Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities

Gianluca Demartini[1], Stefano Mizzaro[2], Damiano Spina[3]
[1]The University of Queensland, Brisbane, Australia, g.demartini@uq.edu.au
[2]University of Udine, Udine, Italy, mizzaro@uniud.it
[3]RMIT University, Melbourne, Australia, damiano.spina@rmit.edu.au

Abstract

The rise of online misinformation is posing a threat to the functioning of democratic processes. The ability to algorithmically spread false information through online social networks together with the data-driven ability to profile and micro-target individual users has made it possible to create customized false content that has the potential to influence decision making processes. Fortunately, similar data-driven and algorithmic methods can also be used to detect misinformation and to control its spread. Automatically estimating the reliability and trustworthiness of information is, however, a complex problem and it is today addressed by heavily relying on human experts known as fact-checkers. In this paper, we present the challenges and opportunities of combining automatic and manual fact-checking approaches to combat the spread on online misinformation also highlighting open research questions that the data engineering community should address.

1 Introduction

As the amount of online information that is generated every day in news, social media, and the Web increases exponentially, so does the harm that false, inaccurate, or incomplete information may cause to society. Experts in fact-checking organizations are getting overwhelmed by the amount of content that requires investigation,[1] and the sophistication of bots used to generate and deliberately spread fake news and false information (i.e., disinformation) is only making the tasks carried out by experts—i.e., identifying check-worthy claims and investigating the veracity of those statements—less manageable.

The aim of this paper is to discuss the main challenges and opportunities of a hybrid approach where Artificial Intelligence (AI) tools and humans—including both experts and non-experts recruited on crowdsourcing platforms—work together to combat the spread of online misinformation.

The remainder of this paper is organized as follows. Section 2 presents an overview of human-in-the-loop AI methods. Section 3 introduces the main challenges in identifying misinformation online. Section 4 summarizes recent work on machine learning methods applied to automatic truthfulness classification and check-worthiness. Section 5 describes recent advances on crowdsourcing one of the key activities in the fact-checking process, i.e.,

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[1]https://www.theverge.com/2020/3/24/21192206/snopes-coronavirus-covid-19-misinformation-fact-checking-staff

judging the truthfulness or veracity of a given statement. Section 6 proposes a hybrid human-AI framework to fact-check information at scale. We conclude in Section 7 by summarizing the main take-away messages.

2 Human-in-the-loop AI

Human-in-the-loop AI (HAI) systems aim at leveraging the ability of AI to scale the processing to very large amounts of data while relying on human intelligence to perform very complex tasks— for example, natural language understanding—or to incorporate fairness and/or explainability properties into the system. Example of successful HAI methods include [8, 9, 15, 30]. Active learning methods [31] are another example of HAI where labels are collected from humans, fed back to a supervised learning model, and used to decide which data items humans should label next [32]. Related to this is the idea of interactive machine learning [2] where labels are automatically obtained from user interaction behaviors [20]. While being more powerful than pure machine-based AI methods, HAI systems need to deal with additional challenges to perform effectively and to produce valid results. One such challenge is the possible *noise* in the labels provided by humans. Depending on which human participants are providing labels for the AI component to learn from, the level of data quality may vary. For example, making use of crowdsourcing to collect human labels from people online either using paid micro-task platforms like Amazon Mechanical Turk or by means of alternative incentives like, e.g., 'games with a purpose' [37] is in general different from relying on a few experts.

There is often a trade-off between the cost and the quality of the collected labels. On the one hand, it may be possible to collect few high-quality curated labels that have been generated by domain experts, while, on the other hand, it may be possible to collect very large amounts of human-generated labels that might be not 100% accurate. Since the number of available experts is usually limited, to obtain both high volume and quality labels, the development of effective quality control mechanisms for crowdsourcing is needed.

Another challenge that comes with HAI systems is the *bias* that contributing humans may create and/or amplify in the annotated data and, consequently, in the models learned from this labelled data [16, 25]. Depending on the labelling task, bias and stereotypes of contributing individuals may be reflected into the generated labels. For example, an image labelling task that requires to identify the profession of people by looking at a picture, may lead to a female individual depicted in medical attire to be labelled as 'nurse' rather than as 'doctor'. For such type of data collection exercises, it becomes important to measure and, if necessary, control the bias in the collected data so that the bias in the AI models trained with such data is managed and controlled as well, if not limited or avoided altogether. Possible ways to control such bias include working on human annotator selection strategies by, for example, including pre-filtering tasks to profile annotators and to then select a balanced set of human annotators to generate labels for an AI to learn from.

Once manually labelled data has been collected, trained AI models may reflect existing bias in the data. An example of such a problem is that of 'unknown unknowns' (UU) [3], that is, data points for which a supervised model makes a high-confidence classification decision, which is however wrong. This means that the model is not aware of making mistakes. UUs are often difficult to identify because of the high-confidence of the model in its classification decision and may create critical issues in AI.[2] The problem of UU is usually caused by having a part of the feature space being under-represented in the training data (e.g., training data skewed towards white male images may result into AI models that are not performing well on images of people from other ethnicities and of other genders). Thus, such AI models are biased because of the unbalanced training data they have been trained on. Possible ways to control for such bias include making use of appropriate data sampling strategies to ensure that training datasets are well balanced and cover well the feature space also for features that may not have been originally identified or used.

[2] A classic example of this is the Google gorilla mistake, see `https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/`.

When incorporating humans into an HAI system, they become the efficiency bottleneck. While purely machine-based AI systems can learn from very many data points and, once trained, perform decisions in real-time, making use of a human component makes the system less scalable and less efficient. For this reason, it becomes important to decide how to best employ these less efficient and limited human resources and, instead, how to best leverage the scalability of machine-based methods in order to get the best out of the two worlds. The problem becomes even more complex when considering different types of human contributors which come with varying quality, availability, and cost levels. We discuss this in more depth in Section 6.2.

Related to the previous problem, deciding what data points should be manually labelled by human annotators is another challenge. Given a usually very limited manual annotation budget, it becomes important to select the best data items to label in order to maximise their value with respect to the improvements of the trained AI models. Questions of this type are in particular relevant to systems relying on active learning strategies. Such improvements, however may relate not only to effectiveness, but also to other model properties like, for example, *fairness*. Another benefit of involving humans in HAI system is the ability to leverage their skills to improve the *interpretability* and *explainability* of AI models. Human contributors may be leveraged to, for example, add natural language explanations about *why* a certain supervised classifier decision has been made.

Thus, in order to design and develop an high-quality HAI system, researchers have to look at a multi-dimensional problem which includes aspects like efficiency, accuracy, interpretability, explainability, and fairness. Human and machine components of an HAI system can contribute and possibly threaten each of these dimensions. Based on these issues, the overarching question in HAI systems is about deciding *what should humans do* and *what should AIs do* in order to optimally leverage the capabilities of both methodologies. In the remainder of this paper we discuss these challenges and opportunities in the context of fighting online misinformation. We use this problem as a showcase of HAI methods and discuss the potential of such methodology when applied to this context.

3 The Problem of Online Misinformation

3.1 An Interdisciplinary Challenge

The spread of misinformation online is a threat to our safety online and risks to damage the democratic process. For instance, bots and trolls have been spreading disinformation across a number of significant scenarios, such as the election of US President Donald Trump in 2016 [5], the debate in the UK over Brexit [19], and, more recently, exaggerating the role of arson to undermine the link between bushfires in Australia and climate change.[3] The World Health Organization (WHO) has referred to the problem of large amount of misinformation spreading during the COVID-19 pandemic as an "infodemic"[4] [1]. Therefore, fact-checking information online is of great importance to avoid further costs to society.

Because of the importance, impact, and interdisciplinarity of the issue, a number of different research areas have focused on understanding and stopping misinformation spreading online. This includes research in political sciences [22], communication science [40], computational social science [7], up to computer science including the fields of human-computer interaction [33], database [21], and information retrieval [28]. While different research methodologies are being applied, the overarching goal is to understand how misinformation is spreading, why people trust it, and how to design and test systems and processes to stop it.

From a data engineering point of view, online misinformation poses some of the same common challenges observed in modern data management: i) *volume*: large amounts of data to be processed efficiently and in a scalable fashion; ii) *velocity*: processing data and making misinformation classification decisions in a timely

[3] https://theconversation.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556

[4] https://www.who.int/dg/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020

fashion also in conditions when data to be checked comes as a stream (e.g., Twitter propaganda bots generating and propagating misinformation in social networks; iii) *variety*: misinformation comes in multiple formats, from textual statements in news articles, to images used in social media advertising, to deep-fake videos artificially generated by AI models; iv) *veracity*: the core question of truthfulness classification often translates in deciding which data source can be trusted and which not. Thus, the data engineering community not being new to dealing with such challenges, can surely provide solutions, systems, and tools able to support the fight to online misinformation. We however still believe that this is an interdisciplinary challenge, and in the remainder of this paper we present a framework that goes beyond data engineering by including humans in the loop and by considering human factors as well.

3.2 Misinformation Tasks

From the existing scientific literature about misinformation, we can see that there are a number of more specific tasks that need to be addressed to achieve the overarching goal of fighting online misinformation. The first task that comes to mind is *truthfulness classification*, that is, given a statement decide its truth level, in a scale from completely true to completely false. Fully automated approaches [23] as well as crowdsourcing-based approaches [28] have been proposed to address this task. However, other tasks related to online misinformation exist. For example, it is also important to decide about the *check-worthiness* of online content. As there are way too many statements and claims that could possibly be fact-checked, before expert fact checking can take place, a pre-processing filtering step needs to be completed to identify which statements should be going through a complete fact-checking process, out of a large collection of potential candidates. Criteria to be considered for such a selection process include: the potential harm that a certain statement being false could create, the reach of that statement, the importance and relevance of the topic addressed by the statement, etc. Automated methods for check-worthiness have been proposed in the literature [11], but are far from being effective enough to be deployed in practice and replace expert fact-checkers on this task.[5] Another task related to misinformation is *source identification*. Being able to detect the origin of online information can provide additional evidence to information consumers about its level of trustworthiness. More than just either manual or automatic approaches to address these tasks, an additional way is to combine them together in order to optimize processes and leverage the best properties of each method.

4 Machine Learning for Fighting Online Misinformation

For each of the misinformation tasks described in the previous section, there have been attempts to develop machine learning methods to tackle them. In this section we provide a summary of such research. For the problem of truthfulness classification, benchmarks on which to compare the effectiveness of different approaches have been developed. A popular benchmark for truthfulness classification is the LIAR dataset [39] that makes use of expert fact-checked statements from the PolitiFact website. More than 12K expert-labeled statements are used as ground truth to train and evaluate automatic classification systems effectiveness, so that system quality can be compared. Even larger than that is the FEVER dataset [34] that contains 180K statements obtained by altering sentences extracted from Wikipedia. Other earlier and smaller truthfulness classification benchmark datasets include [36, 14].

A lot of effort has been made within the AI research community not only to obtain accurate classification decision, but also to provide explainable results. Supervised methods for this task have looked at which features are the most indicative of truthfulness [27]. Recent approaches have designed neural networks that aim at combining evidence fragments together to inform the truthfulness classification decision [43]. Such evidence

[5] https://www.niemanlab.org/2020/07/a-lesson-in-automated-journalism-bring-back-the-humans/

can then be used to explain the automatic classification decisions. Other studies looking at the explainability dimension of this problem have observed that different features may be indicators for different types of fake news and can be used to cover different areas of the feature space [26]. Adversarial neural networks have shown to improve the effectiveness in identifying distinctive features for truthfulness classification [42].

Methods to automatically decide on check-worthiness [11] have looked at how to assign a score to a sentence and to predict the need for it to be checked by experts using supervised methods and training data. While some methods make use of contextual information, that is, of the surrounding text, to decide on the check-worthiness of a sentence [13], the most effective ones consider each sentence in isolation and use domain specific word embeddings within an LSTM network [17].

Metadata about information sources presented to social media users have an effect on the perceived truthfulness of the information [24]. Providing news source and contextual metadata may help users to make informed decisions [12]. Related to this, the New York Times R&D group has started a project to provide provenance metadata around news using blockchain technology to track the spread of news online and to provide contextual information to news readers.[6]

5 Crowdsourcing Truthfulness

More than just machine learning-based methods, crowdsourcing can be used as a way to label data at scale. In the context of misinformation, crowdsourcing is a methodology that can provide, for example, truthfulness classification labels for statements to be fact-checked. While experts may not be directly replaced by crowd workers (see work by Bailey et al. [4]), by deploying appropriate quality control mechanisms, crowdsourcing can provide reliable labels [10]. In a recent research on crowdsourcing truthfulness classification decisions we have looked at how to scale the collections of manual labels and at the impact of the annotators' background on the quality of the collected labels specifically looking for the impact of the annotator political bias with respect to the assessed statement and of the scale used to express the truthfulness judgment [28]. In another follow-up study, we have then looked at the impact of the *timeliness* of the assessed statements on the quality of the collected truthfulness labels. Results show that even more recent statements can still reliably be fact-checked by the crowd [29]. More in detail, we looked at how the crowd assessed the truthfulness of COVID-19 true and false statements during the pandemic, finding an agreement with expert judgments comparable to that in the previous study.

Another common challenge for expert fact-checkers, due to the limited available resources, is deciding which items should be fact-checked among very many candidates. More than just leveraging crowdsourcing to decide on truthfulness, the crowd may also be able to support expert fact-checkers in performing the task of deciding about the 'check-worthiness' of content, that is, asking the crowd to decide whether or not a given piece of content would benefit from being fact-checked by experts. Several factors affect the decision of selecting a statement to undergo a fact-checking process. The crowd may be involved in validating these factors which include, for example, the level of public interest of the assessed content, the possible impact of such content not being true, and the timeliness of the content. In this way, it would be possible to manually filter more content for fact-checking (the effectiveness of fully automated check-worthiness approach is still very low [11]) thus allowing expert fact-checkers to focus on actual fact-checking rather than on filtering and deciding what needs to be fact-checked.

[6] https://open.nytimes.com/introducing-the-news-provenance-project-723dbaf07c44

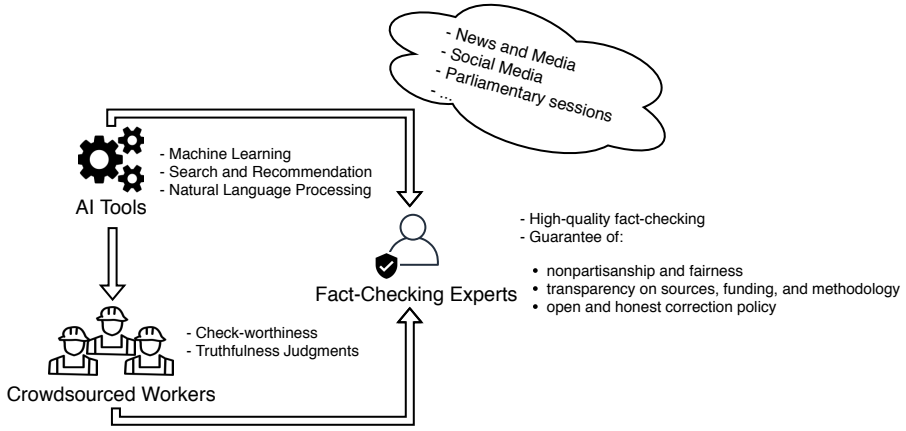

Figure 1: Human-in-the-loop AI framework for fighting online misinformation.

6 A Hybrid Human-AI Framework for Fighting Online Misinformation

6.1 Combining Experts, AI, and Crowd

Given the limitations of both automated and human-based methods for fact checking, we rather envision a hybrid human-AI approach to fight online misinformation. Such an approach has the benefit of leveraging the positive aspects of each of the different approaches, that is, the scalability of AI to efficiently process very large amounts of data, the ability of expert fact-checkers to correctly identify the truthfulness level of verified statements in a transparent and fair way, and the ability of crowdsourcing to manually process significantly large datasets. We are starting to see the appearance of hybrid approaches for fact-checking, like, for example, the work presented by Karagiannis et al. [21]. The proposed system is an example of how to efficiently use human fact checking resources by having a machine-based system supporting them to find the facts that need to be manually checked out of a large database of possible candidates.

The combination of these methods may not only result in more efficient and effective fact-checking processes, but also lead to improved trust on the outcomes over purely AI-based methods and may also leverage the embedded human dimension to increase the level of transparency of the truthfulness labels attached to news (i.e., explaining *why* a certain piece of news has been labelled as fake, like fact-checkers do already, but something that AI-based methods still struggle to provide). Such an approach may also lead to resource optimization, where the more expensive and accurate expert fact checkers may be intelligently deployed only on the few most important and challenging verification tasks, while the crowd and AI can work together to scale-up the execution of very large amounts of fact-checking tasks. We thus envision a waterfall model where different levels of cost/quality trade-offs can be applied at different stages by means of appropriate task allocation models.

6.2 The Framework

The existence of numerous challenges and constraints that need to be resolved concurrently leads us to the proposal of a solution that not only combines humans and machines, but that in doing so leverages different types and levels of engagement in the process of fighting misinformation. Our proposed framework consists of three main actors: fact-checking experts, AI methods, and crowdsourcing workers (see Figure 1).

Fact-checking experts are the protagonists of the framework and are the ones who make use of the other two components to optimize the efficiency of the fact-checking process and maintain high-quality standards. Also,

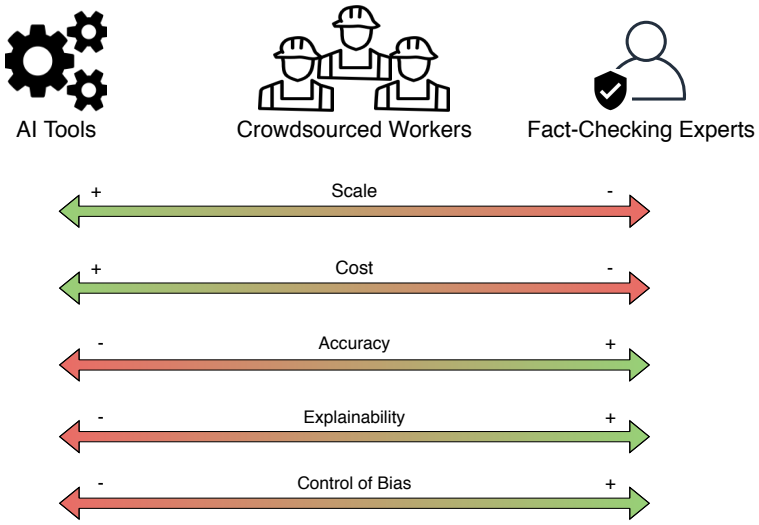

Figure 2: Trade-offs between the actors of the framework.

they are the only ones who can guarantee that this HAI system meets the three principles[7] of (i) non-partisanship and fairness, (ii) transparency on sources, funding, and methodology; and (iii) open and honest correction policy.

AI tools consist of automatic methods that fact-checkers can use to deal with the large amount of (mis)information produced through different channels such as news and media, parliamentary sessions, or social media [6]. Although AI tools are able to process data at scale, automatic predictions are typically not free from errors. For instance, machine learning methods used in systems such as ClaimBuster [18] or check-worthiness systems for the CheckThat! Lab at CLEF [11] are far from being 100% accurate. Moreover, it is not clear whether these tools would perform at the same level of accuracy in other scenarios, e.g., predicting check-worthiness of statements related to non-American politics. In summary, although state-of-the-art machine learning can compete—and even surpass—experts when data scale and costs are measured, as of today they are far from reaching human experts when considering the level of accuracy, explainability, and fairness.

Crowd workers somehow lie in between experts and AI on all the five above mentioned dimensions (scale, cost, accuracy, explainability, and control of bias) and can be deployed on-demand based on changing requirements and trade-offs. Figure 2 summarizes the strong and weak points of the actors involved in the proposed framework.

The proposed framework comes with several benefits:

- **Cost-quality trade-offs**: it comes with the ability to trade-off and optimize between required cost and quality of the label collection process where human experts (i.e., fact checkers) come with the highest quality and cost and AI comes with the lowest cost;
- **Load management**: it allows to deal with peaks of fact-checking tasks that may be otherwise impossible to deal with for expert fact-checkers working under constrained resource conditions. In such situations, they may be able to leverage the more scalable crowd and AI tools to deal with a sudden increase in annotation workload;
- **Trustworthiness**: it can serve as a way to make AI technology accepted in well-established traditional journalistic environments that would not see positively an 'AI taking over their job'.

In such an intertwined framework, the key question becomes *who should do what*. Given a workload of misinformation tasks, a deadline, and required constraints like a minimum level of quality and a maximum

[7] https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles

cost, the problem becomes to identify a task allocation solution that satisfies the constraints with maximum value. This can be addressed with a cascade model [35, 41, 38] with humans-in-the-loop, where AI tools, crowd workers and fact-checking experts cooperate to maximize value. For example, looking at the trade-off between *urgency* and quality, as soon as a statement is identified as requiring fact-check, an AI model can first be adopted to very efficiently provide a truthfulness label which could then possibly be replaced later on once a team of expert fact-checkers has concluded their forensic investigation of the available evidence in favour or against the statement being true. Such *cascade of annotation tasks* where many (or all) labels are quickly estimated automatically, only a small subset of those is sent to the crowd for a quick (but non-real time) validation of their truthfulness, and then only very few remaining statements are sent to experts to investigate in depth is the core idea of the proposed framework that leverages different levels of the size-quality-cost trade-offs that the different methodologies provide.

One dimension that impacts task allocation decisions is the cost and scale of the annotation problem. In order to leverage the best of the automated and manual methods, AI and crowdsourcing can be used to scale up the annotation effort to very many statements thus being able to possibly provide truthfulness labels for every single statement being published online. Expert fact-checkers can then be parsimoniously deployed on statements that are either difficult to label by AI or crowdsourcing methods (e.g., selected by means of low algorithmic confidence or low annotator agreement within the crowd), or important to label accurately due to the possibly wide implications of the statement being false or due to the importance of the speaker who made the statement and its potential reach.

Another open research question is on understanding how experts would actually work when embedded in this new framework: they would need to change consolidated and validated fact-checking processes and, instead, adapt to an environment in which their work is being complemented by AI and non-experts. This would necessarily require a certain level of trust in the HAI system that, on its side, is making decisions on what expert fact-checkers should do and on which statements they should work on. This translates into experts giving up a certain level of control on the process to the HAI system that has to decide what they do not get access to. For this to work, there needs to be a certain level of trust in the system that could possibly be achieved by the employment of self-explainable AI tools. This is also critical as as the fact-checking experts need at the end to be able to guarantee transparency on the process and methods used for fact-checking.

7 Take-Away Messages

In this paper we discussed the problem of online misinformation and proposed a hybrid human-AI approach to address it. We proposed a framework that combines AI, crowdsourcing, and expert fact-checkers to produce annotations for statements by balancing annotation cost, quality, volume, and speed thus providing information consumers (e.g., social media users) with timely and accurate fact-checking results at scale.

The proposed HAI approach aims at combining different methods to leverage the best properties of both AI and human-based annotation. Moreover, involving humans in the loop allows to better deal with the interdisciplinary nature of the misinformation problem by also providing human support on issues like explainability, trust, and bias.

The model presented in this paper envisions a complex collaborative scheme between different humans and different AIs where the open research question moves to the optimization of these complementary resources and on how to decide which task should be allocated to which element of the HAI system. A human-in-the-loop solution to misinformation can also provide increased transparency on fact-checking processes leveraging together algorithms and AI and, in the end, provide more evidence and power to the end users to make informed decisions on which online information they should and which they should not trust.

References

[1] F. Alam, S. Shaar, A. Nikolov, H. Mubarak, G. D. S. Martino, A. Abdelali, F. Dalvi, N. Durrani, H. Sajjad, K. Darwish, et al. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. *arXiv preprint arXiv:2005.00033*, 2020.

[2] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[3] J. Attenberg, P. Ipeirotis, and F. Provost. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1):1–17, 2015.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter? In *Proceedings of SIGIR*, pages 667–674, 2008.

[5] A. Bovet and H. A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1):1–14, 2019.

[6] A. Cerone, E. Naghizade, F. Scholer, D. Mallal, R. Skelton, and D. Spina. Watch 'n' Check: Towards a social media monitoring tool to assist fact-checking experts. In *Proceedings of DSAA*, 2020.

[7] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *PNAS*, 113(3):554–559, 2016.

[8] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478, 2012.

[9] G. Demartini, B. Trushkowsky, T. Kraska, M. J. Franklin, and U. Berkeley. Crowdq: Crowdsourced query understanding. In *CIDR*, 2013.

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

[11] T. Elsayed, P. Nakov, A. Barrón-Cedeno, M. Hasanain, R. Suwaileh, G. Da San Martino, and P. Atanasova. Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims. In *Proceedings of CLEF*, pages 301–321, 2019.

[12] N. Evans, D. Edge, J. Larson, and C. White. News provenance: Revealing news text reuse at web-scale in an augmented news search experience. In *Proceedings of CHI*, pages 1–8, 2020.

[13] L. Favano, M. J. Carman, and P. L. Lanzi. TheEarthIsFlat's submission to CLEF'19 CheckThat! challenge. In *CLEF (Working Notes)*, 2019.

[14] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of NAACL-HLT*, pages 1163–1168, 2016.

[15] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 61–72, 2011.

[16] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of KDD*, pages 2125–2126, 2016.

[17] C. Hansen, C. Hansen, S. Alstrup, J. Grue Simonsen, and C. Lioma. Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. In *Proceedings of TheWebConf*, pages 994–1000, 2019.

[18] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al. ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12): 1945–1948, 2017.

[19] P. N. Howard and B. Kollanyi. Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. *Available at SSRN 2798311*, 2016.

[20] T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *Computer*, 40(8):34–40, 2007.

[21] G. Karagiannis, M. Saeed, P. Papotti, and I. Trummer. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *Proceedings of the VLDB Endowment*, 13(11):2508–2521, 2020.

[22] J. H. Kuklinski, P. J. Quirk, J. Jerit, D. Schwieder, and R. F. Rich. Misinformation and the currency of democratic citizenship. *Journal of Politics*, 62(3):790–816, 2000.

[23] S. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, and Z. Marinho. Automated fact checking in the news room. In *Proceedings of TheWebConf*, pages 3579–3583, 2019.

[24] A. Oeldorf-Hirsch and C. L. DeVoss. Who posted that story? processing layered sources in facebook news posts.

Journalism & Mass Communication Quarterly, 97(1):141–160, 2020.

[25] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[26] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto. Explainable machine learning for fake news detection. In *Proceedings of WebSci*, pages 17–26, 2019.

[27] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[28] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, and G. Demartini. Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of SIGIR*, pages 439–448, 2020.

[29] K. Roitero, M. Soprano, B. Portelli, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, and G. Demartini. The covid-19 infodemic: Can the crowd judge recent misinformation objectively? In *Proceedings of CIKM*, 2020. In press. arXiv preprint arXiv:2008.05701.

[30] C. Sarasua, E. Simperl, and N. F. Noy. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International semantic web conference*, pages 525–541. Springer, 2012.

[31] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[32] D. Spina, M.-H. Peetz, and M. de Rijke. Active learning for entity filtering in microblog streams. In *Proceedings of SIGIR*, pages 975–978, 2015.

[33] K. Starbird, A. Arif, and T. Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *PACMHCI*, 3(CSCW):1–26, 2019.

[34] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: A large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[35] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of CVPR*, volume 1, pages I–I. IEEE, 2001.

[36] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, 2014.

[37] L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[38] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings of SIGIR*, pages 105–114, 2011.

[39] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pages 422–426, 2017.

[40] C. Wardle. Fake news. It's complicated. *First Draft*, 16, 2017.

[41] D. Weiss and B. Taskar. Structured prediction cascades. In *Proceedings of AISTATS*, pages 916–923, 2010.

[42] L. Wu, Y. Rao, A. Nazir, and H. Jin. Discovering differential features: Adversarial learning for information credibility evaluation. *Information Sciences*, 516:453–473, 2020.

[43] L. Wu, Y. Rao, X. Yang, W. Wang, and A. Nazir. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of IJCAI*, pages 1388–1394, 2020.