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 alliances, 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 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 desigeng 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:

1. **Level of AI**: Overview of recent developments of disinformation detection for different modalities,

2. **Level of IDSS**: Overview of important requirements which need to be considered in the design process of an IDSS for the collaborative disinformation detection task,

3. **Level of the human**: Overview of different roles of human intelligence and how it can be incorporated in an IDSS to improve the overall performance.

Furthermore, we propose and 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, 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 analyzing 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 (Szczechpański 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 datasets 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., 2019), 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 (Schütter 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, con-
text, and accuracy evaluation; therefore, further research is necessary to standardize disinformation detection for the text domain.

**Fake Detection for Images and Videos** Advances in image and video editing and manipulation tools have made it significantly easier to create fake video content and imagery, highlighting the need for better visual forensics algorithms (Hub 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 (Thyagarajan and Kalairasi, 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 transferring 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 VOAKG2txt dataset, which consists of 15k real news arti-
Fake news detection is an ambiguous task with respect to the various use cases, where human intelligence is needed. 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 (Mehrabiet 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 (Vö lone 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, for the domain of fake news detection,
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). (Baena-Yates, 2022). (3) Human bias: several human level biases could be identified in previous research (Mehrabi et al., 2021), whereas social biases (Gumusel 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 this cannot be easily influenced 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 (Meert 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 analyzed (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 claims 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 processing chain of a fake news detection application 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 counterfactuals. 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.](image)

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 presented 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 include some biases 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 of the fake news detection task.

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 calling for both crowds and expert collaboration. 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 multimodal 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


news content, social context and dynamic information for studying fake news on social media. ArXiv, abs/1809.01286.


