Combining Human and Machine Confidence in Truthfulness Assessment

YUNKE QU, The University of Queensland, Australia
KEVIN ROITERO and DAVID LA BARBERA, University of Udine, Italy
DAMIANO SPINA, RMIT University, Australia
STEFANO MIZZARO, University of Udine, Italy
GIANLUCA DEMARTINI, The University of Queensland, Australia

Automatically detecting online misinformation at scale is a challenging and interdisciplinary problem. Deciding what is to be considered truthful information is sometimes controversial and also difficult 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 superviseds 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.

CCS Concepts: • **Information systems** → **Information retrieval**; • **Human-centered computing** → **Collaborative and social computing**

Additional Key Words and Phrases: Misinformation, crowdsourcing, hybrid intelligence

ACM Reference format:
Yunke Qu, Kevin Roitero, David La Barbera, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2022. Combining Human and Machine Confidence in Truthfulness Assessment. *J. Data and Information Quality* 15, 1, Article 5 (December 2022), 17 pages.
https://doi.org/10.1145/3546916

5

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

This work is supported by a Facebook Research award, the Australian Research Council (Grants No. DP190102141 and DE200100064), by the ARC Training Centre for Information Resilience (Grant No. IC200100022), and by the ARC Centre of Excellence for Automated Decision-Making and Society (Grant No. CE200100005).
Authors' addresses: Y. Qu and G. Demartini, The University of Queensland, Brisbane, Australia; emails: yunke.qu@uq.net.au, demartini@acm.org; K. Roitero, D. La Barbera, and S. Mizzaro, University of Udine, Udine, Italy; emails: {kevin.roitero, david.labarbera, mizzaro}@uniud.it; D. Spina, RMIT University, Melbourne, Australia; email: damiano.spina@rmit.edu.au.

1936-1955/2022/12-ART5 $15.00
https://doi.org/10.1145/3546916

media users [31, 38]. Fighting the spread of online misinformation is a multidisciplinary issue that requires both technical advances to process large amounts of false digital information and understanding of the societal context in which such spreads happen. In order to best deal with the need to both scale to a 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 [1, 8, 25].

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 [7]. Examples of successful human-in-the-loop methods include ZenCrowd [6], CrowdQ [9], CrowdDB [13], and Crowdmap [32]. Active learning methods [34] are another example in which 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 interactive machine learning (ML) [2], in which labels are automatically obtained from user interaction behaviors [19].

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 microtask platforms such as Amazon MTurk [14] or by means of alternative incentives such as, for example, "games with a purpose" [39] generally is 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 a few high-quality curated labels that have been generated by domain experts. 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 [28–30, 36].

Rather than seeing these data collection approaches as mutually exclusive, in this article 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*—that is, the estimate of the reliability of the prediction—given by either a machine or a human annotator.

In this article, we focus on the following research questions in depth:

- 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 in which 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 article is organized as follows. Section 2 discusses the related work. Section 3 details the methodology used in our study. We report our results in Section 4. We analyze the results in Section 5. Section 6 concludes the article by summarizing our findings and describing

future work. All of the code and models used in this work are available at: https://github.com/KevinRoitero/human-and-machine-confidence-in-truthfulness.

2 RELATED WORK

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

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

Poggi et al. [27] consider a complete overview of 76 state-of-the-art confidence measures for ML. Mandelbaum and Weinshall [24] discuss distance-based confidence scores in the case of neural network–based classifiers. Guo et al. [16] 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 labeled data have been collected, trained models may reflect existing bias in the data. An example of such a problem is that of "unknown unknowns"(UUs) [4], that is, data points for which a supervised model makes a high-confidence classification decision that turns out to be incorrect. 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 [17] discusses the role of confidence in the "wisdom of the crowd" paradigm. Hertwig points out how human confidence may be influenced by social interaction and the presence of others' annotations. Joglekar et al. [20] describe methods to generate confidence intervals in order to capture crowd workers' confidence and bound accuracy scores. Jarrett et al. [18] consider workers' self-assessment, investigate 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. [15]. Related to this observation, Li and Varshney [22] show that workers' annotation performance does not increase when considering the confidence scores to weight their contribution. Song et al. [35] consider worker confidence in the setting of a labeling task performed with active learning techniques. Difallah et al. [12] 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 [26] study inter-annotator agreement and show how annotation quality can be improved when considering agreement scores to aggregate labels. Aroyo and Welty [3] study the relationships between gold questions and workers' agreement stating that agreement metrics do not necessarily correlate with quality but may uncover alternative ways to label data. Checco et al. [5] discuss agreement measures applied to crowdsourcing and propose an alternative measure that is able to deal with sparse and incomplete data. Maddalena et al. [23] 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

In this work, we rely on the datasets collected by two fact-checking websites: PolitiFact and ABC. PolitiFact [40] is a collection of thousands of statements about US politicians that are manually

labeled by expert fact-checkers using a six-level truthfulness scale with values (from lower to higher): *pants-on-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*, while ABC[1] is a collection of about 500 statements about Australian politics labeled by experts using a 30-level fine-grained semantic scale and then mapped into a 3-level scale with the labels *negative*, *in-between*, and *positive*. We make use of these datasets in two different ways. We rely on the whole PolitiFact dataset to train a machine learning classifier using, for each statement, the claim and its truthfulness level. We also make use of manual truthfulness labels obtained from a crowdsourcing experiment as presented by Soprano et al. [36]. The crowdsourcing task is 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, 3 from ABC, and 2 are used as quality checks. For each statement, according to the design defined by Roitero et al. [28], workers are asked to provide a truthfulness label. In addition to the design by Roitero et al. [28], 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 6-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 3-level truthfulness scale used by the expert assessors to evaluate the statements.

3.2 Machine Learning for Truthfulness Classification

We develop a binary ML classifier for truthfulness using BERT (Bidirectional Encoder Representations from Transformers) [37], a language representation model based on performing bidirectional training of a transformer model. The core part of the model is the encoder/decoder architecture [11], which consists of different steps: the tokenization and numericalization of the input sequence followed by a set of embedding layers that 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 learning a meaningful language model. In the BERT architecture, multiple encoder-/-decoder blocks are stacked together to form the final model. This architecture allows BERT to encode the entire input sequence at once and perform two training tasks simultaneously: Masked Language Model and Next Sentence Prediction.

We relied on the HuggingFace framework[2] to fine-tune the `bert-base-uncased`[3] model introduced by Devlin et al. [10]. To fine tune the `bert-base-uncased` model for a truthfulness classification task, we rely on the PolitiFact dataset described in Section 3.1, using the text of each truthfulness statement to predict its truthfulness label on a 2 level scale. We used a stratified sampling strategy to select 1,000 statements as a validation set. Finally, we used the statements selected by Soprano et al. [36] as a test set to compare the BERT classifier with the results obtained by the crowd and used the remaining set of statements as a training set. We fine-tuned each model for 3 epochs to avoid over-fitting and using the recommended parameters as detailed in Devlin et al. [11].

To use both PolitiFact (six-level) and ABC (three-level) ground truth as binary labels for training and testing, we need to map them as binary scales. Therefore, we binarize the PolitiFact scale by mapping *Pants-on-Fire*, *False*, and *Barely-True* to *False*, and *Half-True*, *Mostly-True*, and *True* to *True*. We call this *Binarization 012-345*, in which the numbers from 0 to 5 represent the six labels from PolitiFact with increasing truthfulness value. The part before the "-" are the values mapped into 0 (False); the others are mapped into 1 (True). Concerning the ABC scale, we keep the positive and

[1] https://apo.org.au/collection/302996/rmit-abc-fact-check.
[2] https://huggingface.co/.
[3] https://github.com/google-research/bert.

negative class (we map them respectively into *True* and *False)* and remove the statements labeled with the *In Between* expert label, which is not meaningfully transformable into a two-level scale.

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. To compare against BERT, we use crowd labels aggregated by the average of the scores given by the 10 different workers who judged the same statements and we binarize human aggregated labels (originally collected on a 5-point $[-2, 2]$ Likert scale) by considering negative values as the `False Statements` class and positive values as the `True Statements` class. Since no considered statement has a mean truthfulness value of 0, we do not drop any statement (a statement with aggregated truthfulness mean equal to 0 would not indicate a binary classification decision). We remove the 20 ABC labels with an in-between value because, as before, they do not represent a binary classification decision. Here, we aggregate using the average as an aggregation function relying on previous works [28–30, 36] that showed how other aggregation functions do not improve accuracy.

Thus, we generated a dataset that contains, for a total of 160 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 BERT and Crowd Confidence

To compute crowd and machine learning confidence, we proceed as follows. For crowdsourced labels, we consider both the self-reported confidence scores for each statement and the standard deviation among the 10 crowd labels collected for each statement. We refer to these two scores, respectively, as *self-reported confidence* and *estimated confidence* scores. In this work, we focus on the comparison between human and machine learning confidence scores since an in-depth study on how crowd workers' scores correlate with each other has already been carried out in the literature [21, 28–30, 36].

Regarding the machine learning approaches, as shown by Guo et al. [16], it is well known that neural networks tend to be over-confident and that the soft-max scores computed in the last layer of the neural network should not be interpreted as a reliable confidence score. Moreover, recent work also has shown that the outcome of a neural network is dependent on the seed used for training [33]. Therefore, the soft-max scores cannot be treated as confidence scores, and bootstrapping of soft-max scores is not the best possible option to compute machine learning confidence scores. Thus, we inspect the calibration of the BERT model employed according to the procedure detailed in Guo et al. [16]. That is, we first compute the *Expected Calibration Error (ECE)* (i.e., the difference in expectation between model confidence and accuracy) and the *Maximum Calibration Error (MCE)* (i.e., the estimate of the deviation between confidence and accuracy). We summarize these results in Figure 1. As we can see by inspecting the left plot, the BERT model is originally not well calibrated, with higher *ECE* (i.e., higher red bins) and *MCE*. We recall that a model ideally perfectly calibrated should plot an identity function ($x = y$ dashed line) between confidence (x-axis) and accuracy (y-axis). Therefore, the blue bins are expected to grow following the dotted line. We computed the temperature scaling to calibrate the network such that it outputs reliable confidence scores following the methodology detailed in Guo et al. [16]. As shown in the right plot of Figure 1, after temperature scaling, the confidence scores represented by the blue bars are better distributed following the dotted line since both *ECE* and *MCE* are lower and, in general, the relation between confidence and accuracy approximates better the identity function. Thus, the calibrated soft-max scores can be used as reliable confidence scores.

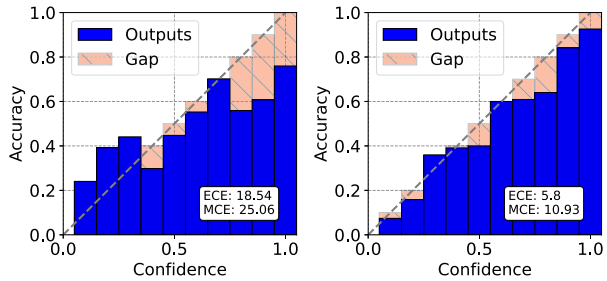

Fig. 1. Reliability diagram of the BERT model for the 012-345 binarization: the x-axis shows model confidence; the y-axis shows accuracy. Left: before temperature scaling. Right: after temperature scaling. The dashed line is $x = y$.

4 RESULTS

4.1 BERT and Crowd Accuracy

First, we report on the truthfulness classification accuracy of both BERT and crowd-based methods to label the truthfulness of statements in the dataset. As compared with expert ground-truth labels, the BERT model and crowd workers (truthfulness labels for a statement aggregated by means of sum as raw labels are in $[-2, 2]$) perform at a similar level of accuracy (BERT: 59.37%, crowd: 60.62%).

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 BERT and Crowd Confidence

To understand whether the algorithmic and self-reported human confidence scores can be used to reliably estimate the quality of truthfulness decisions, we investigate and compare their behavior when evaluating the same statements. Figure 2 shows both the BERT confidence and the self-reported crowd confidence for each statement with a breakdown on the correctly and not correctly classified statements. On the left plot, we show a pairwise breakdown between the correctness of the classification given by the crowd and ML. On the right plot, we consider the correctness only for BERT (first row) and for the crowd (second row). As we can see from Figure 2, ML confidence scores are almost always lower on average for statements in which BERT decisions are wrong and higher when BERT correctly classifies them (i.e., easy statements), even if such differences are small and not statistically significant. Moreover, both plots in the first row show a higher confidence than BERT when correctly labeling a statement. On the contrary, we see that crowd confidence is almost equally distributed across the truthfulness classifications. Also, as expected, we can see in both plots of the second row a higher confidence over incorrectly labeled statements by the crowd. Thus, answering **RQ1**, we can see that confidence scores from BERT and those self-reported by the crowd behave differently. The BERT confidence seems to be a better indicator of the correctness of the classification in respect to the crowd self-reported confidence scores. This may be a very weak signal indicating accurate classification decisions, thus, leading to risks of undetectable classification errors (i.e., *unknown unknowns*).

We have also found evidence that aggregated confidence scores are a very weak signal indicating accurate classification decisions that should not be used as it may lead to undetectable classification errors when analyzing the confidence score for BERT confidence and for both self-reported crowd confidence and estimated crowd confidence with a breakdown on statement truthfulness rather

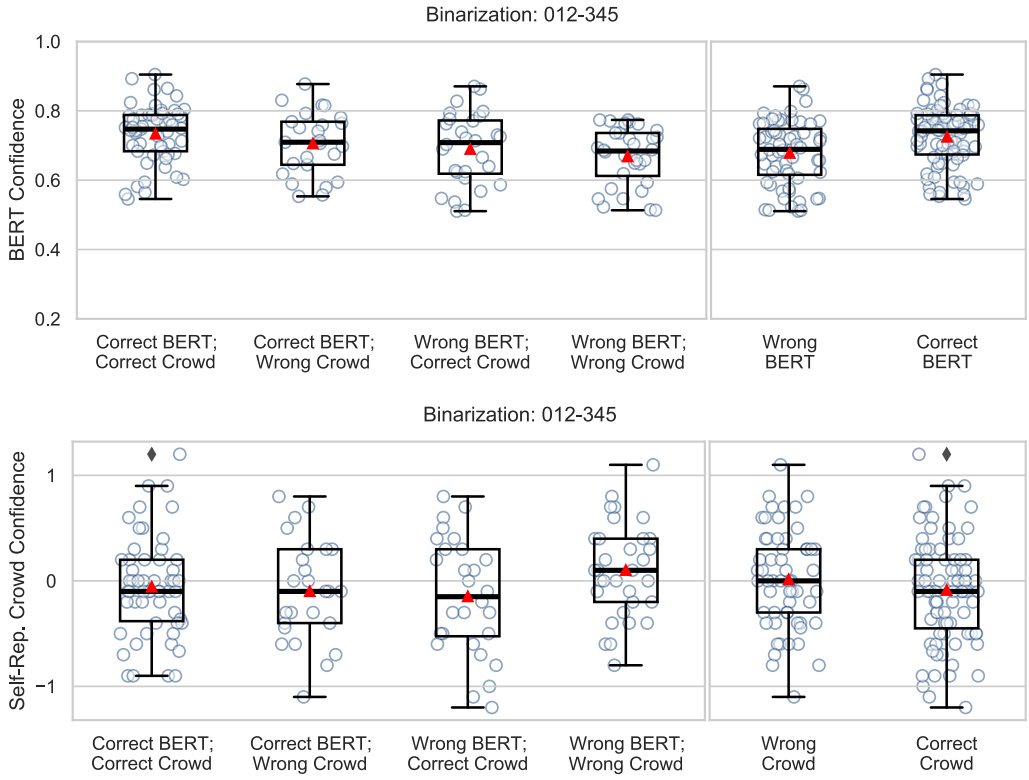

Fig. 2. ML and mean aggregated crowd confidence scores over correct and incorrect truthfulness classifications for the 012-345 binarization.

than the correctness of its classification. We found that both BERT and self-reported and estimated crowd confidence behave similarly when compared with the statements' ground truth. Thus, we do not report the plots since we found no interesting signals.

We now turn to studying the relationship between BERT and both of the considered aggregated crowd confidence scores to see whether they are correlated and whether one confidence score can act as a proxy for the other. The first two plots of Figure 3 show on the x-axis the aggregated crowd confidence scores and the BERT confidence on the y-axis. Each dot is a statement; the different colors in the plot highlight a breakdown on either correctly and incorrectly classified statements by both the BERT and the crowd. As we can see by considering both plots as a whole, both estimated and self-reported crowd confidence show a similar behavior when compared with the BERT 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. We also investigate the correlation between the confidences for each breakdown using Pearson ρ and Kendall τ. We found that a significant correlation holds only for the self-reported and BERT confidences when considering the correctly classified by both BERT and by the crowd breakdown, with a negative correlation value of -0.33 and significance of $p = 0.006$ for Pearson, and of -0.199 and 0.021, respectively, for Kendall. Since for the majority of the breakdown there is no evidence of statistically relevant correlation, we conclude that we found no clear signal on the correlation between the considered confidences. This is a further confirmation that trusting both BERT and crowd confidence scores can lead to classification errors.

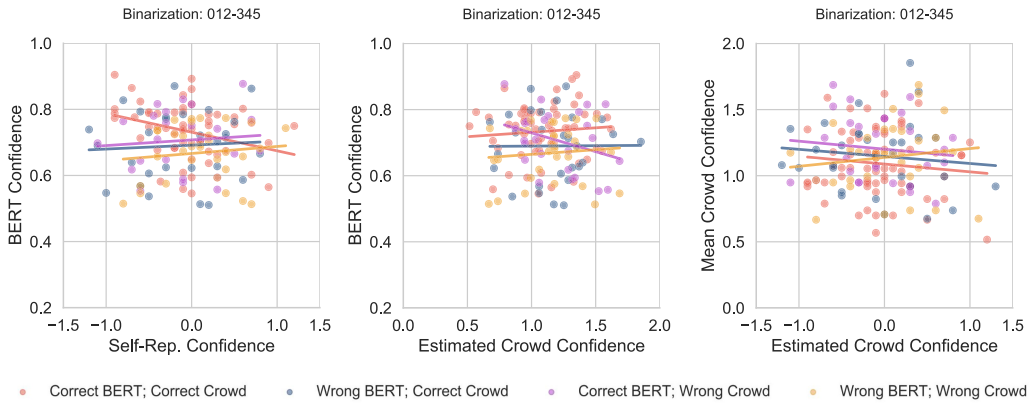

Fig. 3. BERT versus both self-reported (left plot) and estimated (center plot) crowd confidence, and estimated versus self-reported crowd confidence (right plot) for the 012-345 binarization. Breakdown on classification errors.

Summarizing the results observed so far, we can conclude that both BERT 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 mean self-reported (i.e., the scores submitted by the workers) and estimated (i.e., the ones automatically derived by considering the standard deviation of the truthfulness labels as submitted by the workers) confidence scores show very similar behavior when compared with BERT confidence scores. Thus, this set of preliminary results hints that estimated confidence scores can act as a proxy for self-reported scores if the aim is to compare them with BERT scores. Therefore, researchers and practitioners can avoid asking for self-reported confidence scores if their focus is on accuracy and comparison with BERT confidence scores, reducing the effort required by the crowd workers when performing the task.

To verify whether this conjecture holds in general, we turn to the last plot of Figure 3 to compare the mean self-reported (y-axis) and estimated (x-axis) crowd confidence scores. As we can see from the plot, while mean self-reported and estimated crowd confidence scores show a very similar behavior when compared with BERT confidence (see the first two plots of Figure 3), we can see that the two measures are not correlated and each statement shows a different self-reported and estimated score. Thus, if the focus of research and practitioners is purely on crowd confidence scores, estimated and self-reported 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 regarding effectiveness.

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., BERT or crowd) results in correct classification decisions, but the other method does not.

While these negative results hint that it appears challenging to make use of confidence scores to increase the effectiveness of such methods and identify the cases in which one of the two methods (i.e., BERT 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 that combines non-expert human and machine truthfulness classification to obtain better quality decisions. We will discuss this approach in the following.

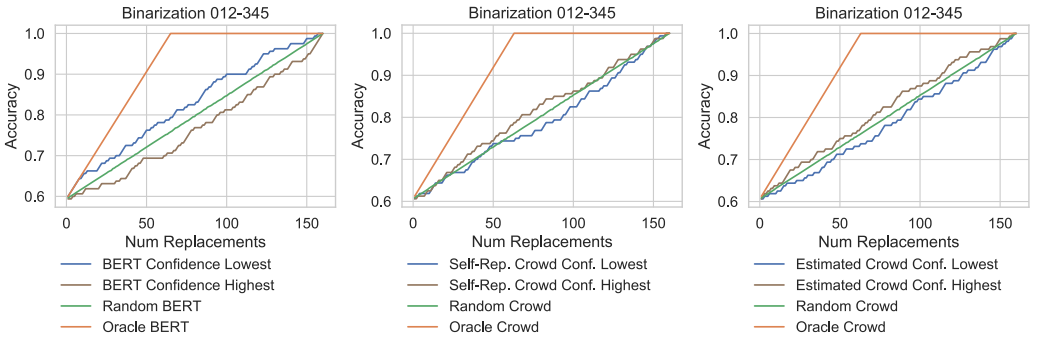

Fig. 4. BERT (left) and crowd (self-reported, center; estimated, right) accuracy after replacing their labels with expert labels for statements (i) selected by an oracle (maximizing accuracy on each replacement), (ii) with lowest and highest confidence, or (iii) uniformly at random. Binarization 012-345.

4.3 Can Confidence Be Leveraged?

Having studied the signal provided by both the BERT and crowd confidence scores, we now investigate whether such signals can be leveraged to improve classification accuracy and 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 BERT and crowd, we replace the labels (i.e., the classification decisions for statements) that have the lower/higher 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 BERT 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. With the oracle strategy, we replace each time n wrongly classified statements with the same number of correctly classified statements (i.e., we correct their classification with the correct value). With this method, we employ a strategy that maximizes accuracy. While the former baseline represents the average random case, the latter represents the optimal replacement selection strategy.

Figure 4 shows in the x-axis the number of statements that have been replaced in the original dataset and in the y-axis either the BERT or crowd accuracy scores. The four series represent the oracle, the random choice, and both of our strategies based on replacing the statements according to their confidence scores, replacing the ones with the lowest/highest confidence first, respectively. As we can see in the leftmost plot in Figure 4, the BERT accuracy is always on average more effective than the random selection strategy when the replacements are done by removing the statements with lower confidence. Moreover, we can see that the lowest confidence replacement strategy generally leads to a higher accuracy with regard to the highest confidence one. Also, all of the aforementioned series are substantially less effective than the oracle. This result suggests that there is room for improvement, which can be seen as an opportunity to study and develop novel methods to leverage confidence scores with the aim of identifying misclassified statements and improving the overall model effectiveness. We leave for future work the analysis of more sophisticated approaches based on confidence or other signals. As we consider the other two plots of Figure 4 (center, right), we can see a different type of behavior in respect to the BERT confidence. In fact, when considering both self-reported and estimated crowd confidence, the replacement

strategy (lower/higher) is more similar to the random strategy (green line) in respect to the BERT confidence case. These results are consistent with our previous observation on the lack of signal in BERT confidence scores and that of previous work [22] indicating that self-reported reliability is not accurate in crowdsourcing scenarios.

5 THE EFFECT OF SCALES BINARIZATION

The results presented in the previous sections have been obtained by analyzing a binarization of the considered scales. Let us briefly recap the set of scale transformations employed: to produce the binary BERT classifier, we binarize the PolitiFact training labels from a six-level scale to a binary {True, False} scale. Furthermore, to compare the confidence against such binary classifier, we binarize human labels from a 5-point $[-2, 2]$ Likert scale to a {True, False} scale by considering $\{-2, -1\}$ as the `False Statements` class and $\{1, 2\}$ as the `True Statements` class.

This process might introduce some artifacts in the results presented earlier in the article. In particular, given that we are comparing human and machine confidence, the most impactful binarization deals with the one employed to the training data used to produce the binary BERT classifier. To investigate this effect, in this section we train many BERT models to consider all possible binarizations of PolitiFact labels and report the effect on the results. As previously done, we use the temperature scaling technique to derive the BERT confidence scores from the soft-max layer of the BERT algorithm.

We consider the following binarizations in more detail:

- *Binarization 0–12345*: *Pants-on-Fire* is mapped to *False*; the other classes are mapped to *True*.
- *Binarization 01–2345*: *Pants-on-Fire* and *False* are mapped to *False*; *Barely-True*, *Half-True*, *Mostly-True*, and *True* are mapped to *True*.
- *Binarization 012–345*: *Pants-on-Fire*, *False*, and *Barely-True* are mapped to *False*; *Half-True*, *Mostly-True*, and *True* are mapped to *True*. This binarization is the one used in the first part of the article.
- *Binarization 0123–45*: *Pants-on-Fire*, *False*, *Barely-True*, and *Half-True* are mapped to *False*; *Mostly-True* and *True* are mapped to *True*.
- *Binarization 01234–5*: All classes are mapped to *False*, except for *True*.

The following subsections discuss the effect of binarization on the different aspects related to confidence investigated in this work.

5.1 Effect on Confidence Scores

In Figure 5 we show, similar to Figure 2, the BERT (first row) and the crowd self-reported (second row) confidence scores for the aggregated labels with a breakdown on the statements correctly and not correctly classified by either the BERT model or the crowd workers for each binarization. As in Figure 2, we show in the left plot a breakdown on the correctness pairs between crowd and BERT. On the right plot, we show the breakdown when considering only BERT and the crowd in the first and second row, respectively. As we can see, considering the results for the BERT confidence (first row), we have evidence of much higher confidence for each breakdown for the most extreme binarizations. Generally, we have a high confidence variability across binarizations when considering BERT. Therefore, we can conclude that the binarization has a major effect on the confidence scores. Instead, the results for the crowd show a lower per-binarization variability, with similar mean and median values for each boxplot and binarization (as for Figure 2, we do not report the plots for the estimated crowd confidence, since it is very similar to the self-reported confidence plot). Therefore, we can conclude that the binarization has an higher impact on the BERT in respect to the crowd confidence, particularly when considering the most extreme binarizations.

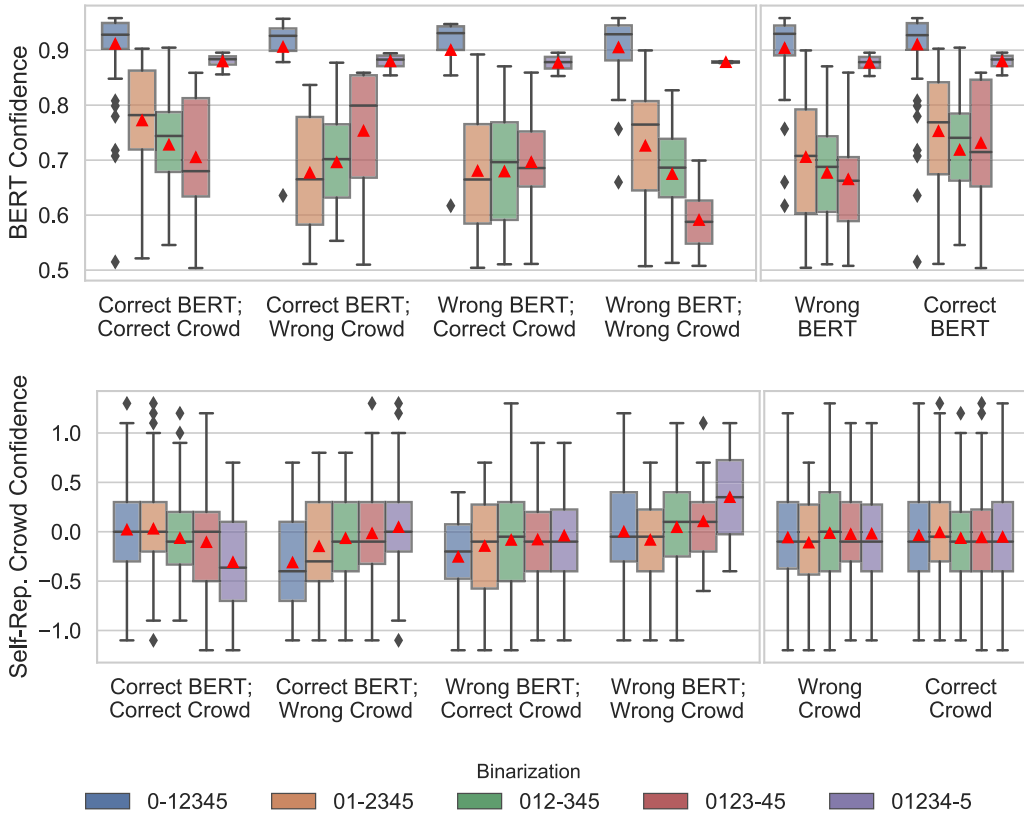

Fig. 5. BERT (first row) and crowd self-reported (second row) confidence scores over correct and incorrect truthfulness classifications for each binarization.

Furthermore, by inspecting the plot as a whole, we see that when considering the most balanced binarizations (01–2345 in orange, 012–345 in green and 0123–45 in red) the BERT confidence scores are almost always lower on average for statements in which BERT classification is wrong and higher when BERT classification is correct, even if such differences are again small and not statistically significant. This is a confirmation of what previously noted, that is, confidence scores appear to be a weak signal indicating accurate classification decisions. Thus, they should not be trusted as there is the risk that they might lead to undetectable classification errors (i.e., *unknown unknowns*).

5.2 Effect on Correlations Between Confidence Scores

Figure 6 shows, similar to the first two plots of Figure 3, the effect of binarization on estimated (first row) and self-reported (second row) crowd confidence with a breakdown on classification errors.

As we can see by inspecting the plots in Figure 6, as a whole, the effect of binarization is to limit the variance of the confidence generated by the BERT model (y-axis), whereas the self-reported (and estimated, not shown) crowd confidence remains unchanged. As it was for the original binarization, we see that the relative amount of statements correctly and incorrectly classified by both human and machine methods changes particularly for the two most extreme binarizations. Moreover, the two central binarizations (01–2345 and 0123–45) show the statements spread across

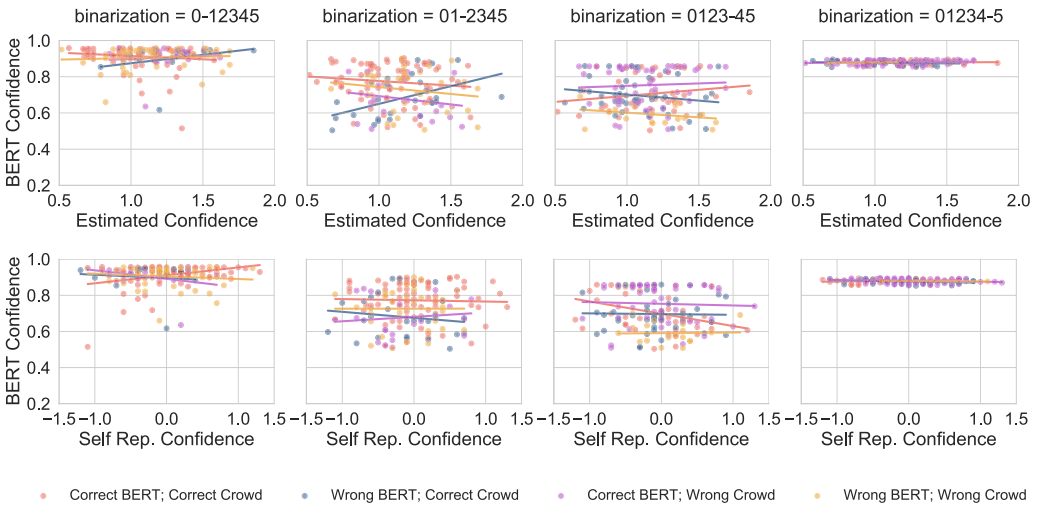

Fig. 6. BERT versus estimated (first row) and self-reported (second row) crowd confidence with a breakdown on classification errors. Each column is a binarization. Compare with Figure 3 (left and center).

the plot without a clear trend or pattern, similar to what we have observed in Figure 3 for the original binarization. Again, this is yet another confirmation that trusting both BERT and crowd confidence scores can lead to classification errors and that this result is not dependent on the binarization applied to the data. The trend is similar when inspecting the statements for which both methods have either a high (top-right) or low (bottom-left) confidence score.

As for the original binarization, we have inspected the confidences for each breakdown using Pearson ρ and Kendall τ. We found some statistically significant results:

- Binarization 0–12345 (first column): For BERT confidence versus self-reported confidence (second row) $\rho = 0.328, p = 0.002$ for the breakdown with statements correctly classified by both the crowd and BERT (orange), and for estimated confidence (first row) $\tau = 0.464, p = 0.021$ when statements are correctly classified only by the crowd (blue).
- Binarization 01–2345 (second column): For BERT confidence versus estimated confidence (first row) for statements correctly classified only by the crowd (blue), we found statistical significance for both $\rho = 0.484, p = 0.012$ and $\tau = 0.358, p = 0.011$.
- Binarization 012–345 (third column): For BERT confidence versus self-reported confidence (second row) and for statements correctly classified by both the crowd and BERT (red), we found statistical significance for both $\rho = -0.353, p = 0.018$ and $\tau = -0.253, p = 0.017$.
- Binarization 0123–45 (fourth column): For BERT confidence versus self-reported confidence (second row) and for statements correctly classified only by the crowd, we found a correlation $\tau = -0.225, p = 0.044$.

While these results are indeed significant since, as before for the majority of the breakdowns for each binarization there is no statistically sound evidence of correlation, we can conclude that the crowd and BERT confidences have no correlation and that this behavior is not affected by the considered binarization.

5.3 Effect on Accuracy

We now study the effect of the binarization on the accuracy of both BERT and the crowd. As summarized in Table 1, the different binarizations have a high impact on the accuracy for both BERT

Table 1. Accuracy for Both BERT and the Crowd for the Considered Binarizations

	Binarization				
	0–12345	**01–2345**	**012–345**	**0123–45**	**01234–5**
Accuracy for Crowd (%)	60.6	66.8	60.6	55.6	45.6
Accuracy for BERT (%)	62.5	63.7	59.4	60.0	73.7

and the crowd. Particularly, we can see the lowest accuracy for BERT in the original binarization (012–345) and the highest for the binarization 01234–5 (i.e., when the majority of the labels are set to false), whereas the crowd obtains the worst accuracy among all. These results hint at a different behavior for BERT and the crowd across the considered binarizations when evaluating truthfulness statements. We aim to investigate this in future work.

As we already did in Section 4.3 for the original binarization, we now investigate the effect on the correlation between the effectiveness of the methods used to predict the truthfulness of statements and their accuracy for all of the newly considered binarizations.

In Figure 7 we show, similar to Figure 4, the number of replaced statements in the original dataset in the x-axis and, in the y-axis, the accuracy scores for BERT (first column), self-reported crowd confidence (second column) and estimated crowd confidence (third column). The four series represent the oracle, the random choice, and our strategies (replace the statements with lower/higher confidence first). Each row is a binarization. As we can see by focusing on the plot as a whole, while the behavior of the random and oracle series is almost unchanged as expected, by focusing on the series representing our strategies, we can see some small differences. More specifically, for these series, we see that our strategies perform more like the random strategy, particularly for the BERT confidence (first column) when considering the two most extreme binarizations (0–12345 in the first row and 01234–5 in the last row), whereas the results for the other two binarizations (01–2345 and 0123–45) are more similar to the patterns observed in Figure 4. In considering both crowd confidences (second and third column) we can generally see a lower distance to the average line for both of our strategies, with the exception of the binarization 01–2345 for the estimated crowd confidence, where the difference from the random line is more evident for both strategies.

Consistent with what has already been observed in Figure 4, we can see that for BERT accuracy the strategy to replace the statements with lower confidence first performs in general worse than the random case. As before, the highest confidence replacement strategy generally performs worse than the random case. This holds across all binarizations with the exception of the central part of the plot for the binarization 0–12345. In considering both crowd confidences, we can see that the lowest confidence replacement strategy performs slightly better than the random case, as noted for the original binarization 012–345. However, there are some exceptions, for example, binarization 0–12345 and 01–2345 for self reported crowd confidence and 01234–5 for estimated crowd confidence. Finally, the highest confidence replacement strategy is always on average less effective than the random selection strategy, again, with some exceptions (estimated crowd confidence for binarization 0–12345, and both plots for binarization 0123–45).

These results point to a negative correlation between the replacements done considering the self-reported crowd confidence and the random series. It appears that replacing statements based on the self-reported confidence scores is harmful and, thus, should be avoided.

We investigated the replaced statement for such binarization, but we could not identify any peculiar feature or pattern that could justify such behavior. We tried the following features: number of words, statement length, percentage of Republican/Democratic statements, date, average time spent to assess the statement, statement source, and statement truthfulness. We leave for future work an in-depth analysis of the features of high-confidence statements.

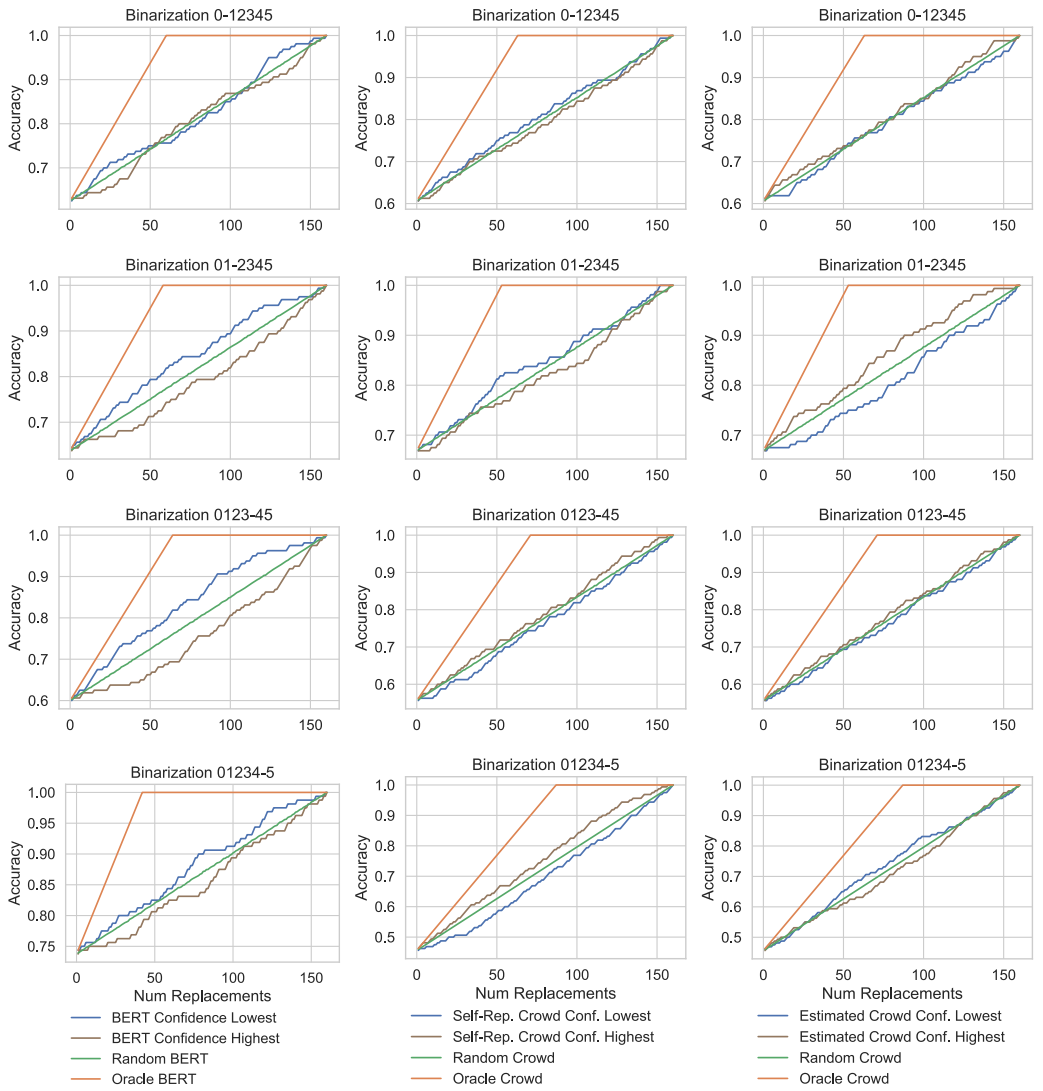

Fig. 7. BERT (left) and crowd (self-reported, center; estimated, right) accuracy after replacing their labels with expert labels for statements selected by an oracle (maximizing accuracy on each replacement, orange line) with lowest (blue) and highest (brown) confidence, or uniformly at random (green). Each row is a binarization. Compare with Figure 4.

These results together with the previous findings suggest that BERT confidence cannot act as a proxy for effectiveness and that it cannot be leveraged to increase BERT accuracy (at least not in a naïve way). More precisely, these results show that this effect is not dependent on the binarization process employed to the data used to train the BERT model.

6 CONCLUSIONS AND FUTURE WORK

In this article, we studied how BERT 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 that involves machines, non-experts (crowd workers), and

experts (fact-checkers). We focused on both accuracy and confidence of the different approaches. We looked at both the accuracy and confidence signals alone; 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 BERT and crowd confidence scores (both self-reported and estimated) are not related to effectiveness, they can be leveraged to increase the effectiveness of the misinformation detection system. In this respect – and consistent with other crowdsourcing studies in the literature [22] – estimated 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 that their predictions do not agree in general. This result opens up 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 — in particular, their confidence scores — can be leveraged to increase the effectiveness of systems when expert fact-checkers are brought into the loop in cases in which automatic ML or non-expert crowd workers are not confident on the submitted labels.

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

REFERENCES

[1] 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 (2021), eabf4393. https://doi.org/10.1126/sciadv.abf4393

[2] 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 (2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[3] Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *The 5th International ACM Conference on Web Science in 2013 (WebSci'13)*. ACM, Paris, France.

[4] 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 (2015), 1–17. https://doi.org/10.1145/2700832

[5] 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*. AAAI Press, Quebec City, Canada, 11–20. https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15927

[6] 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*. ACM, New York, NY, 469–478. https://doi.org/10.1145/2187836.2187900

[7] 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 (2017), 1–87. http://dx.doi.org/10.1561/1800000025

[8] 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 (Sep 2020), 65–74. http://sites.computer.org/debull/A20sept/p65.pdf

[9] 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'13)*. www.cidrdb.org, Asilomar, CA, USA. http://cidrdb.org/cidr2013/Papers/CIDR13_Paper137.pdf

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[11] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] 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*. ACM, New York, NY, 855–865.

[13] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: Answering queries with crowdsourcing. In *Proceedings of SIGMOD*. ACM, New York, NY, 61–72. https://doi.org/10.1145/1989323.1989331

[14] 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 (2015), 81–85.

[15] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Trans. Comput.-Hum. Interact.* 24, 4, Article 30 (Aug. 2017), 26 pages. https://doi.org/10.1145/3119930

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[17] Ralph Hertwig. 2012. Tapping into the wisdom of the crowd-with confidence. *Science* 336, 6079 (2012), 303–304. https://doi.org/10.1126/science.1221403

[18] 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 3rd IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb'15)*. IEEE, Washington, DC, 85–90. https://doi.org/10.1109/HotWeb.2015.9

[19] Thorsten Joachims and Filip Radlinski. 2007. Search engines that learn from implicit feedback. *Computer* 40, 8 (2007), 34–40. https://doi.org/10.1109/MC.2007.289

[20] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2013. Evaluating the crowd with confidence. In *Proceedings of KDD*. ACM, Chicago, IL, 686–694. https://doi.org/10.1145/2487575.2487595

[21] David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. In *Proceedings of ECIR*. 207–214.

[22] 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)*. ACM, Pittsburgh, PA, 49–54. https://doi.org/10.1145/3055601.3055607

[23] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering assessor agreement in IR evaluation. In *Proceedings of ICTIR*. ACM, Amsterdam, NL, 75–82. https://doi.org/10.1145/3121050.3121060

[24] Amit Mandelbaum and Daphna Weinshall. 2017. Distance-based confidence score for neural network classifiers. (2017). https://arxiv.org/abs/1709.09844

[25] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto BarrÃşn-CedeÃśo, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of IJCAI*. IJCAI, 4551–4558. https://doi.org/10.24963/ijcai.2021/619 Survey Track.

[26] 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'10)*. ACM, Philadelphia, PA, 557–566. https://doi.org/10.1145/1743384.1743478

[27] 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'17)*. Piscataway, NJ, IEEE 2017, Venice, IT, 5228–5237. http://doi.org/10.1109/ICCV.2017.559

[28] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of SIGIR*. ACM, Xi'an, China, 439–448. https://doi.org/10.1145/3397271.3401112

[29] 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. (2021), 31 pages. https://doi.org/10.1007/s00779-021-01604-6

[30] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland*. ACM, New York, NY, 1305–1314. https://doi.org/10.1145/3340531.3412048

[31] Lauren L. Saling, Devi Mallal, Falk Scholer, Russell Skelton, and Damiano Spina. 2021. No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. *PLOS ONE* 16, 8 (8 2021), 1–13. https://doi.org/10.1371/journal.pone.0255702

[32] Cristina Sarasua, Elena Simperl, and Natalya F. Noy. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the International Semantic Web Conference (ISWC'12)*. Springer, Boston, MA, 525–541. https://doi.org/10.1007/978-3-642-35176-1_33

[33] Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander Nicholas D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The MultiBERTs: BERT reproductions for robustness analysis. In *ICLR 2022*. https://arxiv.org/abs/2106.16163

[34] Burr Settles. 2009. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences. http://digital.library.wisc.edu/1793/60660

[35] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244–258. https://doi.org/10.1016/j.knosys.2018.07.010

[36] 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 (2021), 102710. https://doi.org/10.1016/j.ipm.2021.102710

[37] 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*. Curran Associates, Inc., Long Beach, CA, 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web* 13, 2, Article 10 (Mar 2019), 22 pages. https://doi.org/10.1145/3316809

[39] Luis Von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (2006), 92–94. https://doi.org/10.1109/MC.2006.196

[40] William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*. ACL, Vancouver, Canada, 422–426. https://aclanthology.org/P17-2067

Received 23 November 2021; revised 31 March 2022; accepted 25 May 2022