Survey on Explainable Fake News Detection MAI4CAREU Project

Alessandro Lombardini, Memoona Shah, Giorgos Adamides
Master's Degree in Artificial Intelligence

alessandr.lombardin3@studio.unibo.it
memona.shah@studio.unibo.it
adamides.giorgos@ucy.ac.cy

Abstract

In recent years, a vast amount of papers have covered the topic of online fake news. Fake news detection is a critical challenging problem emphasized in recent years by the exponential proliferation of untrue information, The rapid rise of social networking platforms yielded a vast increase in information accessibility but has also, as a side effect, accelerated the spread of false news. Thus, the effect of fake news has been growing, sometimes extending to the offline world and threatening public safety. Given the massive amount of Web content, automatic detection has become essential, in order to reduce the human time and effort to detect and prevent the spread. In recent years several techniques have been proposed, however, these detection systems lack explainability i.e., providing the reason for their prediction, whose critical advantage is the identification of bias and discrimination in detection algorithms. This paper will focus on looking at existing explainable AI methods and highlights the current state of the art in explainable fake news detection. Based on our insights, we outline promising research directions, including more fine-grained, detailed, fair, and practical detection models.

1 Introduction

Nowadays **fake news** tends to be intrusive and diverse in terms of topics, styles, and platforms, and it is not easy to construct a generally accepted definition for it. Stanford University provides the definition as: "*the news articles that are intentionally and verifiably false, and could mislead readers*". According to Wikipedia, is: "*a type of journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media.*". In this paper, we proposed the definition of the Treccani vocabulary, which affirms that fake news *"designates information that is partially or wholly not true, intentionally or unintentionally disclosed through the media, and characterized by an apparent plausibility"*.

Fake news has always been spread throughout history: the sensationalism of not-so-accurate eye-catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted throughout all kinds of information broadcast. For example, propaganda played an important role in Octavian and Mark Antony's civil war: Octavian ran a campaign of misinformation against his rival Mark Antony, portraying him as a drunkard, a womanizer, and a mere puppet of the Egyptian queen Cleopatra VII. It would be a piece of fake news that was to be the proverbial straw that broke the camel's back: Octavian managed to get hold of a document that he claimed was Antony's official will and testament, which claimed that Mark Antony, upon his death, wished to be entombed in the mausoleum of the Ptolemaic pharaohs. Whether it was real or not the will was considered the most atrocious in the Roman eye, setting the Roman people against Antony. The document played on many of the anti-eastern (and anti-Cleopatra) prejudices of the ancient Romans, who have been convinced that Antony had lost his head and given himself over to the allure and despotism of Cleopatra, queen of Egypt. (Scott, 1933)

However, even if the authenticity of Information is a long-term issue affecting society both for printed and digital media, only recently we have begun to truly appreciate the effects they can have in an intercommunicated world, in particular since the birth of social media.

2 Background

The exploding development of the World Wide Web after the mid-1990s has significantly advanced the way that people communicate with each other. Social media in particular has facilitated the distribution of real-time information among people from all over the world: users share information, connect

with other people and stay informed about trending events. However, much of the appearing information is dubious and, in some cases, intended to mislead. It can be said (Zhang and Ghorbani, 2020) the birth of social media has had the side effect of having transformed the Internet into the ideal breeding ground for spreading misleading information, fake reviews, fake advertisements, rumors, fake political statements, satires, etc.

Since disinformation sometimes involves spreading false information with harmful intent and is sometimes generated and propagated by hostile foreign actors, it has arisen as a threat to individuals and society, extending its effect to the offline world and threatening public safety. The reach and effects of information spread occur at such a fast pace and are so amplified, that distorted, inaccurate or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. It has the potential to cause serious problems, and for these reasons, it requires the utmost attention.

The issue of online fake news has gained more attention, especially after the 2016 U.S. presidential election (Horne and Adali, 2017), which was such a controversial case that let the term *fake news* in the mainstream vernacular. In that political campaign, hundreds or thousands of Russian fake accounts posted anti-Clinton messages, such as "Hillary was sick", "Hillary was a criminal", "Obama had a secret army", and so on, to influence soft Hillary Clinton supporters. There is a view that Donald Trump's victory in the 2016 U.S. presidential election is somehow regarded as the outcome of fake news (Balmas, 2012). Fake news has been accused of increasing political polarization and partisan conflict, influencing the voters by misleading political statements and claims (Riedel et al., 2017).

With the characteristics of ease of use, low cost, and rapid rate, social platforms are becoming increasingly popular across different generations and slices of the population. They are already considered the most relevant tools for information transmission (Shu et al., 2017a), as also the largest medium at all for sharing false information (Balmas, 2012). The main problem is that this growth of users, jointly with a massive amount of misleading information created and displayed every day, has the potential to destroy people's faith and beliefs in authorities, experts and the government. In fact, by creating a vast competition for real news, fake news is able to reduce the impact of real ones and undermine the trust in serious media coverage, creating a general erosion of the traditionally authoritative voices of the news industry. BuzzFeed News analysis (Silverman, 2016) found that the top fake news stories about the 2016 U.S. presidential election received more engagement on Facebook than top stories from major media outlets.

3 Fake news detection

Simply speaking, fake news detection is the task of assessing the truthfulness of a certain piece of news (Vlachos and Riedel, 2014). In the news industry, in particular, but also in society at large, detecting when online content is untrue and intended to mislead has become a central discussion topic and a critical interest, as the need to permanently assess the veracity of digital content has been raised by the constant spread of false information.

Societal issues are being raised (Álvaro Figueira and Oliveira, 2017) about the ability of individuals to tell apart what is fake and what is authentic while surfing and actively engaging in information-overloaded networks. On an individual scale, the ability to actively confront false narratives and take care when sharing information can reduce the prevalence of falsified information. However, it has been noted (Anderson, 2017) that this is vulnerable to the effects of confirmation bias, motivated reasoning and other cognitive biases that can seriously distort reasoning, particularly in dysfunctional and polarised societies. Youngsters are known to be tech-savvy when compared to their parents, but when it comes to the ability to tell if a news piece is fake or not, they seem as confused as the rest of society and 44% have confirmed it in research conducted by Common Sense Media (Robb, 2017). The same research also indicates that 31% of kids aged 10 to 18 have shared online at least one news story that they later found out was inaccurate or fake. This situation raises a whole new dimension of concerns related to digital literacy that goes beyond the mere ability to access and manage technology.

There exist a considerable number of worldwide entities, organizations and initiatives aimed at stopping the spread of fake news. These include, for example, online fact-checking systems, such as FactCheck.org and PolitiFact.com, which are based on manual detection approaches by professionals. Human resources-based methods are involved in

the majority of this kind of websites, however, this approach faces scalability and latency issues. A large amount of real-time information is created, commented and shared every day, and this makes online real-time fake news detection with this approach very difficult. Online information also is very diverse, covering a large number of subjects, so the practical applicability of those systems is limited, due to the high variety of news types and formats. Thus, the difficulties imposed by manual fact-checking approaches paved the way towards research-based approaches or the automatic detection of fake news.

4 Automatic fake news detection

Using algorithms to fight algorithms: since algorithms are part of what spreads fake news they can also be part of the solution, by identifying fake content and validating the information sources (Álvaro Figueira and Oliveira, 2017). Identifying credible social information from millions of messages, however, is anyway technically challenging: content is easily generated and quickly spread, leading to a large volume of content to analyse. Online information is also very diverse, covering a large number of subjects, which contributes complexity to this task. In the end, it is difficult to distinguish online truthful signals from fake and anomalous information, because they are usually intentionally written to mislead readers.

The features used for fake news detection usually are divided into three main categories (Shu et al., 2017a)(A.B. et al., 2023):

- *user features*: include information in the user profiles, credibility-related information, and user behavior patterns. Common implicit user features include age, gender, personality, etc.
- *news content features*: are derived from the analysis of words, sentences, and news content. These are derived analyzing words and sentences of news content. These are referred also to as linguistic-based and syntactic-based features. There is another class of news content features with focuses on style-based and visual-based features.
- *social context features*: is derived from the propagation pattern of the news and the interactions of social media users. It includes temporal, distribution, and network-based features.

The linguistic-based features extracted from the news content are in fact not sufficient for revealing the in-depth underlying distribution patterns of fake news (Shu et al., 2017a) (Zhao et al., 2020). Auxiliary features such as the credibility of the news author and the spreading patterns of the news, play more important roles in online fake news prediction. In *single-modal* fake news detection methods, a single type of feature is analyzed. The addition of auxiliary information (combining different features) aids in a *multi-modal* system. In that case, various feature combinations are examined. By now news content features are used in most of the single-modal methods.

5 Model Evaluation

Evaluating a fake news detection model involves carefully assessing how well it performs in distinguishing between real news and fabricated articles. This evaluation is crucial to determine the model's reliability, strengths and weaknesses. To perform this evaluation, researchers typically use a dataset containing labeled news articles, ensuring a balanced representation of both genuine and fake news, in order to have a fair assessment of the model's ability to differentiate between the two categories.

The dataset is then tested out, using the fake news detection model, to make predictions. These predictions indicate whether the model classified the articles as genuine or fake. Researchers then analyze the model's performance based on these generated predictions. At the evaluation step of such fake news detection models, the designers and developers tend to examine the outcomes using some quantitative instances. One of such instances are the so-called false positives (FP), which represent the genuine articles that were mistakenly labeled as fake. Another instance are the false negatives (FN), which represent the fake articles that were incorrectly classified as genuine. The analysis of those misclassifications leads to a more clear view of the capabilities of the model and some of its the aspects that may need further improvement. This examination helps researchers understand the causes behind those misclassifications, identify patterns, detect wrongly overlooked features and refine the model to enhance its performance in a more accurate detection of fake news. Additionally, researchers may assess the model's generalizability by testing it on diverse datasets containing news articles from vari-

ous sources and time periods. This helps determine how well the model can detect fake news in different contexts.

Evaluating a fake news detection model is a complex and iterative process that requires researchers to carefully consider various factors. This includes curating the dataset, analyzing predictions, and assessing the model's ability to generalize its findings. Through this thorough evaluation, researchers gain a deeper understanding of how the model performs and its effectiveness in combating the spread of fake news. These insights enable them to make improvements and refine the model's capabilities, ensuring it becomes more reliable and efficient in distinguishing between genuine and fabricated news articles.

5.1 Confusion Matrix

The confusion matrix is a valuable tool used in evaluating the performance of fake news detection models. It provides a comprehensive overview of how well the model classifies news articles into different categories based on their true and predicted labels. The matrix consists of four important metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics represent specific outcomes of the classification process. In the case of fake new classification, the true positives correspond to the correct detection of fake news, while true negatives represent accurate detection of genuine news. On the other hand, false positives occur when genuine news is mistakenly classified as fake, and false negatives arise when fake news is incorrectly classified as genuine.

By analyzing the values in the confusion matrix, researchers can derive various evaluation metrics such as accuracy, precision, recall, and specificity. These metrics provide insights into the model's performance, allowing for a comprehensive assessment of its ability to classify fake and genuine news articles. The confusion matrix serves is a vital tool in assessing the model's performance and understanding its strengths and weaknesses, as it enables the researchers to have a better understanding on the performed procedures to refine the model and improve its accuracy and effectiveness in combating the dissemination of fake news.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 1: Confusion Matrix

5.2 Accuracy

Accuracy is one of the most common metrics and it is often used to evaluate machine learning models' performance. It provides a general assessment of the model's ability to accurate classify the input given. Accuracy metric refers to the ratio of the total number of correct classifications, which are the true positive and negatives, and the total number of articles in the dataset. Although accuracy metric provides instant feedback on correctness of the models, it is crucial to keep in mind that relying solely on accuracy may fail to provide the full image of the model's effectiveness, that is why it fails to capture potential imbalances in the dataset and to take into account the consequences of any misclassifications. Thus, it is important to complement the accuracy metric with other evaluation measures to obtain a more comprehensive understanding of the model's capabilities in detecting fake samples accurately.

$$Accuracy = \frac{TP + TN}{TotalNumberOfPredictions} \quad (1)$$

5.3 Precision

The precision metric is a fundamental measure used to indicate the model's ability in identifying fake samples, such as fake articles, by examining the proportion of articles classified as fake that are actually genuine. Precision is the division of the number of true positive classifications by the sum of true positive and false positive classifications. A high precision value indicates a low rate of misclassification, as the model demonstrates a strong ability to accurately label articles as fake. However, it's once again important to consider precision in conjunction with other evaluation metrics, as an

emphasis on precision alone may result in overlooking genuine samples falsely identified as fake.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

5.4 Recall

The recall metric is a crucial measure used in assessing the performance of binary classification models. Recall metric aims to provide an instant feedback on how well the model detects the actual fake samples. Recall, also known as sensitivity or true positive rate, is calculated by dividing the number of true positive classifications by the sum of true positive and false negative classifications. A high recall value indicates that the model is effective at capturing a significant proportion of the actual fake news articles. It is important, as always, to consider recall in conjunction with other evaluation metrics, because an emphasis on recall alone may lead to an increased number of false positive misclassifications.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

5.5 F1 Score

The F1 Score metric is a widely used measure and it provides a balanced assessment by considering both precision and recall, offering a comprehensive understanding of the model's effectiveness. The F1-score is calculated as the harmonic mean of precision and recall, ensuring that both metrics contribute equally to the final score. This metric is particularly useful in scenarios where precision and recall are both crucial and need to be weighted equally. A high F1-score indicates a model that achieves a good balance between correctly identifying fake news articles and minimizing false positives. Combining the F1-score with other evaluation metrics gives a holistic view of the model's performance.

$$F1Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4}$$

5.6 ROC and AUC

The Receiver Operator Characteristic (ROC) curve is used to provide a visual representation of the trade-off between the true positive rate, or sensitivity and recall as it is also known, and false positive rate, or specificity, at different classification thresholds, demonstrating the model's ability to distinguish between genuine and fake news articles. The AUC, computed as the area under the ROC curve, serves as a comprehensive measure of the model's discriminatory power. A higher AUC value indicates better performance in correctly ranking fake samples higher than genuine samples. Nevertheless, it is important to consider that the ROC curve and AUC metrics assume a fixed classification threshold and may not fully capture the nuanced costs associated with misclassifications.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + FN} \tag{6}$$

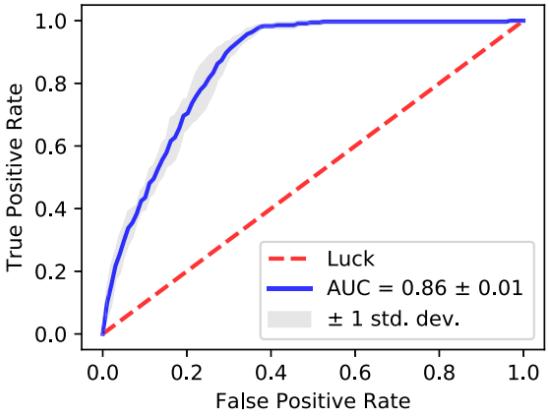

Figure 2: ROC and AUC paradigm

6 Explainability

Explainability is crucial for the development and evaluation of fake news detection models, as it refers to the ability to understand and interpret the model's decision-making process, providing details and information on how and why certain predictions are made. Explainability in fake news detection models is vital for several reasons. Firstly, it enhances transparency, allowing researchers and users to comprehend the factors influencing the model's classifications. This transparency fosters trust and accountability, as it enables the designers to validate the model's decisions and identify potential biases or limitations. Secondly, explainability facilitates model improvement and refinement by understanding the features, patterns, or linguistic cues that contribute to the classification process. Thus, researchers can identify areas for enhancement, refine the model's decision boundaries and address any potential vulnerabilities. Explainability promotes regulatory and ethical considerations and

enables compliance with legal and ethical frameworks by providing justifications for the model's predictions, ensuring accountability towards the propagation of biased information.

Achieving explainability in fake news detection models can be challenging because lot of these models employ algorithms, such as deep learning, which are often described as "black boxes" (Mishima and Yamana, 2022) due to their intricate inner workings. This opacity limits explainability as it becomes difficult to trace and interpret the model's decision-making process. Efforts are being made to develop techniques and methodologies that enhance explainability in these models. This includes using interpretable machine learning algorithms, such as decision trees or rule-based models, which offer embodied transparency at a higher degree (Shu et al., 2019). Additionally, researchers are exploring post-hoc interpretability methods (Shu et al., 2019), such as feature importance analysis or attention mechanisms, to shed light on the specific features or regions of input data that contribute to the model's predictions. Overall, achieving explainability in fake news detection models is a vital area of research that holds the potential to enhance transparency, trust, and effectiveness in combating the spread of misinformation.

6.1 Intrinsic Explainability

The goal of intrinsic explainability is to ensure that the model's predictions and classifications can be understood and justified without relying on additional tools and techniques. To achieve the intrinsic explainability, researchers explore different kind of approaches. These approaches include developing models based on transparent algorithms, like decision trees and rule-based systems (Meesad, 2021), where the decision rules are explicitly visible and understandable. These models allow researchers to examine and interpret the logic behind the classifications. Furthermore, techniques like feature importance analysis and attention mechanisms are employed (A.B. et al., 2023) to highlight the input features or regions of the data that significantly contribute to the model's predictions. By understanding the intrinsic explainability of the model, researchers can gain insights into how the model discerns between genuine and fake news, providing transparency, accountability, and a basis for further model improvements.

6.2 Post-hoc explainability

Post-hoc explainability techniques can be used to reveal the decision-making process of complex models that lack inherent interpretability and to do so they involve the analysis of model predictions after they have been generated. Researchers employ various approaches (Mishima and Yamana, 2022), including the use of external models or tools, to unravel the black-box nature of these models. For instance, researchers may employ surrogate models or simplified rule-based models that mimic the behavior of the original model to provide more interpretable explanations (Le et al., 2020) (Szczepański et al., 2021). Additionally, visualization techniques and feature importance analysis can be leveraged to uncover the salient features or patterns contributing to the model's predictions. The utilization of those external models or tools enables a deeper understanding of the underlying mechanisms and reasoning employed by the initial complex model.

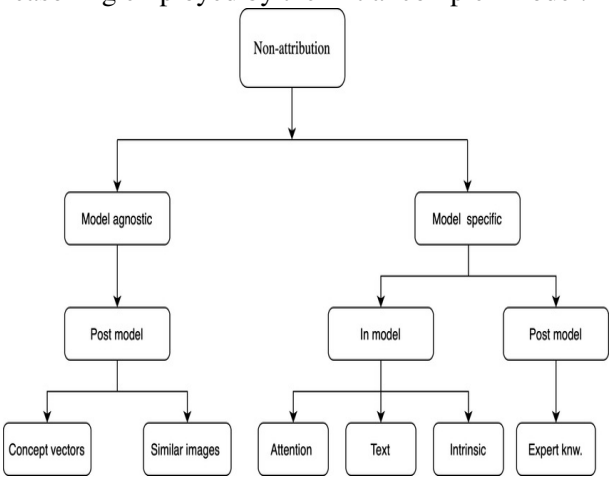

Figure 3: Explainability approaches

7 Available Datasets

Plenty of fake news and medical misinformation datasets are made public and used to train models. However, only a small number of datasets can evaluate interpretability. The following datasets, as shown in Figure 4, are frequently used in research on fake news detection and its explainability. If one notices the names of the column i.e. *visual content* and *Social context*, it is clear that different data sets have different kinds of features available. Therefore, the choice of data set depends on the feature selection. As discussed earlier in section 4, and in the papers ((A.B. et al., 2023), (Kaliyar et al., 2021) and others) there are three kinds of features that are used for fake news detection

which are text-based, image-based, and social context features. So naturally, the first criterion in the choice of a dataset is based on what features the dataset contains. Once this is clear another important factor is the size of the data set because the larger is the data the more efficient is the training of any model. In this regard, as per 4, the news bag dataset seems like a good choice. However, it is also vital to check if the dataset is well balanced with respect to real vs fake news articles included in it. As is shown in Figure 4, the *News Bag dataset* is very unbalanced with *200,000* to *15000* records of real and fake news articles respectively (Jindal et al., 2019). This is because fake news tends to be fewer in number in comparison to real news and thus it is difficult to scrape those. That is why the creators of the *News Bag dataset* extended it by using data augmentation techniques to create a more balanced and larger dataset i.e., *News Bag +++* (Jindal et al., 2019).

Dataset	No. of real news articles	No. of fake news articles	Visual Content	Social Context	Public Availability
BuzzFeedNews	826	901	No	No	Yes
BuzzFace	1,656	607	No	Yes	Yes
LIAR	6,400	6,400	No	No	Yes
Twitter	6,026	7,898	Yes	Yes	Yes
Weibo	4,779	4,749	Yes	No	Yes
FacebookHoax	6,577	8,923	No	Yes	Yes
TI-CNN	10,000	10,000	Yes	No	Yes
FakeNewsNet	18,000	6,000	Yes	Yes	Yes
NewsBag Test	**11,000**	**18,000**	Yes	No	Yes
NewsBag	**200,000**	**15,000**	Yes	No	Yes
NewsBag++	**200,000**	**389,000**	Yes	No	Yes

Figure 4: Publicly available datasets

7.1 FakeNewsNet

The FakeNewsNet dataset (Shu et al., 2018) is a substantial dataset made up of both authentic news articles from trusted sources and false news pieces from shady sources. It contains a variety of data kinds, such as news content, structural data (such as URL and title), and social context elements (such as retweet count and user information). This dataset enables researchers to investigate several facets of false news detection, such as explainability.
These are the optional Feature contents that can be chosen by the user according to their own flexibility.

- **news_articles:** This option downloads the news articles for the dataset. images: is a list of the URLs of all the images in the news article web page.
- **tweets:** This option downloads tweet objects posted sharing the news on Twitter. This makes use of Twitter API to download tweets.
- **retweets:** This option allows to download the retweets of the tweets provided in the dataset.
- **user_profile:** This option allows to download of the user profile information of the users involved in tweets. To download user profiles, tweet objects need to be downloaded first in order to identify users involved in tweets.
- **user_timeline_tweets:** This option allows to download up to 200 recent tweets from the user timeline. To download user's recent tweets, tweet objects need to be downloaded first in order to identify users involved in tweets.
- **user_followers:** This option allows to download the user followers ids of the users involved in tweets. To download user followers ids, tweet objects need to be downloaded first in order to identify users involved in tweets.
- **user_following:** This option allows to download of the user following ids of the users involved in tweets. To download the user's following ids, tweet objects need to be downloaded first in order to identify users involved in tweets.

However, this dataset is highly unbalanced with respect to the sources. The number of True News, as well as Fake news from GossipCop, is way higher than that of PolitiFact fact-checking sources, which is evident in the Figure 5.

Platform	PolitiFact	GossipCop
# Users	68,523	156,467
# Comments	89,999	231,269
# Candidate news	415	5,816
# True news	145	3,586
# Fake news	270	2,230

Figure 5: Statistical Analysis of Fake News Net Dataset

7.2 LIAR DATA

Fact-checking and fake news detection are the main objectives of the benchmark dataset LIAR (Wang, 2017a). It contains 12.8 thousand simple statements that have been manually labeled and checked

for authenticity by the editors of Politifact.com's API 5 (Wang, 2017b). It includes claims classified according to their degree of veracity (true, mostly true, half true, barely true, false, and pants on fire) as well as the news sources mentioned. For the purpose of explaining how fake news works, this dataset serves as a foundation for research on its linguistic and content-based characteristics. However, it does not contain any features based on images or social context which makes it limited. Figure 6 shows and example taken from this dataset.

Statement: *"The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero."*
Speaker: Donald Trump
Context: presidential announcement speech
Label: Pants on Fire
Justification: According to Bureau of Economic Analysis and National Bureau of Economic Research, the growth in the gross domestic product has been below zero 42 times over 68 years. Thats a lot more than "never." We rate his claim Pants on Fire!

Figure 6: Example sample from Liar dataset

7.3 NewsBag: A Multimodal Benchmark Dataset for Fake News Detection

Fake news articles use text and images to manipulate and spread false information. To address this issue, the authors proposed two new benchmark datasets that combine text and images, aiming to improve the quality of fake news detection (Jindal et al., 2019). The first dataset comprises manually collected real and fake news data from various online sources, while the second dataset explores data augmentation techniques using a Bag of Words approach to increase the amount of fake news data. These datasets are larger than existing ones. Extensive experiments were conducted, training both uni-modal and multi-modal fake news detection algorithms on the proposed data sets and comparing the results with those obtained from existing data sets. The findings demonstrate the effectiveness of the proposed data sets, indicating that augmenting data to increase the quantity of fake news does not increase detection accuracy. Furthermore, utilizing multi-modal data significantly outperforms uni-modal algorithms in detecting fake news.

7.4 FakeHealth

FakeHealth is a new explainable dataset about health misinformation developed by (Dai et al., 2020). The dataset includes an explanation of the ground truth in addition to common elements like news content, social engagement, and user information.

8 Research directions

The research directions listed below can be utilized as inspiration for future work:

8.1 Analysing the importance of visual contents used to spread news and explainability of fake news detection methods

A promising area for examination is to understand the importance of visual features in detecting fake news. Researchers can draw up methods for detecting and understanding fake news more easily, by analyzing elements such as image composition and context (A.B. et al., 2023). The research is designed to combat misinformation and empower citizens and platforms with tools that will improve their ability to respond effectively to social media news.

8.2 Explainability fake News Detection: Investigating Feature Selection for Improved Explainability"

It can be important to study the explainability of fake news detection by combining different elements of the news articles, it is possible to provide accurate identification of their authenticity (A.B. et al., 2023). An emerging area of research is aggregation techniques, which include models with different features in order to improve interpretation. This can include linguistic-based, content-based, source-based, and cross-referencing features. This approach aims at enhancing understanding and explanation of the detection of false news, using a wide range of features and their combined intelligence.

8.3 Investigating Multi-Class explainable detection of Fake news

There is currently no agreed-upon definition of false news, which makes it difficult to identify it. To this end, it can be beneficial to create a Multi-Class Explanation of Fake News Detection System

(A.B. et al., 2023) which may distinguish real news from other types and give explanations for its predictions. The aim of such a system is to make it possible to identify fake news from different points of view, overcoming limitations on the use of one definition. The system can efficiently categorize fake news into various categories, enabling a more sophisticated understanding of the misinformation landscape by means of its multi-class approach.

8.4 Exploring the correlation of textual and visual features of news content for explainable fake news detection

Another research area aimed at the development of explanatory models for the detection of false news is exploring the combination of text and image, as well as analyzing their correlation in Social Media posts (A.B. et al., 2023). Researchers can develop strong models to give clear explanations of the detection of fake news if they combine these features as this integrated approach not only leverages the power of both textual and visual information but also explores how they interact with each other, enhancing the accuracy and interpretability of the detection process.

8.5 Dataset Creation for explainable fake news detection

One important research area in explainable fake news detection is the creation of a structured and robust dataset (A.B. et al., 2023). Such a dataset can play a vital role in training and evaluating models for fake news detection while ensuring explainability. The dataset can include various attributes that contribute to the understanding and classification of fake news. These attributes may involve the type of fake news, distinguishing between satire, misinformation, or propaganda. Additionally, considering the writing style, such as formal, colloquial, or technical, can provide insights into the language patterns used in fake news articles. Political bias is another crucial attribute that can be included, classifying news articles as neutral, right-leaning, or left-leaning. Lastly, author information, such as past published articles and years of experience, can offer additional context for assessing the credibility of news sources. By incorporating these attributes into a comprehensive dataset, researchers can facilitate the development of more accurate and interpretable models for explainable fake news detection.

9 Conclusion

The worldwide spread of false information via social networking platforms has drawn considerable attention to the issue of online fake news. Its rapid expansion puts public safety at risk, necessitating the development of automatic detection systems. However, the lack of explainability in existing detection techniques makes detecting bias and discrimination difficult. This paper focused on explainable fake news detection and summarized some features, models, and evaluation techniques from the literature for this task based on explainability and fake news detection. Some light on a few available data sets was shed but it was observed that well-collected datasets with respect to the diverse features of news content are lacking. Moreover, a few research areas have been highlighted by this study where further research work can be carried out including new dataset creation through web-scraping and other techniques. Moving forward, there is a need for more advanced, detailed, fair, and practical detection models to address the intrusive and diverse nature of fake news across various topics, styles, and platforms.

References

Athira A.B., S.D. Madhu Kumar, and Anu Mary Chacko. 2023. A systematic survey on explainable ai applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122:106087.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research.

Jenny Anderson. 2017. Even social media-savvy teens can't spot a fake news story. 05.11.2023.

Meital Balmas. 2012. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation and cynicism. *Communication Research*, 41.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):853–862.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, Tanmoy, and Chakraborty. 2019. Newsbag: A multimodal benchmark dataset for fake news detection.

Rohit Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Deepfake: improving fake news detection using tensor decomposition-based deep neural network. *The Journal of Supercomputing*, 77.

Thai Le, Suhang Wang, and Dongwon Lee. 2020. Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.

Phayung Meesad. 2021. Thai fake news detection based on information retrieval, natural language processing and machine learning. *SN Computer Science*, 2(6):425.

Ken Mishima and Hayato Yamana. 2022. A survey on explainable fake news detection. *IEICE Trans. Inf. Syst.*, 105-D(7):1249–1257.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6086–6093. European Language Resources Association.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

Michael Robb. 2017. Our new research shows where kids get their news and how they feel about it. 05.11.2023.

Kenneth Scott. 1933. The political propaganda of 44-30 b. c. *Memoirs of the American Academy in Rome*, 11:7–49.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. Fake news detection on social media: A data mining perspective. *CoRR*, abs/1708.01967.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017b. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2017c. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on facebook. 05.11.2023.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

William Yang Wang. 2017a. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

William Yang Wang. 2017b. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. volume 57, page 102025.

Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ data science*, 9(1):7.

Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825. CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017.