHAP, DeepSHAP

4.1.2 SHAP in Action

4.1.3 Introducing Unseen Data

4.1.4 Results

4.2 Explainability of Social Context Models

4.2.1 GNNExplainer

4.2.2 GNNExplainer in Action

4.2.3 Introducing Unseen Data

4.2.4 Results

5 Conclusion

List of Figures

2.1 Fake News and Fake News Detection Publications by Year 5
2.2 Market Reaction to Fake Tweet 7
2.3 Total Facebook Engagements for Top 20 Election Stories 8
2.4 Characterization of Fake News Detection Models 13

3.1 The preprocessing pipeline for a textual input. 34
3.2 Units and layers of an FCN 35
3.3 The Transformer Model Architecture 38
3.4 News content dataset distribution by label and train/validation/test split 42
3.5 WordCloud visualiations for most frequently occurring 500 tokens in both classes 43

List of Tables

3.1 Notation 32
3.2 Parameters of news content model 43
3.3 The performance metrics for news content model. 43

Bibliography

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

Addawood, A., Schneider, J., & Bashir, M. (2017). Stance classification of twitter debates: The encryption debate as a use case. *Proceedings of the 8th International Conference on Social Media & Society*. https://doi.org/10.1145/3097286.3097288

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

Agrawal, A. (2016). Clickbait detection using deep learning. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 268–272. https://doi.org/10.1109/NGCT.2016.7877426

ALDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, *58*(4), 102597. https://doi.org/https://doi.org/10.1016/j.ipm.2021.102597

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

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393. https://doi.org/10.1126/sciadv.abf4393

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine bias.

Antunes, P., Herskovic, V., Ochoa, S. F., & Pino, J. A. (2008). Structuring dimensions for collaborative systems evaluation. *ACM Comput. Surv.*, *44*(2). https://doi.org/10.1145/2089125.2089128

Arras, L., Montavon, G., Müller, K.-R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. https://doi.org/10.48550/ARXIV.1706.07206

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.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, 722–735.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. https://doi.org/10.48550/ARXIV.1607.06450

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.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, *11*(61), 1803–1831.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. https://doi.org/10.48550/ARXIV.1409.0473

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

Bansal, T., Belanger, D., & McCallum, A. (2016). Ask the gru: Multi-task learning for deep text recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*. https://doi.org/10.1145/2959100.2959180

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., & Ali, Z. S. (2020). Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, & N. Ferro (Eds.), *Experimental ir meets multilinguality, multimodality, and interaction* (pp. 215–236). Springer International Publishing.

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

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*, *5*(2), 157–166.

Bennetot, A., Laurent, J.-L., Chatila, R., & Díaz-Rodríguez, N. (2019). Towards explainable neural-symbolic visual reasoning. https://doi.org/10.48550/ARXIV.1909.09065

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, *abs/1607.04606*.

Braşoveanu, A. M. P., & Andonie, R. (2019). Semantic fake news detection: A machine learning perspective. In I. Rojas, G. Joya, & A. Catala (Eds.), *Advances in computational intelligence* (pp. 656–667). Springer International Publishing.

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

Bridle, J. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In D. Touretzky (Ed.), *Advances in neural information processing systems*. Morgan-Kaufmann.

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. https://doi.org/10.48550/ARXIV.2104.13478

Buciluâ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 535–541. https://doi.org/10.1145/1150402.1150464

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. *Proc AMIA Symp*, 212–215.

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

Cauchy, L. A. (1847). M´ethode générale pour la résolution des systémes d'équations simultanées. *Compte Rendu á l'Académie des Sciences*.

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

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. https://doi.org/10.48550/ARXIV.1406.1078

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

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

Denil, M., Demiraj, A., & de Freitas, N. (2014). Extraction of salient sentences from labelled documents. *CoRR*, *abs/1412.6815*.

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

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*.

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

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

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

Du, J., Xu, R., He, Y., & Gui, L. (2017). Stance classification with target-specific neural attention. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 3988–3994. https://doi.org/10.24963/ijcai.2017/557

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.

English wikipedia. (2022).

Escalante, H. J., Escalera, S., Guyon, I., Baro, X., Gucluturk, Y., Guclu, U., & van Gerven, M. (2018). *Explainable and interpretable models in computer vision and machine learning* (1st). Springer Publishing Company, Incorporated.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134. https://doi.org/https://doi.org/10.1016/j.artint.2005.03.001

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

Gage, P. (1994). A new algorithm for data compression. *C Users J.*, *12*(2), 23–38.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, *75*(7), 450–451.

Gammerman, A., Vovk, V., & Vapnik, V. (1998). Learning by transduction. *In Uncertainty in Artificial Intelligence*, 148–155.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. https://doi.org/10.48550/ARXIV.1705.03122

Gers, F., & Schmidhuber, J. (2000). Recurrent nets that time and count. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, *3*, 189–194 vol.3. https://doi.org/10.1109/IJCNN.2000.861302

Ghanem, B., Rosso, P., & Rangel, F. (2018). Stance detection in fake news a combined feature representation. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 66–71. https://doi.org/10.18653/v1/W18-5510

Gokaslan, A., & Cohen, V. (2019). Openwebtext corpus.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, *51*(5). https://doi.org/10.1145/3236009

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

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, *10*, 178–206. https://doi.org/10.1162/tacl_a_00454

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *CoRR*, *abs/1706.02216*.

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

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods [PMID: 30332290]. *American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. *Proceedings of the 27th International Conference on Computational Linguistics*, 1859–1874.

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

Hassan, N., Li, C., & Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. *Proceedings of the 24th ACM International on Conference on*

Information and Knowledge Management, 1835–1838. https://doi.org/10.1145/2806416.2806652

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. https://doi.org/10.48550/ARXIV.1512.03385

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

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). https://doi.org/10.48550/ARXIV.1606.08415

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. https://doi.org/10.48550/ARXIV.1503.02531

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (1st ed.). Springer.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Jin, Z., Cao, J., Jiang, Y.-G., & Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. *2014 IEEE International Conference on Data Mining*, 230–239. https://doi.org/10.1109/ICDM.2014.91

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.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, *abs/1607.01759*.

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

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A., & Kavukcuoglu, K. (2016). Neural machine translation in linear time. https://doi.org/10.48550/ARXIV.1610.10099

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., & Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 829–839. https://doi.org/10.18653/v1/S19-2145

Kim, B. (2015).

Kim, B., Glassman, E. L., Johnson, B., & Shah, J. A. (2015). Ibcm: Interactive bayesian case model empowering humans via intuitive interaction.

Kim, B., Rudin, C., & Shah, J. (2015). The bayesian case model: A generative approach for case-based reasoning and prototype classification. https://doi.org/10.48550/ARXIV.1503.01161

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980

Kowalski, P. A., & Kusy, M. (2018). Sensitivity analysis for probabilistic neural network structure reduction. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1919–1932. https://doi.org/10.1109/TNNLS.2017.2688482

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. *2013 IEEE 13th International Conference on Data Mining*, 1103–1108. https://doi.org/10.1109/ICDM.2013.61

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. https://doi.org/10.1145/2939672.2939874

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*, e0130140. https://doi.org/10.1371/journal.pone.0130140

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

Liang, B., Li, H., Su, M., Li, X., Shi, W., & Wang, X. (2021). Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, *18*(1), 72–85. https://doi.org/10.1109/tdsc.2018.2874243

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

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319–330. https://doi.org/https://doi.org/10.1002/asmb.446

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266. https://doi.org/10.1017/S1358246100005130

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. https://doi.org/10.48550/ARXIV.1907.11692

Lopez-Paz, D., Nishihara, R., Chintala, S., Schölkopf, B., & Bottou, L. (2016). Discovering causal signals in images. https://doi.org/10.48550/ARXIV.1605.08179

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

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/ARXIV.1705.07874

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3818–3824.

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

Magdy, A., & Wanas, N. (2010). Web-based statistical fact checking of textual documents. *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, 103–110. https://doi.org/10.1145/1871985.1872002

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

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. *CHI'06 extended abstracts on Human factors in computing systems*, 1097–1101.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://doi.org/10.48550/ARXIV.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. https://doi.org/10.48550/ARXIV.1706.07269

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. https://doi.org/10.48550/ARXIV.1712.00547

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2016). Stance and sentiment in tweets. *CoRR*, *abs/1605.01655*.

Mohseni, S., Block, J. E., & Ragan, E. D. (2018). A human-grounded evaluation benchmark for local explanations of machine learning. https://doi.org/10.48550/ARXIV.1801.05075

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

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011

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*.

Nagel, S. (2016). Cc-news.

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

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., & Sohl-Dickstein, J. (2018). Sensitivity and generalization in neural networks: An empirical study. https://doi.org/10.48550/ARXIV.1802.08760

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

Oesper, L., Merico, D., Isserlin, R., & Bader, G. D. (2011). Wordcloud: A cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, *6*(1), 7.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: A large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, *10*(1), 36.

Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. In *Recent trends in learning from data* (pp. 155–196). Springer International Publishing. https://doi.org/10.1007/978-3-030-43883-8_7

Palau-Sampio, D. (2016). Reference press metamorphosis in the digital context: Clickbait and tabloid strategies in elpais.com. *Communication & Society*, *29*, 63–79. https://doi.org/10.15581/003.29.2.63-79

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

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training recurrent neural networks. https://doi.org/10.48550/ARXIV.1211.5063

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

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd). Cambridge University Press.

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

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

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*.

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1589–1599.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rao, Z., He, M., & Zhu, Z. (2019). Input-perturbation-sensitivity for performance analysis of cnns on image recognition. *2019 IEEE International Conference on Image Processing (ICIP)*, 2496–2500. https://doi.org/10.1109/ICIP.2019.8803012

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.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. https://doi.org/10.1214/aoms/1177729586

Rony, M. M. U., Hassan, N., & Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 232–239. https://doi.org/10.1145/3110025.3110054

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.

Rudin, C. (2018). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. https://doi.org/10.48550/ARXIV.1811.10154

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

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

Santia, G., & Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), 531–540.

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.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. https://doi.org/10.48550/ARXIV.1508.07909

Shapley, L. S. (2016). 17. a value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (am-28), volume ii* (pp. 307–318). Princeton University Press. https://doi.org/doi:10.1515/9781400881970-018

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. https://doi.org/10.48550/ARXIV.1704.02685

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.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, workshop track proceedings*.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1966). *The general inquirer: A computer approach to content analysis* (Vol. 4). https://doi.org/10.2307/1161774

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sundararajan, M., Taly, A., & Yan, Q. (2016). Gradients of counterfactuals. https://doi.org/10.48550/ARXIV.1611.02639

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

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. https://doi.org/10.48550/ARXIV.1409.3215

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

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. *CoRR*, *abs/1806.07687*.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. https://doi.org/10.18653/v1/N18-1074

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Trabelsi, A., & Zaiane, O. R. (2014). Finding arguing expressions of divergent viewpoints in online debates. *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, 35–43. https://doi.org/10.3115/v1/W14-1305

Trinh, T. H., & Le, Q. V. (2018). A simple method for commonsense reasoning. https://doi.org/10.48550/ARXIV.1806.02847

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph attention networks. https://doi.org/10.48550/ARXIV.1710.10903

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

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. https://doi.org/10.3115/v1/W14-2508

Walker, M., Anand, P., Abbott, R., & Grant, R. (2012). Stance classification using dialogic properties of persuasion. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 592–596.

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.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *CoRR*, *abs/1610.05256*.

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

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. https://doi.org/10.48550/ARXIV.1903.03894

Yu, B. (2013). Stability. *Bernoulli*, *19*(4). https://doi.org/10.3150/13-bejsp14

Yuan, X., He, P., Zhu, Q., & Li, X. (2017). Adversarial examples: Attacks and defenses for deep learning. https://doi.org/10.48550/ARXIV.1712.07107

Zang, C., Cui, P., & Faloutsos, C. (2016). Beyond sigmoids: The nettide model for social network growth, and its applications. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015–2024. https://doi.org/10.1145/2939672.2939825

Zhang, Z., Cui, P., & Zhu, W. (2018). Deep learning on graphs: A survey. https://doi.org/10.48550/ARXIV.1812.04202

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2018). Graph neural networks: A review of methods and applications. https://doi.org/10.48550/ARXIV.1812.08434

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

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. https://doi.org/10.1109/CIG.2018.8490433

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*.