ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Role of Statistics in Detecting Misinformation: A Review of the State of the Art, Open Issues, and Future Research Directions

Zois Boukouvalas and Allison Shafer

Department of Mathematics and Statistics, American University, Washington, DC, USA; email: boukouva@american.edu

Annu. Rev. Stat. Appl. 2024. 11:27–50

First published as a Review in Advance on October 13, 2023

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040622-033806

Keywords

misinformation, multimodal learning, imbalance, explainability, fairness

Abstract

With the evolution of social media, cyberspace has become the default medium for social media users to communicate, especially during high-impact events such as pandemics, natural disasters, terrorist attacks, and periods of political unrest. However, during such events, misinformation can spread rapidly on social media, affecting decision-making and creating social unrest. Identifying and curtailing the spread of misinformation during high-impact events are significant data challenges given the scarcity and variety of the data, the speed by which misinformation can propagate, and the fairness aspects associated with this societal problem. Recent statistical machine learning advances have shown promise for misinformation detection; however, key limitations still make this a significant challenge. These limitations relate to using representative and bias-free multimodal data and to the explainability, fairness, and reliable performance of a system that detects misinformation. In this article, we critically discuss the current state-of-the-art approaches that attempt to respond to these complex requirements and present major unsolved issues; future research directions; and the synergies among statistics, data science, and other sciences for detecting misinformation.

1. INTRODUCTION

With the evolution of social media technologies, there has been a fundamental change in how information is accessed, shared, and propagated. Propagation of information, particularly misinformation, becomes especially important during high-impact events such as pandemics; natural disasters; terrorist attacks; and periods of political transition, unrest, or financial instability. During these periods, systems and entire societies become increasingly vulnerable to misinformation, and the spread of misinformation, even without malicious intentions, can easily cause significant damage. For instance, this has been recently observed during the coronavirus disease 2019 (COVID-19) pandemic, when misinformation such as harmful health advice, hoaxes, wild conspiracy theories, and racism was dangerously spreading.

Another example of misinformation spread through social media is "sharks on the freeway" posts, in which platform users circulate digitally manipulated images of sharks swimming on a flooded freeway. The same hoax has resurfaced over the course of a decade, including during Hurricane Irene in 2011 and Hurricane Harvey in 2017. Similar images showing sharks in suspicious locations emerged during Hurricane Sandy in 2012 and Hurricane Florence in 2018. Misinformation like this can impact evacuation efforts and keep aid from reaching those in need. While identical or similar images are produced for similar events, the impact of these posts demonstrates that the reach and consequences of even small bits of relatively easily identifiable misinformation are hard to predict. Malicious intentions or not, misinformation in social media can rapidly spread, affecting decision-making, communications, social harmony, and markets. It has also been claimed that spreading misinformation is a tactic actively used against democracies by state actors to push wedge issues, expand social fault lines, and sow discord.

Recent machine learning advances have shown significant promise for detecting misinformation (see Gupta et al. 2013, Sharma et al. 2019, Islam et al. 2020, Abdali 2022); however, the problem remains a significant challenge due to several key limitations. One such limitation relates to the use of multimodal data, i.e., information collected about the same phenomenon using different types of modalities (Abdali 2022, Damasceno et al. 2022). Multimodality has not been fully leveraged in intelligent systems, which traditionally use a single modality, typically text or images. Examples of multimodal data relevant for the detection of misinformation include, but are not limited to, textual information; images; network topology; and other content information or metadata such as hashtag topics (#), user references (@), and social content information such as user interactions, commenting, and reposting habits. To become more effective in detecting misinformation, machine learning algorithms must be able to understand content holistically. To address this challenge, the research community is focused on building intelligent systems that take the different modalities present in a particular post and then fuse them in a sophisticated way to enable the system to analyze them together, as humans do.

Another major limitation in misinformation detection is explainability—i.e., the ability of the model to summarize the causes of its decisions efficiently and hence gain the trust of its users (Hansen & Rieger 2019, Swartout & Moore 1993). The need for explainability becomes even more pronounced in high-impact events as it is key for an analyst to understand the significance of predictions and suggest mitigation. In addition, during such events, explainability is extremely important in the context of bias and ethical use of artificial intelligence since understanding the reasons behind certain predictions will enable users to identify potential discrimination against certain groups and demographics. It is important to underline that explainability is a broad umbrella term that includes interpretability while emphasizing that solutions need to be reliable and can be audited.

Last, a significant limitation in misinformation detection is the presence of labeled training data for building models. In an ideal world, to build a reliable model to detect misinformation,

we would need a large dataset of reliable posts and a dataset of posts containing misinformation. Unfortunately, finding or creating practical datasets for this problem is difficult. In practice, doing so is not feasible since detecting posts containing misinformation is inherently a class imbalanced problem (Branco et al. 2016). Indeed, the majority of posts are reliable, whereas a tiny minority contain misinformation. This, as well as other sources of biases, can cause the induced detector of misinformation to make decisions unfairly and, in so doing, to restrict freedom of expression unduly.

This article critically discusses the current state-of-the-art approaches that attempt to respond to these complex requirements and presents major unsolved issues; future research directions; and the synergies among statistics, data science, and other sciences for detecting misinformation. The article is organized as follows. In Section 2, we define the problem of misinformation detection and present the significant challenges related to this problem. In Section 3, we present current literature and approaches related to the problem of misinformation detection. In Section 4, we present open issues and future research directions. The article concludes with a broader discussion about the synergies among statistics, data science, and other sciences for detecting misinformation.

2. PROBLEM FORMULATION AND CHALLENGES IN MISINFORMATION DETECTION

Misinformation spreads more uncontrollably on social media than in traditional media outlets such as print media and television. One of the major differences between social media and these traditional media outlets is that the information on social media is crowdsourced. In contrast to television, print, or news websites where the sources of information are few and known (i.e., usually credible), users on Twitter act like its censors and fill in information gaps about an event. This difference is especially prevalent during high-impact events, which can be defined as newsworthy events by virtue of their timing, significance, prominence, and human interest. Examples of such events include, but are not limited to, forest fires, earthquakes, hurricanes, terrorist incidents, financial/market events, riots, and political unrest.

2.1. Problem Formulation

Rumors, clickbait, propaganda, humor, satire, fabricated content, unverified information, manipulated content, and imposter content are all forms in which misinformation can manifest. In the context of examining high-impact events, we follow the current practice in the literature to define misinformation as an umbrella term to include all false or inaccurate information that is spread in social media and use the terms misinformation and disinformation interchangeably.

This is a valuable heuristic because, on social media platforms where any user can publish anything, it is difficult to determine whether a piece of misinformation was deliberately created or not. In addition, more specific categories (e.g., fake news, rumor, misinformation) often overlap and are not exclusive (Wu et al. 2019). Following this definition allows for establishing a scope or boundary of the problem, which is crucial for designing a machine learning algorithm to effectively detect misinformation in a fairness-aware fashion enabling freedom of expression. Statistics and data science are intertwined throughout the life cycle of a typical misinformation detection pipeline, from the creation of data to the methods utilized in the algorithms detecting misinformation to the evaluation measures taken to determine the accuracy or performance of the algorithm. **Figure 1** illustrates the high-level idea of a traditional machine learning pipeline focused on detecting misinformation. We use the example mentioned above of misinformation spread during the initial spread of COVID-19, which resulted in serious negative consequences. As we see in the upper-left corner of **Figure 1**, each modality of the data collected from a social media platform,

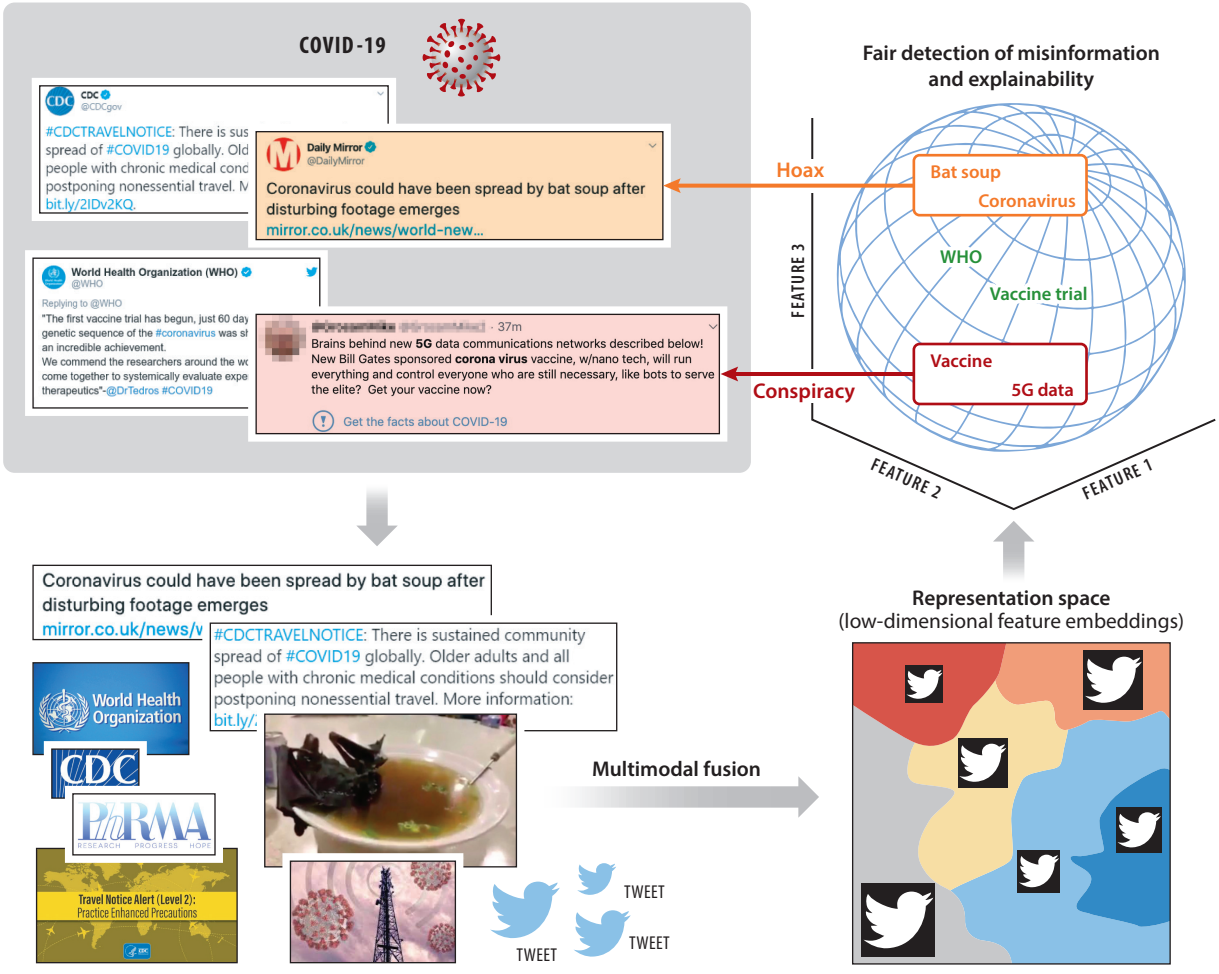

Figure 1

Graphical illustration of a traditional misinformation detection pipeline. The pipeline consists of raw data during a high-impact event; data preprocessing, including data cleaning and normalization; the construction of a low-dimensional representation space using multimodal learning fusion techniques; the fair detection of misinformation; and explainability.

such as the text and images from a post, as well as metadata about the user or post, represents different characteristics of the information being shared.

An example of a typical raw multimodal dataset is the MediaEval 2016 image verification corpus[1] (Damasceno et al. 2022), which includes separately labeled training and test tweet text and multimedia datasets collected in 2016. The training dataset consists of 9,140 tweet records associated with 352 different images and representing 15 unique events. Of the training data tweets, 5,127 are considered fake and 4,013 are considered real. Five of the 15 events in the training dataset include both real and fake tweets, while 10 of the events include only fake tweets. The test dataset consists of 796 tweet records associated with 92 different images and representing 23 unique events. The events represented in the training data and testing datasets are disjoint.

[1]The dataset is available at **https://github.com/MKLab-ITI/image-verification-corpus**.

Of the test data tweets, 467 are considered fake and 329 are considered real. Seven of the 23 events have both real and fake tweets associated with them, one event has only real tweets associated with it, and 15 of the events include only fake tweets. These raw data serve as the input data to the pipeline, and before training a machine learning model, they need to be preprocessed via tasks such as data cleaning and normalization. Normalization in natural language processing (NLP) refers to the process of transforming text into a standardized or canonical form, which can help improve the quality of downstream NLP tasks, such as text classification. Some techniques for text normalization include stop word removal, spell checking, lemmatization and stemming, and case normalization. Since the pipeline in **Figure 1** deals with multiple modalities, it is worth noting that the type of normalization depends on the data modality that is processed.

Since each modality provides different insights on how reliable a post may be, it is beneficial to perform multimodal data fusion to extract meaningful information from all the modalities in a joint manner, as demonstrated in the bottom half of **Figure 1**. There are multiple methods for multimodal data fusion, which are discussed later in this article. Additionally, when considering the use of multiple modalities, it is important to keep in mind that for each source, there are many features or extractions that can be used to represent the data, increasing the size or dimensionality of the data, which makes it difficult for humans and computers alike to determine which information is most impactful when determining the validity of the information in a post.

Often, methods to reduce the dimensionality of the data are utilized with the goal of obtaining as much information as possible from the high-dimensional data and to aid in identifying which features of the data are most meaningful for differentiating a reliable post from an unreliable one. The lower-right side of **Figure 1** depicts the data post–fusion and dimensionality reduction, showing the low-dimensional representation space spanned by the joint features. The low-dimensional data representations, along with different statistical modeling techniques such as logistic regression, decision trees, Bayesian networks, and neural networks, to name a few, can then be used to build models that classify the data as either misinformation or truthful information. These models are trained using labeled data and are then used to classify new data as either misinformation or genuine information, as depicted in the image in the upper right side of **Figure 1**. In this case, we see that the input tweet from the World Health Organization that featured text about a "vaccine trial" had characteristics that led the model to label the post as factual, while the tweet from a tabloid-like resource mentioning "bat soup" and "coronavirus" and the unverified source mentioning "vaccine" and "5G data" were labeled as misinformation. Finally, statistics are used to evaluate the performance of machine learning models in detecting misinformation. Metrics such as precision, recall, and the F1 score are commonly used to measure the model's accuracy. Statistical approaches such as cross-validation and hypothesis testing can be used to assess the model's generalizability.

2.2. Challenges in Misinformation Detection

Determining if a social media post is false or accurate is challenging for a number of reasons. We address five important challenges in the interdisciplinary area at the junction of statistics, data science, and multimodal machine learning, such that a trustworthy cyberspace can be maintained during high-impact events. The main challenges in misinformation detection include the abilities (*a*) to work with reliable, representative, and free-of-bias data; (*b*) to learn mutual relationships from data across multiple modalities in a robust manner; (*c*) to provide explainable predictions at a reasonable cost; (*d*) to establish reliable performance in a fairness-aware fashion; and (*e*) to assess the system's performance properly. These challenges become even more pronounced if the detection system needs to address all of those challenges in a joint manner.

2.2.1. Data-related challenges. Detecting misinformation poses several data-related challenges that must be addressed to build effective detection algorithms. One of the primary challenges in training machine learning models for detecting misinformation is the scarcity of high-quality labeled datasets that are large enough to train models effectively. In addition, the distribution of misinformation and genuine information is often imbalanced in real-world datasets, where the number of accurate posts present usually far outweighs the number of false or misinforming posts. Additionally, this imbalance can present bias to the majority class, making it more difficult to identify false information. This makes it challenging for machine learning algorithms to identify and classify misinformation accurately. Another significant challenge in misinformation detection for high-impact events is that misinformation continually evolves and adapts to new contexts. It is challenging to keep up with the new forms of misinformation, which can make it difficult to develop models that can detect new and emerging forms of misinformation. Last, misinformation does not impact only a single group of people; having detection models that can produce results for multiple languages is important. Datasets for misinformation detection need to be free of biases, including biases related to socioeconomic groups, etc. The training data used to develop machine learning models can often reflect underlying biases in the society or culture in which they were created. Misinformation is highly context dependent, and its meaning can vary depending on the context and cultural background of the audience. This can make it difficult to build models that can accurately detect misinformation across different contexts and languages and can lead to biased models that perform poorly on data from different contexts.

As we see in later sections, addressing these data-related challenges requires careful consideration of the data collection, annotation, and preprocessing techniques and the selection of appropriate machine learning models and evaluation metrics. Additionally, it may require interdisciplinary collaboration between machine learning researchers, domain experts, and social scientists to develop more robust and effective approaches for detecting misinformation.

2.2.2. Multimodal fusion in misinformation detection. Many social media posts are multimodal in nature—i.e., the data consist of information collected about the same phenomenon using different modalities, such as text, image, or video, to present the information. Multimodal data have yet to be fully leveraged in intelligent systems, which traditionally use a single modality, typically text or images. Machine learning algorithms must be able to understand content holistically to become more effective in detecting misinformation.

Traditional machine and deep learning algorithms have difficulties classifying multimodal content due to their limited ability to efficiently process and analyze the complex and interconnected information that exists across different modalities. Therefore, the research community has proposed different multimodal learning approaches to address this challenge. As presented by Atrey et al. (2010), Baltrušaitis et al. (2018), Hori et al. (2017), Zhao et al. (2017), and Kumari & Ekbal (2021), popular and suitable approaches for fusing multimodal data include early, late, and hybrid fusion approaches as well as attention-based fusion approaches.

Early fusion is a popular technique used in multimodal learning, where information from multiple modalities, such as images and text, is combined into a single representation early in the learning process. In early fusion, the input data from different modalities are merged together at the input layer of the deep learning model. For example, consider a multimodal learning task that involves classifying a social media post with an image and a textual description. In early fusion, the image and textual data are concatenated or stacked together, and this combined input is fed into the machine learning or deep learning model. The model then learns the joint representation of both the image and the text, which helps to improve the classification accuracy. Early fusion has several advantages. It allows the model to capture the interactions between the different modalities at an early stage, which can help the model learn more robust and discriminative features. It can

also reduce the computational cost of the model by sharing parameters across modalities, making the training process faster and more efficient. However, early fusion also has some limitations. It assumes that all modalities are equally important, which may not always be true. It may also lead to overfitting if the input features from different modalities are not well aligned or the input data are noisy. In addition, when combining multiple modalities through concatenation, the dimensionality of the feature space can significantly increase. This has been observed by Damasceno et al. (2022). This can pose challenges for traditional statistical approaches, including penalized regression or logistic regression, which may struggle to handle high-dimensional data efficiently.

By contrast, late fusion is a technique used in multimodal learning, where information from multiple modalities, such as images and text, is combined later in the learning process. In late fusion, each modality is processed separately, and the resulting features are combined at a later stage, typically after the output of each modality is computed. Considering the same example we used for early fusion, in late fusion, the image and textual data are processed separately by two independent machine learning or deep learning models. The output from each network is then combined at a later stage, such as by concatenating the feature vectors, and this combined feature representation is fed into a final classifier. Late fusion has several advantages. It allows the model to process each modality independently, which can help to capture the unique characteristics of each modality. It also allows the model to combine the features later, which can help reduce overfitting and improve generalization. However, late fusion also has some significant limitations. It may be less effective when the modalities are highly dependent on each other, such as when the textual description of an image is necessary to identify the image content. Although late fusion can reduce the dimensionality of the data in the early stages of modeling, it can still incur additional computational costs during the separate processing of each modality. Specialized models or algorithms may need to be applied to each modality, which can introduce computational complexity and resource requirements.

Hybrid fusion is a technique used in multimodal learning that combines the advantages of both early and late fusion. In hybrid fusion, some modalities are fused early, while others are fused later in the learning process. For example, consider a multimodal learning task that involves classifying an image with both textual and audio information. In hybrid fusion, the image and textual data may be fused early, while the audio data may be fused later. This can be achieved by processing the image and textual data together in the early layers of a neural network while the audio data are processed independently. The resulting feature representations from the early fusion and late fusion branches can then be combined and fed into a final classifier. Hybrid fusion has several advantages. It allows the model to capture the interactions between some modalities early on while preserving the independence of other modalities to be processed later. This can help to improve the accuracy and robustness of the model, especially when some modalities are highly dependent on each other while others are not. However, hybrid fusion also has some limitations. It can increase the computational cost and complexity of the model, as it requires multiple branches of a neural network. It may also require careful tuning and selecting the modalities to fuse early and late, as the optimal configuration may vary depending on the specific task and data.

Attention-based fusion uses attention mechanisms to learn the importance of each modality at different stages of the model. For example, in an image and text classification task, the model can use attention to focus on the relevant regions of the image and the important words in the text. Attention-based fusion provides a natural way to highlight the most relevant features or modalities contributing to a prediction. This makes it easier to interpret the model's decision-making process and identify potential areas for improvement. Moreover, attention-based fusion can help to reduce the impact of noisy or irrelevant features by giving more weight to the most informative modalities or features. This can improve the robustness of the model and make it less susceptible

to overfitting. However, attention-based fusion requires the selection of an appropriate attention mechanism, which can be challenging and computationally expensive, especially when dealing with large datasets or complex models.

The choice of a multimodal fusion approach depends on the specific task and the nature of the multimodal data. Each approach has strengths and limitations, and selecting the appropriate approach requires careful consideration of the data and the underlying model.

2.2.3. Explainability in misinformation detection. Explainability in machine learning refers to the ability of a model to provide a clear and interpretable explanation of its decision-making process. Overall, explainability is a critical aspect of machine learning systems geared toward detecting misinformation, as it enables users to understand and evaluate the system's decision-making process and identify potential biases and fairness aspects that need to be corrected in the pipeline before it is deployed at scale.

In the context of misinformation detection, one of the key components of explainability is to successfully demonstrate the existence of interpretable representations such that a detection system will be in full agreement with most of the desiderata for useful explanations of artificial intelligence as they have been defined by Hansen & Rieger (2019) and Swartout & Moore (1993). Some of the most relevant desiderata in this context refer to the system's ability to provide explanations that should be understandable to humans, including domain experts and nonexperts. The system should be able to explain its decision-making process and the factors that influenced the decision in a clear and meaningful way. In addition, the system should be able to justify its decisions based on the available evidence and data, and the explanations should be consistent with human knowledge and expectations. The explanations provided by an artificial intelligence system should not cause harm or unintended consequences. The system should be able to identify and mitigate any potential risks associated with its decision-making process and provide explanations that do not compromise privacy or security. Finally, the explanations provided by an artificial intelligence system should be easy to use and integrate into existing workflows and applications. This implies that the system should be able to provide explanations promptly and efficiently without requiring significant additional resources or expertise.

Two relevant explainability approaches for misinformation detection are local interpretation and global interpretation. Local interpretation refers to explaining an individual prediction made by a machine learning model. It aims to understand why a particular prediction was made by examining the features or variables most influential in that particular instance. Local interpretation is often used for tasks such as image recognition, where it is important to know which features are most relevant in classifying a particular image. Global interpretation, on the other hand, aims to explain the overall behavior of a machine learning model across multiple predictions. It looks at the patterns and trends in the model's output and the most influential features across the entire dataset. Global interpretation is often used for tasks such as fraud detection, where it is important to understand which factors are most indicative of fraud across the entire dataset. For a more detailed discussion of explainable artificial intelligence, we refer readers to Linardatos et al. (2021).

There are various methods and techniques for achieving explainability in machine learning, such as model-agnostic methods like Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016), SHapley Additive exPlanations (SHAP) (Lundberg & Lee 2017), and Testing with Concept Activation Vectors (TCAV) (Kim et al. 2018), as well as model-specific methods like decision trees and rule-based models. Although LIME, SHAP, and TCAV have successfully explained black-box models, they are not a panacea and have some limitations. For instance, they generate an explanation for each instance separately, which can result in complex explanations that are difficult to understand. In addition, their performance is sensitive to options the user selects as

they rely on several hyperparameters that must be chosen carefully to ensure that the explanation is accurate. Last, both LIME and SHAP may generate a simple and interpretable explanation that is accurate for the instance being explained but does not accurately represent the behavior of the model as a whole. This trade-off between interpretability and accuracy can be challenging to navigate. Additionally, TCAV comes with significant computational complexity in construction, dominating the cost of designing an intelligent system.

Statistical methods such as decision trees and rule-based models are popular approaches for building explainable machine learning models. A decision tree is a hierarchical model that represents a set of decisions and their possible consequences. The tree is built by selecting the feature that best splits the data into the most homogeneous subsets based on a chosen metric, such as information gain. Once the tree is built, it can be visualized and interpreted by following the path from the root to a specific leaf node. This path represents the set of rules that are used to classify a specific instance. Rule-based models are models that are constructed from a set of logical rules, where each rule consists of a set of conditions that are used to make a decision. For example, a rule-based model might say, "if feature X is greater than 20 and feature Y is less than 10, then classify the instance as positive." Rules can be generated using various techniques, including decision tree induction, association rule mining, or expert knowledge. Once the rules are generated, they can be used to classify new instances and can be easily interpreted by humans. Both decision trees and rule-based models have the advantage of being highly interpretable, as they explicitly represent the rules used to make decisions. However, they suffer from limitations such as difficulty handling large feature spaces and poor generalization performance. Additionally, these models may not be suitable for complex tasks requiring many feature interactions.

2.2.4. Reliable performance in a fairness-aware fashion. Misinformation detection is a challenging task due to the class imbalance problem, where a vast majority of posts are reliable while only a small fraction contains misinformation. For example, Twitter removed 2,230 misleading tweets between March 16 and April 18, 2020, out of the approximately 6,000 tweets posted per second.[2] This creates a significant class imbalance, with a ratio of 0.000014%, posing a considerable challenge for automatic detection. The class imbalance problem has been widely recognized as a persistent issue in machine learning over the past two decades. However, addressing this problem alone is not sufficient for creating a reliable misinformation detector, as it may overlook sensitive features and lack human supervision. Moreover, the class imbalance problem can lead to unfairness in the model, disproportionately impacting certain groups through false positives or false negatives.

In the past decade, significant progress has been made in addressing the class imbalance problem, as discussed by Branco et al. (2016) and Leevy et al. (2018). Resampling strategies like the Synthetic Minority Oversampling TEchnique (SMOTE) and its variations have proven to be effective in supervised settings. However, supervised approaches have notable drawbacks. First, they require labeled instances for training and testing, which can be costly and unfeasible in certain domains like detecting misinformation. Second, they struggle to accurately detect radical deviations of misinformation from previously encountered examples. Third, addressing the class imbalance problem itself can be challenging and ineffective in certain situations. These disadvantages stem from the need for labeled data, limited generalization to rapidly evolving misinformation, and the difficulty of mitigating class imbalances.

In addition, addressing class imbalance ensures fairness in the model and can be achieved through techniques such as oversampling the minority class, undersampling the majority class,

[2]For more details, we refer readers to Peters (2020).

or using data augmentation techniques. However, fairness in detecting misinformation is a significant challenge due to several factors. First, machine learning algorithms are only as unbiased as the data used to train them. Suppose the data used to train the algorithm are biased, for example, containing a disproportionate amount of misinformation from certain demographics or sources—in that case, the algorithm will learn and perpetuate these biases in its decision-making process. This is known as algorithmic bias, and it can result in certain groups being unfairly targeted as sources of misinformation or being unfairly impacted by the consequences of being labeled as purveyors of misinformation.

Another challenge in ensuring fairness in detecting misinformation using machine learning techniques is that the definition of what constitutes misinformation is subjective and can vary depending on cultural, political, and social contexts. This means that what is considered misinformation in one context may not be considered as such in another. This creates a challenge in training machine learning models that can accurately identify and distinguish between true and false information without perpetuating existing biases.

2.2.5. Evaluation. A critical step in the misinformation detection pipeline is to measure the model's performance on unseen data. Success in this phase depends on (*a*) using the right metric for evaluation and (*b*) following the proper evaluation process. Evaluations are of two types: intrinsic and extrinsic.

2.2.5.1. Intrinsic evaluation. Intrinsic evaluation in misinformation detection refers to the process of evaluating the performance of machine learning models on a specific dataset or task without considering external factors or use cases. Intrinsic evaluation is typically used during the development and testing of machine learning models to determine their accuracy, precision, recall, and other performance measures on a given task, such as identifying false information in news articles or social media posts. Accuracy is used when the output variable is categorical or discrete, and it denotes the fraction of times the model makes correct predictions compared with its total predictions. Precision measures the proportion of true positive results out of all positive results, while recall measures the proportion of true positive results out of all actual positive instances. Precision and recall can be used to evaluate the effectiveness of a machine learning model in identifying false information while minimizing false positives or false negatives. The F1 score is the harmonic mean of precision and recall and can provide a more balanced evaluation of the overall performance of a machine learning model.

2.2.5.2. Extrinsic evaluation. Extrinsic evaluation in machine learning refers to evaluating a model's performance in the context of a real-world task or application rather than simply evaluating its performance on a specific benchmark dataset. A popular extrinsic evaluation technique for the detection of misinformation is fact-checking, which involves verifying the accuracy of claims or statements made in news articles or other sources of information. Machine learning models can be evaluated on their ability to identify false or misleading claims and to provide evidence-based information to support their conclusions. In addition, machine learning models can be evaluated based on their ability to engage users with accurate and informative content and to reduce the spread of misinformation in online communities. This can be measured through metrics such as user engagement, social sharing, and click-through rates. In addition, machine learning models can be evaluated based on their ability to recommend news articles or other sources of information that are accurate and reliable while avoiding those that contain false or misleading information. This can help to promote informed decision-making and reduce the spread of misinformation.

Crisis management metrics refer to the performance measures used to evaluate the effectiveness of machine learning models in detecting and responding to crises or emergencies, such as

natural disasters or public health emergencies. For instance, response time is a crisis management metric that measures the time a machine learning model takes to detect a crisis or emergency and generate an appropriate response. A shorter response time can help reduce the crisis's impact and provide timely assistance to those affected. Another metric is user engagement, which measures interactions with the machine learning model during the high-impact event, such as the number of people who access or share the information provided by the model. A higher level of user engagement can help to increase the reach and impact of the machine learning model.

3. RELATED WORK

We now summarize the existing literature on data availability for misinformation detection and statistical machine and deep learning models.

3.1. Data for Misinformation Detection: Related Work

Generating datasets for the task of misinformation detection to simulate the real-world challenges discussed in Section 2.2.1 is crucial to improving techniques to detect misinformation. We can only advance misinformation detection models if we have a collection of reliable and practical data and can generate reliable high-level features. To provide datasets for use in developing models and algorithms to detect misinformation, data scientists have worked to compile such data. Murayama (2021) surveyed 118 datasets related to fake news on a large scale from three perspectives: fake news detection, fact verification, and datasets for other tasks. D'Ulizia et al. (2021) systematically reviewed the main characteristics of 27 popular datasets available for misinformation detection, including news domain, type of disinformation, content type, language, medium platform, size, etc. They concluded that there is still a need to develop benchmark datasets for fake news detection that are multimedia, multilingual, cross-domain, and large-scale. Readers are referred to this article for example datasets and related discussions.

3.2. Detection of Misinformation: Related Work

Statistical methods play a substantial role in the machine learning models utilized to classify a resource as reliable or unreliable. Data representations, hand-crafted or created with deep learning techniques, using approaches in the following subsections, often serve as input to machine learning algorithms using statistical methods such as random forests, support vector machines, naive Bayes-based models, K-nearest neighbors, or decision trees to analyze the features and determine to which class—real or fake—a social media post belongs. Furthermore, while evaluating misinformation detection algorithms and models, various statistics can be employed to determine the classification accuracy of the dataset and to gauge the significance and impact of various features in such models. In this section, we discuss several widely used approaches for feature extraction and disinformation detection and their limitations.

3.2.1. Approaches based on hand-crafted features. Features are the variables fed into machine learning models, and they house the properties or information of the analyzed dataset. The idea behind the hand-crafted features approach is to define features from textual and visual information to capture specific characteristics of misinformation. Then, a separating hyperplane is trained using a selected classifier (Gupta & Kumaraguru 2012, Zhao et al. 2015, Shu et al. 2018, Vosoughi et al. 2018). Although these approaches provide interpretable results to a certain degree, deciding the best features for a given problem is a nontrivial task. In most cases, the selection of the features is tied to the particular application affecting the generalization ability of the trained model.

Statistically derived information and NLP-based outputs are often used as hand-crafted features in a dataset. These include quantities such as cosine similarity, term frequency–inverse document frequency, word count, modeled topics, n-gram bag-of-words, and the number of particular punctuation marks. Other hand-created features include readability scores and parts of speech tagging (Bhatt et al. 2018, Chen et al. 2020, Fayaz et al. 2022). Additionally, metadata or user features such as user name, dates of posts, number of followers, age of the account, etc., have also been used as features in research (Amador et al. 2017, Hoang & Mothe 2018, Paka et al. 2021).

3.2.2. Approaches based on deep neural networks. Deep neural networks use algorithms and multiple hidden layers to recognize patterns in data and predict outputs. Deep neural networks are also used to automatically learn and extract features for misinformation detection (Shu et al. 2017, 2019; Wang 2017).

Classification tasks for misinformation detection can also be performed using deep neural networks, especially by a few well-known examples. Convolutional neural networks provide a final fully connected layer to output the probability of the classification label. Recurrent neural networks allow information to persist and use sequential or time series data; they can be used to classify sentiment, which can be used in misinformation detection. Long short-term memory (LSTM) can also process entire sequences of data, such as a sentence or textual work. Similarly, other deep learning networks with different architectures also offer classification: Inception-v3; Xception, which provides a final logistic regression layer; residual neural networks (ResNet); EfficientNet; and VGG. Studies that use neural networks to derive features or classify resources for misinformation detection include those of Tuan & Minh (2021), Kirchknopf et al. (2021), Kaliyar et al. (2020, 2021), Azri et al. (2021), Agarwal & Meel (2021), and Anusha et al. (2022).

These deep neural network approaches can learn the latent representation in a dataset and have shown great promise in classification performance. However, given the hidden layers and data transformations throughout the network, the interpretability of these methods is not direct, making the discovery and assessment of the connections between the high-level features and low representation space a significant challenge (Hansen & Rieger 2019). Even with the use of computationally costly and indirect tools such as LIME or TCAV, we cannot directly discover and assess the connections between the high-level features and low representation space, making it extremely difficult to identify potential biases in the data that could lead to unethical decisions.

3.2.3. Approaches based on Bayesian techniques. Most machine learning applications for misinformation detection require supervised or labeled data to train a classifying algorithm. However, as mentioned elsewhere in this article, datasets are somewhat sparse, and the creation of labeled data for misinformation detection is very time consuming and is subject to error due to the deceitfulness of misinformation. As an alternative to annotated data, some researchers have utilized Bayesian techniques to actively learn or model the credibility of resources and users (Yang et al. 2019, Sahan et al. 2021). Furthermore, companies have turned to crowdsourcing to flag activity or misinformation to minimize the spread of fake news. Tschiatschek et al. (2018) implemented Bayesian techniques to learn about the accuracy of the users' labels to determine which news was fake with higher confidence. Another limitation when conducting misinformation detection using specific methods is the inability to represent the uncertainty of the prediction due to incomplete or finite information about the claim. To address this, Zhang et al. (2019) propose a Bayesian deep learning model that outputs a distribution to represent the prediction and its uncertainty.

Other approaches employing Bayesian techniques include using a Bayesian network trained on a dataset of fake news to provide a probabilistic level of the truthfulness of the information (Casillo et al. 2020), using naive Bayes when comparing multiple classifiers for misinformation detection to study the performance of various methods (Khan et al. 2021, Han & Mehta 2019, Kumar &

Singh 2022, Abdullah-Al-Kafi et al. 2022, Mandical et al. 2020), and using a hybrid of naive Bayes and LSTM for fake news prediction (Rohera et al. 2022).

3.2.4. Approaches based on network topology. Network topology maps different nodes and is connected by edges or lines to give the data a shape, depicting the connections within a given network. In a social media network, network topology can visualize connections between users and the distribution of information, including the spread of misinformation. Using network topology and graph neural networks (GNNs) can allow us to leverage this relational information to help us understand the patterns of spreading on social media and even identify communities where misinformation persists or the origins of the spread of misinformation. Experiments indicate that social network structure is an important feature for allowing highly accurate fake news detection (Monti et al. 2019).

Network topology and GNNs have been employed for misinformation detection (see, e.g., Monti et al. 2019, Han et al. 2020, Nguyen et al. 2020, Rath et al. 2020, Jeong et al. 2022).

3.2.5. Approaches based on deep multimodal learning. Multiple approaches have been taken to integrate deep multimodal learning for misinformation detection (summarized by Hangloo & Arora 2022). One shortcoming of many of the current approaches for the detection of fake news is their inability to learn a shared representation of multimodal information, which would allow a misinformation detector to utilize the multimodal representations obtained to classify posts as fake. Some recent work in this domain includes that of Glenski et al. (2019), Agrawal et al. (2017), Khattar et al. (2019), Singhal et al. (2019), Cui et al. (2019), Jin et al. (2017), Wang et al. (2018), and Zhang et al. (2020). While these advances represent steps towards detecting misinformation using multimodal deep learning, in the methods described in these studies, multimodal features are generated through either simple concatenation or implicit modeling of multimodal associations. As a result, the synergistic nature of different modalities is not fully leveraged (Alam et al. 2021).

Enhancements to the methods and algorithms used to create joint representations of multimodal data in the same space to allow interactions or dependencies between the modalities to be explored for misinformation detection are sought after and have been demonstrated in the research of Damasceno et al. (2022), Kumari & Ekbal (2021), Wang et al. (2022), and Song et al. (2020).

3.2.6. Approaches based on latent variable methods. A recent alternative approach for detecting misinformation is based on latent variable analysis using matrix and tensor decompositions. Here, the relationships between articles and terms are captured through different tensor decompositions (Guacho et al. 2018, Hosseinimotlagh & Papalexakis 2018, Abdali et al. 2020), leading to groups of articles that belong to different categories of false news. In addition, nonnegative matrix factorization methods are used to cluster false and true documents based on the factorization of word–word cooccurrence matrices (Wu et al. 2017). More recently, we demonstrated the potential of independence by using independent component analysis (ICA) to detect unreliable tweets while enabling knowledge discovery through estimated interpretable feature embeddings (Boukouvalas et al. 2020, Moroney et al. 2021). Although these latent variable solutions have been shown to be reliable, efficient, and explainable, they either operate on single modalities or implicitly model associations among the multimodal data (see, e.g., Abdali et al. 2020, Moroney et al. 2021).

4. OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

The problem of disinformation detection is complex and multifaceted; therefore, we can claim with certainty that it is far from being completely solved. There are several open issues that

researchers and practitioners must address, including data-related issues, integration of sophisticated multimodal methods into intelligent systems, proper formulation of misinformation detection as a class imbalance problem, fairness issues, and human-centered approaches for effective misinformation detection and proper evaluation. Without underestimating the difficulty and importance of multimodality and class imbalance formulations, we strongly believe that human-centered approaches and data-related issues are the most significant. In this section, we address various open issues from distinct perspectives and suggest diverse directions for future research.

4.1. Data-Related Open Issues

There are several research directions related to data in misinformation detection. One important direction is the development of better methods for collecting and annotating data for misinformation detection. This involves creating high-quality labeled datasets that can be used to train and test machine learning models. This process is vital because, as previous work presented by Boukouvalas et al. (2020) highlighted, an unbalanced class problem exists. However, creating human annotations for a large input dataset collected from various social media sources during a high-impact event is very challenging. Hence, transfer learning, in which a machine learning model trained on one task is fine-tuned for another related task, can be employed. This can be useful for misinformation detection when limited labeled data are available for a specific domain or task. Unsupervised learning techniques can also be used to identify patterns and anomalies in large datasets without needing labeled data. Research in this area focuses on developing unsupervised methods for detecting misinformation and evaluating their effectiveness on real-world data. In addition, active learning, a semisupervised machine learning technique that involves iteratively selecting a subset of the dataset with human annotation, can be used. This approach is beneficial when the initial dataset is too large to be annotated entirely by humans or when the distribution of misinformation in the dataset is unknown.

Last, in practice, researchers can collect data from various social media platforms like Twitter, Reddit, or 4chan by creating a 24/7 data collection setup. Various publicly accessible application programming interfaces (APIs) and endpoints must be utilized to collect data from these social networks. In addition to the social media APIs, network-specific crawlers can be extracted to collect information from these sources. RSS (really simple syndication) feeds generated by various news sources like the BBC, CNN, and so on are also good sources of information for an intelligent system. However, the use of APIs for data collection faces limitations as social media platforms often restrict access to their data. These restrictions include limiting the amount of data that can be retrieved, imposing strict requirements on researchers, and restricting access to certain features like user comments. Additionally, social media platforms may limit access to historical data and impose restrictions on data sharing and usage. These API restrictions pose challenges for researchers in collecting and analyzing data for detecting misinformation.

4.2. Fusion Methods for Detection of Disinformation

Fusion methods play a crucial role in detecting misinformation. True fusion, as defined by Damasceno et al. (2022), involves learning the mutual relationships among modalities and generating joint representations. This is highly desirable in the context of multimodal learning. Damasceno et al. (2022) present an example of true fusion in which a novel multivariate data fusion framework is introduced. The framework utilizes pretrained deep learning features and a parameter-free joint blind source separation method called independent vector analysis (IVA) to detect misinformation. Specifically, the proposed multimodal framework employs the efficient multimodal fusion algorithm IVA-M-EMK (independent vector analysis using semiparametric

density estimation via multivariate entropy maximization), which captures complex, nonlinear relationships between textual and visual modalities.

To mathematically formulate true fusion as defined by Damasceno et al. (2022), let $\mathbf{X}^{[k]} \in \mathbb{R}^{d\times V}$ be the kth observation matrix from kth modality, where d denotes the number of initial high-level feature vectors in the kth modality and V denotes the total number of social media posts. The noiseless IVA model is given by

$$\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}, \;\; k = 1, \ldots, K, \qquad 1.$$

where $\mathbf{A}^{[k]} \in \mathbb{R}^{d\times N}$ is the kth mixing matrix, and $\mathbf{S}^{[k]} \in \mathbb{R}^{N\times V}$ are latent variable estimates, i.e., the kth set of source estimates, which in our setting, correspond to the features. The estimates of the features span the joint low-dimensional representation space and will be used to train a machine learning algorithm for the detection of misinformation. It is worth noting that when $K = 1$, Equation 1 reduces to a simple blind source separation problem with one modality, and the most popular way to achieve blind source separation is by using ICA (Moroney et al. 2021).

IVA provides a smart connection across multiple datasets through the definition of a source component vector (SCV), which enables one to take full statistical information across the multimodal datasets. Using the random vector notation (as opposed to the one written using observations in Equation 1), we write $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}$, $k = 1, \ldots, K$, where $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times N}$, $k = 1, \ldots, K$, are invertible mixing matrices; $\mathbf{s}^{[k]} = [s_1^{[k]}, \ldots, s_N^{[k]}]^\top$ is the vector of features for the kth dataset; and $(\cdot)^\top$ denotes the transpose of a vector/matrix. In the IVA model, dependence across corresponding components of $\mathbf{s}^{[k]}$ is taken into account through the SCV, which is obtained by vertically concatenating the nth source from each of the Kth dataset as $\mathbf{s}_n = [s_n^{[1]}, \ldots, s_n^{[K]}]^\top$. The goal in IVA is to estimate K demixing matrices to yield source estimates $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$, such that each SCV is maximally independent of all other SCVs. We note that while we consider the noiseless IVA model, in real-world applications the effect of noise is taken into account through dimension reduction, such as principal component analysis (PCA). Thus, we start with an overdetermined problem where $d > N$ and use PCA to project the data to a lower-dimensional space where $d = N$. This simple step is critical for multimodal data fusion since each modality might exhibit different levels of noise, and thus identifying the optimal signal subspace would help improve generalization abilities of the solution.

The IVA optimization parameter is defined as a set of demixing matrices $\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]}$, which can be collected into a three-dimensional array $\mathcal{W} \in \mathbb{R}^{N\times N\times K}$ and can be estimated through the minimization of the IVA objective function given by

$$J_{\text{IVA}}(\mathcal{W}) = \sum_{n=1}^{N} H(\mathbf{y}_n) - \sum_{k=1}^{K} \log\left|\det\left(\mathbf{W}^{[k]}\right)\right| + C. \qquad 2.$$

Here, $H(\mathbf{y}_n)$ denotes the (differential)[3] entropy of the estimated nth SCV that serves as the term for modeling the complex relationships among the different modalities. By definition, the term $H(\mathbf{y}_n)$ can be written as $\sum_{k=1}^{K} H(y_n^k) - I(\mathbf{y}_n)$, where $I(\mathbf{y}_n)$ denotes the mutual information within the nth SCV.

Therefore, it can be observed that minimization with respect to each demixing matrix $\mathbf{W}^{[k]}$ of Equation 2 automatically increases the mutual information within the components of an SCV, revealing how IVA exploits statistical dependence across different modalities. Hence, as shown by Equation 2, the ability to explicitly learn the mutual relationships among the multiple modalities depends on the development of flexible and efficient models for differential entropy and their

[3] We consider continuous-valued random variables and, in the following, refer to differential entropy as simply entropy for simplicity.

estimation similar to IVA-M-EMK presented by Damasceno et al. (2021, 2022). In addition, true fusion for misinformation detection (Damasceno et al. 2022) raised several interesting future research directions. For instance, ensuring fairness and reducing bias in true fusion is challenging, as different modalities may contain different levels of bias and be subject to different forms. Therefore, developing fair and unbiased true fusion models requires careful consideration of these issues. In addition, developing scalable true fusion algorithms is vital since these models could handle large volumes of data in real time.

4.3. Explainability

Explainable artificial intelligence is an emerging branch of artificial intelligence that aims to provide transparent and understandable models to improve trust, accountability, and decision-making. A potential research direction in explainability for misinformation detection is the development of more sophisticated and effective explainability techniques that can help to identify the specific features or patterns that are indicative of misinformation and to explain how the model has made its predictions. In addition, another future research direction includes exploring ways to combine different explainability methods to improve the accuracy and interpretability of the model, such as combining rule-based models with neural networks or using attention mechanisms to identify important features. Moreover, designing interactive artificial intelligence systems that can allow users to provide feedback and guidance to the model, improving its accuracy and reducing the risk of false positives and false negatives, would be an interesting and currently unexplored research direction. Last, developing quantitative ways to measure explainability is another potential research direction.

4.4. Disinformation Detection as a Class Imbalanced Problem

As previously discussed, misinformation detection is often formulated as a class imbalance problem, which can lead to challenges in developing accurate and effective machine learning models, as the models may be biased toward the majority class and may struggle to detect the minority class. Some of the open issues in this area include the following:

- Real-time detection: Real-time detection of misinformation is critical for addressing the spread of misinformation on social media platforms. Future research should explore techniques for developing real-time detection models that can effectively address the class imbalance in the data while maintaining high accuracy and low latency.
- Developing new models: The development of new models that can effectively address the class imbalance problem is a key research direction. These models should be able to handle the data imbalance and capture the problem's complexity. Examples include ensemble models, deep learning models, and transfer learning models.
- Bias and fairness: Imbalanced datasets can lead to biased models that are more likely to classify examples from the majority class. This can result in unfair and inaccurate models that disproportionately penalize the minority class.
- Overfitting and underfitting: Imbalanced datasets can make it more difficult to prevent overfitting or underfitting machine learning models. Overfitting can occur when the model becomes too specialized to the training data and fails to generalize to new data, while underfitting can occur when the model is too simple and fails to capture the complexity of the problem.
- Active learning: The use of active learning techniques to address the class imbalance problem is an area of ongoing research. Active learning allows the model to select the most informative examples to learn from, which can improve its performance on imbalanced datasets.

4.5. Fairness in Detection of Disinformation

Fairness aims to identify and reduce biases of people or groups, such as gender, race, political party, demographic, or sexual orientation. From data available to the models and algorithms utilized, machine learning for misinformation detection faces major challenges in fairness and unbias. Researchers have been working to address fairness by focusing on many factors to avoid human or societal biases.

Bias can enter the workflow of misinformation detection from the beginning with the datasets created and made available for use, even from reputable sources. A significant challenge is the need for and availability of data in which one can train algorithms to detect misinformation, and finding training data for misinformation detection is a significant challenge. Input data can lead to unfairness if there is an unequal representation or not enough positive outcomes for one of the groups. Detecting bias, which is often subtle, embedded in news articles is challenging due to the lack of suitable training data to use for bias identification in social media news sources. Lim et al. (2020) proposed a novel news dataset facilitating approaches for detecting bias in a news article. They obtained labeling through crowdsourcing, similar to the origin of many datasets available for developing and evaluating methods for fake news detection. Raza et al. (2022) developed Dbias, a Python package available for developers and practitioners to help mitigate biases in textual data like news articles.

However, crowdsourcing can also introduce bias in misinformation detection problems. Like the study mentioned above, crowdsourcing is often utilized to aid in creating these datasets through fact-checking to identify misinformation reliably. However, methods like this are burdened by the potential for cognitive biases of the crowd workers, which can reduce the quality of truthfulness judgments. This is explored by Draws et al. (2022), who perform a systematic analysis of publicly available crowdsourced data to identify potential systematic biases that may occur when crowd workers perform fact-checking tasks. Additionally, fairness can be addressed at the algorithmic level. Automatic misinformation detection poses ethical and societal risks that are not yet fully understood. Ensuring algorithms aimed at detecting misinformation are fair regarding representation, distribution of benefits and burdens, and credibility is critically important. Failing to address algorithmic bias risks compounding inequalities and historical injustices (Neumann et al. 2022).

There is a need to identify biased algorithms used for misinformation detection and for transparency to explain why a particular recommendation is made, as biased algorithms may cause unintentional discrimination that could result in the loss of opportunities or social stigmatization at a large scale (Mohseni et al. 2019). Neumann et al. (2022) developed a framework and guidance to help researchers, policymakers, and practitioners with the design of algorithmic fairness audits in the domain of misinformation detection. Other researchers state the need for an audit of machine learning-based misinformation detection algorithms and ways to mitigate any found bias (Park et al. 2022).

4.6. Human-Centered Approaches for Enhancing Intelligence and Improving Evaluation Strategies for Detection of Disinformation

Human-centered approaches are crucial for boosting intelligence and refining evaluation strategies in the detection of disinformation. These approaches emphasize human insight, adaptability, and understanding in combating the challenges posed by deceptive information.

4.6.1. Human-centered machine learning approaches. Another open issue is the role of humans in disinformation detection. While algorithms can be effective at identifying patterns and anomalies, they often lack the ability to understand a situation's broader context and nuances.

Future research could explore ways to combine human and machine intelligence to improve the accuracy and effectiveness of disinformation detection, such as by using crowdsourcing or expert panels.

Human-centered approaches to detecting misinformation involve incorporating human expertise and judgment into the detection process rather than relying solely on automated algorithms or machine learning models. Crowdsourcing involves leveraging the collective intelligence of a large group of people to detect and verify the information. This approach can be beneficial for identifying misinformation in real time, such as during breaking news events. Crowdsourcing platforms like Reddit or Twitter can solicit input and feedback from a diverse group of individuals who can provide different perspectives and expertise. In addition to crowdsourcing, expert panels involve bringing together a group of individuals with specialized knowledge and expertise in a particular domain to assess the veracity of information. For example, during an election, a panel of political scientists, journalists, and fact-checkers could be convened to evaluate political ads or social media posts for accuracy and bias. Expert panels can provide more nuanced and context-specific assessments of information that automated algorithms may not capture.

Another human-centered approach is collaborative verification, which involves leveraging the expertise of multiple individuals and organizations to verify the accuracy of information. For example, journalists can work together across different news outlets to verify information during breaking news events or election campaigns. Collaborative verification can help increase transparency and accountability in the verification process while facilitating the sharing of knowledge and expertise across different domains. Finally, education and media literacy initiatives can help individuals to develop the skills and knowledge needed to identify and evaluate information critically. This can include teaching individuals how to fact-check information, evaluate sources, and recognize different types of misinformation. When individuals are empowered with these skills, they can detect and respond to misinformation in their communities more effectively.

4.6.2. Human-centered evaluation. Human-centered evaluation is an essential component in misinformation detection. Ensuring that the models are accurate and reliable and align with human perceptions of what constitutes disinformation is vital. Some future directions for human evaluation in the detection of disinformation using machine learning include but are not limited to the following:

- Measuring agreement between humans and models: Following a similar strategy to Chang et al. (2009), one can evaluate the explainability of a model by using the fraction of subjects agreeing with the model as well as measure the topic log odds to quantify the agreement between the model and the human judgments on this task.
- Developing better evaluation protocols: There is a need for standard evaluation protocols for disinformation detection models that incorporate human judgments. These protocols should consider aspects such as the choice of annotators, the number of annotations per instance, and the definition of what constitutes disinformation.
- Studying the impact of bias and diversity: Research should investigate the impact of different types of bias on human evaluation. For example, annotators may exhibit biases related to their demographics, cultural background, or political beliefs. It is also essential to ensure that the evaluation sample is diverse to avoid overrepresenting certain groups.
- Exploring the role of explainability: Explainability is an important component of disinformation detection models, as it enables humans to understand how the models arrived at their decisions. Future research should investigate how explainability affects the human evaluation of disinformation detection models.

- Studying cross-cultural differences: Disinformation can vary across cultures and contexts, and what constitutes disinformation in one culture may not be the same in another. Future research should investigate how cross-cultural differences can affect the human evaluation of disinformation detection models.

5. DISCUSSION AND SYNERGIES AMONG DISCIPLINES

Besides the impact of addressing misinformation, it is important for researchers and practitioners in this field to develop workflows and processes that can be applied beyond the immediate subject of interest. This includes ensuring that these methods are transferable and applicable to other domains such as computational social and financial sciences, as well as healthcare. By aiming for generalizability, the knowledge and techniques developed can have broader applications and benefits.

Moreover, work in such an active field needs to include a set of synergistic activities, including prototype development and dissemination and community building between computer scientists, mathematicians, statisticians, psychologists, legal experts, and industry practitioners. In particular, statistics and data science are essential for detecting patterns and trends in large datasets. They can be used to preprocess and analyze data, identify relevant features, and build models that can classify information as either misinformation or genuine information. Data science can also be used to develop algorithms that automatically extract information from sources, such as social media, that are commonly used to spread misinformation. Computer science and engineering can provide expertise in developing efficient algorithms and software that can process large amounts of data in real time. They can also design user interfaces and visualizations that allow users to interact with the data and gain insights into the spread of misinformation.

At the same time, psychology and social sciences can provide insights into the motivations and behaviors of individuals who create and spread misinformation. They can also help identify the social and cultural factors contributing to spreading misinformation and develop strategies to combat it. Journalism and communication can provide expertise in identifying reliable sources of information and verifying the accuracy of information. They can also help develop communication strategies to promote accurate information and combat the spread of misinformation. Ethics and law can provide guidance on the ethical and legal implications of detecting and combating misinformation. They can help develop guidelines and policies for handling sensitive information and protecting the privacy of individuals. Therefore, collaborative efforts among these fields can provide synergies that enhance the accuracy and reliability of misinformation detection models and strategies.

To sum up, the problem of misinformation detection is complex and far from being completely solved. As discussed in this article, there are several challenges and open issues related to the detection of misinformation, and if researchers do not address those adequately, misinformation can have significant negative impacts on individuals and society as a whole, especially for the current generation growing up in a world where digital media is ubiquitous. Therefore, it is imperative to solve the problem of misinformation, especially for the coming generations who will shape the future of our society. Misinformation can undermine the democratic process, erode trust in institutions, and contribute to a breakdown of civil society. By addressing the problem of misinformation, we can help ensure that young people have access to accurate information, can make informed decisions, and are equipped to engage in meaningful civic discourse. This can contribute to a more equitable and informed society.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abdali S. 2022. Multi-modal misinformation detection: approaches, challenges and opportunities. arXiv:2203.13883 [cs.LG]

Abdali S, Shah N, Papalexakis EE. 2020. Hijod: Semi-supervised multi-aspect detection of misinformation using hierarchical joint decomposition. arXiv:2005.04310 [cs.SI]

Abdullah-Al-Kafi M, Tasnova IJ, Wadud Islam M, Banshal SK. 2022. Performances of different approaches for fake news classification: an analytical study. In *Advanced Network Technologies and Intelligent Computing*, ed. I Woungang, SK Dhurandher, KK Pattanaik, A Verma, P Verma, pp. 700–14. Cham, Switz.: Springer

Agarwal A, Meel P. 2021. Stacked Bi-LSTM with attention and contextual BERT embeddings for fake news analysis. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1, pp. 233–37. Piscataway, NJ: IEEE

Agrawal T, Gupta R, Narayanan S. 2017. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1045–49. Piscataway, NJ: IEEE

Alam F, Cresci S, Chakraborty T, Silvestri F, Dimitrov D, et al. 2021. A survey on multimodal disinformation detection. arXiv:2103.12541 [cs.MM]

Amador J, Oehmichen A, Molina-Solana M. 2017. Characterizing political fake news in Twitter by its metadata. arXiv:1712.05999 [cs.CL]

Anusha G, Praveen G, Mounika D, Krishna US, Cristin R. 2022. Detection of fake news using recurrent neural network. In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pp. 1–5. Piscataway, NJ: IEEE

Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.* 16:345–79

Azri A, Favre C, Harbi N, Darmont J, Noûs C. 2021. Calling to CNN-LSTM for rumor detection: a deep multi-channel model for message veracity classification in microblogs. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, ed. Y Dong, N Kourtellis, B Hammer, JA Lozano, pp. 497–513. Cham, Switz.: Springer

Baltrušaitis T, Ahuja C, Morency LP. 2018. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(2):423–43

Bhatt G, Sharma A, Sharma S, Nagpal A, Raman B, Mittal A. 2018. Combining neural, statistical and external features for fake news stance identification. In *WWW '18: Companion Proceedings of the Web Conference 2018*, pp. 1353–57. Geneva: Int. World Wide Web Conf. Steer. Comm.

Boukouvalas Z, Mallinson C, Crothers E, Japkowicz N, Piplai A, et al. 2020. Independent component analysis for trustworthy cyberspace during high impact events: an application to Covid-19. arXiv:2006.01284 [cs.LG]

Branco P, Torgo L, Ribeiro RP. 2016. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* 49(2):31

Casillo M, Colace F, Conte D, De Santo M, Lombardi M, et al. 2020. A multi-feature Bayesian approach for fake news detection. In *Computational Data and Social Networks*, ed. S Chellappan, KKR Choo, N Phan, pp. 333–44. Cham, Switz.: Springer

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. 2009. Reading tea leaves: how humans interpret topic models. In *NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems*, ed. Y Bengio, D Schuurmans, JD Lafferty, CKI Williams, A Culotta, pp. 288–96. Red Hook, NY: Curran

Chen M, Wang N, Subbalakshmi KP. 2020. Explainable rumor detection using inter and intra-feature attention networks. arXiv:2007.11057 [cs.SI]

Cui L, Wang S, Lee D. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 41–48. Piscataway, NJ: IEEE

Damasceno LP, Cavalcante CC, Adal T, Boukouvalas Z. 2021. Independent vector analysis using semi-parametric density estimation via multivariate entropy maximization. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3715–19. Piscataway, NJ: IEEE

Damasceno LP, Shafer A, Japkowicz N, Cavalcante CC, Boukouvalas Z. 2022. Efficient multivariate data fusion for misinformation detection during high impact events. In *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings*, ed. P Pascal, D Ienco, pp. 253–68. Cham, Switz.: Springer

Draws T, La Barbera D, Soprano M, Roitero K, Ceolin D, et al. 2022. The effects of crowd worker biases in fact-checking tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pp. 2114–24. New York: ACM

D'Ulizia A, Caschera MC, Ferri F, Grifoni P. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Comput. Sci.* 7:e518

Fayaz M, Khan A, Bilal M, Khan SU. 2022. Machine learning for fake news classification with optimal feature selection. *Soft Comput.* 26(16):7763–71

Glenski M, Ayton E, Mendoza J, Volkova S. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. arXiv:1909.05838 [cs.SI]

Guacho GB, Abdali S, Shah N, Papalexakis EE. 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 322–25. Piscataway, NJ: IEEE

Gupta A, Kumaraguru P. 2012. Credibility ranking of tweets during high impact events. In *PSOSM '12: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, artic. 2. New York: ACM

Gupta A, Lamba H, Kumaraguru P, Joshi A. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *WWW '13 Companion: Proceedings of the 22nd International Conference on World Wide Web*, pp. 729–36. New York: ACM

Han W, Mehta V. 2019. Fake news detection in social networks using machine learning and deep learning: performance evaluation. In *2019 IEEE International Conference on Industrial Internet (ICII)*, pp. 375–80. Piscataway, NJ: IEEE

Han Y, Karunasekera S, Leckie C. 2020. Graph neural networks with continual learning for fake news detection from social media. arXiv:2007.03316 [cs.SI]

Hangloo S, Arora B. 2022. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia Syst.* 28(6):2391–422

Hansen LK, Rieger L. 2019. Interpretability in intelligent systems—a new concept? In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. W Samek, G Montavon, A Vedaldi, LK Hansen, K-R Müller, pp. 41–49. Cham, Switz.: Springer

Hoang TBN, Mothe J. 2018. Predicting information diffusion on Twitter—analysis of predictive features. *J. Comput. Sci.* 28:257–64

Hori C, Hori T, Lee TY, Zhang Z, Harsham B, et al. 2017. Attention-based multimodal fusion for video description. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4193–202. Piscataway, NJ: IEEE

Hosseinimotlagh S, Papalexakis EE. 2018. *Unsupervised content-based identification of fake news articles with tensor decomposition ensembles*. Paper presented at Workshop on Misinformation and Misbehavior Mining on the Web (MIS2), Aug. 14, virtual

Islam MR, Liu S, Wang X, Xu G. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Soc. Netw. Anal. Min.* 10:82

Jeong U, Ding K, Cheng L, Guo R, Shu K, Liu H. 2022. Nothing stands alone: relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 596–605. Piscataway, NJ: IEEE

Jin Z, Cao J, Guo H, Zhang Y, Luo J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816. New York: ACM

Kaliyar RK, Goswami A, Narang P. 2021. FakeBERT: fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools Appl.* 80:11765–88

Kaliyar RK, Goswami A, Narang P, Sinha S. 2020. FNDNet—a deep convolutional neural network for fake news detection. *Cogn. Syst. Res.* 61:32–44

Khan JY, Khondaker MTI, Afroz S, Uddin G, Iqbal A. 2021. A benchmark study of machine learning models for online fake news detection. *Mach. Learn. Appl.* 4:100032

Khattar D, Goud JS, Gupta M, Varma V. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *WWW '19: The World Wide Web Conference*, ed. L Liu, R White, pp. 2915–21. New York: ACM

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, et al. 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). *Proc. Mach. Learn. Res.* 80:2668–77

Kirchknopf A, Slijepčević D, Zeppelzauer M. 2021. Multimodal detection of information disorder from social media. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 207–10. Piscataway, NJ: IEEE

Kumar S, Singh TD. 2022. Fake news detection on Hindi news dataset. *Glob. Trans. Proc.* 3(1):289–97

Kumari R, Ekbal A. 2021. AMFB: attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Syst. Appl.* 184:115412

Leevy J, Khoshgoftaar T, Bauder R, Seliya N. 2018. A survey on addressing high-class imbalance in big data. *J. Big Data* 5:42

Lim S, Jatowt A, Färber M, Yoshikawa M. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1478–84. Paris: Eur. Lang. Resour. Assoc.

Linardatos P, Papastefanopoulos V, Kotsiantis S. 2021. Explainable AI: a review of machine learning interpretability methods. *Entropy* 23(1):18

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ed. U von Luxburg, I Guyon, S Bengio, H Wallach, R Fergus, pp. 4768–77. Red Hook, NY: Curran

Mandical RR, Mamatha N, Shivakumar N, Monica R, Krishna AN. 2020. Identification of fake news using machine learning. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6. Piscataway, NJ: IEEE

Mohseni S, Ragan E, Hu X. 2019. Open issues in combating fake news: interpretability as an opportunity. arXiv:1904.03016 [cs.SI]

Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM. 2019. Fake news detection on social media using geometric deep learning. arXiv:1902.06673 [cs.SI]

Moroney C, Crothers E, Mittal S, Joshi A, Adal T, et al. 2021. The case for latent variable versus deep learning methods in misinformation detection: An application to COVID-19. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings*, pp. 422–32. Cham, Switz.: Springer

Murayama T. 2021. Dataset of fake news detection and fact verification: a survey. arXiv:2111.03299 [cs.LG]

Neumann T, De-Arteaga M, Fazelpour S. 2022. Justice in misinformation detection systems: an analysis of algorithms, stakeholders, and potential harms. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1504–15. New York: ACM

Nguyen VH, Sugiyama K, Nakov P, Kan MY. 2020. FANG: leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, pp. 1165–74. New York: ACM

Paka WS, Bansal R, Kaushik A, Sengupta S, Chakraborty T. 2021. Cross-SEAN: a cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. *Appl. Soft Comput.* 107:107393

Park J, Ellezhuthil R, Arunachalam R, Feldman LA, Singh VK. 2022. Toward fairness in misinformation detection algorithms. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, ed. C Budak, M Cha, D Quercia. Palo Alto, CA: AAAI

Peters J. 2020. Twitter will remove misleading COVID-19-related tweets that could incite people to engage in 'harmful activity.' *The Verge*, April 22. **https://www.theverge.com/2020/4/22/21231956/twitter-remove-covid-19-tweets-call-to-action-harm-5g**

Rath B, Salecha A, Srivastava J. 2020. Detecting fake news spreaders in social networks using inductive representation learning. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 182–89. Piscataway, NJ: IEEE

Raza S, Reji DJ, Ding C. 2022. Dbias: detecting biases and ensuring fairness in news articles. *Int. J. Data Sci. Anal.* **https://doi.org/10.1007/s41060-022-00359-4**

Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. New York: ACM

Rohera D, Shethna H, Patel K, Thakker U, Tanwar S, et al. 2022. A taxonomy of fake news classification techniques: survey and implementation aspects. *IEEE Access* 10:30367–94

Sahan M, Smidl V, Marik R. 2021. Active learning for text classification and fake news detection. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pp. 87–94. Piscataway, NJ: IEEE

Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. 2019. Combating fake news: a survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.* 10(3):21

Shu K, Bernard HR, Liu H. 2018. Studying fake news via network analysis: detection and mitigation. arXiv:1804.10233 [cs.SI]

Shu K, Sliva A, Wang S, Tang J, Liu H. 2017. Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explor. Newsl.* 19(1):22–36

Shu K, Wang S, Liu H. 2019. Beyond news contents: the role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 312–20. New York: ACM

Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S. 2019. SpotFake: a multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 39–47. Piscataway, NJ: IEEE

Song C, Ning N, Zhang Y, Wu B. 2020. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Inform. Proc. Manag.* 58:102437

Swartout WR, Moore JD. 1993. Explanation in second generation expert systems. In *Second Generation Expert Systems*, ed. J-M David, J-P Krivine, R Simmons, pp. 543–85. Berlin: Springer

Tschiatschek S, Singla A, Gomez Rodriguez M, Merchant A, Krause A. 2018. Fake news detection in social networks via crowd signals. In *WWW '18: Companion Proceedings of the Web Conference 2018*, pp. 517–24. Geneva: Int. World Wide Web Conf. Steer. Comm.

Tuan NMD, Minh PQN. 2021. Multimodal fusion with BERT and attention mechanism for fake news detection. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 43–48. Piscataway, NJ: IEEE

Vosoughi S, Roy D, Aral S. 2018. The spread of true and false news online. *Science* 359(6380):1146–51

Wang J, Mao H, Li H. 2022. FMFN: fine-grained multimodal fusion networks for fake news detection. *Appl. Sci.* 12(3):1093

Wang WY. 2017. "Liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv:1705.00648 [cs.CL]

Wang Y, Ma F, Jin Z, Yuan Y, Xun G, et al. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 849–57. New York: ACM

Wu L, Li J, Hu X, Liu H. 2017. Gleaning wisdom from the past: early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, pp. 99–107. Philadelphia, PA: SIAM

Wu L, Morstatter F, Carley KM, Liu H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explor. Newsl.* 21(2):80–90

Yang S, Shu K, Wang S, Gu R, Wu F, Liu H. 2019. Unsupervised fake news detection on social media: a generative approach. *Proc. AAAI Conf. Artif. Intel.* 33:5644–51

Zhang Q, Lipani A, Liang S, Yilmaz E. 2019. Reply-aided detection of misinformation via Bayesian deep learning. In *WWW '19: The World Wide Web Conference*, pp. 2333–43. New York: ACM

Zhang T, Wang D, Chen H, Zeng Z, Guo W, et al. 2020. BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. Piscataway, NJ: IEEE

Zhao J, Xie X, Xu X, Sun S. 2017. Multi-view learning overview: recent progress and new challenges. *Inform. Fusion* 38:43–54

Zhao Z, Resnick P, Mei Q. 2015. Enquiring minds: early detection of rumors in social media from enquiry posts. In *WWW '15: Proceedings of the 24th International Conference on World Wide Web*, pp. 1395–405. Geneva: Int. World Wide Web Conf. Steer. Comm.