Combining Crowd and Machine Intelligence to Detect False News on Social Media[1]

Forthcoming at *MIS Quarterly*

Xuan Wei
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China {weix@sjtu.edu.cn}

Zhu Zhang
State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, CHINA {zhu.zhang@ia.ac.cn}

Mingyue Zhang
School of Business and Management, Shanghai International Studies University, Shanghai, China {zhangmy@shisu.edu.cn}

Weiyun Chen
School of Management, Huazhong University of Science and Technology, Wuhan, China {chenweiyun@hust.edu.cn}

Daniel Dajun Zeng
State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, and School of Economics and Management and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, CHINA {dajun.zeng@ia.ac.cn}

[1] Zhu Zhang and Mingyue Zhang are the corresponding authors.

Xuan Wei is an assistant professor in the Department of Information, Technology and Innovation, Antai College of Economics and Management, Shanghai Jiao Tong University. He received his B.S. degree from the Shanghai Jiao Tong University in 2014 and his Ph.D. degree in management information systems at the University of Arizona in 2020. His research interests include crowd intelligence and crowdsourcing, social media analytics, statistical machine learning, probabilistic modeling and interference, and deep learning. His work has been published in major information systems journals such as *Nature Human Behaviour, MIS Quarterly, INFORMS Journal on Computing*, etc.

Zhu Zhang is an associate professor at the Institute of Automation, the Chinese Academy of Sciences. He received his B.S. degree in automation from Zhejiang University in 2008 and his Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences in 2015. His current research interests include hybrid intelligence, data mining, and business analytics. He has published more than 20 articles in major conferences and journals such as *MIS Quarterly, INFORMS Journal on Computing, Journal of Management Information Systems, Journal of Medical Internet Research, ACM Transactions on Management Information Systems*, etc.

Mingyue Zhang is an associate professor in the School of Business and Management, Shanghai International Studies University. She received her Ph.D. degree in management science and engineering from the School of Economics and Management, Tsinghua University, in 2017. Her current research interests include online communities, incentive mechanisms, and social media analysis. Her work has been published in journals such as *MIS Quarterly, Decision Sciences, Decision Support Systems, Information Sciences, ACM Transactions on Knowledge Discovery from Data*, etc.

Weiyun Chen is an associate professor in the Department of Management Science and Information Management, School of Management at Huazhong University of Science and Technology, Wuhan, China. His current research focuses on the wisdom of crowds and social computing. He received his Ph.D. in control theory and control engineering from the Chinese Academy of Sciences, and his bachelor's and master's degrees from the Department of Automation at Huazhong University of Science and Technology, Wuhan, China. His research has appeared in *MIS Quarterly, IEEE Intelligent Systems*, and *British Journal of Educational Technology*, etc.

Daniel Dajun Zeng received the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University. He is a professor at the Institute of Automation, Chinese Academy of Sciences. He was a Gentile Family Professor of MIS in the Department of Management Information Systems, the University of Arizona. His research interests include social computing, recommender systems, intelligence and security informatics, infectious disease informatics, applied operations research, and game theory. He has published one monograph as well as more than 300 peer-reviewed articles, and has co-edited 22 books and proceedings. His work has been published in major information systems journals such as *Nature Human Behaviour, MIS Quarterly, Management Science, INFORMS Journal on Computing, IEEE Transactions on Knowledge and Data Engineering*, etc. He currently serves as the editor-in-chief of ACM Transactions on Management Information Systems. He is a fellow of the IEEE and the AAAS.

Combining Crowd and Machine Intelligence to Detect False News on Social Media

Abstract

The explosive spread of false news on social media has severely affected many areas such as news ecosystems, politics, economics, and public trust, especially amid the COVID-19 infodemic. Machine intelligence has met with limited success in detecting and curbing false news. Human knowledge and intelligence hold great potential to complement machine-based methods. Yet they are largely underexplored in current false news detection research, especially in terms of how to efficiently utilize such information. We observe that the crowd contributes to the challenging task of assessing the veracity of news by posting responses or reporting. We propose combining these two types of scalable crowd judgments with machine intelligence to tackle the false news crisis. Specifically, we design a novel framework called CAND, which first extracts relevant human and machine judgments from data sources including news features and scalable crowd intelligence. The extracted information is then aggregated by an unsupervised Bayesian aggregation model. Evaluation based on Weibo and Twitter datasets demonstrates the effectiveness of crowd intelligence and the superior performance of the proposed framework in comparison with the benchmark methods. The results also generate many valuable insights, such as the complementary value of human and machine intelligence, the possibility of using human intelligence for early detection, and the robustness of our approach to intentional manipulation. This research significantly contributes to relevant literature on false news detection and crowd intelligence. In practice, our proposed framework serves as a feasible and effective approach for false news detection.

Keywords: false news, fake news, wisdom of crowds, hybrid intelligence, graphical model

1. INTRODUCTION

False news, referring to any false news article or message that is published and propagated through media and has an assertion in it, is now viewed as one of the largest concerns globally, especially amid the COVID-19 infodemic (Shu et al., 2017; The Lancet Infectious Diseases, 2020; Vosoughi et al., 2018). Almost one third of U.S. survey respondents reported that they had been exposed to fake news[2] in the past week, and the rate is as high as 49% in some countries such as Turkey (McCarthy, 2018). The proliferation of social media platforms further facilitates the prevalence of false news from its generation and spread to consumption (Tran et al., 2020a). Such prevalence has seriously affected individuals and society (Oh et al., 2018; Tran et al., 2020b). Individuals are reporting declining trust in the news media. A survey from 2018 revealed that only 44% of people trust the news; social media news is perceived as even more unreliable, with only 23% trusting such sources (Newman, 2018). This phenomenon is envisioned as a growing problem for business, as social media now serves as a critical channel for e-commerce such as advertising. Our economies are not immune to false news either because investors frequently rely on news to make investment decisions. For example, a false tweet saying that Barack Obama was injured in an explosion wiped out $130 billion in stock value in the blink of an eye (Vosoughi et al., 2018). During the current COVID-19 pandemic, false news has become even more concerning because it undermines trust in health institutions and programs (The Lancet Infectious Diseases, 2020).

Firms, ranging from behemoths like Facebook and Google to small startups, have started taking actions to curb the false news epidemic. For example, Facebook, which has suffered nearly incessant false news problems in recent years, allows users to flag potential false news

[2] Fake news is a type of false news that is intentionally false. See Section A of the online supplement for a clarification of false news and related concepts. The online supplement is available at https://osf.io/6svp9/.

and then uses third-party organizations to fact-check the flagged news (Kim et al., 2018). Machine learning algorithms have also been utilized to help detect false news and accounts. In China, Sina Weibo, one of the most popular microblogging services also allows users to report false posts in order to maintain a healthy and sustainable community.

The academic community has also devoted much effort to tackling the false news crisis. Existing research mainly considers two types of data sources: news content and social context (Shu et al., 2017). Relevant features are extracted from these data sources and then fed into machine learning algorithms. Overall, the current effort to detect false news heavily relies on machine intelligence (e.g., machine learning approaches) to achieve automatic detection; human intelligence, which holds great potential to complement machine-based methods, remains underexplored. Since machines and human brains have different problem-solving capabilities, they can symbiotically benefit from each other (Kamar, 2016). Machine algorithms can easily learn certain false news-related patterns from large-scale data, yet largely depend on the training data. Humans are generally more competent than machines at intelligent tasks such as natural language understanding (Demartini et al., 2017), as they understand context and nuance better than machines (Vaughan, 2018). Humans can also access continually updated knowledge, which is critical to detecting ever-changing false news. However, since human intelligence is a valuable resource with high costs and constraints (Kamar, 2016), most existing approaches that exploit human intelligence for false news detection (e.g., expert-oriented and crowdsourcing-oriented fact-checking) are not applicable in large-scale scenarios. Given the potential value of underexplored human intelligence, we propose the following questions: *What types of scalable human intelligence can be exploited? How useful are they? How can such intelligence be efficiently utilized? And how can human and machine intelligence be combined symbiotically?*

Premised on the wisdom of crowds (Surowiecki, 2005) and the online disinhibition effect from the cyberpsychology literature (Suler, 2004), we propose incorporating two types of scalable crowd intelligence. The online disinhibition effect suggests that online users express themselves more openly in the less restrained web environment (Suler, 2004). On social media, when users are exposed to a piece of news with which they personally disagree, some individuals may honestly share such opinions through actions such as commenting or reporting. Aggregating these crowd opinions can be beneficial to detecting false news, as confirmed in prior literature that has successfully taken advantage of the wisdom of crowds in various online scenarios such as microtask crowdsourcing (Wang et al., 2017), prediction markets (Chen et al., 2017), and collective social reporting (Oh et al., 2013).

We demonstrate our idea with a real example from Sina Weibo. A Weibo user posted a piece of news about Typhoon Mangkhut, as shown in the left panel of Figure 1. Other users can click the "report" button in the upper right of the post to flag it as false information, a personal attack, or many other types of problematic information. Weibo records how many times this post is reported as false information. The example in Figure 1 was reported by seven crowd users in four days (data as of December 7, 2020) and then verified as false news by the Sina community management center.[3] The right panel lists some of its responses, where many users debunked the false news via responses (marked with underlines). This example indicates that crowd users may contribute their intelligence (i.e., judgment as to the veracity of news) by posting responses or reporting posts on social media. Such human intelligence could potentially be used to help us recognize false news. In addition, such information is scalable because reporting and commenting often occur voluntarily alongside posting. However, the use of human intelligence

[3] https://service.account.weibo.com/

for such purposes is not completely reliable because not every reader will report or negate the post in a comment when the post is false; further, readers may also mistakenly report or debunk a piece of true news. In this research, we combine two types of less reliable yet scalable human intelligence with machine intelligence to help detect false news. Specifically, we propose a novel framework called *Crowd-powered fAlse News Detection* (CAND), which first extracts machine, human, and hybrid judgments from news features, reports, and received responses, and then aggregates the extracted judgments with an unsupervised Bayesian result aggregation model to obtain the final prediction. By evaluating the proposed framework with two real-world datasets from Weibo and Twitter, this research demonstrates the effectiveness of crowd intelligence in fighting false news and the superior performance of our framework to utilize human intelligence in comparison with the benchmark methods. Our analysis also generates many valuable insights, such as the complementary value of human and machine intelligence, the possibility of using human intelligence for early detection, and the robustness of our approach to intentional manipulation.

The rest of this paper is organized as follows: The following section reviews the relevant literature. Then, we detail the proposed CAND framework and evaluate the proposed framework by describing the experimental design and showing the empirical results. We conclude the paper by presenting contributions, implications, limitations, and future research directions.

Figure 1. An Example of Human Intelligence in Responses and Reports

Note: We have blurred all private information. "[bless]" means an emoji and "[loc: Shenzhen]" refers to a hyperlink about the location. In the right panel, the debunking responses are underlined.

2. RELATED WORK

In this section, we survey several streams of relevant literature. First, we review related literature that serves as the theoretical foundations of this work. Second, we summarize existing studies about false news detection on social media. Third, we review false news studies that are related to crowd intelligence. Since the proposed framework needs to aggregate the extracted judgments, we conclude by reviewing the literature on information aggregation.

2.1 Theoretical Foundations

Our proposed approach builds on the online disinhibition effect from the cyberpsychology literature (Suler, 2004) and the wisdom of crowds (Surowiecki, 2005). The online disinhibition

effect demonstrates that some people feel less restrained and express themselves more openly online than in-person (Suler, 2004). This effect is triggered by the characteristics of the online environment such as anonymity and asynchronous communication and can manifest in both positive and negative directions (Lapidot-Lefler & Barak, 2012; Suler, 2004). In our scenario, we propose utilizing crowd opinions to help detect false news. The online disinhibition effect ensures that at least some users will post their judgments honestly (e.g., by commenting and reporting) when reading a piece of news that is counter to their beliefs; when communicating in-person, individuals may refuse to assert their own opinions in opposition to others.

The wisdom of crowds (a.k.a., collective intelligence, crowd wisdom) refers to the theory that large groups of individuals are often collectively smarter than any single member and even expert individuals for many tasks such as problem solving, decision-making, and predicting[4] (Surowiecki, 2005). Crowdsourcing is the most prominent and successful practice for unleashing the wisdom of crowds.[5] Crowdsourcing is an online and distributed problem-solving and production model that leverages the collective intelligence of online communities to finish specific tasks (Brabham, 2013). This model (and therefore the wisdom of crowds) has been successfully applied in a wide array of applications such as crowdsourcing marketplaces (e.g., Amazon Mechanical Turk), crowdfunding (e.g., Kickstarter), and user-generated content (e.g., Yelp). Information systems (IS) researchers have also devoted much effort to studying the wisdom of crowds and its applications (Atanasov et al., 2017; Bayus, 2013; Lee et al., 2018; Lukyanenko et al., 2014; Wang et al., 2017). In this research, we tap into the wisdom of crowds

[4] For a single task, crowds are not necessarily smarter than individuals, as the task may be completed by a group of unskilled or biased individuals. However, when evaluated on all tasks, the crowds are often averagely smarter than individuals.

[5] Crowd wisdom and crowdsourcing are sometimes treated the same (Doan et al. 2011). We distinguish them in that crowd wisdom is a conceptual-level idea, while crowdsourcing is a problem-solving practice that leverages crowd wisdom.

and propose utilizing scalable crowd judgments (i.e., responses and reports on social media) for the task of false news detection. Note that a single user's intelligence may be unreliable. For example, not all users will debunk the news in their responses; even if they do so, their debunking may be unreliable for various reasons including political affiliation, ambiguous news, intentional or unintentional mistakes, etc. In our proposed approach, this issue is mitigated by modeling the credibility of humans and aggregating the judgments from many users. The premise of our approach is that based on the wisdom of crowds, false news and true news have different debunking and reporting patterns after aggregating individuals' opinions.

2.2 Computational False News Detection on Social Media

In the existing literature about false news detection on social media, two major types of information, i.e., news content (e.g., headline and news text) and social context of news (e.g., users' information and social engagement with news), are leveraged to detect false news (Zhang & Ghorbani, 2020; Zhou & Zafarani, 2020). Based on the data source, approaches for false news detection can be divided into two categories: news content-based and social context-based (Shu et al., 2017).

In news content-based approaches, useful linguistic or visual features are extracted from news content such as the news source, headline, body text, images, and videos. Linguistic features, including lexicon-level features (e.g., word number and frequency of special words), syntax-level features (e.g., parts-of-speech tagging), and other language features, serve as signals to identify certain writing styles in false news. Visual features (e.g., clarity score and coherence score) are also used for false news detection (Gupta et al., 2013; Jin et al., 2017) since visual cues have been shown to be an important manipulator in fake news propaganda (Castillo et al., 2014). Based on the extracted features, several categories of models are designed to detect false

news: knowledge-based, style-based, and deep learning-based. Knowledge-based approaches use external sources (e.g., knowledge graph) to fact-check proposed claims in news content. Style-based approaches capture false news-related manipulators from the perspective of writing style, such as deception (Wang, 2017) and nonobjectivity (Potthast et al., 2018). In addition, deep learning has recently been used to detect multisource or multimodal fake news (Karimi et al., 2018; Singh et al., 2021; Wang et al., 2018), as deep learning enables automatically extracting latent features from data in place of laborious and time-consuming manual feature extraction.

In social context-based approaches, four major types of features (i.e., user-based, post-based, network-based, and flagging-based) are extracted from social context. First, user-based features extracted from user profiles can measure user characteristics and credibility (Shu et al., 2018), since user profiles include some clues to infer social bots or cyborgs that create or spread false news (Shu et al., 2019). For instance, an unsupervised Bayesian approach is proposed to simultaneously model the truth of news and user credibility (Yang et al., 2019).

Second, post or response-based features are useful information extracted from relevant social media postings, such as posts about a news event or responses to these posts. This information is useful because people often express their emotions or opinions about news in their social media postings (Shu et al., 2017). To efficiently extract social stances from unstructured text data, deep learning-based approaches are widely used. To identify rumors, recurrent neural network-based methods are used to learn the hidden representations that capture the social responses in a series of relevant posts over time (Ma et al., 2016). More advanced deep learning techniques, e.g., the attention mechanism, have also been used in recent literature (Liu & Wu, 2020). For example, Guo et al. (2018) showed that a hierarchical neural network combined with social information and attention mechanism has outstanding performance in rumor detection. When user responses

are rare, e.g., at the early stages of news propagation, the neural generative model can be used to generate user responses to news articles in order to enable the early detection of false news (Qian et al., 2018). False news detection can also be jointly trained with stance/opinion classification using a multitask learning framework in order to take advantage of shared task-invariant features (Kochkina et al., 2018; Ma et al., 2018).

Third, network-based features are often extracted from the interactions between users and news by building specific networks in terms of users' stances (Jin et al., 2016; Tacchini et al., 2017), co-occurrence (Ruchansky et al., 2017), and so forth. For instance, in a hybrid model for fake news detection, Ruchansky et al. (2017) constructed a weighted co-occurrence network where an edge denotes the number of articles with which two users have both engaged. After extraction, these three types of features are further fed into false news detection models. Existing models include propagation-based and stance-based. Propagation-based models exploit some propagation methods (e.g., PageRank-like) to predict the veracity of a news event via the credibility of its relevant social media posts (Shu et al., 2020). In stance-based models, users' stances on relevant social media posts are utilized to infer the veracity of news articles (Tacchini et al., 2017).

Fourth, in the literature, news reporting/flagging is mainly used for early identification and containment of false news (Sharma et al., 2019). For example, the marked temporal point process is used to model the trade-off between the number of received flags and the exposure of fake news (Kim et al., 2018). Based on the framework, an online algorithm was developed to decide which stories to send for fact-checking and when to do so in order to efficiently reduce the spread of fake news. In the same direction, a Bayesian approach was developed to jointly detect fake news and learn about users' flagging accuracy over time (Tschiatschek et al., 2018).

Currently, human responses and reports are still largely underexplored, especially in terms of how to efficiently utilize such information. Compared with the existing literature using responses and reports, our work is significantly different in terms of how such data sources are used. Human responses are often treated as features of black-box methods in the literature (Guo et al., 2018; Ma et al., 2016); in contrast, we incorporate more interpretable structures into our Bayesian approach. For example, in our model, intuitively false news receives more debunking responses, compared with true news. In black-box models, responses are fed into the model with the hope that the algorithms will learn such patterns from the data. However, in our design, we explicitly model the part that contains an interpretable structure and leave the other parts to black-box models. Specifically, we assume that false and true news items have different debunking patterns, which are captured by the logistic-normal assumption. For unstructured data (e.g., news body and news responses), we use deep learning methods to handle it. In the literature, human reports are often used separately to explore the early detection of false news (Kim et al., 2018; Tschiatschek et al., 2018). Our work uses an interpretable structure by assuming different reporting patterns for false and true news items. The learned pattern is then combined with feature-based machine judgments and human intelligence in responses to generate final predictions. In our conclusion, we described the several advantages that our approach has over the existing literature.

2.3 False News Studies Using Crowd Wisdom

Currently, there are three major ways to utilize crowd wisdom in the fight against false news. First, crowdsourcing-oriented fact-checking uses crowdsourcing platforms to recruit ordinary people to fact-check potential false news (Shu et al., 2017). For example, Fiskkit, an online commenting platform, allows users to discuss and annotate the veracity of a news article by

rating or tagging it. Although such a fact-checking method is scalable, it cannot identify false news in real time. Second, crowd wisdom is used in computational studies regarding false news detection. The two most frequently explored data sources are crowd responses and reporting/flagging. The relevant literature was introduced in the last subsection regarding post-based and flagging-based features. Finally, crowdsourcing marketplaces, primarily Amazon Mechanical Turk (AMT), are widely used in explanatory social science studies regarding false news (Pennycook & Rand, 2018, 2019a). For example, to verify whether laypeople's judgments are reliable indicators of news source quality, thousands of people were recruited from AMT and Lucid to participate in two preregistered experiments where individuals rated familiarity with and trust in 60 news sources (Pennycook & Rand, 2019b).

Our work belongs to the second category: we aim to develop a practical computational method to detect fake news by combining two types of crowd wisdom (i.e., user responses and reporting/flagging information) with feature-based machine learning methods.

2.4 Information Aggregation

In our proposed approach, we attempt to aggregate judgments from humans and machines in order to predict the veracity of news. Information aggregation techniques have been widely studied in many scenarios under different names. One well-documented scenario is classifier combination, where each classifier outputs a result for the same machine learning task and then the results are aggregated (Tulyakov et al., 2008). In this stream, many ensemble methods (e.g., bagging and boosting) and their variants consider both how to construct the classifiers and how to aggregate the results. We only review the literature on result aggregation where a fixed set of classifiers are given, since the goal of our setting is to combine human and machine judgments. In addition, we focus on relevant literature where the goal of the task classification and the

classifier outputs are prediction scores or labels. To combine prediction scores, simple rules (e.g., sum rule, product rule, max rule, and min rule) are frequently used and have shown good performance in many tasks (Mohandes et al., 2018). To aggregate prediction classes, voting-based strategy and its variants are widely used. Majority Voting (MV) selects the candidate class that has a majority (e.g., more than half of the votes in binary classification). It is extended by assigning different weights to each classifier (Mohandes et al., 2018). Some other studies take a probabilistic perspective to solve the problem by introducing parameters to govern the behaviors of classifiers and ground truth. For example, a Bayesian model called independent Bayesian Classifier Combination (iBCC) is proposed to model the generative process of classifiers' predictions by taking into account each classifier's reliability (Kim & Ghahramani, 2012).

The second major scenario is crowdsourced answer aggregation in microtask crowdsourcing, where a huge amount of microtasks are assigned to potentially unreliable crowd workers through some web-based crowdsourcing platforms (e.g., AMT) and then the results are aggregated (Brabham, 2008; Wang et al., 2017). The most relevant scenario is data labeling, where workers are requested to assign one (single-label) or multiple (multi-label) labels to data instances in each microtask (Moreno et al., 2015; Wei et al., 2017). Actually, when the results to be aggregated are labels, classifier combination and crowdsourced answer aggregation boil down to the same technical problem; however, some application-specific features, e.g., worker grouping in the crowdsourcing setting (Moreno et al., 2015), can be exploited under different scenarios. In the seminal work, the DS (Dawid-Skene) model uses a confusion matrix to model workers' behavior; and an EM algorithm is developed to estimate the parameters (Dawid & Skene, 1979). The above-mentioned iBBC model is a Bayesian extension of the DS model. These two models are further extended by considering various related factors such as task easiness (Kim &

Ghahramani, 2012), instance attribute (Welinder et al., 2010), label correlation (Wei et al., 2017), and worker grouping (Moreno et al., 2015).

Our scenario is different from the aforementioned information aggregation literature because only humans or machines are involved in the literature, whereas we need to combine the extracted judgments from humans and machines in our scenario. Hence, we need to design a specific model for our setting to combine such mixed judgments.

3. A CROWD-POWERED FRAMEWORK FOR FALSE NEWS DETECTION

In this section, we present our proposed framework for false news detection—*crowd-powered false news detection* (CAND). Figure 2 shows the overall design, which consists of two stages: information extraction and result aggregation. Without loss of generality, we represent each piece of news by its content and context features (Shu et al., 2017). For content features, we consider the news text and whether each piece of news has external links, images, or videos.[6] Context features include posting time and author profiles (e.g., whether an author is a verified author and whether an author has an avatar). Author profiles are incorporated because prior work has found that features about the news source (i.e., news author in our context) play a critical role in identifying false news (Oh et al., 2013; Sharma et al., 2019). For example, news posts from organizational Weibo accounts are less likely to be considered false compared with those from personal accounts. See Section B1 of the online supplement for all news features and their descriptions. Given a piece of news, humans may contribute their intelligence regarding the veracity of news in responses and reports. Based on content and context features, human responses, and human reports, the information extraction stage extracts machine, human, and

[6] We do not consider images and video content because: (1) considering them necessitates designing multimodal models for the classifiers in Task T1 and hence goes beyond the scope of this work, and (2) it does not influence our evaluation, as it increases the performance of our approach and the benchmark methods at the same time.

hybrid human-machine judgments. To symbiotically aggregate the extracted judgments, we build an effective and interpretable Bayesian result aggregation model called the *credibility-based logistic-normal aggregation model* (CLNAM). In the following, we elaborate each stage in detail.

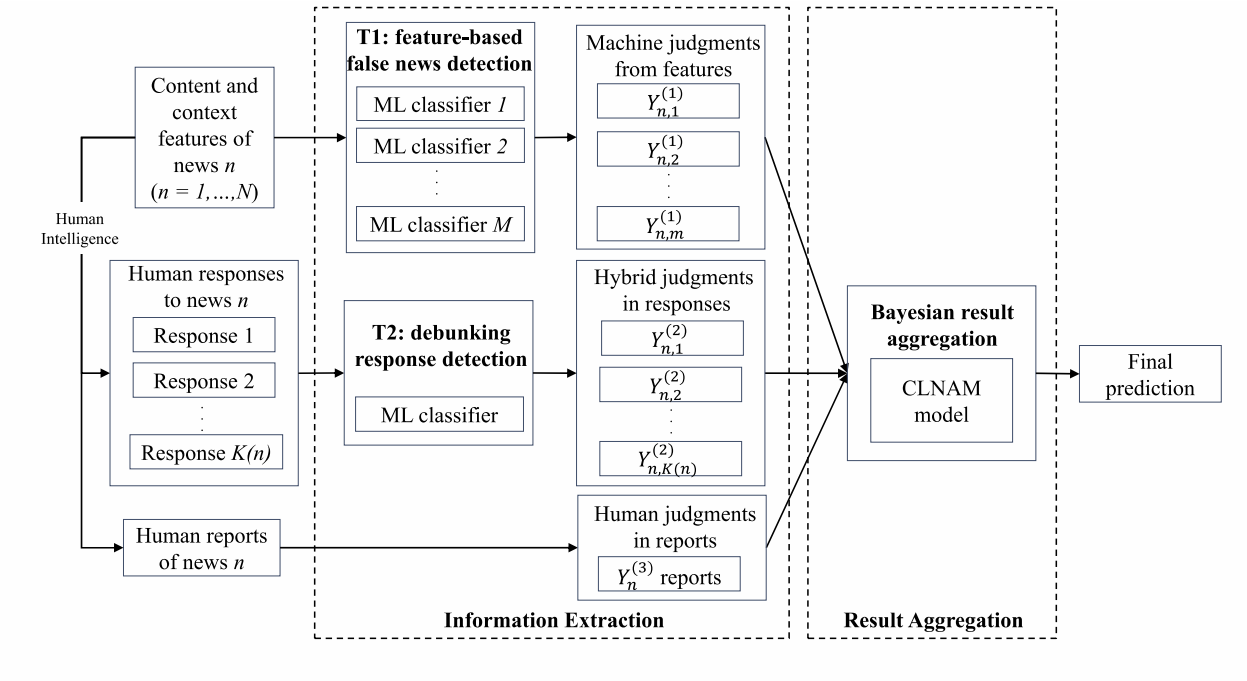


Figure 2. The Crowd-Powered Framework for False News Detection (CAND)

3.1 Information Extraction

Suppose there are N pieces of news whose veracity we want to assess (hereafter, "test set"), where news n receives $K(n)$ response(s) and a certain number of reports. We take news n as an example to illustrate the procedure. For news features, one or more (denoted as M) machine learning (ML) algorithms are trained on a dataset that is independent of the test set. The trained M classifiers are then used to predict the veracity of news n in the test set. Let $Y_{n,m}^{(1)} \in (0,1)$ denote the predicted probability generated by classifier m. The results are called machine judgments from features. We formulate this task as T1: feature-based false news detection.

Similarly, to extract useful knowledge from responses, we train another classifier on a response dataset that is independent of the test set. The classifier then predicts whether each response to news n is a debunking response or not. Let $Y_{n,k}^{(2)} \in (0,1)$ be the predicted probability for response k. The output is a hybrid of human and machine intelligence. We formulate this process as Task T2: debunking response detection. For the human reports, the number of received reports of news n is readily available without further extraction. We denote it as $Y_n^{(3)} \in \{0,1,\dots\}$.

3.2 Unsupervised Bayesian Result Aggregation Model CLNAM

After information extraction, three types of extracted information are available for each piece of news in the test set: machine judgments from news features, hybrid judgments in responses, and human judgments in reports. The goal of CLNAM is to aggregate them and obtain a final prediction as to the veracity of each piece of news.

3.2.1 Challenges in Result Aggregation

Although information aggregation has been intensively studied in the fields of classifier combination (Mohandes et al., 2018) and crowdsourced answer aggregation (Wei et al., 2020; Zhang et al., 2016), the proposed aggregation scenario possesses unique characteristics because of the hybrid human-machine nature, and hence poses new challenges. The first challenge arises from the potentially low credibility of humans and machines. Specifically, when extracting machine judgments from news features, ML algorithms may make mistakes. Also, in their response to a piece of news, human users may express doubt about the veracity of true news. Further, when a piece of news is false, only a small portion of users will question the veracity of the false news in their response; most users post responses that are unrelated to the veracity of the news. In addition, even if a user debunks the news, the ML algorithm may fail to recognize it. A similar problem exists for user reporting/flagging of news stories: false news may receive no

user reports and true news may be reported as false by users. All these potential cases make the aggregation problem more sophisticated compared with the relevant previous literature, where only humans or machines are involved. With humans and machines working together, it is harder to pinpoint the source of the unreliability, which necessitates a more carefully crafted scheme for result aggregation. The second challenge relates to the need to design a specific model to combine multiple data sources with mixed data types (i.e., continuous and discrete values). Machine judgments from news features and hybrid judgments in user responses are given continuous values between 0 and 1; and the number of received reports are represented by discrete values. Compared to the existing information aggregation literature, which does not use mixed data types and involves humans or machines only, our scenario is obviously different. Therefore, we needed to design a specific model for our scenario to combine mixed data types.

3.2.2 **Technical Insights**

We introduce the major technical insights we use to solve the above challenges. First, we take a probabilistic perspective to modeling the credibility of humans and machines. As the results of users' debunking behaviors are binary, we use the Bernoulli distribution to depict them. Specifically, we assume that users will write a debunking response to false or true news with the probability η or $\bar{\eta}$, respectively. Intuitively, a higher η and lower $\bar{\eta}$ indicate more reliable human intelligence in responses. To model the credibility of human intelligence in reports, we assume that the number of received reports follows a Poisson distribution, which is popular for modeling the frequency of an event within a specific time interval. Specifically, for false or true news, the number of received reports follows a Poisson distribution parameterized by ϕ or $\bar{\phi}$ respectively. Intuitively, higher ϕ and lower $\bar{\phi}$ indicate more reliable human intelligence in reports. In the proposed scenario, modeling the behavior of ML algorithms (in both Tasks T1 and T2) is

equivalent to answering what the distribution of the prediction is, given the ground truth of 0 or 1. The most widely used solution is the two-coin model, which assumes that the classifier predicts correctly with a certain probability when the ground truth is 0 or 1 (Bragg & Weld, 2013; Raykar et al., 2010). However, this incurs information loss, as it treats the predicted probability of 0.51 and 1 as the same. To maximize the value of the predictions generated by classifiers, we assume that when the ground truth is 0 (or 1), the predicted probability follows a logistic-normal distribution (Atchison & Shen, 1980), which means the logit of the predicted probability follows a normal distribution. The intuition is that many ML algorithms, including deep learning models and logistic regression, often end with a logistic function. We defend this choice by presenting some empirical evidence in Figure 3, which plots a histogram of the logit of the predicted probability (blue bar) and the fitted normal distribution (orange line) under various tasks and methods. Figure 3 demonstrates that logistic-normal distribution approximates the histogram quite well, especially for SVM and BERT. Note that we do not theoretically claim that the prediction follows a logistic-normal distribution. However, compared with modeling the prediction as a Bernoulli distribution, the logistic-normal assumption retains more information. This is also empirically verified below in the Baseline Methods and Evaluation Metrics section by the superior performance of CAND (logistic-normal assumption) in comparison with BAM (Bernoulli assumption).

To combine multiple data sources with mixed data types, we propose a generative Bayesian model for the underlying mixed data generation process. Specifically, all three types of judgments from the extraction stage are generated based on an unknown ground truth (i.e., whether the news is false or not). With the generative process, we infer the posterior distribution of the ground truth based on the generated data. The details are elaborated in the following

section.

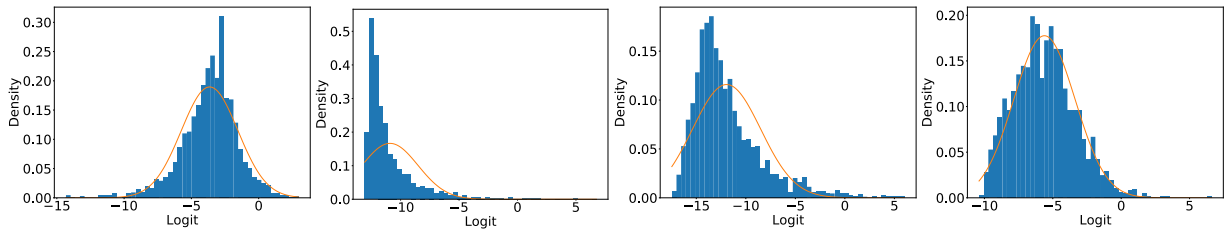


Figure 3. Empirical Evidence for the Logistic-Normal Assumption

Note: From left to right: SVM, CNN, Bi-LSTM, and BERT in task T1. See details of these methods in Baseline Methods and Evaluation Metrics section.

3.2.3 **CLNAM Model**

Building on the proposed techniques for modeling the credibility of humans and machines, we define the generative process of the extracted mixed information. Take news n as an example, to generate machine judgments, we first generate the ground truth z_n for news n. We use $\gamma \in [0,1]$ to denote the probability of a piece of news being false. Conditioning on γ, the ground truth $z_n \in \{0,1\}$ follows Bernoulli(γ). Depending on the value z_n, the classifiers in Task T1 will behave differently (i.e., generate different probability values). When the news is false, the prediction of classifier m will follow a logistic-normal distribution:

$$Y_{n,m}^{(1)} \sim LogisticNormal\left(\mu_m^{(1)}, \left(\sigma_m^{(1)}\right)^2\right), \quad (1)$$

where $\mu_m^{(1)}$ and $\sigma_m^{(1)}$ are the mean and standard deviation of the predicted probability's logit respectively. When the news is true, the prediction will follow another logistic-normal distribution parameterized by $\bar{\mu}_m^{(1)}$ and $\bar{\sigma}_m^{(1)}$. Finally, following previous Bayesian modeling literature (Blei et al. 2003; He et al. 2019), we place conjugate beta and Normal-Inverse-Gamma (NIG) priors over the Bernoulli and logistic-normal distributions respectively:

$$\gamma \sim Beta(e_0, f_0), \quad (2)$$

$$\mu_m^{(1)}, \left(\sigma_m^{(1)}\right)^2 \sim NIG\left(\omega_0^{(1)}, \nu_0^{(1)}, \alpha_0^{(1)}, \beta_0^{(1)}\right), \tag{3}$$

$$\bar{\mu}_m^{(1)}, \left(\bar{\sigma}_m^{(1)}\right)^2 \sim NIG\left(\bar{\omega}_0^{(1)}, \bar{\nu}_0^{(1)}, \bar{\alpha}_0^{(1)}, \bar{\beta}_0^{(1)}\right), \tag{4}$$

Compared with the generative process of machine judgments, generating hybrid judgments from responses differs in that: (1) The classifier here attempts to predict whether each response is a debunking response, rather than attempting the predict the veracity of the news. Hence, we introduce a latent variable $d_{n,k}$ to represent whether response k to news n is a debunking response; (2) Only one classifier was trained to predict the debunking probability. Formally, given a piece of false or true news, users will debunk the news with probability η or $\bar{\eta}$ respectively, namely,

$$d_{n,k} \sim Bernoulli(z_n\eta + (1 - z_n)\bar{\eta}). \tag{5}$$

Then, the classifier generates a prediction depending on the value of $d_{n,k}$. Specifically, when $d_{n,k}$ is 1,

$$Y_{n,k}^{(2)} \sim LogisticNormal\left(\mu^{(2)}, \left(\sigma^{(2)}\right)^2\right), \tag{6}$$

and when $d_{n,k}$ is 0, the logistic-normal distribution is parameterized by $\bar{\mu}^{(2)}$ and $\bar{\sigma}^{(2)}$. Similarly, we impose conjugate priors over Bernoulli and logistic-normal distributions:

$$\eta \sim Beta(a_0, b_0), \quad \bar{\eta} \sim Beta\left(\bar{a}_0, \bar{b}_0\right) \tag{7}$$

$$\mu^{(2)}, \left(\sigma^{(2)}\right)^2 \sim NIG\left(\omega_0^{(2)}, \nu_0^{(2)}, \alpha_0^{(2)}, \beta_0^{(2)}\right) \tag{8}$$

$$\bar{\mu}^{(2)}, \left(\bar{\sigma}^{(2)}\right)^2 \sim NIG\left(\bar{\omega}_0^{(2)}, \bar{\nu}_0^{(2)}, \bar{\alpha}_0^{(2)}, \bar{\beta}_0^{(2)}\right) \tag{9}$$

The generation of human judgments in reports is straightforward. After drawing the ground

truth z_n from a Bernoulli distribution parameterized by γ, the model generates the number of reports based on the veracity of news n:

$$Y_n^{(3)} \sim Poisson(z_n\phi + (1 - z_n)\bar{\phi}), \tag{10}$$

which means the average numbers of received reports for false and true news are ϕ and $\bar{\phi}$, respectively. Last, we impose conjugate gamma priors over Poisson distributions to complete our Bayesian model:

$$\phi \sim \text{Gamma}(g_0, h_0), \quad \bar{\phi} \sim \text{Gamma}(\bar{g}_0, \bar{h}_0). \tag{11}$$

The graphical representation of CLNAM is presented in Figure 4. For clarity, the priors are not listed. At the top, the large outer plate represents that there are N pieces of news to assess; the inner plate on the left represents that M classifiers generate probabilities as to the veracity of each piece of news. The bottom and right plates follow similar meanings. The shaded circles denote observable variables, while the empty circles represent latent variables that we want to infer.

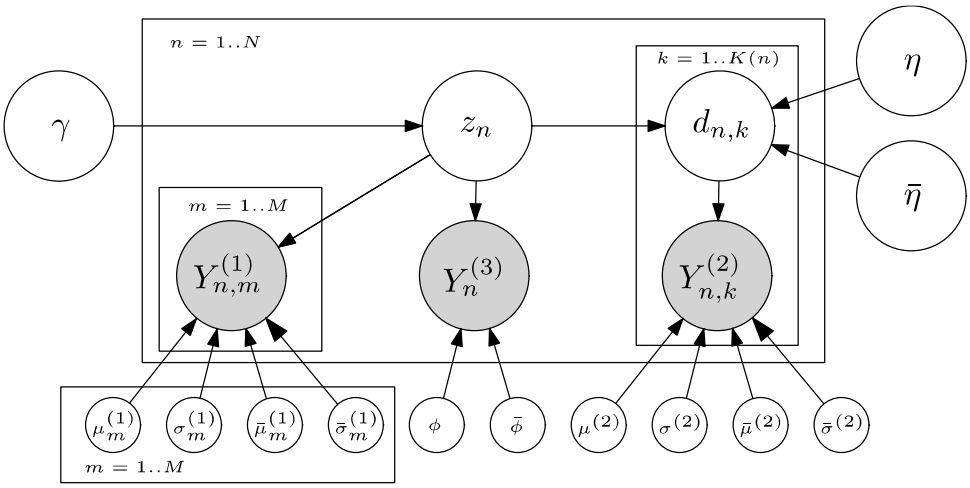


Figure 4. Graphical Representation of the CLNAM Model

3.2.4 **Model Inference**

The goal of the model inference is to compute the posterior of latent variables, primarily the ground truth of each news, i.e., $\boldsymbol{z}$, conditioning on the observable data $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$, and $\boldsymbol{Y}^{(3)}$ **(**let θ_o denote all these observable data**)**. Unfortunately, this conditional density is intractable for exact inference because it necessitates the computation of marginal density of the observations (i.e., $p(\theta_o)$), which requires exponential time (Wainwright & Jordan 2008). To address this issue, we used approximate inference and took a variational inference approach, which is a faster alternative to Markov chain Monte Carlo (MCMC) and scales to large-scale data more easily (Blei et al. 2017). Specifically, we used coordinate ascent mean-field variational inference (Bishop 2006; Blei et al. 2017) to perform approximate inference under CLNAM model. The key idea is to approximate the posterior distribution by a more tractable variational family called mean-field family and hence cast the conditional inference into an optimization problem; then, coordinate ascent algorithm is applied to solve it.

For convenience, we denote all the latent variables as θ_l and denote all the hyperparameters (i.e., parameters of the prior distributions) as θ_h. In the proposed inference algorithm, we specify a fully factorized mean-field variational family $\mathcal{Q}$ over the latent variables:

$$\begin{aligned} q(\theta_l) = {} & q(\boldsymbol{z}|\boldsymbol{\lambda})q(\gamma|e,f)q(\boldsymbol{d}|\boldsymbol{\tau})q(\eta|a,b)q\left(\bar{\eta}\middle|\bar{a},\bar{b}\right)q(\phi|g,h)q\left(\bar{\phi}\middle|\bar{g},\bar{h}\right) \\ & \cdot q\left(\boldsymbol{\mu}^{(1)},\boldsymbol{\sigma}^{(1)}\middle|\boldsymbol{\omega}^{(1)},\boldsymbol{\nu}^{(1)},\boldsymbol{\alpha}^{(1)},\boldsymbol{\beta}^{(1)}\right)q\left(\bar{\boldsymbol{\mu}}^{(1)},\bar{\boldsymbol{\sigma}}^{(1)}\middle|\bar{\boldsymbol{\omega}}^{(1)},\bar{\boldsymbol{\nu}}^{(1)},\bar{\boldsymbol{\alpha}}^{(1)},\bar{\boldsymbol{\beta}}^{(1)}\right) \\ & \cdot q\left(\mu^{(2)},\sigma^{(2)}\middle|\omega^{(2)},\nu^{(2)},\alpha^{(2)},\beta^{(2)}\right)q\left(\bar{\mu}^{(2)},\bar{\sigma}^{(2)}\middle|\bar{\omega}^{(2)},\bar{\nu}^{(2)},\bar{\alpha}^{(2)},\bar{\beta}^{(2)}\right), \end{aligned} \tag{12}$$

where each latent variable is governed by its own variational distribution: $\boldsymbol{z}$ and $\boldsymbol{d}$ follow Bernoulli distributions; γ, η, and $\bar{\eta}$ follow beta distributions; ϕ and $\bar{\phi}$ follow gamma distributions; all pairs of μ and σ are NIG distributions. For convenience, we denote all variational parameters as θ_v.

In the variational family $\mathcal{Q}$, each member serves as a candidate approximation to the exact posterior $p(\theta_l|\theta_o, \theta_h)$. Hence, our aim is to find the one that is closest in Kullback-Leibler (KL) divergence to the exact posterior. Formally, the inference problem is casted as the following optimization problem:

$$q^*(\theta_l) = \underset{q(\theta_l) \in \mathcal{Q}}{arg\,min}\, KL(q(\theta_l)||p(\theta_l|\theta_o, \theta_h)). \quad (13)$$

It is well-known that minimizing KL divergence in the variational approach amounts to maximizing the evidence lower bound (ELBO) $\mathcal{L}(\theta_v)$, which is equivalent to the KL divergence up an added constant (Blei et al. 2017). The form of the ELBO is as follows:

$$\mathcal{L}(\theta_v) = \mathbb{E}_q[log\, p(\theta_o, \theta_l|\theta_h)] - \mathbb{E}_q[log\, q(\theta_l)]. \quad (14)$$

To maximize the ELBO, we develop a coordinate ascent algorithm. Specifically, we iteratively optimize the variational parameters for each latent variable while holding the others fixed. Because all the complete conditionals (i.e., distribution of each latent variable conditioned on other latent variables and observable variables) are in the exponential family and the priors are conjugate, each coordinate update in the proposed CLNAM model is available in closed form and the coordinate ascent algorithm is guaranteed to climb the ELBO to a local optimum (Blei et al. 2017). Because of space limitation, we present the updated formulas in Section B2 of the online supplement. In practice, we iteratively update the variational parameters of each latent variable until the change in ELBO falls below some small threshold (e.g., 0.0001) or the algorithm has iterated for a certain number of times (e.g., 200 iterations). Once the algorithm converges, the final prediction is obtained by examining λ_n, the posterior probability of news n being false.

3.2.5 **Learning Prior Beliefs from Data**

In practice, the performance of Bayesian models is sensitive to prior assumptions; consequently, it's necessary to choose proper prior distributions based on prior knowledge (Liu & Aitkin, 2008). We observe that we can take advantage of the two-stage characteristics and learn prior beliefs from the information extraction stage. For example, when we train a classifier at the first stage, we can foresee how the classifier will perform in an unknown test dataset by examining its behavior in the validation dataset, which is used for model selection. To leverage such information, we propose learning prior beliefs from the data. See Section B3 of the online supplement for more details.

4. EMPERICAL EVALUATIONS

In this section, we comprehensively evaluate the proposed CAND framework using two real-world datasets from the social media platforms Weibo and Twitter. First, we describe the experimental design including datasets, benchmark methods, evaluation metrics, and experimental procedure. Then, we report and discuss the experimental results.

4.1 Experimental Design

4.1.1 **Datasets and Preprocessing**

The first dataset was collected from Sina Weibo. After scrutinizing the reported posts, Sina Weibo announces all officially fact-checked false news in the Sina community management center under the false news category. We retrieved all the false posts between February 2012 and August 2018. To obtain a set of true news, we retrieved posts that were not reported from general threads, following the practice in the previous literature (Ma et al., 2016). We then collected relevant information for each post, including post content, comments, number of received reports, user profiles, and context information (e.g., timestamp). For Chinese sentences, we

segmented them into meaningful words using jieba, a popular Chinese word segmentation tool (Peng et al., 2017). In this dataset, the false and true news items were not collected from the same time period. To learn the intrinsic characteristics of false news and prevent the model from learning news event-specific features, we used an effective event detection algorithm, i.e., a single-pass clustering method, to discover unique news events (Wang et al., 2018). After clustering, we selected the most popular post (i.e., the one receiving the most responses) from each cluster to represent the news event, because the most popular post usually contains the most complete information regarding this event, which makes the task of false news detection easier. In the Real-World Application of CAND section below, we show that clustering the posts does not impact the real-world application of our framework. In total, the dataset consisted of 2,186 pieces of false news and 9,455 pieces of true news. Table 1 shows some examples of false and true Weibo posts. In our collected dataset, we record the number of received reports for false news. Unfortunately, the following information is not available: true news that was mistakenly reported and false news that was not reported. To remedy this data unavailability issue, we assume a misreport rate of true news of 0.01—namely, the number of received reports for true news follows a Poisson distribution with a parameter of 0.01. In addition, we assume a report rate of false news of 50%. To achieve this, we randomly chose 50% of the false news items and set the number of received reports as 0. Although this is one limitation, we conduct stress testing for these two rates in the Stress Testing of the Simulation Parameters section below. We also verify the performance of our proposed framework with and without the partially simulated data.

We collected another dataset from Twitter based on three reference datasets: Twitter (Ma et al., 2016), Twitter15, and Twitter16 (Ma et al., 2017), where the ground truth of the news is confirmed through an expert-oriented fact-checking website, i.e., Snopes. For these tweets, we

then retrieved relevant information including all responses to the tweets, user profiles, and context information. Note that reporting/flagging is not available on the Twitter platform. Hence, we did not take into account human intelligence in reports when evaluating the proposed approach using the Twitter dataset. Unlike the Weibo dataset, no event detection was needed, as the datasets from the literature were already at the news-event level. In total, the dataset contained 943 pieces of false news and 1,007 pieces of true news. See Table 1 for some examples. Note that, in this work, we focus on identifying false news (regardless of the intention and the means of falsifying information) rather than other specific types of false news such as fake news or partisan fake news (Pennycook & Rand, 2019a). When using these two datasets, we made an implicit assumption that each dataset we collected was a set of false and true news.

Table 1. Examples of Posts and Responses with Different Labels

Task	**Dataset**	**Label**	**Examples**
T1: feature-based false news detection	Weibo	False news	#Typhoon Mangkhut# On the highway of Shenzhen along the river, a minibus was blown over. Hope everyone in the car is safe [bless] [loc: Shenzhen]
		True news	My favorite movie is coming. Looking forward...
	Twitter	False news	42 Million Dead In Bloodiest Black Friday Weekend On Record [link]
		True news	Senate approves $600 million in border security funds [link]
T2: debunking response detection	Weibo	Debunking	Reported, it's false!
		Not debunking	It's so scaring.
	Twitter	Debunking	@[user] ig but many people saying its not true
		Not debunking	Thanks u i wish your services weren't needed but i am thankful you are here to balance the scales.

Note: All examples are from our datasets. In the content, "[bless]" means an emoji, "[loc: Shenzhen]" refers to a hyperlink about the location, [link] is a link to another website, and [user] is the username of a Twitter user.

In the previous literature, datasets are frequently constructed as balanced (Ma et al., 2015, 2016; Zhang et al., 2015). However, there are many more true posts than false ones in real-world scenarios. To rigorously evaluate our proposed approach, we focused on unbalanced settings and constructed the test bed with different imbalance ratios (i.e., ratio between false and true news, denoted as IR). Specifically, for the Weibo dataset, we varied the IR from 1:1 to 1:100; for the Twitter dataset, we set the IR from 1:1 to 1:2.5. Note that because of size limits, we were unable to use a drastic IR in the Twitter dataset because a drastic IR would have made the constructed dataset too small to train our deep learning-based methods. In the following, we use Weibo as an example to illustrate how the data was prepared for Tasks T1 and T2. For each IR, we randomly sampled from the dataset and constructed the largest possible set that satisfied the IR. Given that the IR can be as drastic as 1:100, we used stratified sampling in our experiment. Figure 5 shows the case when IR is 1:7. Specifically, we randomly sampled 6,483 posts (60%) as training posts, 2,161 (20%) as validation posts, and the remaining 2,161 (20%) as test posts. The test posts were intended to be the set of Weibo postings for assessing the veracity when using our framework on real-world platforms. To get the dataset for Task T2, we randomly sampled 12,000 responses from the responses in the non-test dataset. For training the model in T2, four annotators with the necessary background independently labeled each post-response pair based on whether the response was a debunking response or not. The averaged inter-annotator Cohen's kappa score is 0.816, indicating very good annotation consistency (McHugh, 2012). The final label was obtained using the iBCC model (introduced above), which is an effective crowdsourced answer aggregation model. In the annotated dataset, 17.51% of the responses were debunking responses. See Table 1 for positive and negative examples of responses. Because the aim of Task T2 was to extract information from responses to the test posts, we adopted a different data split scheme. We

randomly sampled 9,000 responses (75%) as the training responses and the remaining 3,000 responses (25%) as the validation responses, while treating all the responses to the test posts as test responses. Preprocessing of the Twitter dataset was similar. In Appendix A1, we present additional details of the data preprocessing procedure.

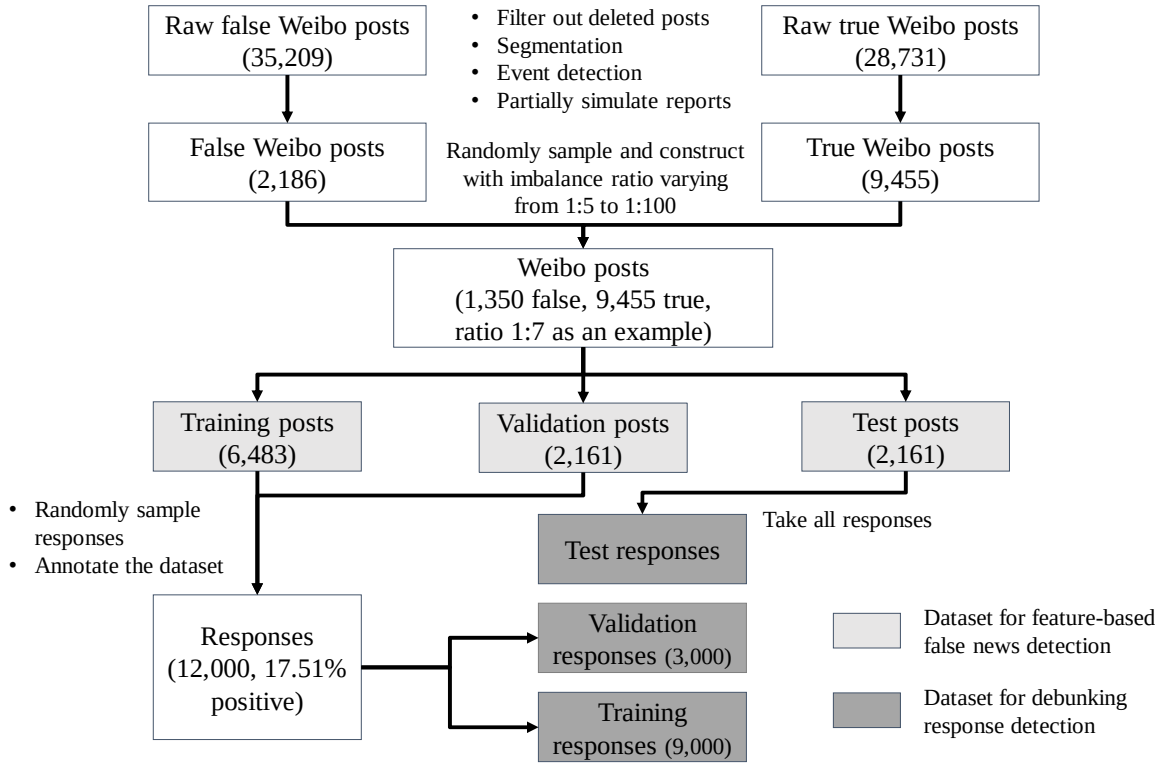


Figure 5. Data Preprocessing for Weibo Dataset

Table 2 presents the statistical summary of the datasets. For the Weibo dataset, false posts tended to be shorter and received more responses compared with true posts. This is consistent with the literature that false news usually receives more attention (Vosoughi et al., 2018). However, such differences were not salient for the Twitter dataset. We also report the average debunking probability because the key assumption of our crowd-powered framework is that, compared with true news, false news tends to be debunked more by crowd users in their responses. For the Weibo dataset, the average debunking probability of false news (i.e., 26.3%)

was significantly higher than that of true news (i.e., 4.5%). For the Twitter dataset, although the difference was smaller, the crowd intelligence in responses still significantly contributed to the detection of false news, as demonstrated by the experimental results in the Twitter dataset (Appendix B2). Since the reports are partially simulated in the Weibo dataset and not available in the Twitter dataset, the average number of reports is not presented in Table 2.

Table 2. Statistical Summary of Datasets

	False News				**True News**			
	# of posts	**Avg. length**	**Avg. # of responses**	**Avg. debunking prob.**	**# of posts**	**Avg. length**	**Avg. # of responses**	**Avg. debunking prob.**
Weibo	2,186	57.07 (53.30)	49.80 (58.69)	0.263	9,455	88.95 (123.84)	34.73 (49.04)	0.045
Twitter	943	17.64 (6.58)	19.73 (32.14)	0.107	1,007	17.53 (6.00)	18.94 (30.28)	0.083

Note: Values in parentheses represent standard deviations. Length of Weibo posts refers to the number of words after segmentation. The debunking probabilistic is predicted based on the CNN classifier in T2.

4.1.2 **Baseline Methods and Evaluation Metrics**

Recall that the major innovation of the proposed CAND framework lies in incorporating crowd judgments and developing the CLNAM result aggregation model. Hence, we evaluate the framework from the following two aspects: (1) whether combining crowd judgments in responses and reports contributes to false news detection, and (2) whether the proposed CLNAM model is effective compared to other end-to-end or aggregation benchmark methods. To demonstrate the contribution of crowd judgments, we incorporated the data sources into the CAND framework in an incremental manner and named it based on the index of the data sources. For example, CAND-12 integrates machine judgments from news features and hybrid judgments in responses. For methods using content and context features, we designed five benchmark methods based on widely used methods from the literature of false news detection

and deep learning: support vector machine (SVM), convolutional neural network (CNN), long-short term memory (LSTM), bidirectional LSTM (Bi-LSTM), and bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019; Kim, 2014; Ma et al., 2016; Nguyen et al., 2017; Wang, 2017; Wang et al., 2018). These benchmark methods also serve as the base classifiers in Task T1. The reason we designed these benchmark methods for Task T1 is to show the value of the base classifiers, ranging from shallow to deep structures. We present the experimental results and the values of these classifiers below. Unfortunately, when crowd intelligence is taken into account, no benchmark methods use exactly the same set of inputs (i.e., content and context features, user responses, and user reports). We designed two types of benchmark methods, i.e., end-to-end and aggregation, based on the existing literature. For end-to-end models, we mainly consider models that can handle news text and user responses; see a recent survey for all candidates (Sharma et al., 2019). The first benchmark, Concat, is a concatenation-based deep learning approach, which takes advantage of the superior performance of deep learning in fusing multiple data views (Qian et al., 2018; Wang, 2017). The second benchmark, HSA, relies on a more advanced deep learning structure for representation learning, i.e., the hierarchical attention network (Guo et al., 2018; Yang et al., 2016).

There is no readily available aggregation model in the literature for our setting because we need to aggregate mixed data types; the relevant literature uses only one type of data. To rigorously evaluate the CLNAM aggregation model, we implemented two benchmark methods: a trivial majority voting (MV) aggregation model and a nontrivial binary aggregation model (BAM). The BAM model is similar to the CLNAM model except that it depicts the credibility of classifiers in Tasks T1 and T2 based on the widely used two-coin assumption, where the predictions conditioned on the ground truth are binary and follow Bernoulli distributions (Bragg

& Weld, 2013; Raykar et al., 2010; Wei et al., 2017). Details of these benchmark methods are presented in Appendix A2. Finally, for both end-to-end and aggregation models, we also incorporated data sources in an incremental manner. Table 3 summarizes the full list of methods for benchmarking.

Table 3. Summary of Methods

Data sources	Method type	Benchmark methods	Proposed methods
Source 1	End-to-end	SVM, CNN, LSTM, Bi-LSTM, BERT	/
	Aggregation	MV-1, BAM-1	CAND-1
Source 1 & 2	End-to-end	Concat-12, HSA-12	/
	Aggregation	MV-12, BAM-12	CAND-12
Source 1, 2, & 3	End-to-end	Concat-123, HSA-123	/
	Aggregation	MV-123, BAM-123	CAND-123

To evaluate the performance of these methods, we adopted widely accepted evaluation metrics, including PR AUC (area under the precision-recall curve), F1 score, recall, and precision. We paid the most attention to PR AUC because, with highly unbalanced datasets, the PR curve gives an accurate picture of an algorithm's performance (Davis and Goadrich, 2006) whereas the precision-recall pair or F1 score represents only one point on the PR curve. After PR AUC, we focused more on recall rather than precision because, given the context of real-world false news detection scenarios, we preferred to recall a larger percentage of false news in exchange for a reasonable sacrifice in precision. Accuracy is not reported because it does not provide adequate information on a classifier's functionality in unbalanced datasets (He & Garcia, 2008). In the following, unless otherwise specified, AUC refers to PR AUC.

4.1.3 **Experimental Procedure**

In the information extraction stage, five feature-based benchmark methods (i.e., SVM, CNN,[7] LSTM, Bi-LSTM, and BERT) were trained as the base classifiers for Task T1. For Task T2, given a post and one of its responses, we used two CNN modules to learn the hidden representations, respectively. The learned representations were then concatenated and fed into a fully connected layer, followed by a softmax layer. All methods were implemented in Python. The SVM was implemented using Python's scikit-learn library (Pedregosa et al., 2012). The HSA was coded using TensorFlow. Other deep learning methods were implemented using Python's Keras library with TensorFlow as the backend. For all deep-learning methods, sentences were represented as a sequence of vectors using word embedding, which is a commonly used language modeling and feature learning technique in text classification (Mikolov et al., 2013; Zhang et al., 2019). The embedding vectors were initialized with open source embeddings trained using Weibo data (Li et al., 2018) or Twitter data (Pennington et al., 2014). For fair comparison, we conducted a random grid search of 40 trials to tune the hyperparameters for all methods, including the benchmark methods. We report these hyperparameters in Section C1 of the online supplement.

The results of Tasks T1 and T2, together with the partially simulated number of reports, were fed as the input to the aggregation models (i.e., MV, BAM, and CAND). Regarding CAND, we limited the time to collect responses to one day and truncated the number of reports to 20 for each piece of news. We defend this choice from the perspective of early detection below. Next, we learned the prior beliefs from the data and finally inferred the latent variables. All methods were evaluated for 10 runs with a different randomization seed in each run. Lastly, the whole

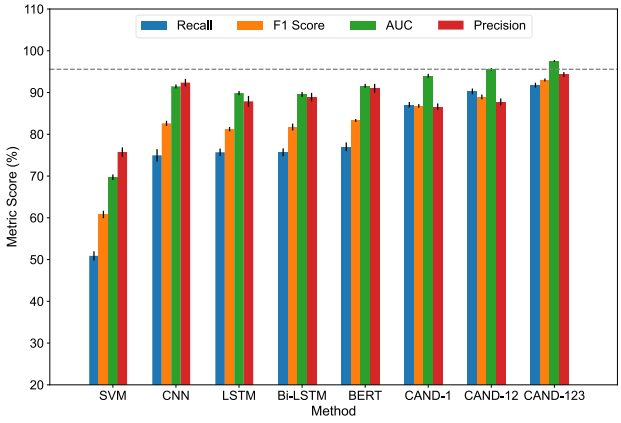
[7] All CNN networks in this paper follow a classical architecture for text classification (Kim 2014).

procedure was repeated for all IRs in both datasets.

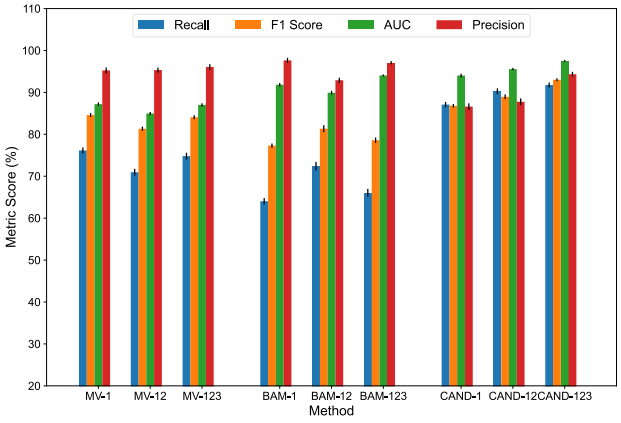
4.2 False News Detection Performance: CAND vs. the Benchmark Methods

In this section, we report the experimental results in the form of bar charts and line charts. Because of space limitations, we mainly report the Weibo dataset results and report the rest of the results in Appendix B. For better visualization, the y axis of the following graphs might not start at 0 or end at 1, although all metrics range from 0 to 1. Whenever an error bar is reported, the result is averaged over 10 runs and the standard error of the mean (SEM) is used.

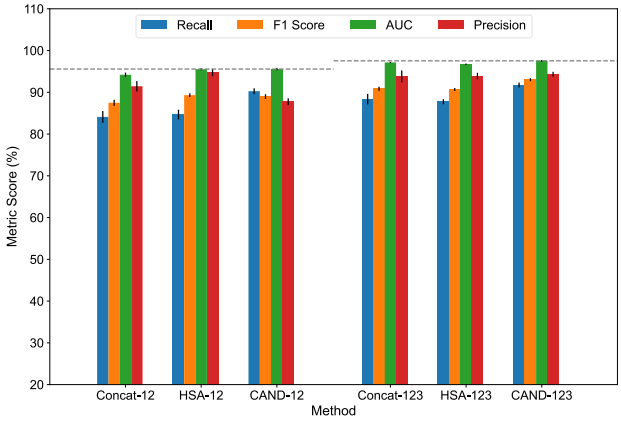
Figure 6 compares the performance of CAND with the benchmark methods under certain IRs. First, given that human intelligence (e.g., responses and reports) is still largely underexplored in the current literature, we examine the performance of CAND compared with end-to-end benchmark methods without crowd intelligence (i.e., feature-based classifiers in Task T1). As shown by Figure 6a, our proposed CAND-123 approach performs significantly better than all benchmark methods across all metrics. For example, compared with the best benchmark, BERT, CAND-123 increases the AUC score from 91.62% to 97.54%. Even if we evaluate our CAND framework without the partially simulated human judgments in reports, CAND-12 still outperforms all benchmark methods across all major metrics by significant margins. In addition, when no human intelligence is considered, CAND-1, as an ensemble of five feature-based classifiers, still performs better than any individual method in AUC score. Overall, the results suggest the superior performance of our framework in comparison with feature-based false news detection methods and the effectiveness of our framework as an aggregation method.



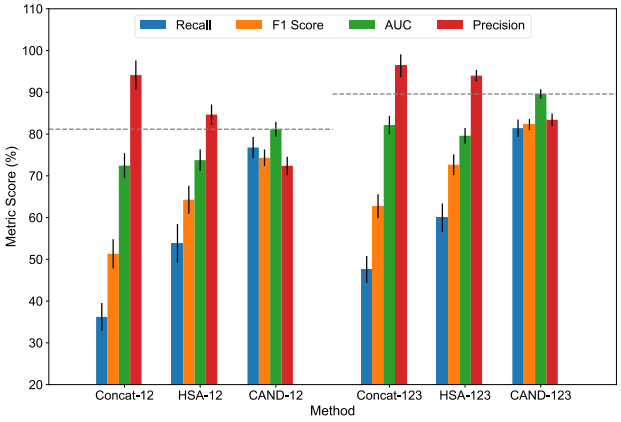
(a) CAND vs. End-to-End Benchmark Methods without Crowd Intelligence (IR=1:7, Ref. line: AUC of CAND-12)



(b) CAND vs. Aggregation Benchmark Methods (IR=1:7)



(c) CAND vs. End-to-End Benchmark Methods with Crowd Intelligence (IR=1:7, Ref. lines: AUC of CAND)



(d) CAND vs. End-to-End Benchmark Methods with Crowd Intelligence (IR=1:50, Ref. lines: AUC of CAND)

Figure 6. Performance of CAND and Benchmark Methods in the Weibo Dataset

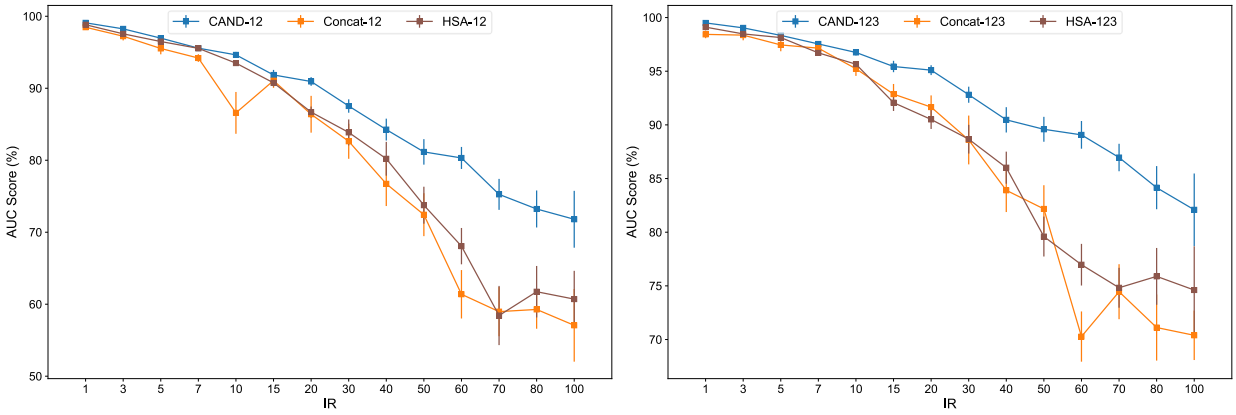
The comparison of the aggregation models in Figure 6b demonstrates the effectiveness of incorporating human intelligence and the CLNAM result aggregation model. For CAND, as more data sources are taken into account, the performance across all metrics increases significantly. For example, from CAND-1 to CAND-12 to CAND-123, the F1 score increases

from 94.02% to 95.57% to 97.54%. However, MV and BAM benefit less from human intelligence because they do not model (or use simple assumptions to model) the credibility of humans and machines. The comparison suggests the usefulness of human intelligence and the necessity of modeling the credibility of humans and machines. To evaluate the effectiveness of the proposed CLNAM aggregation model, we compared all aggregation methods by fixing the combination of data sources. Figure 6b shows that across all choices of data sources, especially when human intelligence is considered, CAND has the best performance in AUC, F1 score, and recall. The comparison clearly demonstrates the superior performance of the proposed CLNAM result aggregation model. In addition, the gap between CAND and BAM is attributed to using the logistic-normal assumption in place of the Bernoulli assumption when modeling the classifiers' credibility.

Last, we compare CAND with end-to-end benchmark methods that consider crowd intelligence (i.e., Concat and HSA) in Figure 6c (IR=1:7) and Figure 6d (IR=1:50). Figure 6c indicates that when the dataset is relatively balanced (e.g., IR=1:7), HSA performs comparably to CAND and they both perform slightly better than Concat. However, as IR becomes more drastic (e.g., IR=1:50), Concat and HSA suffer from the data imbalance problem and tend to predict with high precision but low recall. As a result, our CAND method significantly outperforms the benchmark methods across all major metrics.

To further explore the robustness of our method and the benchmark methods when the data is unbalanced, we compare their performance by varying the IR. The results using AUC as the metric are reported in Figure 7 (vs. end-to-end benchmark methods with crowd intelligence) and Figure B1 of Appendix B1 (vs. feature-based and aggregation benchmark methods). As expected, when the dataset becomes more unbalanced, the performance of CAND and all

benchmark methods decreases. Meanwhile, the performance gap between CAND and the benchmark methods increases, and our method is more stable than the benchmark methods (i.e., smaller SEM). The results suggest that our method, as an ensemble of machine intelligence and human intelligence, is more robust against the data imbalance. This makes our method more applicable and effective, as the actual news in real-world scenarios is often highly unbalanced.



(a) Data Source: 1 and 2

(b) Data Source: 1, 2, and 3

Figure 7. CAND vs. End-to-End Benchmark Methods with Crowd Intelligence under Different IRs (Weibo, AUC Score as Metric)

We conducted the same set of experiments in the Twitter dataset and the results are reported in Appendix B2. The experimental results lead to the same conclusions. In the above experiments, we randomly split the dataset into training/validation/test sets to avoid overfitting and fixed the training set percentage to 60%. To show that such specifications did not influence our experimental results, we conducted two robustness checks. First, under the training/validation/test split scheme, we chronologically split the data and varied the training set percentage from 40% to 90%. The chronological split also demonstrates the efficacy of our method as if it were working in real time. Second, we tested how k-fold cross-validation (with k ranging from 3 to 25) and leave-one-out cross-validation (LOOCV) affect the performance of

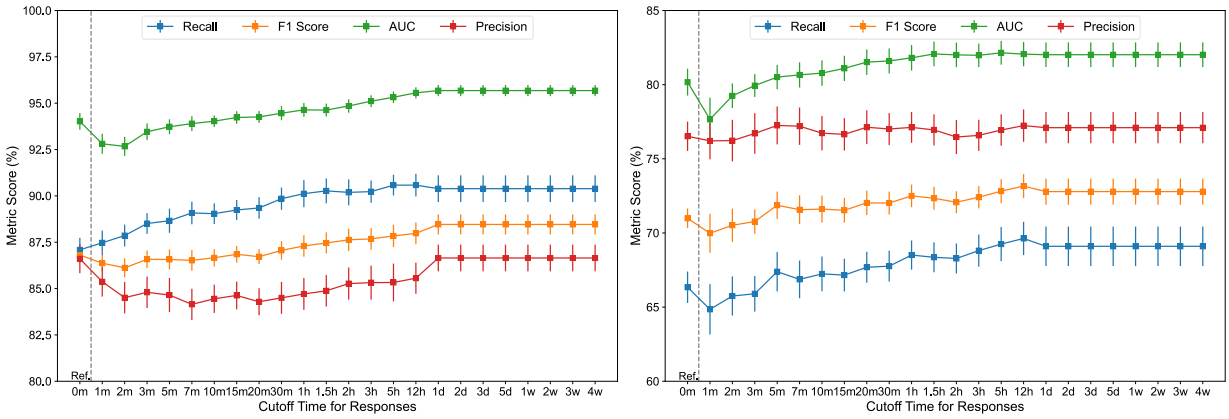
CAND. The results suggest that our evaluation is robust against the dataset split scheme. Because of space limitations, we present the results in Section D1 of the online supplement.

4.3 Analysis and Discussion

4.3.1 Early Detection

In real-world applications, the number of responses and reports usually increases daily. Also, in the development of false news events, debunking messages often appear at a later stage. However, it is almost impossible to collect all human intelligence (i.e., responses and reports) in real-world scenarios. When fighting false news, the false news needs to be identified as soon as possible to stop the propagation process quickly (Tschiatschek et al., 2018). Hence, a practical question is whether our approach enables the early detection of false news. To this end, we set different cutoff times for the responses in CAND-12. Figure 8 plots the performance of CAND-12 in the Weibo dataset (IR = 1:7) and the Twitter dataset (IR = 1:1.5) with the maximum time ranging from one minute to four weeks. As shown by the figures, responses, as one type of human intelligence, start taking effect even in the first few minutes. The performance continues increasing significantly on the first day and stabilizes thereafter, because what matters to the CAND framework is the rate of the debunking responses rather than the number of responses. A moderately long time (e.g., 12 hours to one day) is enough, as long a representative sample of the responses has been collected. Although crowd opinions at a later stage might contain more debunking messages, our experiment suggests that aggregating early human intelligence is enough to enable false news detection. Overall, our experiments suggest that our proposed framework is capable of detecting false news at the early stages. Note that we were unable to conduct similar analysis using human intelligence in reports because the reporting time stamps were unavailable in our datasets. However, the number of responses and reports usually

increases over time. In Section D2 of the online supplement, as a heuristic way to explore early detection, we examine the effect of the number of responses and reports on the performance of our proposed approach.



(a) CAND-12 (Weibo, IR=1:7) (b) CAND-12 (Twitter, IR=1:1.5)

Figure 8. Testing Early Detection by Varying Cutoff Times for Responses

Note: CAND-1 (i.e., "Ref." in the graph) is listed as a reference.

4.3.2 **Real-World Application of CAND**

In our datasets, posts were clustered at the news-event level and we chose the most popular post to represent each event. However, in real-world platforms, the posts usually come one by one, rather than in the form of clusters. In this section, we illustrate how our framework can be applied.

When a new post arrives, we first judge whether this belongs to an existing cluster (i.e., existing news event), based on the clustering standard we use above in the Empirical Design section. If so, the cluster grows; otherwise, the new post is treated as the seed of a new cluster. For a new news event, we keep collecting relevant posts and their responses before feeding them as an input into our algorithm for fact-checking. Figure 9 shows the timeline of a new news event. After the news event starts at $t1$, the number of relevant posts continually increases on

social media, and crowd users' responses and reports gradually accumulate. When we have collected enough human intelligence at $t2$, we use our framework to predict whether this event is false. Note that, our crowd-powered framework does not try to fact-check in real time because no human intelligence is available at $t1$. Instead, we significantly improve the performance by slightly sacrificing timeliness. As our experiments in the Early Detection section show, waiting for about 12 hours is enough to collect a representative sample of human intelligence in responses. We emphasize that this is a necessary trade-off for any algorithm that taps into human intelligence. Last, our framework also allows real-time detection of false news, in which case CAND-12 and CAND-123 reduce to CAND-1.

t_1: Beginning of a news event
t_2: When we run the algorithm to classify the news event

Figure 9. Timeline of a New News Event

4.3.3 Complementary Strengths of Humans and Machines

Figure 10 illustrates an example of the relative contributions of the three types of judgments (i.e., machine judgments from features, hybrid judgments in responses, and human judgments in reports) and prior beliefs when calculating the posterior of ground truth $\boldsymbol{z}$. This is an example of CAND-123 with an imbalance ratio of 1:7. For a certain post (fixed x), we obtain a thin bar where the ratio of each color represents relative contributions from the three sources and prior beliefs. Overall, machine judgments and hybrid judgments contribute the most. However, the

graph shows that no single type of judgment always dominates the others; on the contrary, they complement each other in classifying more than half of the posts. This demonstrates the complementary strengths of human and machine intelligence. Visualizing the examples in the Twitter dataset leads to the same conclusion. This graph also suggests that the proposed Bayesian model is highly transparent and interpretable—we know exactly how the prediction of each post is obtained and how much each data source contributes.

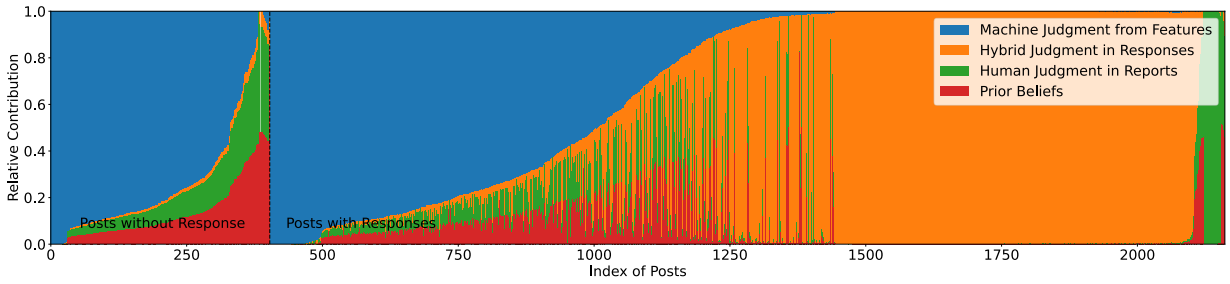


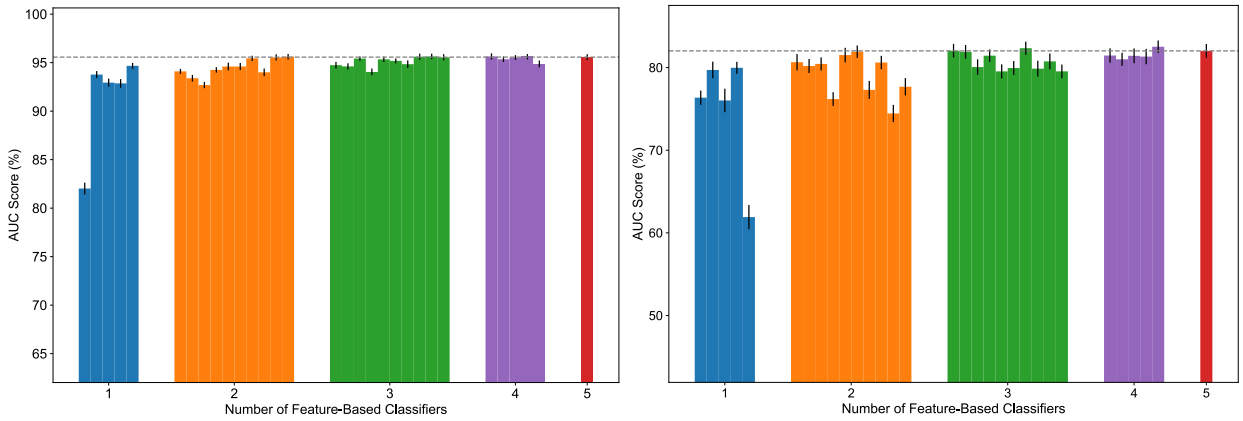
Figure 10. A Stacked Bar Graph Showing the Relative Contributions of the Three Types of Judgments and Prior Beliefs (Weibo, IR=1:7)

Note: When inferring the ground truth of each piece of news, the odds of correctly predicting the ground truth is updated as $\exp\{s_1 + s_2 + s_3 + s_p\}$, where s_* refers to the contributions from the three sources and prior beliefs respectively (see Equation B5 in Section B2 of the online supplement). The relative contributions are computed based on value $\exp\{s_*\}$. The posts are divided into two groups based on the existence of response. Within each group, the posts are reordered for better visualization. Note that hybrid judgments in responses still contribute even when no response exists because the contribution is computed based on the exponential and $\exp\{0\} = 1$.

4.3.4 Value of Multiple Feature-Based Classifiers

In our experiment, we use multiple classifiers for Task T1 to extract machine judgments. To make our approach practically applicable, we explore the value of these classifiers and answer how many classifiers to use. To this end, we varied the combination of five feature-based classifiers (i.e., SVM, CNN, LSTM, Bi-LSTM, and BERT) and examined the performance of CAND-12. The results are presented in Figure 11. Each group of bars of the same color

represents the same number of classifiers. For example, ten orange bars correspond to the results of CAND-12 with combinations of two out of five classifiers. Rather than giving the name of each bar, we focus on how CAND-12 performs against the number of classifiers. In general, as more feature-based classifiers are included, CAND-12 delivers better and more stable performance. The marginal effect, however, is decreasing. The results suggest the significant value of including multiple classifiers—in practice, a moderate number of classifiers (e.g., five) are sufficient. In addition, the performance of the red bars indicates that the inclusion of a weak classifier (i.e., SVM) does not influence the performance. This is because we modeled the credibility of each classifier in our Bayesian result aggregation model and the algorithm assigns less weight to weak classifiers during the parameter estimation.



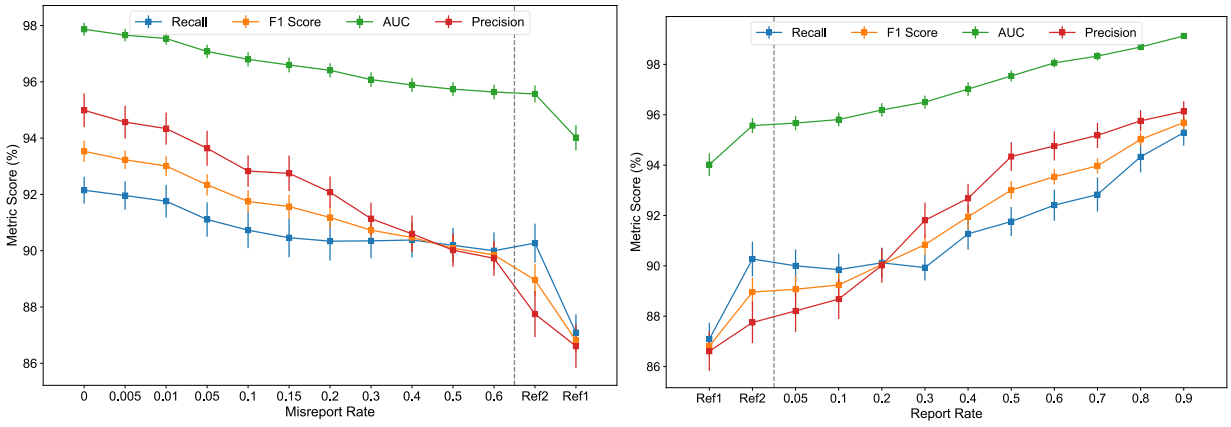
(a) Weibo, IR=1:7 (b) Twitter, IR=1:1.5

Figure 11. AUC Score of CAND-12 under Different Combinations of Feature-Based Classifiers (Ref. lines: red bar)

4.3.5 Stress Testing of the Simulation Parameters

In our experiment, we partially simulated the reporting data to remedy the data unavailability issue. In order to verify whether human intelligence in reports can consistently contribute to false news detection, we conducted stress testing for the simulation parameters. Specifically, we

varied the misreport rate from 0 to 0.6 and examined the performance of CAND-123. The results are reported in Figure 12a. As expected, when more true news posts are misreported, the performance of CAND-123 keeps decreasing. In particular, even when up to 60% of true news posts are misreported, CAND-123 still performs better than CAND-12, in which no report information is considered. Similarly, we varied the report rate of false news from 0.05 to 0.9 and present the results in Figure 12b. We observe that the performance of CAND-123 increases as a function of the report rate. Even when only 5% of the false news posts are reported, CAND-123 still performs better than CAND-12. In summary, although we partially simulate the human reports because of data unavailability, our experiment demonstrates the unique value of crowd reports under various simulated settings, which may even go beyond the normal case.



(a) Varying Misreport Rates of True News

(b) Varying Report Rates of False News

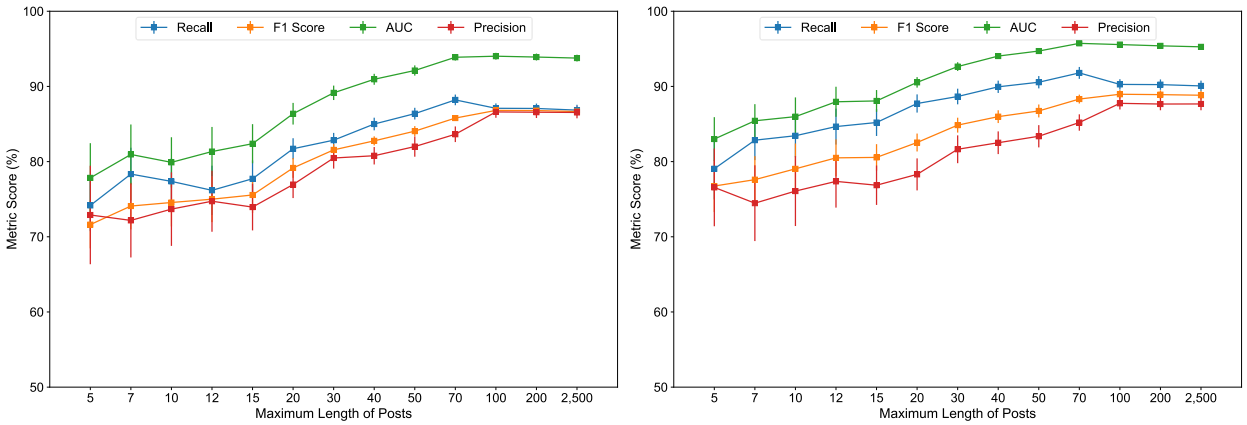
Figure 12. Performance of CAND-123 on Different Misreport Rates of True News and Report Rates of False News (Weibo, IR=1:7)

Note: CAND-1 and CAND-12 (i.e., "Ref1" and "Ref2" in the graph) are listed as references.

4.3.6 Sensitivity to News Characteristics

News length: In the task of false news detection, we would like our approach to be feasible to news items of different lengths, especially short ones containing limited information. In the two

datasets we used, the news item length in the Weibo dataset was more varied (ranging from 1 to 2,484), compared to the Twitter dataset (ranging from 1 to 39). We therefore used the Weibo dataset as a test bed to test the sensitivity of our approach to news length. Specifically, we set the maximum length of news items at between 5 to 2,500 and examined the performance of CAND-1, CAND-12, and CAND-123 (IR=1:7). See the results of CAND-1 and CAND-12 in Figure 13. As expected, performance increased as a function of news length because longer news articles usually contain richer information. Compared to CAND-1, CAND-12 was less affected (lower slope in the graphs) by the length of news items, because CAND-1 only considers news content and its context features while CAND-12 also takes crowd responses and reports into account, making them it dependent on news features. The results suggest another advantage to considering crowd intelligence—the false news detection algorithm will be less impacted by news features such as news length.



(a) Performance of CAND-1 with Varying Maximum Lengths of News Items

(b) Performance of CAND-12 with Varying Maximum Lengths of News Items

Figure 13. Sensitivity of CAND to News Length (Weibo, IR=1:7)

News type: In this paper, we adopted a straightforward classification of false versus true news in the types of posts. However, there could be more granularity in news types, and examining the

sensitivity of our proposed method and the benchmark methods could potentially deliver valuable insights. To this end, we adopted a classification scheme wherein false information is divided into seven potentially overlapping types: satire or parody, fabricated content, misleading content, imposter content, manipulated content, false connection, and false context (Sharma et al., 2019). See Section A of the online supplement for details of the classification scheme. Considering the size of the dataset, we used the Weibo dataset to conduct the sensitivity analysis. We selected 1,350 Weibo posts and annotated them with the above-mentioned seven types. By examining the performance of CAND and other benchmark methods (i.e., SVM, CNN, Bi-LSTM, BERT, Concat-12, and HSA-12), we find that the performance of these methods exhibits similar distribution over the type of news. Overall, false news with false context is the easiest to identify, whereas false news with manipulated content and false connections is among the hardest. The details are presented in Section D3 of the online supplement. Although we only scratch the surface of this interesting topic and show that each method performs differently on different types of news, we plan to further explore this direction to enhance the performance or draw other insightful conclusions.

4.3.7 **Robustness to Intentional Manipulation of Responses and Reports**

In our proposed approach, we utilize crowd intelligence to help detect false news. However, there exist many malicious user accounts (e.g., social bots, cyborg users, and trolls) on social media, which automatically post, share, and even comment (Shu et al., 2017). A critical question is how robust our approach is to such crowd manipulation. We show above that under various misreport rates of true news and report rates of false news, our method that includes human intelligence in reports (i.e., CAND-123) consistently performs better than the one without human reports (CAND-12); further, the performance gap is significant when these two rates are within a

reasonable range. This experiment illustrates that, to some degree, our approach is robust to the intentional manipulation of reports. In the following, we test whether our approach is robust to maliciously manipulated crowd responses. Specifically, we simulated six types of malicious users who blindly manipulated responses in a certain way. We varied the percentage of manipulated responses and compared the performance of CAND-12 and CAND-1. The experimental results show that our probabilistic framework is not vulnerable to these types of adversarial attacks, and integrating crowd responses will contribute to, or at least not jeopardize, the false news detection task. We also analyzed and illustrated why the proposed approach is robust to such maliciously manipulated crowd responses. Because of space limitations, the details are presented in Section D4 of the online supplement.

4.3.8 **Debunking Response Detection as a Multi-Class Classification Problem**

In our proposed CAND framework, we treat the task of debunking response detection as a binary classification problem (i.e., debunking or not). However, the literature frequently categorizes responses into multiple stances or opinions, such as supporting, denying, commenting, and querying (Kochkina et al., 2018; Ma et al., 2018; Sharma et al., 2019). Compared with binary labels, multiple labels may offer more information for false news detection. In the following, we explore whether such a multiclass formulation contributes to our proposed framework.

In our scenario, we were unable to directly extend the proposed CAND framework to utilize multiclass labels because the logistic-normal distribution cannot be naturally generalized to its multivariate form. Its natural generalization is a softmax-normal distribution, which is not mathematically well-defined. Hence, we developed two alternative models that tap into the multiclass information. The first model, the categorical aggregation model (CAM), takes advantage of multiclass information but not the logistic-normal assumption. The second model,

extended CAM (E-CAM), only partially exploits multiclass information in order to take advantage of the logistic-normal assumption. Experimental evaluation shows that E-CAM performs comparably to CAND, and they both outperform CAM. The results illustrate the importance of the logistic-normal assumption in our task. In addition, although the literature has shown the benefit of the multiclass formulation (Kochkina et al., 2018; Ma et al., 2018), the performance gain from treating debunking response detection as a multiclass classification problem in our scenario is quite marginal. We explore the potential reasons behind this and present them in Section D5 of the online supplement. The full details regarding the models and the experimental results are also presented therein.

4.3.9 **Dataset Bias**

Despite the proliferation of AI applications in the past decade, there has been an alarming rise in reports of fairness issues because of the dataset bias (Mehrabi et al., 2019). For example, annotator bias can lead to racial bias in hate speech detection models. The literature has summarized more than 20 potential sources of dataset bias including selection bias (i.e., sampling bias), annotator bias, negative set bias, measurement bias, and many others (Mehrabi et al., 2019; Tommasi et al., 2017; Torralba & Efros, 2011). These biases can lead to unfairness in different downstream learning tasks, e.g., collecting annotated data for the task of intention mining (Zhang et al., 2021a; Zhang et al., 2021b). In this research, we introduce how to mitigate two of the most relevant biases, i.e., annotator bias and selection bias. The details are presented in Section D6 of the online supplement.

5. **CONCLUSION**

Recently, false news has plagued social media and the academic community has devoted much effort (Valecha et al., 2020, 2021). Unlike the existing literature where crowd intelligence is

often treated as a feature of black-box methods, we propose to incorporate more interpretable structures in our Bayesian approach. In comparison with the existing literature, our approach has several advantages. First, it achieves better performance. Experimental results show that our approach performs better than end-to-end black-box models, especially when the data is highly unbalanced. Second, our approach is more interpretable and generates more technical insights. To name a few, combining human intelligence not only improves performance but also enables early detection; we also show that combining human intelligence is actually robust to several types of intentional manipulation. Finally, given the flexibility of our Bayesian model, we can easily extend it to incorporate more data sources. When interpretable structures exist, we combine them into the Bayesian model; if not, we have the necessary black-box methods in Task T1 to handle them.

Our research is of both theoretical and practical value. Theoretically, we show that our approach, which has black-box methods to handle unstructured data and uses interpretable structures to combine multiple judgments, performs better than the end-to-end benchmark methods. As mentioned above, our approach has several advantages and generates many valuable insights. Furthermore, understanding the value of scalable crowd judgments (e.g., responses and reports) is of great interest to false news detection research. Although responses are used in some literature to detect false news, the value of such scalable crowd judgments is still unclear. Our interpretable framework illustrates the complementary value of human and machine intelligence in the task of false news detection. Such a conclusion could also contribute to the broader literature on hybrid human-machine intelligence and other crowd intelligence applications, such as prediction markets (Chen et al., 2017). Finally, we propose modeling the credibility of both humans and machines in a hybrid human-machine system. Compared with the

relevant existing literature involving only humans or machines, the proposed hybrid setting is more complicated. We take a probabilistic perspective and carefully craft a Bayesian scheme for result aggregation. In addition, when modeling machine credibility, we tentatively use the logistic-normal assumption rather than the widely used two-coin assumption (Bragg & Weld, 2013; Raykar et al., 2010; Wei et al., 2017). The superior performance empirically verifies this assumption's efficacy. Our proposed techniques to model credibility have significant implications for future research where potentially unreliable humans and machines are involved.

Our research has many practical implications and actionable insights for relevant stakeholders. For social media platforms, the proposed CAND framework serves as a feasible and effective approach for false news detection on social media platforms. As suggested by the experimental results, the proposed framework combines crowd judgments and thus significantly improves the detection of false news in comparison with the benchmark methods. In practical use, this improvement will immediately translate into large differences in cost, as it usually costs a social media platform millions of dollars to curb the spread of false news.[8] Given the unique value of crowd judgments, a platform could encourage its users to actively contribute their intelligence even though their judgments may be unreliable. In addition, social media platforms that do not support user flagging may want to consider designing such a function. For social media platform users, the proposed research provides valuable insights into how they can personally help curb the false news epidemic—namely, by responding to or reporting social media posts when they doubt the veracity of their content.

We conclude our paper by presenting its limitations and future directions. First, in the Weibo dataset, we treat posts that are not reported as true news (Chen et al., 2018; Guo et al., 2018; Ma

[8] https://www.bbc.com/news/technology-40287399

et al., 2016). It is possible that some of these "true news" items are actually undetected false news. To mitigate this issue, we manually checked a small set of random samples and found that most of these "true news" items are true. Hence, we believe our current strategy of obtaining a set of true news had little impact on the validity of our evaluation. Second, Weibo attaches warning tags to fact-checked false news (see Figure 1). This creates the "implied-truth" effect, meaning that users tend to regard other untagged posts as more objective and verified (Pennycook et al., 2020), leading to fewer people debunking or reporting those untagged posts. In the future, we could further improve the usefulness of crowd intelligence by working to mitigate this effect. Third, we plan to explore assigning different weights to debunking behaviors from different users, as users with specific patterns of debunking (e.g., verified users) may be more reliable. Lastly, as a proof of concept, we represent each piece of news by its content and context features and propose using two types of human intelligence (i.e., responses and reports). Given the flexibility of the proposed Bayesian CLNAM model, it would be straightforward to extend it by incorporating more news information.

Acknowledgements

The authors thank the senior editor, the associate editor, and the anonymous reviewers for their constructive comments. Zhu Zhang and Mingyue Zhang are the corresponding authors of the paper. Most of the work was done when Xuan Wei was a Ph.D. student and Zhu Zhang was a postdoc at the University of Arizona. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0103405, the National Natural Science Foundation of China under Grants 71621002, 71802024, 62071467, 72192822, and 71974187, the Shanghai Chenguang Program under Grant 21CGA13, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27030100, and the Innovative Research Team of Shanghai International Studies University under Grant 2020114044.

References

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691-706.

Atchison, J., & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2), 261-272.

Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management Science*, 59(1), 226-244.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859-877.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14(1), 75-90.

Brabham, D. C. (2013). *Crowdsourcing*, MIT Press.

Bragg, J., & Weld, D. S. (2013). Crowdsourcing multi-label classification for taxonomy creation. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*.

Castillo, C., El-Haddad, M., Pfeffer, J., & Stempeck, M. (2014). Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 211-223).

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 40-52).

Chen, W., Li, X., & Zeng, D. D. (2017). Modeling fixed odds betting for future event prediction. *Management Information Systems Quarterly*, 41(2), 645-665.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240).

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error‐rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 20-28.

Demartini, G., Difallah, D. E., Gadiraju, U., & Catasta, M. (2017). An introduction to hybrid human-machine information systems. *Foundations and Trends® in Web Science*, 7(1), 1-87.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Available at http://arxiv.org/abs/1810.04805).

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, 54(4), 86-96.

Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 943-951).

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 729-736).

He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 9, pp. 1263-1284.

He, J., Fang, X., Liu, H., & Li, X. (2019). Mobile app recommendation: An involvement-enhanced approach. *Management Information Systems Quarterly*, 43(3), 827-849.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings 13th AAAI Conference on Artificial Intelligence*.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3), 598-608.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (pp. 4070-4073).

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1546-1557).

Kim, H.-C., & Ghahramani, Z. (2012). Bayesian classifier combination. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics* (pp. 619-627).

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining* (pp. 324-332).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746-1751).

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3402-3413).

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition, *Computers in Human Behavior*, 28(2), 434-443.

Lee, H. C. B., Ba, S., Li, X., & Stallaert, J. (2018). Salience Bias in Crowdsourcing Contests. *Information Systems Research*, 29(2), 401-418.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 138-143).

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52(6), 362-375.

Liu, Y., & Wu, Y.-F. B. (2020). FNED: A deep network for fake news early detection on social media. *ACM Transactions on Information Systems*, 38(3), Article 25.

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2014). The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research*, 25(4), 669-689.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (pp. 3818-3824).

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1751-1754).

Ma, J., Gao, W., & Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 708-717).

Ma, J., Gao, W., & Wong, K.-F. (2018). Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the Web Conference 2018* (pp. 585-593).

McCarthy, Niall (2018). Where Exposure To Fake News Is Highest [Infographic]. *Forbes.* https://www.forbes.com/sites/ niallmccarthy/2018/06/14/where-exposure-to-fake-news-is-highest-infographic/?sh=198944204a4d

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276-282.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. Available at http://arxiv.org/abs/1908.09635.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 3111-3119.

Mohandes, M., Deriche, M., & Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6, 19626-19639.

Moreno, P. G., Artés-Rodríguez, A., Teh, Y. W., & Perez-Cruz, F. (2015). Bayesian nonparametric crowdsourcing. *Journal of Machine Learning Research*, 16, 1607-1627.

Newman, N. (2018). Overview and key findings of the 2018 report. Reuters. https://www.digitalnewsreport.org/survey/2018/ overview-key-findings-2018/

Nguyen, T. N., Li, C., & Niederée, C. (2017). On early-stage debunking rumors on Twitter: Leveraging the wisdom of weak learners. In *Proceedings of the International Conference on Social Informatics* (pp. 141-158).

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *Management Information Systems Quarterly*, 37(2), 407-426.

Oh, O., Gupta, P., Agrawal, M., & Rao, H. R. (2018). ICT Mediated rumor beliefs and resulting user actions during a community crisis. *Government Information Quarterly*, 35(2), 243-258.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825-2830.

Peng, H., Cambria, E., & Hussain, A. (2017). A review of sentiment analysis research in Chinese language. *Cognitive Computation*, 9(4), 423-435.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543).

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944-4957.

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2), 185-200.

Pennycook, G., & Rand, D. G. (2019a). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50.

Pennycook, G., & Rand, D. G. (2019b). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521-2526.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 231-240).

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 3834-3840).

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the 1st Instructional Conference on Machine Learning*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11, 1297-1322.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A Hybrid Deep Model for Fake News Detection. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, pp. 797-806.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3), Article 21.

Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2020). Hierarchical Propagation networks for fake news detection: Investigation and Exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 626-637).

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

Shu, K., Wang, S., & Liu, H. (2018). Understanding user profiles on social media for fake news detection. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval* (pp. 430-435).

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining* (pp. 312-320).

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

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, 7(3), 321-326.

Surowiecki, J. (2005). *The wisdom of crowds*, Anchor.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *Proceedings of the 2nd Workshop on Data Science for Social Good*.

The Lancet Infectious Diseases. (2020). The COVID-19 Infodemic. *The Lancet. Infectious Diseases* (20:8), Article P875.

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2017). A deeper look at dataset bias. In S. B. Kang (Ed.), *Domain adaptation in computer vision applications* (pp. 37-55). Springer.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1521-1528).

Tran, T., Valecha, R., Rad, P., & Rao, H. R. (2020a). An investigation of misinformation harms related to social media during two humanitarian crises. *Information Systems Frontiers*, 23, 931-939.

Tran, T., Valecha, R., Rad, P., & Rao, H. R. (2020b). Misinformation harms: A tale of two humanitarian crises. *IEEE Transactions on Professional Communication,* 63(4), 386-399.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Companion Proceedings of the Web Conference 2018* (pp. 517-524).

Tulyakov, S., Jaeger, S., Govindaraju, V., & Doermann, D. (2008). Review of classifier combination methods. In J. Kacprzyk (Ed.), *Machine Learning in Document Analysis and Recognition* (pp. 361-386). Springer

Valecha, R., Srinivasan, S. K., Volety, T., Kwon, K. H., Agrawal, M., & Rao, H. R. (2021). Fake news sharing: An investigation of threat and coping cues in the context of the Zika virus. *Digital Threats: Research and Practice*, 2(2), Article 16.

Valecha, R., Volety, T., Rao, H. R., & Kwon, K. H. (2020). Misinformation sharing on Twitter during Zika: An investigation of the effect of threat and distance. *IEEE Internet Computing*, 25(1), 31-39.

Vaughan, J. W. (2018). Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research,* 18(193), 1-46.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. NOW

Wang, J., Ipeirotis, P. G., & Provost, F. (2017). Cost-effective quality assurance in crowd labeling. *Information Systems Research*, 28(1), 137-158.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 422-426).

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 849-857).

Wei, X., Zeng, D. D., & Yin, J. (2017). Multi-label annotation aggregation in crowdsourcing. Available at http://arxiv. org/abs/1706.06120.

Wei, X., Zhang, M., & Zeng, D. D. (2020). Learning from crowd labeling with semi-crowdsourced deep generative models. In *Proceedings of the CCF Conference on Computer Supported Cooperative Work and Social Computing* (pp. 101-114).

Welinder, P., Branson, S., Perona, P., & Belongie, S. J. (2010). The multidimensional wisdom of

crowds. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems* (pp. 2424-2432).

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence.*

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480-1489).

Zhang, M., Wei, X., Guo, X., Chen, G., & Wei, Q. (2019). Identifying complements and substitutes of products: A neural network framework based on product embedding. *ACM Transactions on Knowledge Discovery from Data*, 13(3), Article 34.

Zhang, Q., Zhang, S., Dong, J., Xiong, J., & Cheng, X. (2015). Automatic detection of rumor on social network. In J. Li, H. Ji, D. Zhao, & Y. Feng (Eds.), *Natural Language Processing and Chinese Computing* (pp. 113-122). Springer.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: characterization, detection, and discussion. *Information Processing & Management* (57:2), Article 102025.

Zhang, Y., Chen, X., Zhou, D., & Jordan, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1), 3537-3580.

Zhang, Z., Wei, X., Zheng, X., Li, Q., & Zeng, D (2021a). Detecting product adoption intentions via multiview deep learning. *INFORMS Journal on Computing* 34(1), 541-556.

Zhang, Z., Wei, X., Zheng, X., & Zeng, D. D. (2021b). Predicting product adoption intentions: An integrated behavioral model-inspired multiview learning approach. *Information & Management*, 58(7), Article 103484.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), Article 109.

Appendices: Combining Crowd and Machine Intelligence to Detect False News on Social Media

Appendix A: Additional Details about the Experimental Design

A1. Additional Details of Data Preprocessing

We first present more details about the data preprocessing of the Weibo dataset. When collecting data, we ignored the deleted Weibo posts—actually, for more than 25% of reported records, the original Weibo posts were already deleted by the authors when we collected the data. For the event detection algorithm, i.e., single-pass clustering method (Wang et al. 2018), we first represent each Weibo post using the tf-idf weighting scheme (Ramos 2003). This algorithm sequentially processes the posts, one at a time, and grows clusters incrementally. A new post is absorbed by the most similar existing cluster if the similarity score (i.e., cosine similarity) goes beyond a threshold of 0.3; otherwise, the post will be the seed of a new cluster. Note that the threshold value will not influence the evaluation procedure much because both the proposed algorithm and the benchmark methods will perform better when a higher threshold is adopted.

For the Twitter dataset, when combining the tweets of the three reference datasets, we only kept those with labels "true" and "false" and discarded those with labels "non-rumor" and "unverified." The dataset preprocessing for this dataset is similar to that of Weibo, as shown in Figure A1. For the annotation of debunking responses, two authors independently annotated the responses and a consensus was reached for inconsistent annotations after discussion. The Cohen's kappa score is 0.784.

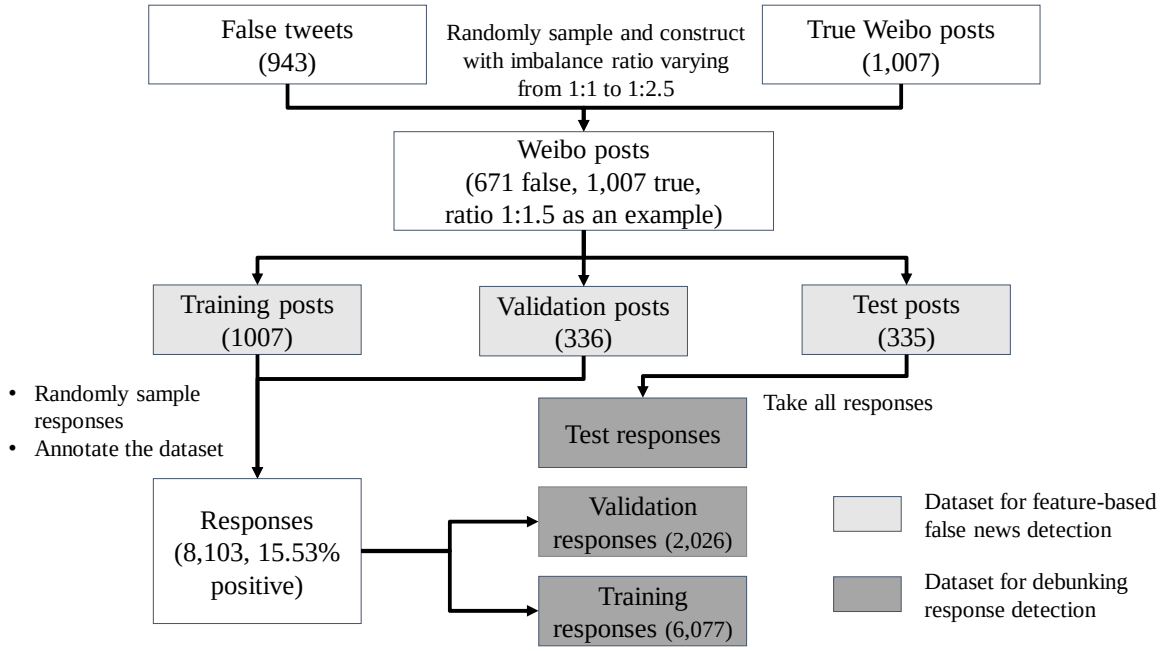


Figure A1. Data Preprocessing for Twitter Dataset

A2. Details of the Benchmark Methods

For SVM, each post text is represented with the tf-idf scheme (Ramos 2003), and all features are directly fed as the input to an SVM classifier. For CNN, the implementation follows a classical architecture (Kim 2014). For LSTM, we have two variants: one-layer LSTM (Ma et al. 2016) and bidirectional LSTM (Wang 2017). These two models are denoted as LSTM and Bi-LSTM respectively. For these deep learning-based benchmark methods, we first learn a hidden representation of the post text. The representation is concatenated with other news features and then fed into a fully connected layer, finally followed by a softmax layer. For BERT, the pre-trained models ("BERT-Base, Chinese" and "BERT-Base, Uncased" from GitHub[9]) are followed by two fully connected layers and a softmax layer, and then fine-tuned for classifying the post.

The first end-to-end benchmark is a concatenation-based deep learning approach (denoted as

[9] https://github.com/google-research/bert

Concat). Directly concatenating various types of inputs after extracting hidden representations is widely used in the false news detection literature (Qian et al. 2018; Wang 2017). To learn hidden representations from text, we build two CNN modules for news text and responses respectively—one is used for news text and the other one is shared by all responses. These representations are later concatenated with other news features and the number of reports. Finally, all these features are fed into a fully connected layer and a softmax layer in sequence to predict the veracity of news. The second end-to-end benchmark, HSA, uses a more advanced deep learning structure, i.e., hierarchical attention network (Guo et al. 2018; Yang et al. 2016). A hierarchical Bi-LSTM model is built for representation learning and the news features are incorporated into the network via the attention mechanism. As the original structure does not consider the number of reports, we adapt it by concatenating the hidden representation with the number of reports before feeding the hidden representation to the fully connected layer.

For MV, the vote from each classifier in Task T1 is positive if the predicted probability is greater than 0.5; the vote from responses is positive when the debunking response rate is higher than the average, and invalid if no response exists; the vote from the last source is positive when the Weibo post receives at least one report. Given the votes, the Majority Voting aggregation model will predict a piece of news to be false if more than half of the received votes are positive.

The BAM model depicts the credibility of classifiers in the information extraction stage based on the widely used two-coin assumption (Bragg and Weld 2013; Raykar et al. 2010; Wei et al. 2017), where the predictions conditioned on the ground truth are binary and follow Bernoulli distributions. Specifically, each classifier (or worker in the crowdsourced answer aggregation literature) predicts the label correctly with a certain probability when the ground truth is 1 and with another probability when the ground truth is 0. As the BAM model is obtained

by replacing the logistic-normal assumption in the CLNAM model with the two-coin assumption, we only introduce the parts that differ. The first two types of information (i.e. $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$) are transformed into their binary prediction results (i.e., 0 or 1). With binary values, the generative processes change accordingly:

$$\arg\max(Y_{n,m}^{(1)})|z_n, \mu_m^{(1)}, \bar{\mu}_m^{(1)} \sim \text{Bernoulli}(z_n\mu_m^{(1)} + (1-z_n)(1-\bar{\mu}_m^{(1)}))$$

$$\arg\max(Y_{n,k}^{(2)})|d_{n,k}, \mu^{(2)}, \bar{\mu}^{(2)} \sim \text{Bernoulli}(d_{n,k}\mu^{(2)} + (1-d_{n,k})(1-\bar{\mu}^{(2)})), \quad \text{(A1)}$$

where arg max(x) equals 1 when x > 0.5 and 0 otherwise. Then, we impose conjugate priors over the parameters that depicts the classifiers' credibility.

$$\mu_m^{(1)}|\alpha_0^{(1)}, \beta_0^{(1)} \sim \text{Beta}\left(\alpha_0^{(1)}, \beta_0^{(1)}\right), \quad \bar{\mu}_m^{(1)}|\bar{\alpha}_0^{(1)}, \bar{\beta}_0^{(1)} \sim \text{Beta}(\bar{\alpha}_0^{(1)}, \bar{\beta}_0^{(1)})$$

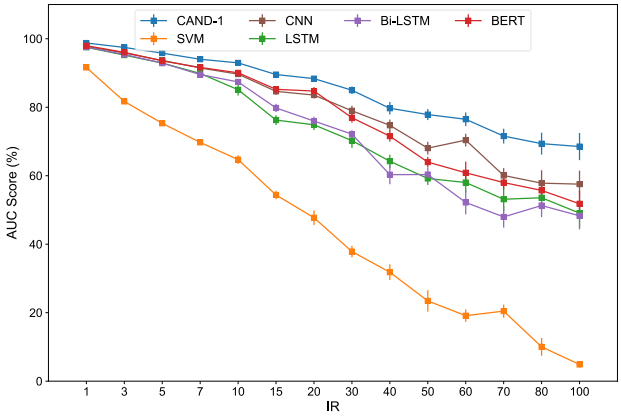
$$\mu^{(2)}|\alpha_0^{(2)}, \beta_0^{(2)} \sim \text{Beta}\left(\alpha_0^{(2)}, \beta_0^{(2)}\right), \quad \bar{\mu}^{(2)}|\bar{\alpha}_0^{(2)}, \bar{\beta}_0^{(2)} \sim \text{Beta}(\bar{\alpha}_0^{(2)}, \bar{\beta}_0^{(2)}), \quad \text{(A2)}$$

Last, the coordinate ascent mean-field variational inference algorithm and prior learning from data can be developed in a similar manner as for the CLNAM model, so we omit the details here.

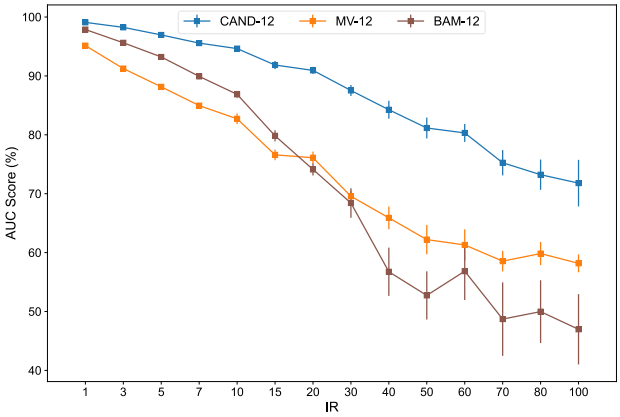
Appendix B: Additional Experimental Results

B1. CAND vs. the Benchmark Methods in the Weibo Dataset

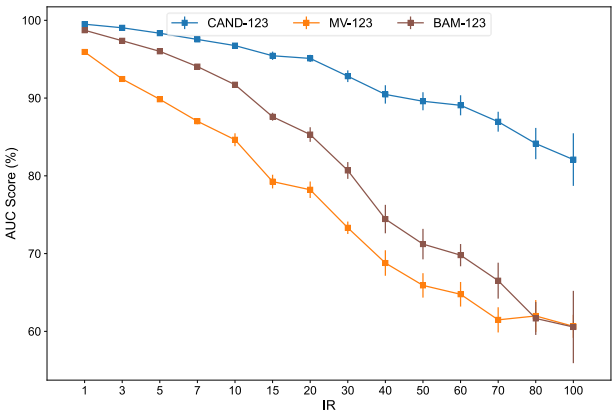
To further explore the robustness of our method and the benchmark methods when the data is unbalanced, we compared their performance by varying the IR. Figure B1 shows the performance of CAND using AUC as the metric in comparison with end-to-end benchmark methods without crowd intelligence and aggregation benchmark methods.



(a) CAND-1 vs. End-to-End Benchmark Methods without Crowd Intelligence



(b) CAND-12 vs. Aggregation Benchmark Methods



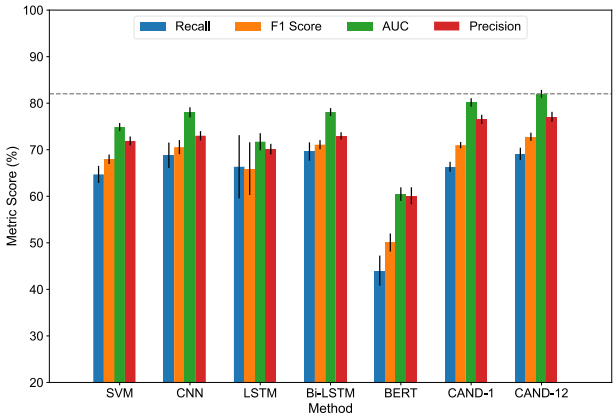
(c) CAND-123 vs. Aggregation Benchmark Methods

Figure B1. CAND vs. End-to-End Benchmark Methods without Crowd Intelligence and Aggregation Benchmark Methods under Different IRs (Weibo, AUC Score as Metric)

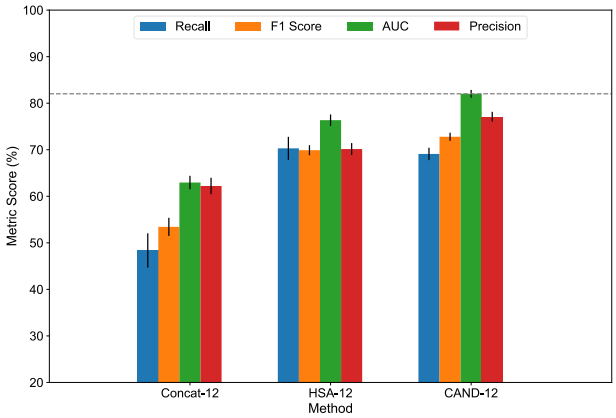
B2. CAND vs. the Benchmark Methods in the Twitter Dataset

Similar to the analysis in the Weibo dataset, we conducted the same set of experiments in the Twitter dataset, including 1) comparing CAND with the benchmark methods across all metrics when IR=1.5; 2) using AUC as the metric and comparing CAND with the benchmark methods by varying the IR from 1:1 to 1:2.5. The graphs are shown in Figure B2 and Figure B3 respectively. Note that CAND-123 is not included because human reports are unavailable in the

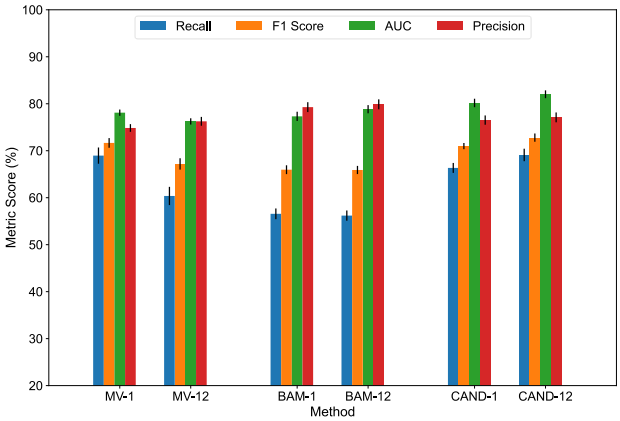
Twitter dataset. By examining the figures, we draw the same conclusions as in the Weibo dataset.



(a) CAND vs. End-to-End Benchmark Methods without Crowd Intelligence (IR=1:1.5)

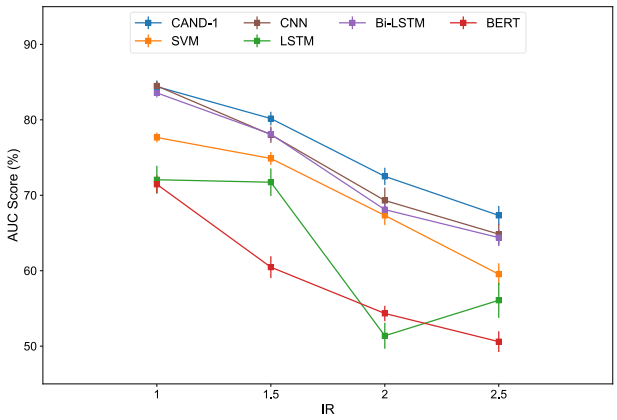


(b) CAND vs. End-to-End Benchmark Methods with Crowd Intelligence (IR=1: 1.5)

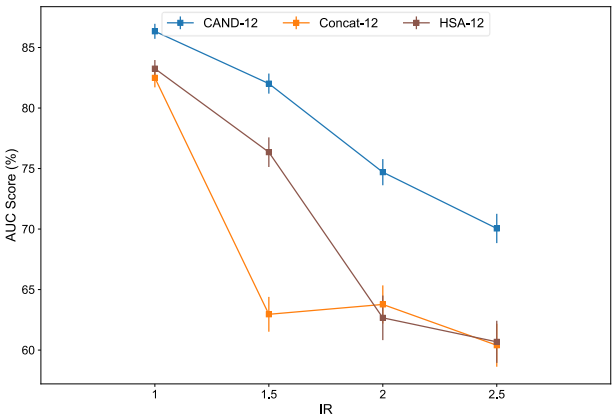


(c) CAND vs. Aggregation Benchmark Methods (IR=1: 1.5)

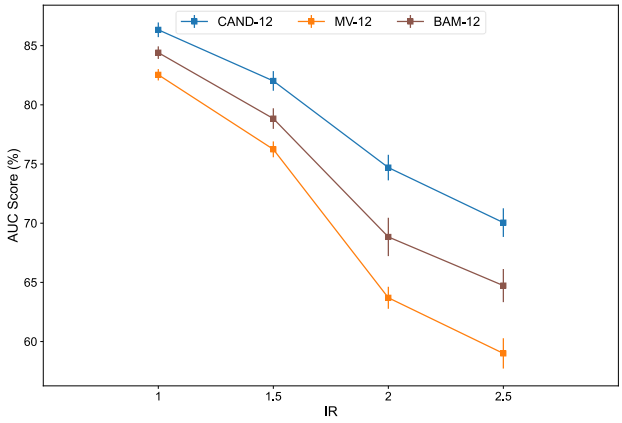
Figure B2. Performance of CAND and the Benchmark Methods (Twitter, Ref. lines: AUC of CAND-12)



(a) CAND-1 vs. End-to-End Benchmark Methods without Crowd Intelligence



(b) CAND-12 vs. End-to-End Benchmark Methods with Crowd Intelligence



(c) CAND-12 vs. Aggregation Benchmark Methods

Figure B3. CAND vs Benchmark Methods under Different IRs (Twitter, AUC Score as Metric)