Fighting Disinformation: Overview of Recent AI-Based Collaborative Human-Computer Interaction for Intelligent Decision Support Systems

Tim Polzehl[1,2], Vera Schmitt[2], Nils Feldhus[1], Joachim Meyer[3] and Sebastian Möller[1,2]
[1]*German Research Center for Artificial Intelligence, Berlin, Germany*
[2]*Technische Universität Berlin, Berlin, Germany*
[3]*Tel Aviv University, Tel Aviv, Israel*

Keywords: Disinformation, Fake Detection, Multimodal Multimedia Text Audio Speech Video Analysis, Trust, XAI, Bias, Human in the Loop, Crowd, HCI.

Abstract: Methods for automatic disinformation detection have gained much attention in recent years, as false information can have a severe impact on societal cohesion. Disinformation can influence the outcome of elections, the spread of diseases by preventing adequate countermeasures adoption, and the formation of allies, as the Russian invasion in Ukraine has shown. Hereby, not only text as a medium but also audio recordings, video content, and images need to be taken into consideration to fight fake news. However, automatic fact-checking tools cannot handle all modalities at once and face difficulties embedding the context of information, sarcasm, irony, and when there is no clear truth value. Recent research has shown that collaborative human-machine systems can identify false information more successfully than human or machine learning methods alone. Thus, in this paper, we present a short yet comprehensive state of current automatic disinformation detection approaches for text, audio, video, images, multimodal combinations, their extension into intelligent decision support systems (IDSS) as well as forms and roles of human collaborative co-work. In real life, such systems are increasingly applied by journalists, setting the specifications to human roles according to two most prominent types of use cases, namely *daily news dossiers* and *investigative journalism.*

1 INTRODUCTION

Artificial Intelligence (AI) technologies promise great opportunities in the fight against disinformation. Essential components of concurrent AI models include the areas of text analysis, audio analysis, image/video analysis, and their combination into a comprehensive and multimodal analysis of media content. In addition, disinformation disguises itself in multiple forms, such as media manipulation, media fabrication, and decontextualization of all media types. Following the recommendation of the *High-Level Expert Group* of the European Commission (EC), the term disinformation can be defined as "verifiable false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm" (HLEG, 2018). In the following, this definition is used to describe disinformation, and fake news interchangeably. The automatic identification of fake news items is inherently difficult for several reasons. News items have no clear, discrete truth value or verifiable evidence, and the truthfulness of items is on a continuum between clearly true and clearly false. Furthermore, the classification of news items depends on the viewer's prior beliefs and knowledge about relevant domains, and items can contain sarcasm and irony, which reverse their meaning. Therefore, detecting fake news still requires the involvement of human expertise, experience, and judgment. The EC further proposes a legal framework for *Harmonised Rules on AI* suggesting human supervision in safety-critical domains affecting human rights, which can be affected when claiming shared content as fake (Schmitt et al., 2021). Moreover, recent research shows that hybrid human-machine systems accomplish tasks that neither can do alone (Glockner et al., 2022). Hereby, Intelligent Decision Support Systems (IDSS) can support human judgment to facilitate the processing and classification of news items. In such systems, human and machine intelligence are joined in a collaborative framework. Often the human decision-maker monitors and

Polzehl, T., Schmitt, V., Feldhus, N., Meyer, J. and Möller, S.
Fighting Disinformation: Overview of Recent AI-Based Collaborative Human-Computer Interaction for Intelligent Decision Support Systems.
DOI: 10.5220/0011788900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 2: HUCAPP*, pages 267-278
ISBN: 978-989-758-634-7; ISSN: 2184-4321

interprets the performance and results of the AI system, which aids in identifying potentially problematic news items. Also, the AI component can actively request human input when the news item is in a predefined fakeness range. IDSS need to be desingend such, that humans can easily interact with the system and understand the provided content. Therefore, adequate Explainable AI (XAI) methods must enable users to understand the provided predictions, fostering trust in IDSS for collaborative fake news detection. Trust further includes the awareness of biases that can distort the predictions, oftentimes emerging from the data, the AI method itself, or the background of humans involved. Overall, a set of adequate design criteria for human-AI collaboration needs to be specified and aligned with the intended purpose of the system and specific use case the system is applied to. Disinformation detection can require domain experts to assess incoming information or media, but also broad human intelligence, e.g. crowdworkers can be incorporated for the collaborative fake news detection task. Humans may act out several roles, such as sensors, data qualifiers, anomaly checkers, context interpreters, or AI teachers, respectively, requiring different skills, knowledge, and availability.

In this work, recent developments on three levels are discussed. On the first level we provide an overview of recent developments of disinformation detection for different modalities. Second, we provide an overview of important requirements which need to be considered in the design process of an IDSS for the collaborative disinformation detection task. Third, we provide an overview of different roles of human intelligence and how it can be incorporated in an IDSS to improve the overall performance. Furthermore, we discuss two distinct use cases which can be observed most often nowadays, i.e., *daily news dossier* and *investigative journalism*. We discuss the roles of humans concerning their relation to AI-based system components and related trust aspects. We present a state-of-the-art literature overview on AI-based models in Sec. 2 first, followed by a discussion of important aspects of IDSS for fake news detection in Sec. 3. Different roles of humans for the joint disinformation detection task are discussed in Sec. 4, and realistic use cases from organizations leading the global fight against disinformation are scrutinized in Sec. 5.

2 AI-based MODELS

Current approaches for automatic credibility assessment of information can be deviated according to the data input they require. Most of the research in the fake news detection domain has been done for textual inputs. Hereby, various types of text items and sources have been used to train AI models, not only news articles, but also social media text messages have been analyzed. Models for fake news detection for images and videos mainly consider analysing deepfakes, which are often shared and used in the social media context. In the domain of fake news detection in speech recognition, there are very few approaches. Furthermore, approaches concerning multimodal fake news detection have emerged recently. Especially in the social media context images are often used in combination with text messages. Some models have been already proposed to handle both input formats at once and achieved reasonable performance. In the following, the state-of-the-art models for the different modalities and multimodal fake news detection are briefly described.

Fake News Detection for Text Items. The numerous approaches for automated textual analysis (Antoun et al., 2020) include dissemination pattern analysis (Liu and Wu, 2018), early disinformation detection and source analysis (Baly et al., 2018), and content-based approaches to disinformation detection, which in turn include methods for extracting lexical or syntactic and linguistic features. Here, disinformation is assumed to use misleading language and certain syntactic styles (Pérez-Rosas et al., 2018). Many approaches combine deep learning (DL) models with handcrafted features (Borges et al., 2019). Most recent results show that pre-trained deep language model classifiers such as BERT-based models (Szczepański et al., 2021), XLNet (Antoun et al., 2020), and GPT-3 (Nakov et al., 2022) perform better than feature-based models. This suggests that a deep understanding of the language is required to detect the subtle stylistic differences in writing disinformation (Antoun et al., 2020). Moreover, analyzing the repetition or reuse of news elements can also be informative in detecting fake news and sometimes combined with unsourced content (Evans et al., 2020). Additionally, NLP methods can be used to pre-process information to facilitate the work of experts identifying false content (Demartini et al., 2020). Accordingly, *truthfulness classification*, *check worthiness*, and *source identification* can be done by DL models and also in a hybrid collaborative setting incorporating crowdworkers (Glockner et al., 2022). This also applies to several datsets which have been published to train and evaluate large language models for the disinformation detection task. Some exemplary datasets for text items are *LIAR* (Wang, 2017), FakeNewsNet (Shu et al., 2018), FEVER (Thorne et al.,

2018), BuzzFeed-Webis (Potthast et al., 2018), RealNews (Zellers et al., 2019), FakeEdit (Nakamura et al., 2020), MultiFC (Augenstein et al., 2019), VitaminC (Schuster et al., 2021), COVID-Fact (Saakyan et al., 2021), and *Mocheg* (Yao et al., 2022), ClaimDiff (Ko et al., 2022), Emergent (Ferreira and Vlachos, 2016), SufficientFacts (Atanasova et al., 2022), RedHOT (Wadhwa et al., 2022), mainly published for the English language only, data on other languages is rare, e.g German *GermanFakeNC* (Vogel and Jiang, 2019). However, most DL methods apply different definitions of disinformation, different domains, context, and accuracy evaluation; therefore, further research is necessary to standardize disinformation detection for the text domain.

Fake Detection for Images and Videos. Advanced image and video editing tools facilitated the creation of fake video content and imagery, highlighting the need for better visual forensics algorithms (Huh et al., 2018). The fast recognition of perceptual image/video partial duplicates for verification purposes, especially for decontextualization analysis, can, in turn, be achieved by perceptual hashing (Thyagharajan and Kalaiarasi, 2021) and partial matching, modified by suitable visual features. One of the most prominent and severe phenomena that are rapidly growing are deepfakes. This term refers to all multimedia content that is somehow synthetically generated or altered by DL approaches. Hereby, DL methods are used to either automatically generate, alter or swap objects, e.g. a person's face in videos or images. Deepfakes are mainly based on autoencoders or Generative Adversarial Networks (GANs), which are becoming more accessible and accurate yearly. The synthesized media is very difficult to distinguish from real images or videos. Hereby, face swapping describes the process of transfering a person's face from a source image to another person in a target image while maintaining photorealism (Nirkin et al., 2018). To mitigate such risks, many deepfake detection approaches have been proposed (Zhao et al., 2021). By using vision transformers (ViT) and convolutional networks (Ding et al., 2020) as well as deep transfer learning (Coccomini et al., 2022), methods for face-swapping detection could be developed that provide high detection rates, including uncertainty estimates (Guarnera et al., 2020). Some further approaches have already been developed as countermeasures to face forgery (Qian et al., 2020; Li et al., 2020) mostly based on GAN-based models. Models to generate and detect deepfakes must be trained on lots of data. Some exemplary datasets which can be used for deepfake detection in images and videos are the DFDC dataset (Dolhansky et al., 2020), containing a large amount of face swap videos, and the WildDeepfake dataset (Zi et al., 2020), containing 7,314 face sequences extracted from 707 deepfake videos.

Fake Detection for Speech Recordings. DL-based speech synthesis has made great progress in recent years, mainly due to the end-to-end learning paradigm: text analysis, acoustic modeling, and speech synthesis are no longer isolated but integrated, trained, and optimized jointly, eliminating the need for expensive expert annotations and achieving ever-improving speech quality. Already methods such as Tacotron 2 (Shen et al., 2018) achieved high speech quality already. Again, GAN-based models recently prevail (Dhar et al., 2022). However, not only is the quality of speech improving but there is also an preference with *attackers* to use GANs because they can more easily be *hardened* against new recognizers that try to detect the fake. So-called *voice cloning* and *voice conversion* systems are freely available, achieve good speech quality in mimicking a target speaker's voice, e.g. (Luong, 2020; Sadekova et al., 2022; Huang et al., 2022), and can directly be used to fake custom speech of almost any target speaker, particularly important for human or automated person identification. Best systems are able to cheat speaker identification systems, with the largest evaluation organized as VoxCeleb Speaker Recognition Challenge (VoxSRC-21) (Kwon et al., 2021), containing over 1.1 million utterances from over 7,300 celebrity speakers to be recognized. In another way, audio manipulation detection tries to localize and falsify information about audio recordings, e.g., regarding recording device, time and location, encoding, etc., i.e. detect recording and processing traces (Aichroth et al., 2021), potentially exploit the fact that many subsequent manipulations of audio material cause inconsistencies concerning natural recording tracks that can be detected.

Comprehensive up to date tools for fake detection are urgently needed.

Multimodal Fake Detection. In practice, disinformation almost always manifests itself in multiple modalities. The development of combined analysis methods that are as diverse as possible is therefore crucial.

Multimodal systems have just recently begun to mature until a degree of practical applicability, e.g., the evaluation of messages and associated images in social media, e.g. SpotFake uses NLP models such as BERT to learn text features, in conjunction with VGG-19 to consider image features (Singhal et al.,

2019), and others (Dhawan et al., 2022; Palani et al., 2022) also for general fake news detection (Singh et al., 2021).

Most recently, (Fung et al., 2021) proposes a fine-grained, knowledge element-level cross-media information consistency checking for fake news detection, where knowledge elements include entities, relations and events extracted from the message body, headlines, images and meta-data of news articles. The authors run experiments on two datasets: (1) The NYTimes-NeuralNews, an established benchmark for multi-media fake news detection with pristine news articles collected by (Biten et al., 2019) fake news generated by Grover in (Tan et al., 2020), as well as (2) a proposed new VOA-KG2txt dataset, which consists of 15k real news articles scraped from Voice of America and 15k machine-generated fake news articles. Comparing against recent baselines of (Tan et al., 2020; Zellers et al., 2019) the authors reached 94.5% and 92.1% detection accuracy, respectively. However, these results must be interpreted with a grain of salt, as current NLP fact-checking benchmark models cannot realistically combat real-world disinformation (Glockner et al., 2022), specifically because it depends on unrealistic assumptions about counter-evidence in the data and/or found evidence may not be strong enough to refute disinformation up to the level required in real life. Finally, the authors also demonstrate that models trained on large-scale fact-checking datasets rely on leaked evidence, casting even more doubt on the interpretability of benchmark results. Overall, one major challenge multimodal fake news detection is the lack of standardization. There are no standards established for rating the fakeness of an item (e.g., binary vs. continuous credibility assessment), how to deal with biases in data and models, how human intelligence can be integrated in the collaborative fake news detection task also with respect to the various use cases, where human intelligence is needed.

3 IDSS FOR FAKE NEWS DETECTION

For the domain of fake news detection, previous research has shown that hybrid human-machine systems outperform settings where only humans or machines are used (Kapantai et al., 2021; Glockner et al., 2022). However, there are different requirements that need to be fulfilled for hybrid human-AI fake news detection (Nasir et al., 2021). Criteria that have been identified to be important are the transparency, usefulness, and understandability of model predictions (Lopes et al., 2022), but also user interface design and user experience criteria (Schulz et al., 2022). Furthermore, to create reliable overall predictions of the hybrid fake news detection task, the consideration of different biases is crucial, as biases in data, the model, or also human biases can influence model predictions and human assessment of the given information (Mehrabi et al., 2021). Therefore, these aspects are described in more detail in the following.

XAI. Fake news detection is an ambiguous task with a lack of consensus on definitions of what can be determined as being true or not. Recent research shows that providing explanations for methods of automatic credibility assessment increases human understanding, trust, and confidence in the AI system for certain tasks (Vilone and Longo, 2021; Lopes et al., 2022). In recent years, much work aimed to develop methods for improving the transparency and personalization of AI-based systems (Schneider and Handali, 2019). Hereby, XAI explanations should answer the questions for a human observer of how models work and why a prediction is made for a particular input (Mohseni et al., 2021; Kotonya and Toni, 2020).

XAI methods can be broadly divided into three different types (Zhou et al., 2021; Lopes et al., 2022): (1) Attribution-based explanations are one of the most common types of explanations and are used to produce importance scores for each input feature based on its relevance for the final prediction (Hase et al., 2021). (2) Rationalization, i.e., textual explanations that are generated by language models. This can either be done in a *post-hoc* fashion where a separate model, e.g., GPT-3, extrinsically tries to *make sense* of the input (and the prediction) (Wiegreffe et al., 2022) or *ad-hoc* with a model that jointly produces both prediction and explanation (Atanasova et al., 2020). (3) Example-based explanations can either manifest as finding very similar examples in the training data (Das et al., 2022) or generating counterfactuals (Dai et al., 2022).

On top of those, interactive tools have been devised for model analysis (Hoover et al., 2020; Tenney et al., 2020; Geva et al., 2022, i.a.), but require a thorough understanding of the explained AI models and thus are mostly aimed at AI model developers. Tools such as dEFEND (Shu et al., 2019), Propagation2Vec (Silva et al., 2021), and XFake (Yang et al., 2019) are more targeted at professional fact checkers.

There are several limitations that have been addressed in previous literature (Lopes et al., 2022): (1) there is still a lack of evaluating XAI approaches on a broad scale and different domains and standardized evaluation procedures (van der Waa et al., 2021),

which is vital to ensure that the integration of XAI methods fulfills the desired goals. (2) Also, a lack of visualization and interaction strategies can be identified. Thus, usability evaluation criteria and context-specific requirements need to be considered (Liao et al., 2022), and the role of dialogue-based explanations needs to be assessed (Feldhus et al., 2022). (3) One of the main shortcomings is the lack of multidisciplinarity (e.g., computer science, HCI, social sciences (Miller, 2019) in the creation and evaluation of XAI methods (Mohseni et al., 2021). As explainability is an inherently human-centric property, research in human-computer interaction can contribute to evaluating objective and subjective useful XAI approaches for different domains and tasks (Lopes et al., 2022). Hence, a multidisciplinary approach to XAI is required to significantly improve the application and display of comprehensible and robust XAI methods and incorporate objective and subjective evaluation metrics (Mohseni et al., 2021).

Bias. AI systems are usually data-driven and the prediction performance, generalizing ability, and usable results depend heavily on the availability of data (Mehrabi et al., 2021). Especially in the domain of fake news detection, where there exist subtle and subjective differences in defining the degree of the fakeness of an item, bias plays an eminent role (Zhu et al., 2022). Thus, for the fake news detection task, different types of biases can be identified, which need to be considered in the collection of data, training of AI models, and presentation of model results. There are many different biases identified in related research so far, which can be broadly classified into three different types (Mehrabi et al., 2021): (1) Data bias: *representation biases* exist in unbalanced datasets and are not representative of the respective population the data is sampled from (Zhu et al., 2022). Moreover, *historic biases* can emerge from human biases and perspectives deeply rooted in different societies (Daneshjou et al., 2021) or *measurement biases* occur when features and labels are used to measure a construct that is not directly encoded, or observable in the data (Suresh and Guttag, 2021). (2) Algorithmic bias: can be caused by the selection of model-specific parameters such as the optimization function or regularization method (Kordzadeh and Ghasemaghaei, 2022; Baeza-Yates, 2022). (3) Human bias: several human level biases could be identified in previous research (Mehrabi et al., 2021; Draws et al., 2022), whereas *social biases* (Gumusel et al., 2022; Čartolovni et al., 2022), *confirmation bias*, *behavioral bias*, and *emergent bias* addressing design biases based on cultural values and societal knowledge which can differ among different user groups (Mehrabi et al., 2021). During data collection, training models, and design of user interfaces, various types of biases need to be considered where also XAI methods can help to highlight the existence of algorithmic and data biases. However, mitigating human biases is more challenging, as some biases are deeply rooted in beliefs, opinions, and social contexts connected to a human user of such systems. Thus, standardized approaches and techniques must be developed to mitigate such biases on different levels.

Usability Aspects. In collaborative decision-making scenarios, several aspects must be considered in the design of the IDSS when human intelligence is integrated to approve AI-based predictions or used as an additional signal for the prediction task, e.g. effectiveness, efficiency, and user experience. Effectiveness can be examined primarily through the Limited Inter-Rater Agreement of humans with the AI classification, for which the Epsilon-Corrected Root Mean Squared Error serves as the primary performance metric. System efficiency can be measured by the time it takes an AI model to classify a news item, but also by the Cognitive Load (Singh et al., 2021) of the user. Another usability aspect is the user experience (Schulz et al., 2022), where pragmatic or functional aspects, but also hedonic aspects (Meel and Vishwakarma, 2020) have to be considered. Moreover, trust can be used as a parameter that can be measured in a hybrid approach to disinformation detection by evaluating user choices [accept, reject, revise] regarding suggestions of analysis procedures (Chancey et al., 2017), but also by the perception of system performance, control over the system and transparency of the system (Mohseni et al., 2020). Preliminary research has been done on the usability aspects in the domain of collaborative fake news detection, but future research is needed to empirically validate the different aspects mentioned above and standardize UI criteria and usability aspects.

4 ROLES OF HUMANS

In the context of disinformation, human in the loop is a methodology that can provide human supervision and judgment. While expert judgments may not be directly replaced by crowd workers' judgments in this respect, naive or trained human online crowdworkers can provide reliable labels (Demartini et al., 2016). Also, dealing with human judgments, the impact of the humans' background, e.g., political bias, and the timeliness of the assessed statements have been ana-

lyzed (Roitero et al., 2020). Results show that also recent statements can still reliably be fact-checked by the crowd. Hereby, crowdworkers can be deployed especially when the effectiveness of fully automated credibility assessment of news items is very low (Elsayed et al., 2019). Especially for tasks such as analyzing public interest in the assessed content, the possible impact of false calims on the formation of opinions, and also to assess the timeliness of the content the crowd can be used (Elsayed et al., 2019).

Several collaborative human-machine systems have been proposed to detect false news in a collaborative setting, e.g. hybrid human-machine systems connected to crowds based on a probabilistic graphical model (Nguyen et al., 2020). A probabilistic model *CURB* was proposed (Kim et al., 2018) deciding when humans should check suspicious claims, i.e. check-worthiness. A Bayesian Inference model was proposed, which integrates crowd flagging for fake news detection (Tschiatschek et al., 2018). In another work, a hybrid detection model, using the text, the source of an article, and the user response as features was proposed (Ruchansky et al., 2017). Similarly, interactive frameworks were developed to determine the credibility of news items, integrating a collaborative learning system for the fast identification of fake news (Bhattacharjee et al., 2017). The *WeVerify* project gathered human-in-the-loop judgments through an open platform to verify the content, the source, and the analysis of disinformation flows (Marinova et al., 2020). Accordingly, the classification of fake news items strengthens the viewers' trust in the items that were not flagged, even if they are of dubious accuracy.

To date, the roles of humans throughout the process of fake news identification seem underspecified wrt. a systematic stratification of human skills, outreach, and knowledge needed in the various steps. We propose the following 5 roles: (1) human as sensor, (2) human as data verifier, (3) human as system anomaly/sanity checker, (4) human as context info and XAI interpreter, and (5) human as AI teacher.

Accordingly, concerning the emergence of any information, humans (individually or organized as crowds) can act as sensors or sensor networks registering and recording data. Taking photos, creating videos, or posting messages are essential examples of this role. Regardless of which way any data has been acquired, both experts and crowds are frequently used to qualify data, i.e., assess noisiness and/or filter out irrelevant bits. Also, annotation and labeling steps and proofing or finishing passes can be seen as qualifications in this regard. Data is now ready to be received by AI components. Monitoring the training as well as monitoring inference by means of learning curves, error margins, and classification results on known and unknown data, AI experts are needed to steer the training of AI models. However, once trained and in inference mode, a sanity check or check for abnormal inference can also be executed by semi-experts, trained laymen or crowds. Overall, this step safeguards high-quality, representative model building that potentially generalizes well. As argued above, the monitoring of AI-System behavior oftentimes requires methods from the field of explainable AI to shed light on inner model parameters. In contrast to the safeguarding role, another role needed to interpret the explanations is the domain expert, who can interpret XAI results like scores and counterfacts. In many cases, this will be a journalist able to verify and double-check model results and XAI explanations with the help of experience and potential further internet research for verification. Ultimately, a final verdict on the helpfulness of any AI-generated suggestion or prediction is essentially and inevitably with the journalists or users that act as domain experts. They can evaluate the findings against any kind of expert world knowledge like historical, processual, and content-related information. Finally, humans can also act as implicit or explicit AI teachers when providing corrections or approval information in return to the system, which then may correct its databases and prepare for retraining.

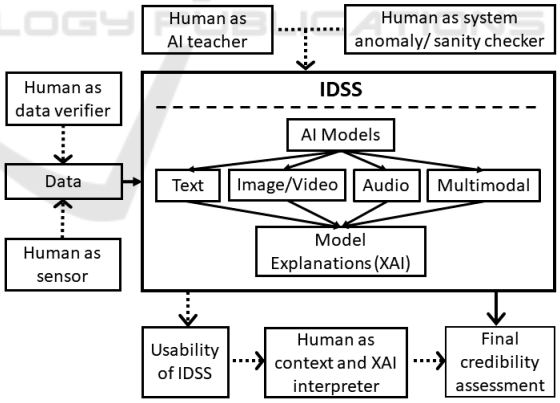

Figure 1: Components of IDSS and Human Roles.

In Figure 1 the IDSS, the requirements for collaborative fake news detection, and the human roles are depicted. Hereby, the human as data verifier and sensor is assigned to the level of the input to the IDSS. The IDSS consists of AI-based models for automatic fake news detection, but also contains the model explanations which are necessary for a useful integration of human intelligence for the collaborative fake news detection task. The IDSS also needs to be designed such, that the content and information is pre-

sented to the respective end user in the most comprehensible way. Hereby the human role as context information and XAI interpreter verifies the output of the XAI and IDSS and can also contribute to the final credibility rating. Furthermore, the human as anomaly detector and sanity checker monitors the overall system performance for anomalies and malfunction. The human as a teacher for AI can interact with the system on multiple levels. Hereby, the human not only receives the overall output and can verify a linkage to data, which can then be used in training again, but also has the opportunity to influence the AI-based model by verifying the training process, hyperparameters used, and training data distribution. Finally, biases affect the process on different levels, e.g., the data can already be biased due to imbalanced datasets or historical biases, biases can emerge from the model implementation and feature distribution itself, but also from human roles influencing the training and prediction process.

5 USE CASES

Daily News Dossiers. In addition to the usual sources such as own investigation results, correspondents' reports, or agency material, user-generated content from the Internet is increasingly being taken into account in creating news items. This content, distributed via social networks, gives media outlets faster access to event content without having journalists on the ground themselves and also puts issues on the agenda that might otherwise be overlooked. To date, this is enormously time-consuming and staff-intensive because automated, AI-based tools are only available in parts and in technological isolation.

Here, the most important roles of human collaboration are as a sensor (1) in order to provide social media clips (mostly naive users), as well as context information and XAI interpreter (mostly expert users) (4). Tight daily editorial deadlines disqualify non-verifiable news candidates by the mere time pressure and a limited number of news in a dossier. Thus, AI-based models may have the function to pre-filter and pre-qualify the abundance and massive information. For the majority of cases, no crowds are included, and the work is done by journalists and professionals. Many media companies have moved to work with fact-checking agencies or set up such units themselves within their corporate structures.

Investigative Journalism. Another equally important use case refers to private and public media organizations and non-governmental organizations (NGOs) striving to conduct comprehensive and in-depth investigative research on socially relevant grievances and malpractices. Thematically, they cover a broad spectrum from politics and business to the environment and society. The ultimate aim here is to provide trustworthy, honest and impartial background reporting. For investigative journalists, this means they have to be especially mindful and careful in their work, which oftentimes spans over several days or weeks. They also face hard-to-find sources and deliberate cover-up and disinformation tactics. In this context, it is essentially and inevitably important that the available information is screened as thoroughly as possible and checked for significance and authenticity down to the smallest detail. The larger well-known organizations in this environment include, for example, *Follow the Money*, *Bellingcat, Correctiv, Netzwerk Recherche, The Intercept, The Center for Investigative Reporting, The Global Investigative Journalism Network* and *EUObserver* as well as *Deutsche Welle*.

Here, non-verifiable information cannot simply be dropped. There is no limitation like a daily editorial deadline, forcing obscure information to be left out, but rather a longer duration of investigative time and efforts to be spent on a specific topic. Truthfulness, check-worthiness, or source reliability analyses will be carried out thoroughly potentially involving collaborative systems including both crowds and experts. The most important roles of human collaboration are thus as data verifier (2), as well as context and XAI interpreter (4). When using online platforms, this can be organized at scale to achieve all of speed, coverage, and high-quality assessments. Being able to consolidate, evaluate, and analyze information with respect to its context and sharing history is crucial to determine the credibility of the information. Still, AI models struggle with this task which is especially crucial for the use case of investigative journalism.

6 CONCLUSION

This paper aims to give an overview about (1) current developments for AI-based fake news detection for text, image/video, speech and mutlimodal fake news detection, (2) requirements, which need to be considered in the design process of an IDSS for collaborative fake news detection, and (3) how human intelligence can be incorporated in the IDSS to improve the overall performance. In addition to the short discussion of recent developments in AI-based automated modeling, we stress the need for human collaboration making the systems applicable for real-life applications. In alignment with the recommendation of the EC calling

to keep humans in the loop in scenarios where human rights (e.g., free speech) are affected, the incorporation of human intelligence also increases the overall performance of uni- or multimodal disinformation detection when extended to hybrid systems such as IDSS. Here, we identify and discuss explainability, bias, and usability aspects to be essentially calling for human collaboration, mostly in form of help and interpretation. Finally, looking more closely at the diverse roles humans resume throughout the explained processes, we identify 5 roles of humans paramount in the respective steps, which are to be seen in the light of realistic use cases for the most important concurrently applied disinformation fight, e.g. compiling daily news dossiers as well as conducting investigative journalistic inquiries.

REFERENCES

Aichroth, P., Cuccovillo, L., and Gerhardt, M. (2021). Audio forensics and provenance analysis: Technologies for media verification and asset management. *Journal of Digital Media Management*, 9(4):348—-366.

Antoun, W., Baly, F., Achour, R., Hussein, A., and Hajj, H. (2020). State of the art models for fake news detection tasks. In *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT)*, pages 519–524. IEEE.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2022). Fact checking with insufficient evidence.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Baeza-Yates, R. (2022). Ethical challenges in ai. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1–2.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. volume 2018-January.

Biten, A. F., Gomez, L., Rusinol, M., and Karatzas, D. (2019). Good news, everyone! context driven entity-aware captioning for news images. volume 2019-June.

Borges, L., Martins, B., and Calado, P. (2019). Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26.

Čartolovni, A., Tomičić, A., and Mosler, E. L. (2022). Ethical, legal, and social considerations of ai-based medical decision-support tools: A scoping review. *International Journal of Medical Informatics*, 161.

Chancey, E. T., Bliss, J. P., Yamani, Y., and Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, 59(3):333–345.

Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection. In *International Conference on Image Analysis and Processing*, pages 219–229. Springer.

Dai, S.-C., Hsu, Y.-L., Xiong, A., and Ku, L.-W. (2022). Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2800–2810, New York, NY, USA. Association for Computing Machinery.

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., and Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology*, 157(11):1362–1369.

Das, A., Gupta, C., Kovatchev, V., Lease, M., and Li, J. J. (2022). ProtoTEx: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2986–2997, Dublin, Ireland. Association for Computational Linguistics.

Demartini, G., Difallah, D. E., Gadiraju, U., and Catasta, M. (2016). An introduction to hybrid human-machine information systems. *Foundations and Trends in Web Science*, 7.

Demartini, G., Mizzaro, S., and Spina, D. (2020). Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Eng. Bull.*, 43(3):65–74.

Dhar, S., Jana, N. D., and Das, S. (2022). An adaptive learning based generative adversarial network for one-to-one voice conversion. *IEEE Transactions on Artificial Intelligence*.

Dhawan, M., Sharma, S., Kadam, A., Sharma, R., and Kumaraguru, P. (2022). Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478*.

Ding, X., Raziei, Z., Larson, E. C., Olinick, E. V., Krueger, P., and Hahsler, M. (2020). Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020(1):1–12.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Draws, T., La Barbera, D., Soprano, M., Roitero, K., Ceolin, D., Checco, A., and Mizzaro, S. (2022). The effects of crowd worker biases in fact-checking tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2114–2124, New York, NY, USA. Association for Computing Machinery.

Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Martino, G. D. S., and Atanasova, P. (2019). Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. volume 11696 LNCS.

Evans, N., Edge, D., Larson, J., and White, C. (2020). News provenance: Revealing news text reuse at web-scale in an augmented news search experience. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Feldhus, N., Ravichandran, A. M., and Möller, S. (2022). Mediators: Conversational agents explaining nlp model behavior. In *IJCAI 2022 - Workshop on Explainable Artificial Intelligence (XAI)*. International Joint Conferences on Artificial Intelligence Organization.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Fung, Y., Thomas, C., Reddy, R. G., Polisetty, S., Ji, H., Chang, S.-F., McKeown, K., Bansal, M., and Sil, A. (2021). Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 1683–1698.

Geva, M., Caciularu, A., Dar, G., Roit, P., Sadde, S., Shlain, M., Tamir, B., and Goldberg, Y. (2022). Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Glockner, M., Hou, Y., and Gurevych, I. (2022). Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. *arXiv preprint arXiv:2210.13865*.

Guarnera, L., Giudice, O., and Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667.

Gumusel, E., Malic, V. Q., Donaldson, D. R., Ashley, K., and Liu, X. (2022). An annotation schema for the detection of social bias in legal text corpora. In *International Conference on Information*, pages 185–194. Springer.

Hase, P., Xie, H., and Bansal, M. (2021). The out-of-distribution problem in explainability and search methods for feature importance explanations. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3650–3666.

HLEG (2018). Report to the european commission on a multi-dimensional approach to disinformation, igh-level expert group on fake news and online disinformation.

Hoover, B., Strobelt, H., and Gehrmann, S. (2020). exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Huang, J., Xu, W., Li, Y., Liu, J., Ma, D., and Xiang, W. (2022). Flowcpcvc: A contrastive predictive coding supervised flow framework for any-to-any voice conversion. In Ko, H. and Hansen, J. H. L., editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2558–2562. ISCA.

Huh, M., Liu, A., Owens, A., and Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117.

Kapantai, E., Christopoulou, A., Berberidis, C., and Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, 23(5):1301–1326.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. volume 2018-Febuary.

Ko, M., Seong, I., Lee, H., Park, J., Chang, M., and Seo, M. (2022). Beyond fact verification: Comparing and contrasting claims on contentious topics.

Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

Kotonya, N. and Toni, F. (2020). Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kwon, Y., Heo, H. S., Lee, B. J., and Chung, J. S. (2021). The ins and outs of speaker recognition: Lessons from voxsrc 2020. volume 2021-June.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. (2020). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010.

Liao, Q. V., Zhang, Y., Luss, R., Doshi-Velez, F., and Dhurandhar, A. (2022). Connecting algorithmic re-

search and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Liu, Y. and Wu, Y.-F. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423.

Luong, H.-T. (2020). Deep learning based voice cloning framework for a unified system of text-to-speech and voice conversion.

Marinova, Z., Spangenberg, J., Teyssou, D., Papadopoulos, S., Sarris, N., Alaphilippe, A., and Bontcheva, K. (2020). Weverify: Wider and enhanced verification for you project overview and tools.

Meel, P. and Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., and Ragan, E. (2020). Trust evolution over time in explainable ai for fake news detection. In *Fair & Responsible AI Workshop at CHI*.

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):1–45.

Nakamura, K., Levy, S., and Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *International Conference on Language Resources and Evaluation*.

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J. M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., et al. (2022). The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *European Conference on Information Retrieval*, pages 416–428. Springer.

Nasir, J. A., Khan, O. S., and Varlamis, I. (2021). Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Nguyen, V. H., Sugiyama, K., Nakov, P., and Kan, M. Y. (2020). Fang: Leveraging social context for fake news detection using graph representation.

Nirkin, Y., Masi, I., Tuan, A. T., Hassner, T., and Medioni, G. (2018). On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE.

Palani, B., Elango, S., Viswanathan K, V., et al. (2022). Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert. *Multimedia Tools and Applications*, 81(4):5587–5620.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer.

Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., and Demartini, G. (2020). Can the crowd identify misinformation objectively?: The effects of judgment scale and assessor's background.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. volume Part F131841.

Saakyan, A., Chakrabarty, T., and Muresan, S. (2021). COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Sadekova, T., Gogoryan, V., Vovk, I., Popov, V., Kudinov, M., and Wei, J. (2022). A unified system for voice cloning and voice conversion through diffusion probabilistic modeling. In *INTERSPEECH*.

Schmitt, V., Solopova, V., Woloszyn, V., and de Pinho Pinhal, J. d. J. (2021). Implications of the new regulation proposed by the european commission on automatic content moderation. In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pages 47–51.

Schneider, J. and Handali, J. (2019). Personalized explanation in machine learning. In *27th European Conference on Information Systems (ECIS 2019)*, volume abs/1901.00770, Stockholm-Uppsala, Sweden.

Schulz, K., Rauenbusch, J., Fillies, J., Rutenburg, L., Karvelas, D., and Rehm, G. (2022). User experience design for automatic credibility assessment of news content about covid-19. *arXiv preprint arXiv:2204.13943*.

Schuster, T., Fisch, A., and Barzilay, R. (2021). Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 624–643, Online. Association for Computational Linguistics.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. volume 2018-April.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *ArXiv*, abs/1809.01286.

Silva, A., Han, Y., Luo, L., Karunasekera, S., and Leckie, C. (2021). Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.

Singh, V. K., Ghosh, I., and Sonagara, D. (2021). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. (2021). New explainability method for bert-based model in fake news detection. *Scientific Reports*, 11(1):1–13.

Tan, R., Plummer, B. A., and Saenko, K. (2020). Detecting cross-modal inconsistency to defend against neural fake news.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., and Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Thyagharajan, K. and Kalaiarasi, G. (2021). A review on near-duplicate detection of images using computer vision techniques. *Archives of Computational Methods in Engineering*, 28(3):897–916.

Tschiatschek, S., Singla, A., Rodriguez, M. G., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Vogel, I. and Jiang, P. (2019). Fake news detection with the new german dataset "germanfakenc". In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, page 288–295, Berlin, Heidelberg. Springer-Verlag.

Wadhwa, S., Khetan, V., Amir, S., and Wallace, B. (2022). Redhot: A corpus of annotated medical questions, experiences, and claims on social media.

Wang, W. Y. (2017). "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.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. (2022). Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. B. (2019). Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, WWW '19, page 3600–3604, New York, NY, USA. Association for Computing Machinery.

Yao, B. M., Shah, A., Sun, L., Cho, J.-H., and Huang, L. (2022). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *arXiv preprint arXiv:2205.12487*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.

Zhu, Y., Sheng, Q., Cao, J., Li, S., Wang, D., and Zhuang, F. (2022). Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2120–2125, New York, NY, USA. Association for Computing Machinery.

Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.-G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390.