A Hybrid Human-In-The-Loop Framework for Fact Checking

David La Barbera[1], Kevin Roitero[1] and Stefano Mizzaro[1]

[1]*University of Udine, Via Delle Scienze 206, Udine, Italy*

Abstract

Online misinformation is posing a serious threat for the modern society. Assessing the veracity of online information is a complex problem which nowadays is addressed by heavily relying on trained fact-checking experts. This solution is not scalable and, also due the importance of the problem the issue gained the attention of the scientific community, which proposed many AI-based automatic solutions. Despite the efforts made, the effectiveness of such approaches is not yet enough to allow them to be used without supervision. In this position paper, we propose a hybrid human-in-the-loop framework for fact-checking: we address the misinformation issue by relying on a combination of automatic AI methods, crowdsourcing ones, and experts. We study the single components of the frameworks as well as their interactions, and we propose an interleaving of the different components which we believe will serve as useful starting point for the future research towards effective and scalable fact-checking.

Keywords
Misinformation, Human-in-the-loop, Artificial Intelligence

1. Introduction

Modern times have highlighted the centrality of the threat for the modern society of fake news and misinformation. Traditionally, misinformation detection is a slow and costly process that is made solely by expert trained fact-checkers, that can not cope with the ever-increasing amount of information shared online everyday. To address this issue, researchers are developing automatic techniques to identify misinformation at scale, and significant efforts have been made to develop fast and scalable state-of-the-art Artificial Intelligence (AI) algorithms [2, 3, 4]. Another less traditional approach to tackle such issue is to take advantage of the wisdom of the crowd [5] and leverage crowdsourcing workers [6, 7, 8, 9, 10, 11, 12, 13]. Both approaches have pro- and contra: while AI is usually cheaper and scalable, crowd-workers can perform more reliable and explainable classifications. To take the best from both worlds, researchers proposed hybrid Human-In-The-Loop (HITL) approaches that integrate AI, crowd, and experts, even though only few implementations exist [14, 15, 16, 17]. Differently from previous work [17], in this paper we propose a concrete architecture for fact-checking, and we inspect the responsibilities of each component as well as their interactions. In particular, we detail a

CEUR Workshop Proceedings (CEUR-WS.org)

pragmatical workflow which should be implemented to effectively classify the veracity of a set of statements at scale.

2. Related Work

There are both numerous examples of AI techniques for misinformation detection [2] as well as of academic interest on their development and evaluation [18]. Many different AI approaches exist: Ozbay and Alatas [3] tested 23 supervised AI algorithms on public datasets, Zhao et al. [4] integrated linguistic, topic, sentiment, and behavioral features to develop a model for health misinformation, Stammbach and Neumann [19] used evidence retrieval techniques and fine-tune a BERT-based model for the FEVER challenge, Konstantinovskiy et al. [20] developed a pipeline to identify misinformation using a multi-task learning approach. Related to that, many approaches addressed the issue of credibility in social media [21].

Focusing on misinformation detection using crowdsourcing, La Barbera et al. [7] first found an effect of judgment scales and evidence of worker assessors' bias on political statements, Soprano et al. [11] used the dataset from Roitero et al. [8] to leverage a multidimensional scale to measure different aspects of a statement, Draws et al. [13] found that workers generally overestimate the truthfulness and that different type of workers show different biases when evaluating a given statement, Pennycook and Rand [6] used the crowd to study effects of reducing social media users' exposure to low-quality news, and Allen et al. [12] compared the accuracy ratings between fact-checkers and crowd-workers.

Finally, some work investigated the combination of AI and humans: Demartini et al. [17] introduced a theoretical hybrid HITL framework for misinformation, Qu et al. [22] used self-reported scores from both AI and crowd to develop a hybrid system, Shabani et al. [14] used humans to provide feedback on news stories about statement contextual information and integrated those features into an AI pipeline, and Yang et al. [15] showed the potential speed up to the fact-checking process by organizing and selecting representative statements.

3. Limitations of Current Approaches

As highlighted by Demartini et al. [17, Figure 2] each of the three state-of-the-art approaches for misinformation detection i.e., experts, AI tools, and crowd has its own advantages and disadvantages in terms of accuracy, scale, cost, explainability, and bias control. We detail these aspects in this section, focusing on the limitations of each approach.

Certainly AI tools outperform both crowd and experts when considering costs[1] and evaluation speed, but despite recent works [26, 27], they provide less or no explainability. More importantly, such models achieve lower accuracy than crowd or experts. To provide some examples, classical machine learning models achieved 74% accuracy on a two-level scale [28], and the best model of this year CLEF CheckThat! Lab reached 54.7% accuracy on a four-level scale [18]. Considering the accuracy from the crowd, experimental results [12] show a high correlation with the experts in terms of agreement, whereas other work reports accuracy values that are lower and

---

[1]while training language models from scratch can cost up to millions of dollars [23], once trained they can be used multiple times leveraging few- or zero-shot learning [24, 25].

comparable to those obtained by AI methods [7, 8, 9, 10, 11, 13]; although further studies are needed to draw definitive conclusions it seems reasonable to assume that crowd accuracy can be higher than automatic AI solutions. The highest accuracy is achieved by the experts, which is always set to the value of 1 for practical reason. Nevertheless, even domain experts need confrontation and discussion phases to reach a final consensus (see for example the process used by PolitiFact[2]).

Bias is also a crucial limitation of current approaches. Experts and crowd-workers being humans are subject to cognitive biases [13, 7], which can be mitigated by the discussion phase in the case of experts, but are difficult to remove for crowd-workers [29]. Moreover, all the aforementioned biases can be propagated from humans to AI models, e.g., when training or fine-tuning a model.

Another limitation of current approaches is given by the specific truthfulness scales used; different scales exist and are used, and such heterogeneity, apart from making a fair comparison difficult, has an impact on the quality of the collected data [7].

We believe that a HITL framework for misinformation detection should address and overcome all of the limitations detailed above by fruitfully combining the capabilities of AI, crowd, and experts.

4. HITL Framework for Misinformation Detection

4.1. Possible Architectures

A natural solution to the task investigated in this paper is to employ a pipeline model where the components are sorted with an increasing accuracy (i.e., first the AI, then the crowd, and finally the experts). Thus, if a statement is not adequately classified by a component, the subsequent pipeline component will perform a more accurate classification. Also, such a pipeline concatenates each component according to their increasing cost and evaluation time. This allows to perform a pipeline of annotation tasks where the majority of the statements are quickly and automatically labeled by AI, only a subset of the statements is sent for a slower evaluation to the crowd, and the few remaining statements are sent to experts for an in-depth investigation. The key advantage of this configuration is that it takes the best from each component, and that it allows to minimize the overall costs. Particularly, this configuration lets the experts (i.e., the more costly component) to evaluate a very small number of statements. Nevertheless, the pipeline model has important limitations as it does not provide feedback among the components: a statement is simply forwarded until it is eventually classified with not much cooperation among the components.

Another possible combination of the components is by means of a blackboard architecture, a common solution in distributed multi-agent settings [30]. Such an approach allows the components to select which statements to evaluate. Each component is an autonomous agent that can access a central repository that contains both the statements and the partial contributions provided by each component. This approach would require both a high synergy between the
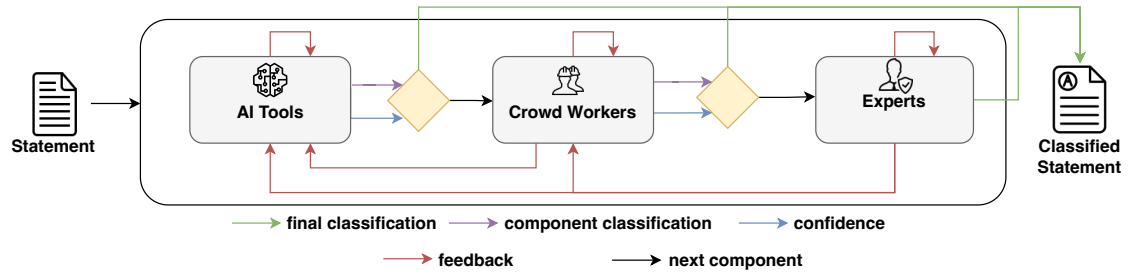

Figure 1: Overview of the proposed framework.

components and to split a classification task in atomic sub-tasks to take advantage from each specific component of the architecture.

4.2. General Framework

An ideal framework should maximize accuracy while minimizing the cost of each component and strengthening the cooperation between and within its modules. Therefore, we propose first a basic framework, where each component provides feedback to, and cooperates with, the others. We then discuss possible variants and extensions.

Our proposal is summarized in Figure 1. Given a statement, each of the three components (AI, crowd, and experts) generates: a classification on a chosen scale and a confidence score for the performed classification. Whenever the component AI or crowd generates a prediction with a high confidence score, the statement is considered as correctly classified. Otherwise if the confidence is low, the statement is forwarded to the subsequent component. If this is the case, the output of the component (such as the confidence score and the classification) can be optionally forwarded along with the statement. This could allow the subsequent component to perform an informed assessment, if necessary. Also samples of statements considered as correctly classified by the component (i.e, with high confidence) should be propagated, to double check their classification score and deal with the problem of unknown-unknowns (i.e., statements for which AI is highly confident about its predictions but is wrong) using humans [31, 32, 33]. This allows each component to provide feedback to the previous ones, thereby improving their classifications.

In the following sections we will detail for each component: its possible internal structure, its specific interactions with other components, and additional outputs that can be added to the general framework.

4.3. First Component: AI

Assuming the use of a state-of-the-art model for misinformation detection [28, 2, 19, 3, 4, 34, 18], the output provided by the AI component should be at least a classification score on a chosen truthfulness scale, and a confidence score. While the classification score is straightforward, the confidence can be reliably calibrated following the methodology by Guo et al. [35]. To provide an adequate classification, AI tool can rely on a Knowledge Base (KB) to perform evidence retrieval.

Examples of such a system are the ones proposed by La Barbera et al. [36] and Stammbach and Neumann [19], who both use a transformer architecture who rely on retrieved evidence. The choice of the Knowledge Base (KB) to use to produce a classification and an explanation is not straightforward, since there is no evidence of a "universally best" KB [37]. Thus, the choice of the specific KB should be performed ad-hoc by leveraging statements and domain specific features, as for example the topic, speaker, year, etc. of the set of statements being processed.

To evaluate the classification score given by the component, we can use optional output. For example, many AI models are able to provide reasons for their predictions [26, 27]. Some implementations are delivered by Kazemi et al. [38] and by Brand et al. [39] who develop models able to generate an explanation for their misinformation assessment. The generation of an explanation could improve the framework by providing additional and human-readable information useful for both the subsequent human-based components and the final classification.

Finally, the AI component could provide self-feedback by using counterfactual explanations [40]: generating instances that the model finds hard to classify or deceiving could improve the model performances, robustness, and generalization abilities.

The output of the AI component is thus made by classification, confidence, and optional information, such as explanation and retrieved evidence. The decision whether the statement has been adequately classified or not can be then performed by relying on the confidence of the model [22] as detailed in Section 4.2. To help this decision, it could be used the optional explanation, for example considering its readability or semantic scores. The decision for some statements might be more critical and not straightforward: a very recent statement made by an important public figure over a highly relevant topic with not much evidence available might be worth further investigation. Hence, it might be worth studying the effectiveness of an importance score using the statement's metadata.

Finally, if the assessment for the statement has a low confidence, the explanation is not satisfactory, or the assessment needs to be refined for any other reason, the statement is sent to the subsequent component: the crowd.

4.4. Second Component: Crowd

As for the AI, the crowd component should perform two tasks: misinformation classification and provide feedback to itself and to the AI component. There are many examples of misinformation classification directly performed by the crowd [6, 7, 8, 11, 12, 13]. It could also be reasonable to perform an informed assessment relying on the output of the AI component [41]. Nevertheless, the use of this additional information could introduce biases into the assessment performed by the crowd, hence further studies in this direction are required. Moreover, to reduce workers cognitive effort, it is possible to design a two steps task using disjoints sets of workers: the first set will search for evidence for a given statement, the second will classify the statement using the provided evidence (and additional data). While all of the different mentioned tasks are indeed reasonable, it is necessary to perform ad-hoc studies to find the best possible setting. Along this line, we can leverage work done in related fields [42] to identify the subset of best workers and exploit their features to be able to minimize the workforce needed and at the same time maximize its effectiveness.

Also, the crowd can be asked to provide additional rationales to motivate their classification

[43, 44]. The classifications can be used to improve the AI component by fine-tuning the models with additional data, or even both workers and AI rationales can be used to adjust the confidence of the final assessment; nevertheless, this should be implemented with caution, as workers rationales might contain bias that can be involuntary injected into AI models. Finally, a subset of crowd-workers should look for counterfactual examples that could highlight AI classification errors with high confidence. While these methodologies still need to be tested in the field of misinformation detection, some work [45] shows the promising results of this approach applied to different domains.

As for the AI component, the output of the crowd component is composed by the default classification and confidence, along with optional additional data such as evidence, explanation, and rationales. Therefore, to decide if a statement is correctly classified or not it is possible to rely not only on the data generated by the crowd, but also to check for agreement and inconsistencies between crowd and AI [22].

At this point of the evaluation, the majority of the statements have been classified by the framework, and only a very small subset will reach the final step of the workflow: the experts.

4.5. Third Component: Experts

The last step of the framework is made by the experts. It is possible to let them evaluate a statement using a pre-defined fact-checking methodology, and ideally to provide to them all the outputs from the previous components to perform an informed assessment. The effects of such a decision need to be studied since, as discussed for the crowd, the use of additional information could introduce bias in the final evaluation. We remark that we believe that critical, important, and difficult statements should always be evaluated or at least checked by the experts. Note that to identify those statements it would be necessary to find a metric to be able to automatically evaluate the importance of a statement in a given context. Also, to increase the robustness of the framework, the experts should be able to directly look at the statements classified by the previous components and to decide whether some of them need to be re-assessed or not. Finally, each classification performed by the experts should be used to re-train the AI models, and used as an example to train the crowd before performing the task. This final aspect could also be performed interactively, following an active-learning scenario.

5. Conclusions

In this work we study the limitations of the current approaches for misinformation detection and propose a hybrid HITL framework that combines AI, crowd, and experts. Our main contributions are the following: we frame the problem and review the related work detailing frameworks for fact-checking; we study possible framework architectures detailing their respective advantages and disadvantages; we propose a solid architecture for performing fact-checking at scale, and we describe each component focusing on its role and outputs, as well as its interactions with other components. The main advantages of our framework are given by an efficient combination of the components in terms of increasing accuracy and evaluation time, decreasing costs, and by the feedback between and within each component.

Future work aims at proving a full framework implementation. More in detail, further study will be done on the synergies between crowd and AI to investigate the effects of an informed assessment made by the crowd leveraging AI outputs, and to set thresholds to decide about statement forwarding among components.

References

[1] D. Nozza, L. Passaro, M. Polignano, Preface to the sixth workshop on natural language for artificial intelligence (nl4ai), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), November 30, 2022, CEUR-WS.org, 2022.

[2] B. Guo, Y. Ding, L. Yao, Y. Liang, Z. Yu, The Future of Misinformation Detection: New Perspectives and Trends, 2019. doi:10.48550/ARXIV.1909.03654.

[3] F. A. Ozbay, B. Alatas, Fake news detection within online social media using supervised artificial intelligence algorithms, Physica A: Statistical Mechanics and its Applications 540 (2020) 123174. doi:10.1016/j.physa.2019.123174.

[4] Y. Zhao, J. Da, J. Yan, Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches, Information Processing & Management 58 (2021) 102390. doi:10.1016/j.ipm.2020.102390.

[5] J. Surowiecki, The Wisdom of Crowds, Anchor, 2005.

[6] G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality, Proceedings of the National Academy of Sciences 116 (2019) 2521–2526. doi:10.1073/pnas.1806781116.

[7] D. La Barbera, K. Roitero, D. Spina, S. Mizzaro, G. Demartini, Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias, in: Proceedings of the 42nd European Conference on Information Retrieval, ECIR, Springer, 2020, pp. 207–214.

[8] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, G. Demartini, Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background, in: Proceedings of the 43rd Conference on Research and Development in Information Retrieval, SIGIR, ACM, 2020, pp. 439–448.

[9] K. Roitero, M. Soprano, B. Portelli, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, G. Demartini, The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?, in: Proceedings of the 29th Conference on Information & Knowledge Management, CIKM, ACM, 2020, p. 1305–1314. doi:10.1145/3340531.3412048.

[10] K. Roitero, M. Soprano, B. Portelli, M. Luise, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, G. Demartini, Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19, Personal and Ubiquitous Computing (2021). doi:10.1007/s00779-021-01604-6.

[11] M. Soprano, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, S. Mizzaro, G. Demartini, The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale, Information Processing & Management 58 (2021) 102710. doi:10.1016/j.ipm.2021.102710.

[12] J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds, Science Advances 7 (2021) eabf4393. doi:`10.1126/sciadv.abf4393`.

[13] T. Draws, D. La Barbera, M. Soprano, K. Roitero, D. Ceolin, A. Checco, S. Mizzaro, The Effects of Crowd Worker Biases in Fact-Checking Tasks, in: Conference on Fairness, Accountability, and Transparency, FAccT, ACM, 2022, p. 2114–2124. doi:`10.1145/3531146.3534629`.

[14] S. Shabani, Z. Charlesworth, M. Sokhn, H. Schuldt, SAMS: Human-in-the-loop Approach to Combat the Sharing of Digital Misinformation, CEUR Workshop Proc. 2846 (2021).

[15] J. Yang, D. Vega-Oliveros, T. Seibt, A. Rocha, Scalable Fact-checking with Human-in-the-Loop, in: IEEE Workshop on Information Forensics and Security, WIFS, 2021, pp. 1–6. doi:`10.1109/WIFS53200.2021.9648388`.

[16] G. Karagiannis, M. Saeed, P. Papotti, I. Trummer, Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification, CoRR abs/2003.06708 (2020).

[17] G. Demartini, S. Mizzaro, D. Spina, Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities, Bulletin of IEEE Computer Society 43 (2020) 65–74.

[18] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, 2022, pp. 495–520.

[19] D. Stammbach, G. Neumann, Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task, in: Proceedings of the 2nd Workshop on Fact Extraction and VERification, FEVER, ACL, 2019, pp. 105–109. doi:`10.18653/v1/D19-6616`.

[20] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection, Digital Threats 2 (2021). doi:`10.1145/3412869`.

[21] M. Viviani, G. Pasi, Credibility in social media: opinions, news, and health information—a survey, WIREs Data Mining and Knowledge Discovery 7 (2017) e1209. doi:`https://doi.org/10.1002/widm.1209`.

[22] Y. Qu, D. L. Barbera, K. Roitero, S. Mizzaro, D. Spina, G. Demartini, Combining Human and Machine Confidence in Truthfulness Assessment, Data and Information Quality (2022). doi:`10.1145/3546916`.

[23] O. Sharir, B. Peleg, Y. Shoham, The Cost of Training NLP Models: A Concise Overview, arXiv (2020).

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[25] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, in: Workshop on Knowledge Retrieval and Language Models, ICML, 2022.

[26] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Generating Fact Checking Explanations, CoRR abs/2004.05773 (2020).

[27] N. Kotonya, F. Toni, Explainable Automated Fact-Checking: A Survey, CoRR abs/2011.03870 (2020).

[28] M. Granik, V. Mesyura, Fake news detection using naive Bayes classifier, in: IEEE First Ukraine Conference on Electrical and Computer Engineering, UKRCON, 2017, pp. 900–903. doi:10.1109/UKRCON.2017.8100379.

[29] T. Draws, A. Rieger, O. Inel, U. Gadiraju, N. Tintarev, A Checklist to Combat Cognitive Biases in Crowdsourcing, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 9 (2021) 48–59.

[30] J. Dong, S. Chen, J.-J. Jeng, Event-based blackboard architecture for multi-agent systems, in: Proceedings of the Conference on Information Technology: Coding and Computing, volume 2 of *ITCC*, 2005, pp. 379–384. doi:10.1109/ITCC.2005.149.

[31] J. Attenberg, P. Ipeirotis, F. Provost, Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns", J. Data and Information Quality 6 (2015). doi:10.1145/2700832.

[32] H. Lakkaraju, E. Kamar, R. Caruana, E. Horvitz, Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI, AAAI Press, 2017, p. 2124–2132.

[33] A. Liu, S. Guerra, I. Fung, G. Matute, E. Kamar, W. Lasecki, Towards Hybrid Human-AI Workflows for Unknown Unknown Detection, in: Proceedings of The Web Conference, WWW, ACM, 2020, p. 2432–2442. doi:10.1145/3366423.3380306.

[34] B. Taboubi, M. A. B. Nessir, H. Haddad, iCompass at CheckThat! 2022: combining deep language models for fake news detection, Working Notes of CLEF (2022).

[35] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th Conference on Machine Learning, volume 70 of *ICML*, JMLR.org, 2017, p. 1321–1330.

[36] D. La Barbera, K. Roitero, J. Mackenzie, D. Spina, G. Demartini, S. Mizzaro, BUM at CheckThat! 2022: A Composite Deep Learning Approach to Fake News Detection using Evidence Retrieval, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF, 2022, pp. 564–572.

[37] D. Stammbach, B. Zhang, E. Ash, The Choice of Knowledge Base in Automated Claim Checking, CoRR abs/2111.07795 (2021). arXiv:2111.07795.

[38] A. Kazemi, Z. Li, V. Pérez-Rosas, R. Mihalcea, Extractive and Abstractive Explanations for Fact-Checking and Evaluation of News, CoRR abs/2104.12918 (2021).

[39] E. Brand, K. Roitero, M. Soprano, A. Rahimi, G. Demartini, A Neural Model to Jointly Predict and Explain Truthfulness of Statements, Data and Information Quality (2022). doi:10.1145/3546917.

[40] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. F. Christiano, G. Irving, Fine-Tuning Language Models from Human Preferences, CoRR abs/1909.08593 (2019).

[41] C. Snijders, R. Conijn, E. Fouw, K. Berlo, Humans and Algorithms Detecting Fake News: Effects of Individual and Contextual Confidence on Trust in Algorithmic Advice, Journal of Human-Computer Interaction (2022) 1–12. doi:10.1080/10447318.2022.2097601.

[42] H. Li, Q. Liu, Cheaper and Better: Selecting Good Workers for Crowdsourcing, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 3 (2015) 20–21.

[43] T. McDonnell, M. Lease, M. Kutlu, T. Elsayed, Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments, in: Proceedings of the 4th Conference on Human Computation and Crowdsourcing, volume 4 of *HCOMP*, 2016, pp. 139–148.

[44] M. Kutlu, T. McDonnell, M. Lease, T. Elsayed, Annotator Rationales for Labeling Tasks in Crowdsourcing, Journal of Artificial Intelligence Research 69 (2020) 143–189. doi:`10.1613/jair.1.12012`.

[45] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, Y. Qi, TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, 2020. doi:`10.48550/ARXIV.2005.05909`.