Human-in-the-Loop Systems for Truthfulness: A Study of Human and Machine Confidence

Yunke Qu
The University of Queensland
Brisbane, Australia
yunke.qu@uq.net.au

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Gianluca Demartini
The University of Queensland
Brisbane, Australia
demartini@acm.org

Abstract

Automatically detecting online misinformation at scale is a challenging and interdisciplinary problem. Deciding what is to be considered truthful information is sometimes controversial and difficult also for educated experts. As the scale of the problem increases, human-in-the-loop approaches to truthfulness that combine both the scalability of machine learning (ML) and the accuracy of human contributions have been considered.

In this work we look at the potential to automatically combine machine-based systems with human-based systems. The former exploit supervised ML approaches; the latter involve either crowd workers (i.e., human non-experts) or human experts. Since both ML and crowdsourcing approaches can produce a score indicating the level of confidence on their truthfulness judgments (either algorithmic or self-reported, respectively), we address the question of whether it is feasible to make use of such confidence scores to effectively and efficiently combine three approaches: (i) machine-based methods; (ii) crowd workers, and (iii) human experts. The three approaches differ significantly as they range from available, cheap, fast, scalable, but less accurate to scarce, expensive, slow, not scalable, but highly accurate.

1 Introduction

The challenge of identifying online misinformation has been rapidly growing given the increase in popularity of online news consumption as well as the ability to profile and micro-target social media users. Fighting the spread of online misinformation is a multi-disciplinary issue which requires both technical advances to process large amounts of false digital information as well as to understand the societal context in which such spreads happen. In order to best deal with the need to both scale to large number of fact-checks and have expert journalists manually checking and evaluating the veracity of posted information, human-in-the-loop systems have been considered (Demartini et al., 2020; Allen et al., 2021; Nakov et al., 2021).

Human-in-the-loop information systems aim at leveraging the ability of machines to scale and deal with 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 hybrid system (Demartini et al., 2017). Example of successful human-in-the-loop methods include ZenCrowd (Demartini et al., 2012), CrowdQ (Demartini et al., 2013), CrowdDB (Franklin et al., 2011), and Crowdmap (Sarasua et al., 2012). Active learning methods (Settles, 2009) are another example where labels are collected from humans, fed back to a supervised learning model, and then used to decide which data items humans should label next. Related to this is the idea of interactive machine learning (ML) (Amershi et al., 2014) where labels are automatically obtained from user interaction behaviors (Joachims and Radlinski, 2007).

While being more powerful than pure machine-based methods, human-in-the-loop 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 non-expert humans. Depending on which human participants are providing labels, 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 MTurk (Gadiraju et al., 2015) or by means of alternative incentives like, e.g., 'games with a purpose' (Von Ahn, 2006) 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 not be 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. Crowdsourcing as a method to collect labels to train veracity classification systems has recently been investigated (Roitero et al., 2020a,b; Soprano et al., 2021; Roitero et al., 2021).

Rather than seeing these data collection approaches as mutually exclusive, in this paper we focus on the possibility of combining machine-based truthfulness classifiers, non-expert annotators, and experts. In particular, we focus on the notion of *confidence*, i.e., the estimate of the reliability of the prediction—given by either a machine or a human annotator.

More in detail, in this paper we focus on the following research questions:

- RQ1: Can algorithmic and self-reported human confidence scores be used to reliably estimate the quality of truthfulness decisions?
- RQ2: Do humans and machines make similar or different mistakes in classifying truthfulness?
- RQ3: Can scarce expert annotator resources be integrated in such human-in-the-loop systems to intervene in cases when both crowd workers and machine-based truthfulness classifiers fail to correctly label an item?

To the best of our knowledge, this is the first attempt to understand the relationship between the effectiveness and confidence of the set including machine-based methods, crowd workers, and experts in a truthfulness classification task.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 details the methodology used in our study. We report and analyze our results in Section 4. Section 5 concludes by summarizing our findings and describing future work.

2 Related Work

In this section we summarize approaches computing and making use of confidence scores generated by ML models or human annotators (either self-reported or implicit).

Different types of ML methods are able to produce not only a classification decision, but to also attach a score that indicates how confident the algorithm is about the made decision. This is possible for a diverse set of methods, from decision trees to deep learning.

Poggi et al. (2017) consider a complete overview of 76 state-of-the-art confidence measures for ML; Mandelbaum and Weinshall (2017) discuss distance based confidence scores in the case of neural network based classifiers; Guo et al. (2017) detail a methodology to correctly interpret and compute confidence scores from ML models.

Trusting classification decisions solely based on algorithmic confidence may be risky. Once manually labelled data has been collected, trained models may reflect existing bias in the data. An example of such a problem is that of ‘unknown unknowns’ (UUs) (Attenberg et al., 2015), 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 ML.

Quantifying decision confidence can also be done when decisions are made by human annotators. Hertwig (2012) discuss the role of confidence in the “wisdom of the crowd” paradigm. They point out how human confidence may be influenced by social interaction and the presence of others’ annotations. Joglekar et al. (2013) describes methods to generate confidence intervals in order to capture crowd workers’ confidence and bound accuracy scores. Jarrett et al. (2015) consider workers’ self-assessment and investigates whether workers confidence correlates with quality and observe that self-evaluation is not indicative of their actual performance. This is consistent with findings by Gadiraju et al. (2017). Related to this observation, Li and Varshney (2017) show that workers annotation performance does not increase when considering the confidence scores to weight their contribution. Song et al. (2018) consider worker confidence in the setting of a labeling task performed with active learning techniques. Difallah et al. (2016) look at how to schedule labeling tasks to optimize their execution efficiency.

More than just human self-reported confidence, it is possible to implicitly measure confidence by, for example, computing inter-assessor agreement metrics. Nowak and Rüger (2010) study inter-annotator agreement and show how annotation quality can be improved when considering agreement scores to aggregate labels. Aroyo and Welty (2013) study the relationships between gold questions and workers agreement stating that agreement metrics do not necessary correlate with quality but may uncover alternative views on possible way to label data. Checco et al. (2017) discuss agreement measures applied to crowdsourcing and propose an alternative measure that is able to deal with sparse and incomplete data. Maddalena et al. (2017) incorporate assessor agreement into information retrieval evaluation metrics. In our work we make use of inter-annotator agreement metrics as a measure of human annotator confidence and quality.

3 Methodology

3.1 Dataset

We make use of manual truthfulness labels obtained from a crowdsourcing experiment as presented by Soprano et al. (2021). The crowdsourcing task was performed as follows. After an initial background survey phase, crowd workers are presented with 11 political statements, one after the other; 6 statements are taken from PolitiFact (Wang, 2017), 3 from ABC,[1] and 2 are used as quality checks. For each statement, according to the design defined by Roitero et al. (2020a), workers are asked to provide a truthfulness label. Additionally to the design by (Roitero et al., 2020a), we ask workers to also provide a confidence score on the expressed truthfulness label on a Likert scale in the $[-2, 2]$ range. The dataset contains a total of 120 statements from PolitiFact: 10 for each of the two political parties and for each level of the six-level truthfulness scale used by the expert assessors to evaluate the statements, and a total of 60 statements from ABC: 10 for each of the two political parties and for each level of the three-level truthfulness scale used by the expert assessors to evaluate the statements.

3.2 Machine Learning for Truthfulness Classification

BERT (Bidirectional Encoder Representations from Transformers) (Vaswani et al., 2017) is a language representation model based on performing a bidirectional training of a transformer based model. The core part of the model is the encoder / decoder architecture (Devlin et al., 2019), which is formed by different steps: the tokenization and numericalization of the input sequence followed by a set of embedding layers, which learn during the training phase a multidimensional embedding for each input token. Then, the learned representation is enriched with the context information represented with the positional encoding of the tokens built using the Multi Head (Self) Attention mechanism, which is fundamental to learn a better language model. In the BERT architecture multiple encoder / decoder blocks are stacked together to form the model. This architecture allows BERT to encode the entire input sequence at once, and perform two training task simultaneously: Masked Language Model and Next Sentence Prediction. The truthfulness classification task has been carried out using the BERT model pre-trained for classification tasks (`bert-base-uncased`[2]) fine-tuned with expert truthfulness labels on political statements. We use the output of the last softmax layer as the ML classification confidence score we use in our analysis.

GloVe (Global Vectors for Word Representation) by (Pennington et al., 2014) is a word vector learning technique which produces a vector space model similar to word2vec. The fundamental idea behind GloVe and word2vec is to learn, given a large corpus, a set of tuples containing a word and its context; then, the model is trained to predict the context given the specific word. Unlike word2vec which captures only the local context of a word, GloVe considers also the global context, implemented through a co-occurrence matrix. A feed-forward architecture with two dense layers (6 and 1 node, respectively), and a soft-max layer at the end. In Section 4 we only report results obtained with BERT for space constraints but results obtained with GloVe were similar.

[1] `https://apo.org.au/collection/302996/rmit-abc-fact-check`

[2] `https://huggingface.co/bert-base-uncased`

3.3 Crowdsourcing for Truthfulness Classification

With the crowdsourcing task design presented in Section 3.1, we collect non-expert labels from Amazon MTurk for 180 statements across different ground-truth truthfulness levels and different sources. In order to compare against supervised binary ML classifiers, we binarize human labels (originally collected on a 5-point $[-2, 2]$ Likert scale) by considering $\{-2, -1\}$ as the `False Statements` class and $\{1, 2\}$ as the `True Statements` class. We also binarize the 6-level Politifact scale and the 3-level ABC scale expert labels.

We use both crowd labels aggregated by the sum of the scores given by the 10 different workers who judged the same statement, as well as using the raw labels and confidence scores provided by individual crowd workers. We remove both the 20 ABC labels with an in-between value and the 5 aggregated crowd labels with a 0 value, as they do not indicate a binary classification decision. We are then left with 159 statements which we use in our analysis.

Thus, we generated a dataset that contains, for a total of 159 statements, truthfulness labels produced by ML models, non-expert crowd workers, and experts (i.e., ground truth labels) together with the respective confidence scores (experts are assumed to have max confidence).

3.4 ML and Crowd Confidence

To compute the crowd and machine learning confidence, we proceed as follows. For crowdsourced labels, we consider both the confidence scores self-reported by individual crowd workers, as well as the standard deviation among the ten crowd labels collected for each document. We refer these two scores respectively as *explicit* and *implicit* confidence scores.

Concerning the machine learning approaches, we cannot directly use the scores returned by the model in their last soft-max layer. Such scores can not be treated as confidence scores as shown in previous studies (Guo et al., 2017). Thus, to compute the machine learning confidence scores, we employed the bootstrap technique (Efron and Tibshirani, 1985): starting from a specific machine learning model, we produced ten different variations of such model obtained by varying the random seeds used in the initialization procedure; then, we run the ten models on the dataset and, similarly to what we do for crowdsourced labels, we compute the standard deviation over the ten scores collected for each document.

4 Results

4.1 ML and Crowd Accuracy

First we report on the truthfulness classification accuracy of both ML and crowd-based methods to label the truthfulness of statements in the dataset. As compared to expert ground-truth labels, ML models and crowd workers (with truthfulness labels for a statement aggregated by means of sum as raw labels are in $[-2, 2]$) perform at a similar level of accuracy (GloVe: 64.5%; BERT: 63.52%; word2vec: 62.9%; crowd: 55.3%). Thus, in the following we only report the results obtained on the most effective ML model.

Next, we explore the opportunity of combining these approaches for truthfulness classification by leveraging confidence-based combinations as well as involving scarce expert annotator resources when most beneficial.

4.2 ML and Crowd Confidence

Figure 1 shows both the ML (i.e., GloVe) and crowd confidence for the non-aggregated labels with a breakdown on the correctly and not correctly classified statements. Note that the ML and crowd confidence scores are shown in two separate plots since they are on two separate and not comparable scales: ML confidence scores are obtained from the bootstrap techniques applied to the soft-max layer of the ML algorithm which returns values in the $[0.5, 1]$ range, while the crowd confidence score is self-reported by each crowd worker on a [-2,2] scale. As we can see from Figure 1, ML confidence scores are almost always slightly lower on average for statements in which ML decisions are wrong and higher when ML correctly classify them (i.e., easy statements), even if such differences are small and not statistically significant. We see that crowd confidence shows the same behavior. Thus, answering **RQ1**, it seems raw confidence scores may be a weak signal indicating accurate classification decisions, thus leading to risks of undetectable classification errors (i.e., *unknown unknowns*) especially for the case of non-expert human annotators.

We now look at the confidence scores for the aggregated crowd labels; these confidence scores

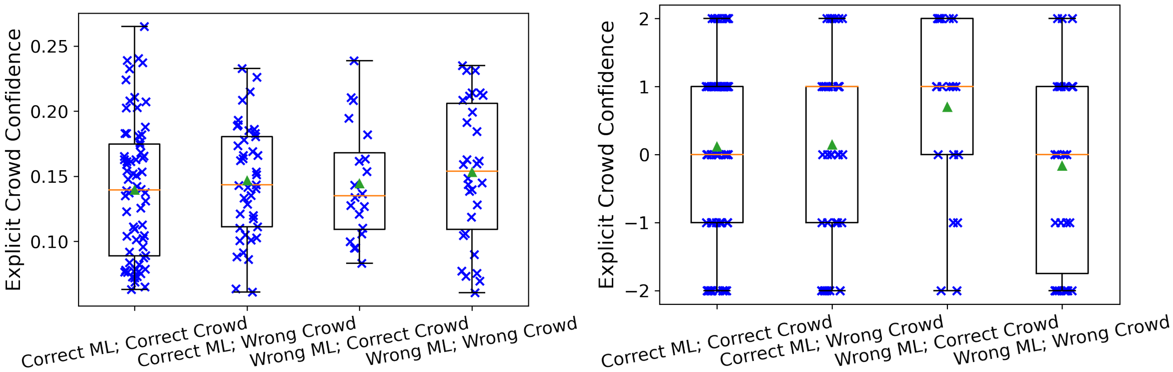

Figure 1: ML and explicit crowd confidence scores for raw crowd labels over correct and incorrect truthfulness classifications.

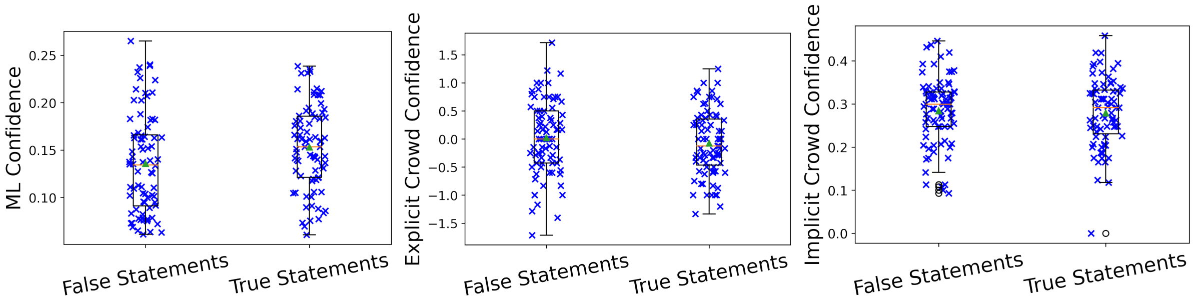

Figure 2: ML (left) and crowd confidence; both explicit (center plot) and implicit (right plot) for aggregated labels over ground-truth classes.

are obtained by taking the average value for each statement over all the workers who assessed it. Figure 2 shows, similarly to Figure 1 but with a breakdown on statement truthfulness rather than the correctness of its classification, the confidence for both ML and crowd truthfulness classification decisions.

As we can see from the plots, the mean confidence score for the 'true' statements is higher (although not significantly different according to a Mann-Whitney test) than the confidence score on the 'false' statements for confidence scores; on the contrary, for ML confidence scores the aggregated confidence scores are slightly higher (although not significantly different either) for the 'false' statements. This indicates that, similarly to what was observed for Figure 1, it seems that aggregated confidence scores are a weak signal indicating accurate classification decisions, and it should not be used as it may lead to undetectable classification errors.

We now move to study the relationship between ML and aggregated crowd confidence scores, to see if they are correlated and if one confidence score can act as a proxy for the other. Figure 3 shows on the x-axis the aggregated crowd confidence scores, on the y-axis the ML confidence; each dot is a statement; the different colors in the plot highlight a breakdown on either correctly and incorrectly classified statements by both the ML and the crowd. As we can see by inspecting the plots as a whole, both implicit and explicit crowd confidence show the same behavior when compared to ML confidence. Moreover, as we can see from inspecting the plots individually, the confidence scores for the statements correctly classified by both human and machine methods are spread across the plot; this is a further confirmation that trusting both ML and crowd confidence scores can lead to classification errors. If we now focus on the top-right and bottom-left part of the plots, we see that it contains dots of different colors; this indicates that even when both methods have either a high (top-right) or low (bottom-left) confidence scores the accuracy is similar. Again, this is a further confirmation of phenomena observed so far which indicates that both ML and crowd confidence scores should not be trusted.

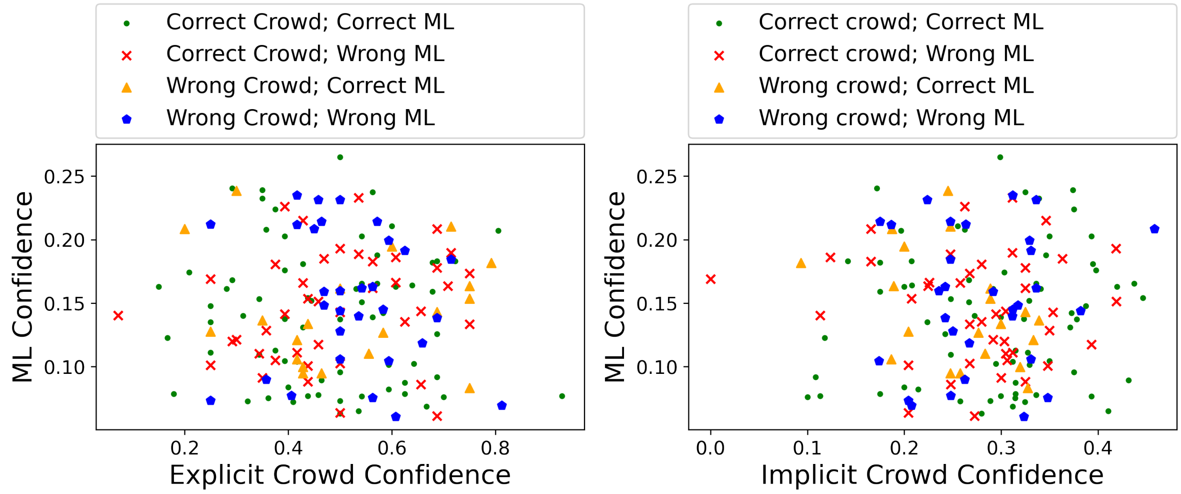

Figure 3: ML versus explicit (left plot) and implicit (right plot) crowd confidence with a breakdown on classification errors.

Summarizing the results observed so far, we can conclude that both ML and crowd confidence scores should be inspected carefully and not blindly trusted, as they can lead to classification errors. Furthermore, we observed a peculiar but interesting behavior for crowd confidence scores; both explicit (i.e., the scores submitted by the workers) and implicit (i.e., the ones automatically derived by considering the standard deviation of the truthfulness labels as submitted by the workers) confidence scores show a very similar behavior when compared to ML confidence scores; thus, this set of preliminary results hints that implicit confidence scores can act as a proxy for explicit scores if the aim is to compare them with ML scores. Thus, researchers and practitioners can avoid asking for explicit confidence scores if their focus is on accuracy and comparison with ML confidence scores, reducing the effort required by the crowd workers when performing the task.

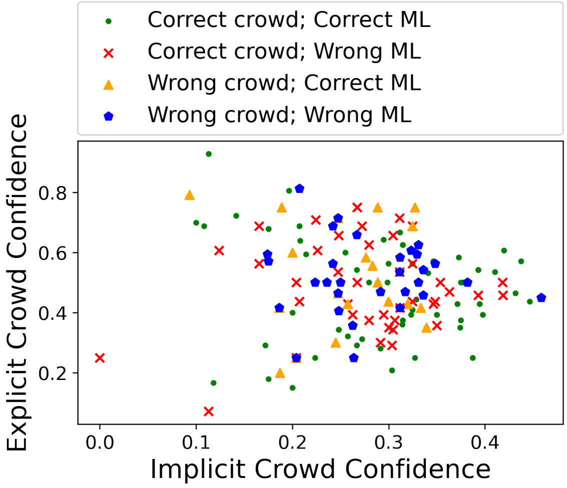

Figure 4: explicit versus implicit crowd confidence with a breakdown on classification errors.

To verify if this conjecture holds in general, we compared the explicit and implicit crowd confidence scores. Similarly to Figure 3, Figure 4 shows on the x-axis the aggregated crowd implicit confidence scores, and on the y-axis the aggregated crowd explicit confidence scores; each dot is a statement; the different colors in the plot highlight a breakdown on either correctly and incorrectly classified statements. As we can see from the plot, while implicit and explicit crowd confidence scores show a very similar behavior when compared to ML confidence (see Figure 3), we can see that the two measures are not correlated, and each statement shows a different implicit and explicit scores. Thus, if the focus of research and practitioners is purely on crowd confidence scores, implicit and explicit ones are substantially different. In the following we will focus on the relationship between effectiveness and confidence of the models, to investigate which crowd confidence scores provide a more informative signal when related to effectiveness.

We now turn to investigate whether the confidence and effectiveness of the methods used to predict the truthfulness of the statements are related. To this aim, we break down the confidence scores into quartiles and for each quartile we plot the accuracy of the considered method. Figure 5

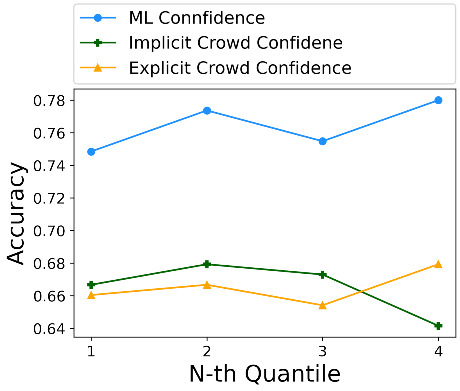

Figure 5: Confidence versus accuracy: group statements by quartiles of confidence scores and plot 4 points; both for ML and crowd.

shows the results, by displaying in the x-axis the confidence quartile, and in the y-axis the corresponding accuracy score; each series represent either the ML or crowd effectiveness scores. As we can see from the plot, there is no apparent clear pattern for all the series, even though it appears that the ML effectiveness scores overall observe a slight increase as the confidence scores itself increases, while the crowd scores, and in particular the implicit ones, observe a slight accuracy decrease while confidence increases.

Answering **RQ2**, we can see from the plots in Figure 3 and focusing on the yellow and blue statements, that there are many statements for which one of the two methods (i.e., ML or crowd) results in correct classification decisions, but the other method does not. Furthermore, Figure 5 shows that there is no clear signal that an increase in confidence is related to an increase in accuracy scores, for both ML or crowd.

While this negative results hint that it appears challenging to make use of confidence scores to increase the effectiveness of such methods and identify the cases where one of the two methods (i.e., ML or crowd) results in correct classification decisions but the other method does not, this set of results suggests the opportunity to investigate those signals in order to build an effective human-in-the-loop system which combines non-expert human and machine truthfulness classification together to obtain better quality decisions. We will discuss such approach in the following.

4.3 Can Confidence Be Leveraged?

Having studied the signal provided by both the ML and crowd confidence scores, we now investigate if such signals can be leveraged to improve the classification accuracy and the label quality when assessing the truthfulness of statements.

To this aim, and to answer **RQ3** about the potential involvement of experts, we perform the experiment as detailed in the following. Starting from the original dataset, for both ML and crowd, we replace the labels (i.e., the classification decisions for statements) that have the lower confidence scores with their corresponding ground truth label (i.e., the label as provided by the experts, which we assume to be always correct). Then, we re-compute the effectiveness of either the ML or crowd approach, measured by accuracy. To ensure a fair comparison, we also report the effectiveness of two baselines to compare against: the replacement with the ground truth label for a random statement in the dataset (repeated 50 times to remove random fluctuations of the series), and the replacement of the statements according to an oracle, which always replaces the statement that lead to obtain the highest increase in effectiveness. While the former baseline represents the average random case, the latter represents the optimal replacement selection strategy.

Figure 6 shows in the x-axis the number of statements which have been replaced in the original dataset, and in the y-axis either the ML or crowd accuracy scores; the three series represent the oracle, the random choice, and our strategy based on replacing the statements according to their confidence scores, replacing the ones with lower confidence first. As we can see focusing on the plot on the left side of Figure 6, the ML effectiveness increases as the replacements are done by removing the statements with lower confidence; we can also see that such strategy is always on average as effective as the random selection strategy, or even worse for same data points; both series are far less effective than the oracle. This results suggests that ML confidence can not act as a proxy for effectiveness, and thus it can not be leveraged (at least not in a naive way) to increase the model accuracy. This is not a definitive result and it suggest that there is room for improvement and it can be seen as an opportunity to study and develop novel methods to leverage confidence scores with the aim of identifying mis-classified statements and improving the overall model effectiveness. We leave for future work the analysis of more sophisticated approaches based on confi-

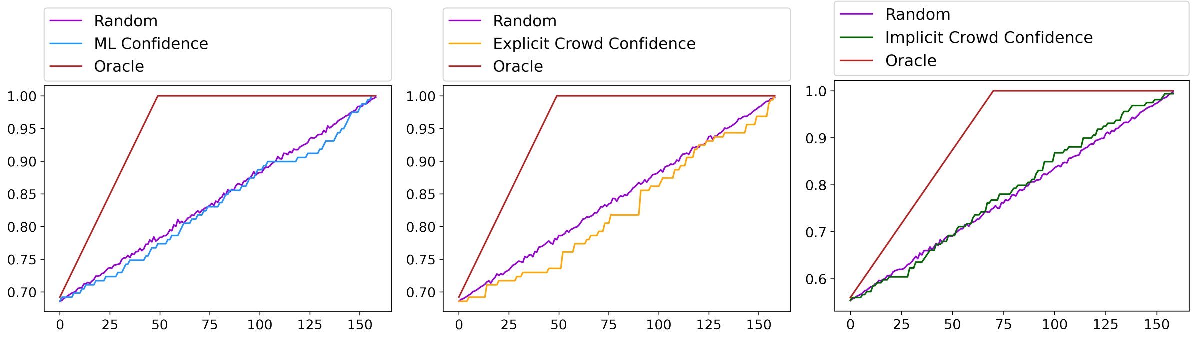

Figure 6: ML (left) and crowd (explicit, center; implicit, right) accuracy after replacing their labels with expert labels for statements (i) selected by an oracle (maximizing accuracy on each replacement), (ii) with lowest confidence, or (iii) uniformly at random.

dence or other signals. As we can see from the plot on the center of Figure 6, the same phenomena can be observed for crowd aggregated scores when explicit confidence scores are used. On the contrary, the situation changes when implicit confidence scores are used, as it can bee seen by inspecting the plot on the right side of Figure 6; such plot shows that, as the number of replacements grows, the accuracy of the methods grows and slightly over-performs the random replacement of statements. This is a positive result as it suggests that implicit confidence signals from crowd workers can be leveraged to increase the effectiveness of such method when employed to classify misinformation statements. These results are consistent with our previous observation on the lack of signal in ML confidence scores and that of previous work (Gadiraju et al., 2017; Li and Varshney, 2017) indicating that self-reported reliability is not accurate in crowdsourcing (i.e., highly confident crowd workers often make mistakes).

5 Conclusions and Future Work

In this paper we studied how ML and non-expert crowd workers classify the truthfulness of statements. To the best of our knowledge, this is the first attempt to study a human-in-the-loop pipeline for truthfulness classification which involves machines, non-experts (crowd workers), and experts (fact-checkers). In particular, we focused on both accuracy and confidence of the different approaches. We looked at both the accuracy and confidence signals alone, and we also studied their combination and their correlation; finally, we looked at identifying potential ways to leverage such signals and to combine them in order to improve the effectiveness of the classification decision process.

Our results show that, while ML and crowd confidence scores are not related to effectiveness, they can be leveraged to increase the effectiveness of the misinformation system. In this respect, implicit crowd confidence is a better indicator of effectiveness than crowd workers' self-reported confidence. We have also observed that ML and non-expert crowd workers make different mistakes, and their predictions do not agree in general. This result opens up to the opportunity of identifying more effective ways to combine these two approaches to increase the effectiveness of misinformation detection systems. Finally, we have shown that crowd workers and in particular their confidence scores can be leveraged to increase the effectiveness of systems when experts fact-checkers are brought into the loop in the cases where automatic ML or non-expert crowd workers are not confident on the submitted labels.

While our preliminary results are promising, there is still large room for improvement in making the most out of limited expert annotator resources; we believe this work is a first step towards the identification of signals for building an effective human-in-the-loop pipeline for misinformation assessment.

Acknowledgments. This work is supported by a Facebook Research award, the Australian Research Council (Grants No. DP190102141 and DE200100064), and by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

References

Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of WebSci*.

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

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of HCOMP*, pages 11–20.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of WWW*, pages 469–478.

Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta. 2017. An introduction to hybrid human-machine information systems. *Foundations and Trends in Web Science*, 7(1):1–87.

Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 43(3):65–74.

Gianluca Demartini, Beth Trushkowsky, Tim Kraska, Michael J Franklin, and UC Berkeley. 2013. CrowdQ: Crowdsourced query understanding. In *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2016. Scheduling human intelligence tasks in multi-tenant crowd-powered systems. In *Proceedings of the 25th international conference on World Wide Web*, pages 855–865.

Bradley Efron and Robert Tibshirani. 1985. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17):1–35.

Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: Answering queries with crowdsourcing. In *Proceedings of SIGMOD*, pages 61–72.

Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4):81–85.

Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. *ACM Trans. Comput.-Hum. Interact.*, 24(4).

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330.

Ralph Hertwig. 2012. Tapping into the wisdom of the crowd—with confidence. *Science*, 336(6079):303–304.

Julian Jarrett, Larissa Ferreira Da Silva, Laerte Mello, Sadallo Andere, Gustavo Cruz, and M Brian Blake. 2015. Self-generating a labor force for crowdsourcing: Is worker confidence a predictor of quality? In *Proceedings of the 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pages 85–90.

Thorsten Joachims and Filip Radlinski. 2007. Search engines that learn from implicit feedback. *Computer*, 40(8):34–40.

Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2013. Evaluating the crowd with confidence. In *Proceedings of KDD*, pages 686–694.

Qunwei Li and Pramod K Varshney. 2017. Does confidence reporting from the crowd benefit crowdsourcing performance? In *Proceedings of the 2nd International Workshop on Social Sensing (SocialSens)*, pages 49–54.

Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering assessor agreement in ir evaluation. In *Proceedings of ICTIR*, page 75–82.

Amit Mandelbaum and Daphna Weinshall. 2017. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv:1709.09844*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of IJCAI*, pages 4551–4558. Survey Track.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pages 557–566.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. 2017. Quantitative evaluation of confidence measures in a machine learning world. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5228–5237.

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020a. Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of SIGIR*, pages 439–448.

Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2021. Can the crowd judge truthfulness? a longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*.

Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020b. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? In *Proceedings of CIKM*, pages 1305–1314.

Cristina Sarasua, Elena Simperl, and Natalya F Noy. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 525–541.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems*, 159:244–258.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management*, 58(6):102710.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pages 422–426.