IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 13, 2020, accepted August 29, 2020, date of publication September 8, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022773

A Robust Consistency Model of Crowd Workers in Text Labeling Tasks

FATTOH ALQERSHI[1], MUHAMMAD AL-QURISHI[2], (Member, IEEE), MEHMET SABIH AKSOY[1], MAJED ALRUBAIAN[2], (Member, IEEE), AND MUHAMMAD IMRAN[3], (Member, IEEE)

[1]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[2]College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[3]College of Applied Computer Science, King Saud University, Riyadh 19676, Saudi Arabia

Corresponding author: Muhammad Imran (dr.m.imran@ieee.org)

This work was supported by the Deanship of Scientific Research through the initiative of DSR Graduate Students Research Support (GSR), King Saud University.

ABSTRACT Crowdsourcing is a popular human-based model to acquire labeled data. Despite its ability to generate huge amounts of labelled data at moderate costs, it is susceptible to low quality labels. This can happen through unintentional or intentional errors by the crowd workers. Consistency is an important attribute of reliability. It is a practical metric that evaluates a crowd workers' reliability based on their ability to conform to themselves by yielding the same output when repeatedly given a particular input. Consistency has not yet been sufficiently explored in the literature. In this work, we propose a novel consistency model based on the pairwise comparisons method. We apply this model on unpaid workers. We measure the workers' consistency on tasks of labeling political text-based claims and study the effects of different duplicate task characteristics on their consistency. Our results show that the proposed model outperforms the current state-of-the-art models in terms of accuracy.

INDEX TERMS Crowdsourcing, reliability, consistency, text labeling, fake news.

I. INTRODUCTION

Crowdsourcing has an open collaborative nature with high availability of ordinary Internet users (crowd workers) [1]. This enable crowdsourcing to provide economical micro-labeling solutions [2]. For example, text labeling of computational linguistics costs $1 million dollar for million label compared to $380k–$430k dollar when leveraging a crowdsourcing platform [3]. Therefore, many researchers resort to crowdsourcing as a labeling choice. Consequently, their research incorporate with the crowdsourcing, for example, responding to Covid-19 pandemic [4] and disasters [5], detecting fake news [6], and deep learning applications [7], [8]. One major issue in crowdsourcing is quality control [9], [10]. This issue is rooted in the human-based nature of crowdsourcing [11]–[13]. Reliability is one quality concern that examine the crowd workers' trustworthiness. The crowd workers can be unintentionally ill-qualified [13], or they may give incorrect answers intentionally to increase their income. The identification of reliable workers is hence a key issue in any crowdsourcing system. This identification is commonly achieved by evaluating worker output using a gold standard [14]–[16] and by using consensus methods such as majority voting [17]–[20]. Other reliability measurements include worker-based ones that mainly depend on monitoring worker behavior indicators such as interaction events [21], [22], eye tracking [23] or time-based activities [24]. Additionally, reliability could be estimated by measuring the worker's effort in a task [25].

Consistency analysis is one of these reliability worker-centric measurements. It concerns of examining the workers' ability to adapt to themselves by assuming the same result when repeatedly given a same task. Research on consistency (intra-annotator reliability) based evaluation, where workers are evaluated on the consistency of their own answers, is ongoing. Such research will open new directions in evaluating crowdsourcing workers and enable further investigations on various factors. These factors include the workers' consistency across their answers, the effects of repeated or difficult tasks on workers' consistency, and the relationship between the accuracy and consistency of workers. Moreover, consistency-based quality control can be compared to

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenyu Zhou.

traditional approaches for quality control. Despite the interest in consistency as a reliability measurement in various fields such as healthcare, pervasive computing, and machine learning [26], consistency-based quality control in crowdsourcing has not yet been investigated extensively.

Inter-annotator reliability, i.e., consistency between workers, is used to evaluate workers by comparing their results with the results from their peers [27]. Targeting intra-annotator consistency, [28] used time limitations and the workers' errors to evaluate the workers in different types of tasks. The authors in [29] studied worker consistency over the long-term. Reference [30] used pattern recognition estimation of the consistency of data annotators based on their annotations on similar images. Exploring the consistency of relevance judgments is studied in [31]. They examined different factors that affect the judgment such as distance between the duplicated documents and the topic of the documents. Other work, [32] explored the consistency of participants in three replicated surveys by asking personal information and motivation. The consistency in [26] is measured using the absolute errors of workers counting objects in duplicate images. Another work applied inconsistency score measure by duplicating randomly set of questionnaire questions twice [33]. They used the weighted Euclidean distance to measure the consistency.

The contributions of this article can be summarized as:

- A main contribution lies in its proposal of a novel reliability model for crowdsourcing based upon the consistency of workers. This novelty is represented by applying the pairwise comparison method instead of using traditional distances calculations. Also, the using of multi duplicates of text labeling tasks compared to only single task duplicate. Moreover, in contrast to previous work, we study a different pool of workers, namely workers with intrinsic motivation rather than paid workers. Furthermore, in term of performance, our model achieves an average accuracy outperforms other competing methods [26], [33].
- Other contributions are the results and findings that reveal the consistency level of the unpaid workers and illustrate the effect of different task factors on the worker consistency. Furthermore, the dataset which is the first available dataset of consistency in crowdsourcing.

This article is organized into eight sections. Section 2 provides related work. Section 3 describes the problem formulation. The proposed model is illustrated in Section 4. Section 5 presents the experimental results. Section 6 analyses the performance of the proposed model. Section 7 discusses future works. The last section is the conclusion.

II. RELATED WORK

A. WORKER RELIABILITY

Traditional methods for measuring crowdsourcing workers include those based on simple human evaluations. Workers are evaluated by normal workers who are independent validators chosen to assess the answers of other crowdsourcing workers [34]. A more common approach is the majority vote where multiple workers work on the same tasks and the correct answer is taken as the one with the majority vote [29], [31], [36].

The use of gold standards (ground truths) is another popular approach where high-quality answers or labels are already known. The reliability of the workers can be evaluated by comparing their answers with the gold standard. These datasets can be created by injecting a few ground truth labels from experts in rich crowd labeling [38] or by gathering a set of experts [14]. The datasets can also be generated automatically from just a few gold unit seeds [15]. Moreover, A real time system can evaluate crowd workers reliability using a collected reference set [16].

A recent approach is worker behavior analysis, where the workers' quality is measured by tracing the behavior of the workers as they perform their tasks. The authors in [21] proposed a task fingerprinting approach based on recording sequential logs of interface events of what the workers did and when. Similar work was performed in [39], which presented an approach called "Application Layer Monitoring" and studied three time aspects (completion, working phases, and consideration). Others works analyzed the behavior of workers at different times [40] or in terms of personality traits [41].

B. CONSISTENCY BASED MODELS

Consistency-based research in crowdsourcing is still limited. Peer-consistency, for example, is used as an alternative to the gold standard for evaluating workers by comparing their results with their peers using a bonus as a motivator [27]. Focusing more on consistency, [28] evaluated the consistency of workers in different types of tasks with time limitations and compared the number of errors made by the workers. Focusing on intra-annotator consistency the time length taken by the workers to complete tasks, [29] found that workers gave consistent answers over long-term settings. The study in [32] found that 30% of participants in a survey were inconsistent when they took the same survey twice. Work [30] used galaxy images annotation data to study the annotators consistency. They compared the labels of workers for the same image as binary scale. They recommended their method to enhance the quality of training data as input for supervised machine learning algorithms.

Investigating how accurately workers judge the relevance of duplicated documents, [31] found a high level of inconsistency. They studied the possible sources of errors such as documents' topics length and distances between documents. They found that less distance and leads to high consistency and assumed that extremely long topics do reduce worker inconsistency.

Another study [26] explored consistency as a reliability measurement in crowdsourcing by using the absolute errors of workers counting objects in duplicated images. They studied the effects of different factors on task consistency. They found that, generally, the difficulty of the task decreased the

consistency, image transformation had no significant effects on consistency, and increasing the offset between duplicate images decreased the consistency. Other work [33] applied inconsistency score measure on psychometric questionnaire. They duplicated a set of questionnaire questions twice and calculated a weighted Euclidian distance of workers' duplicated answers. Their point Likert scale ranging from 1 to 7. Their method detects only 31% of the invalid responses.

Pairwise comparison method extensively considered in various other domains such as operations research, economics, engineering. Its' main application is a multi-criteria decision making tool. It supports in evaluating the decision makers and ranking alternatives [42]. To the best of our knowledge, this is the first work that employs a pairwise comparison to measure the crowd workers' consistency.

Our work differs from prior work in several dimensions. We implemented a more advanced consistency measurement (pair-wise matrix), studied a text labeling task, using multi duplicates of same tasks, and targeted workers with intrinsic motivations rather than paid crowd workers.

III. PROBLEM DEFINITION

The consistency reliability measurement of crowdsourcing workers should be definable. We define the problem of measuring the worker consistency in this section.

The set of workers who participate in the labeling is formally defined by a vector:

$$w = \{w_1, w_2, \ldots, w_n\},$$

where n is the total number of workers. These workers process a set of statements formally defined by a vector:

$$s = \{s_1, s_2, \ldots, s_n\},$$

where n is the number of statements. Each of these statements has three duplicates. Each statement duplicate SD has a set of characteristics:

$$SD = \{SDp, SDd, SDr\},$$

where p is the placement, d the difficulty, and r the rephrasing. The total number of statements is $n * size(SD)$. These statements are queued randomly to the workers who are asked to label the statements. The labels can be binary or, in our case, fall inside a set:

$$l = \{l_1, l_2, \ldots, l_n\}$$

where n is the number of labels. Each statement s should have a label l given by a worker w.

If we assume that the label is binary $0, 1\}$ with just a single duplicate, then the worker consistency is measured as follows:

While worker w_i is still labeling statements, we check if his/her labeling of a statement $l\left(w_i s_j\right)$ matches his/her labeling of the duplicate $l\left(w_i sd_j\right)$. If the labels are matched, then worker w_i is consistent for this statement. Otherwise, he/she is inconsistent. This comparison is repeated until the worker finishes all the statements.

In our case, there is scale of labels and three duplicates. The measurement will be as follows:

While the workers are still labeling statements, for each statement s_j labeled by worker w_i, the *Pairwise Matrix* is calculated using the pair-wise errors (differences) between the labels of the statement and its duplicates $l\left(w_i s_j\right), l\left(w_i sd_{pj}\right), l(w_i sd_{dj}), l(w_i sd_{rj})$. This gives six differences. These differences and their reciprocals are written as matrices. The *ConsistencyRatio* for this worker w_i is then calculated and then compared with the *ConsistencyThreshold*. If the consistency ratio is less than the threshold, then worker w_i is consistent, otherwise he/she is inconsistent. This is repeated for all the workers.

There are some factors that affect the consistency of the workers. We ask a few research questions about these factors. The first factor is the type of the workers. In our study, the workers were unpaid volunteer workers. We thus ask the first Research Question 1:

[RQ1] Will unpaid workers achieve higher consistency results?

Other factors are related to the three duplicates with different characteristics of the task performed by the worker. In contrast to [26], we used a text-based task with three defined factors. The first factor is the placement, which is the position of the statement in the queue. The second factor is the difficulty, where we provided less information about the claim and no clear judging rule. The third factor is the phrasing, where we changed the claim by rephrasing the statement. We ask Research Question 2:

[RQ2]: What is the effect of each of the three factors on the workers' consistency?

IV. PROPOSED MODEL

The proposed model comprises of a few main components that work in an algorithmic manner as explained in Figure 1. The first component is data collection, where we scrapped a fact-checking service and stored the scrapped data in a database. The second component is the design of the tasks that will be delivered to the workers later. The third component is our consistency algorithm. The input to this algorithm is the workers' labels which are already stored in the database. The output of the algorithm is a set of matrices from which the consistency ratios are calculated and then used as input for analysis. These components are described in more detail in the following:

A. DATA COLLECTION

Fake news is a recent phenomenon in social media and requires more fact-checking efforts to counter it [43]. Political data, like statements from politicians, is one type of data which is susceptible to fake news [44]. We scrapped PolitiFact[1] [45]–[47], which is a platform that provides a fact-checking service called Truth-O-Meter presenting truth ratings of claims from politicians based on investigations

[1] https://www.politifact.com/

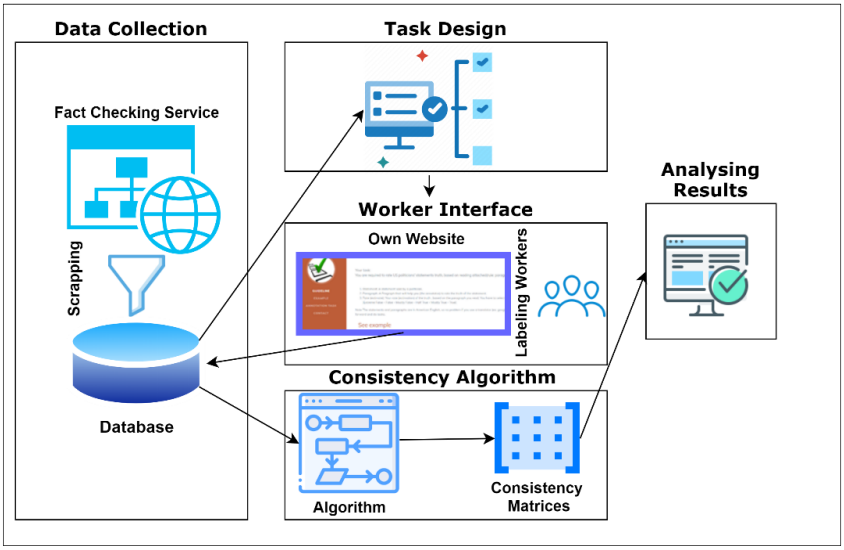

FIGURE 1. Our proposed model.

by journalists. We randomly selected eighteen statements from three politicians (Barack Obama, Hillary Clinton, and Donald Trump) related to different topics such as personnel matters, taxes, healthcare, and the military. These statements along with three duplicates made up a total of seventy-two statements. The total set of labeled statements comprised 792 statements collected from 11 unpaid volunteer workers. We followed the PolitiFact scale for the truth of each statement. This is a six-level scale to represent the degree of truth, namely [Extremely false, False, Mostly false, Half true, Mostly true, and True]. We selected this scale to serve as a ground truth to be used later for accuracy measurements, and for ease of adaption to our pairwise method design. For the original 18 statements, we tried to balance the categories. There are 3 extreme false, 4 false, 3 mostly false, 3 half true, 3 mostly true, 2 true. About the total 792 statements, since each 18 core statements duplicated four times, so each worker of the 11 ones is asked to label 72 statements. They categorized as 12 extreme false, 16 false, 12 mostly false, 12 half true, 12 mostly true and 8 true statements. To share the dataset with the scientific community, we make it publicly available at: https://github.com/fattoh/Politi_Stat.

B. TASK DESIGN

One approach to fact-checking is to fact-check individual claims [43]. Crowdsourcing tasks can be used to classify such claims or statements [44].

The task in the experiment began with a set of guidelines. The presence of such instructions increases the reliability of the workers [48]. An illustrative example was provided as part of these guidelines, since this is a recommended practice [49]. After reading the guidelines, the workers could proceed to label the statements. The statements were shown sequentially with a judgement rule for each statement that was the same as the rule in the PolitiFact service. The rule gave a summary of facts, statistics, or research studies about the statement to provide the worker with evidence to support his labeling.

TABLE 1. Illustrative example of a statement and its duplicates.

Statement type	Statement	Rule
SO Original	Not one of the 17 GOP (Grand Old Party, Republicans) candidates has discussed how they'd address the rising cost of college.	While some Republican hopefuls haven't had much to say on the issue, it's not accurate to say none of them have. Rubio has made higher education spending a major plank of his campaign, and other candidates like Christie and Fiorina have set forth ideas and positions. Other candidates have at least mentioned the subject at times.
SD1 Placement	*Original statement with offset.*	*Same judgment rule as original.*
SD2 Difficulty	Not one of the 17 GOP candidates has discussed how they'd address the rising cost of college.	Republican senator Marco Rubio has been talking about the issue since at least February 2014, when he made a major policy speech about controlling crippling college debt. He repeated his positions in a July speech after he declared his candidacy.
SD3 Rephrasing	About controlling college costs, it is disappointing, but not surprising, that no one of GOP candidate has talked.	*Same judgment rule as original.*

The last page was a set of questionnaire questions to collect feedback on the task and experiment.

We created three duplicates for each of the eighteen original statements. The order of the original statements was

manually seeded and the other duplicates were then randomly distributed. The original statement (SO) was the raw statement with the judgement ruling. The first duplicate (SD1) was the same as the SO but with a position offset that determined the distance between SO and SD1. This offset was determined randomly. The second duplicate (SD2) was a more difficult task. We replaced the judgment rule with some inconclusive clues about the statement by editing some paragraphs from the statement discussion on the PolitiFact service. The third duplicate (SD3) was a rephrased statement. An example of statement with its duplicates is shown in Table 1.

The PolitiFact scale is converted to corresponding numbers as (Extremely false $= -7$, False $= -5$, Mostly false $= -3$, Half true $= 3$, Mostly true $= 5$, True $= 7$). This scale is selected following the reference scale (PolitiFact scale) and with choosing small values according to [50]. We also chose this assignment to ensure larger distances at the scale extremes. This design allows implementing such tasks using the traditional crowdsourcing platforms like Amazon Mechanical Turk (AMT)[2]. The participants in this experiment were unpaid volunteers. They are a PhD candidates in the College of Computer Science at King Saud University. These students were selected from the pool of high graduate students who have reasonable background of crowdsourcing, where they have performed crowdsourcing tasks before. They motivated using the social human interaction in academia as a community internist motivation [51]. About biasness, since that background could affect the truthfulness [52]. We expected an unbiasedness according to their lower interesting in politics as they told in the post-questionnaire.

C. WORKER INTERFACE

We built our own task website [53] to perform several experiments, as shown in Fig 2. The experiment mainly studied the consistency of unpaid workers labeling a set of US politicians' claims and how different factors affect the consistency. The website consists of: (i) a set of guidelines to help the worker in his labeling as shown in Fig. 2(a), (ii) the tasks comprising the claim statement and the guiding rule with the labels given as radio buttons as shown in Fig. 2(b), and (iii) a final post-questionnaire about the difficulties encountered and general comments.

D. PAIR-WISE CONSISTENCY ALGORITHM

Since worker consistency has a more pronounced impact on the annotation task than any other element, it was necessary to inject some random elements into each task before the annotation process was started. Our thorough examination of this research problem revealed that the most impactful factors are the difficulty of the statement, followed by the offset of injected statements, and finally the rephrasing of the statements. The ranking process was further complicated by the fact that some of the statements were qualitative and hence could not be evaluated with fully automated methods. To counter this, a single instance of external knowledge importation was made in which a person with the relevant expertise generated the ground truth matrix that described the impact of every considered statement.

The pairwise comparison method was chosen to evaluate the worker consistency. After the evaluations were made for each pair of statements, the outcome was recorded into the matrix. Because by definition, the difference cannot be taken between each statement and itself, the diagonal dimension of the matrix was populated exclusively with '1' values. Otherwise, the direct comparison indicates if a certain statement was rated by the worker to be more true or false than the statement compared against. For example, if the workers evaluate a statement by giving it the value of S, then this indicates that the statement contributes to only $1/S$ of the predicted value that the second statement can provide. The entire matrix was populated with paired values obtained in this fashion, allowing for precise understanding of the relative comparison for each statement.

Once this matrix was fully filled, a set of priority vectors was determined through the following mathematical procedure: The maximum combined value of the entire set was estimated based on the matrix eigenvectors, after which the matrix was normalized by having each field divided by the summarized value and the priorities were formulated as vectors, as exemplified in Algorithm 1. The process was cyclical, and direct comparisons were made until all possible couplings of the statements have been exhausted. The algorithm assumes the perspective of an unbiased worker who is making rational evaluations based strictly on the outcomes of the pairings. Because such an annotator would have to make precisely defined choices and aim to not contradict himself, a certain number of statements can be reordered, paraphrased or rewritten and added to the dataset to measure his/her consistency easily. This analogy allows us to formulate the ground truth matrix in such a way that its consistency is ensured at a high level, although some contradictions may still occur due to various unintended events. For this reason, an earlier developed metric called the consistency ratio (CR) was introduced into the model and used to optimize the matrix. This variable was calculated starting from a simpler measure known as the index of consistency (CI), which was compared with a random index (RI) to find the appropriate ratio. RI have computed and obtained depending on a simulation of random pair-wised matrices [54]. This approach was instrumental for the identification of the eigenvector with maximal value. In effect, this resulted in the creation of a symmetrical matrix that was guaranteed to have the maximum consistency under the circumstances. A realistic limit of CR>0.1 was implemented according to [54], [55], and any variations that resulted in a value above this limit was eliminated from consideration. The exact process of obtaining CR values from the available data is presented in the Algorithm 1.

Example 1: An illustrative example of Algorithm 1.

[2]https://www.mturk.com/

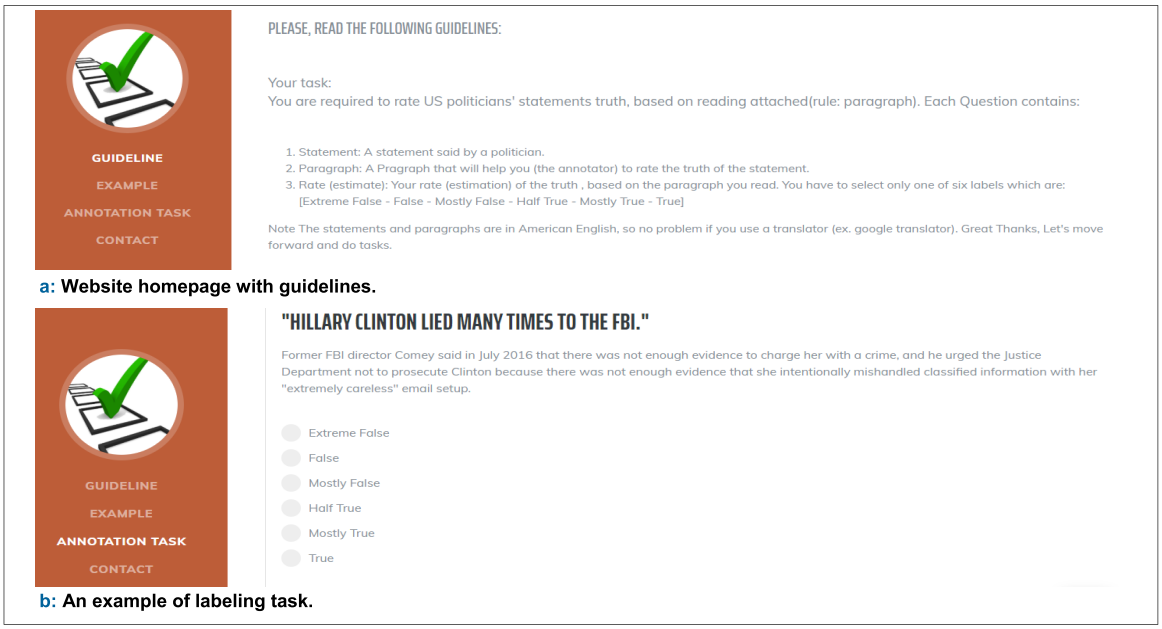

FIGURE 2. Our labeling website interface.

The pair-wise consistency matrix S_{ij} for Worker w and Statement S is

$$S = \begin{pmatrix} 1 & S_{12} & S_{13} & S_{14} \\ S_{21} & 1 & S_{23} & S_{24} \\ S_{31} & S_{32} & 1 & S_{34} \\ S_{41} & S_{42} & S_{43} & 1 \end{pmatrix} \quad (1)$$

As discussed above, the diagonal elements of the matrix must be 1, and the matrix must satisfy the reciprocal relation $S_{ij} = 1/S_{ji}$. To determine the pair-wise consistency values in the matrix, we calculated the absolute distance/difference (i.e. ignoring the sign) between the ratings of the two compared statements. The difference will be stored in the matrix as S_{ij} and its reciprocal location will equal $S_{ji} = 1/S_{ij}$. If the difference is zero, then the value in the matrix will be set 1 to reflect the perfect consistency.

For example, consider a worker who rates the original statement S_1 as 'Mostly False', which corresponds to -3 in our scale, and rates its offset duplicate S_2 as 'Half True' which on the scale corresponds to 3. Then the value $S_{12} = |S_1 - S_2| = |-3-3| = 6$ then $S_{21} = 1/6 = 0.167$. The same worker rates the second duplicate S_3 (the difficult duplicate) as 'Mostly True', i.e. 5 on the scale. So $S_{13} = |-3-5| = 8$, and subsequently $S_{31} = 0.125$. Now S_{23}, the difference between the offset statement and the difficult duplicate $S_{23} = |S_2 - S_3| = |3-5| = 2$, and so $S_{32} = 0.5$. If the last duplicate S_4 (rephrasing) is rated as 'True', then $S_{14} = 10$. This the initialization of W and so on S (line 1). By continuing likewise until the pair-wise matrix values are filled, the matrix according to (1) will be:

$$S = \begin{pmatrix} 1 & 6 & 8 & 10 \\ 0.167 & 1 & 2 & 4 \\ 0.125 & 0.5 & 1 & 2 \\ 0.1 & 0.25 & 0.5 & 1 \end{pmatrix}$$

Following Algorithm 1, (lines 2,3,4) the values in each column of the pair-wise matrix S are summed:

$$S_{ij} = \sum_{i=1}^{n} S_{ij} \quad (2)$$

The summations of the columns according to (2) are 1.39, 7.75, 11.50, and 17 respectively. The (S normalized) matrix named T is calculated by dividing each element in the matrix by the summation of its column:

$$T_{ij} = \frac{S_{ij}}{\sum_{i=1}^{n} S_{ij}} \quad (3)$$

For example, $T_{11} = {}^{1}/_{1.39} = 0.7189$ and (line 5) the resulting T referring to (3) is:

$$T = \begin{pmatrix} 0.7189 & 0.7742 & 0.6957 & 0.5882 \\ 0.1198 & 0.1290 & 0.1739 & 0.2353 \\ 0.0898 & 0.0645 & 0.0870 & 0.1176 \\ 0.0719 & 0.0323 & 0.0435 & 0.0588 \end{pmatrix}$$

To calculate the weighted vector W, we divide the sum of each row in the normalized matrix by the number of statements ($l = 4$) (averaging):

$$W_{ij} = \frac{\sum_{j=1}^{n} T_{ij}}{l} \quad (4)$$

Algorithm 1 Measuring Worker Consistency

Input: W, a list of the annotators $W = w_1, w_2, w_3, \ldots, w_m \in SR^{n \times n}$, where S is Statement Choice Matrix of worker w
β the threshold of consistency level
Output: W_{cons}, a Consistency ratio matrix

1: **Initialize** W and S
2: **for each column** $c \in S$
3: **Calculate** sum of c with respect to each row r
4: **end for**
5: **Average** values of S matrix over sum(c) and store the result in new matrix $T \leftarrow normalised(S)$
6: **Multiply** normalized T by **sum** of average rows and store the result in vector V
7: $l \leftarrow len(V)$
8: **for each** v $\in V$
9: **Calculate**μ^T from T with respect to V
10: **Find** Eigenvalue $\lambda_{\mathbf{max}} \leftarrow$ argmax μ^T
11: **end for**
12: λ_{max} Should be close to l
13: $C_{index} \leftarrow \frac{\lambda_{\mathbf{max}} - \mathbf{l}}{l-1}$
14: **Find** $W_{cons}(C_{ratio}) \leftarrow \frac{C_{index}}{0.9}$
15: **If** $W_{cons}(C_{ratio}) \leq \beta$ **then**
16: Worker w is consistent for S
17: **Else**
18: Worker w is NOT consistent and results should be refined

Starting by the numerator in (4):

$$\sum_{j=1}^{n} T_{ij} = \begin{pmatrix} 2.7766 \\ 0.6580 \\ 0.3589 \\ 0.2064 \end{pmatrix}$$

The weighted vector W_{ij} :

$$W_{ij} = \begin{pmatrix} 0.6942 \\ 0.1645 \\ 0.0897 \\ 0.0516 \end{pmatrix}$$

We store l as length of this vector $= 4$ (line 7) to be used later (line 12). After that, we compute the consistency matrix μT_j by multiplying the pairwise matrix S with the vector W divided by the weighed sum vector w of each row (lines 8,9,11):

$$\mu T_j = \frac{1}{w_j} (S * W) \tag{5}$$

Continuing in example and according to (5):

$$\mu T_j = \begin{pmatrix} 4.199 \\ 4.049 \\ 4.034 \\ 4.012 \end{pmatrix}$$

The λ_{max} is calculated using $argmax\ \mu^T$ (line 10) as

$$\lambda_{max} = \frac{\sum_{j=1}^{n} \mu T}{l} \tag{6}$$

Following (6):

$$\lambda_{max} = 4.074, \text{ which is close to } l.$$

Finally, we compute C_{index} and W_{cons} as:

$$C_{index} = \frac{\lambda_{max} - l}{l - 1} \tag{7}$$

$$W_{cons} = \frac{C_{index}}{0.9} \tag{8}$$

where W_{cons} is the consistency ratio CR that determine if the worker consistent in this statement or not compared to the threshold.

And so (lines 13,14) according to (7) and (8):

$$C_{index} = \frac{4.074 - 4}{3} = 0.0247$$

and

$$W_{cons} = \frac{0.0247}{0.9} = 0.0274$$

where 0.9 is a random index for our case corresponding to a 4×4 matrix, $n = 4$ in [54]. For this statement (Matrix S), the worker has $W_{cons} = 0.0274$, this is ≤ 0.1 which is our threshold for every statement ($\beta =$ realistic limit of CR). This indicating that worker is consistent for this statement (lines 15-18). For the total consistency for this worker, our threshold β is the average W_{cons} of all workers for all 72 statements.

Albeit the time complexity of the proposed algorithm is not a big concern. The summation of columns of matrix S at lines (2,3) is $O(n^2)$. Then getting the matrix T by normalized (averaging) S at line 4 is also $O(n^2)$. Then at line 5 the complexity of multiplying a matrix by Vector W is $O(n^3)$. Finally, the complexity of the eigenvector $\mu^T = O\left(n^4\right)$ at lines (7-9) requires $l * n * n * l$ and $n \approx l$. So, the running time is $2\left(n^2\right) + \left(n^3\right) + \left(n^4\right)$ and consequently, the time complexity $= O\left(n^4\right)$.

To study the effects of placement/offset, difficulty, and rephrasing, we used the pair-wise comparison method in algorithm 1 with n = 3. For each factor investigated, we excluded the factor to study the effect of exclusion. For example, to study the effects of difficulty, we excluded the difficult statements from the matrix and computed the consistency index and ratio and then compared them with the consistency index and ratio of the complete 4×4 matrix. This procedure was repeated with the offset and rephrased statements.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTS SETTING

All experiments were implemented on a PC with Intel Core i7-3770 CPU @3.40GHz and 12GB memory. The development used were Python 3.8 language, Django web framework, and SQLite database.

B. GENERAL OBESRVATIONS

To evaluate the total performance of the workers, we compared the average performance of their labeling against the

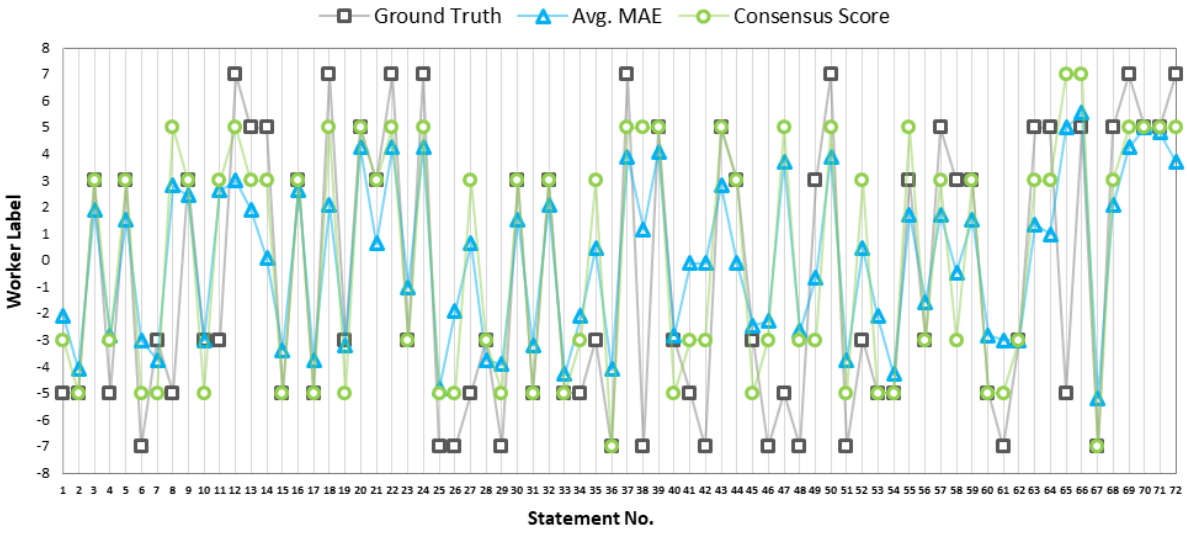

FIGURE 3. Avg. MAE and consensus score for workers compared to the ground truth for all statements.

ground truth. We used the Mean Absolute Errors (MAE) and a consensus-based measurement to measure the worker accuracy. For the MAE measurement, we calculated the mean absolute errors/distances between the ground truth and the label of each statement (original and duplicates) as follows:

$$MAE(s) = \frac{1}{n}\sum_{i=1}^{n} |GT - l\,(s_i)\,| \tag{9}$$

where GT is the ground truth, $l(s)$ is the statement label given by a worker, and n is the number of statement's duplicates $n = 4$. Then for each worker, we calculated an accuracy score from the mean of the MAE. across all core 18 statements via

$$MAE(w) = mean(MAE\,(s_0)\,, MAE\,(s_1)\,, \ldots, MAE\,(s_m)), \tag{10}$$

where m is the number of core statements.

The consensus-based measurements is ranging from simple majority voting by the workers up to complicated statistical and machine-learning models [37]. These methods are mainly helpful in cases where the ground truth is absent [56]. We used the consensus measure proposed by [57].

We scored the workers based on the absolute difference between the worker's label of a statement and the median label of all other workers for the same statement:

$$Consensus_{Score}(w, s) = |l\,(w, s) - median(l\,(w_1, s)\,, l\,(w_2, s)\,, \ldots, l\,(w_m, s)\,|) \tag{11}$$

where m is the number of workers. Subsequently, the worker's score is the median of all statements' scores in (11) as:

$$Consensus_{Score}(w) = mean(Consensus_{Score}(w, s_1), \ldots, Consensus_{Score}(w, s_n)) \tag{12}$$

where n is the total number of statements.

To experiment other measurement, we establish Consistency Baseline measure Cb [26] where Cb calculate the absolute error/difference between the worker label (as scale) of the original statement and worker the label (as scale) of the duplicate statement. In this work, we defined three baseline measures for each worker. We calculated as follows:

$$Cb_p = |label\,(SO) - label\,(SD_p)\,|, \tag{13}$$
$$Cb_d = |label\,(SO) - label\,(SD_d)\,|, \tag{14}$$
$$Cb_r = |label\,(SO) - label\,(SD_r)\,|. \tag{15}$$

The baseline consistency measure is the sum of these three measurements:

$$Cb_{total} = Cb_p + Cb_d + Cb_r \tag{16}$$

We calculated the mean baseline consistency and mean consistency for the three duplicates for each worker.

We used the mean of the (MAE) and consensus score of the workers for all the statements as shown in Fig. 3. We observed that there is an approximate uniformity of the performance in the mean MAE, consensus score, and ground truth across all the 72 statements. This gives a general indication of the quality of their work. There were no random labeling or extreme differences in labeling that affected the average performance. It was expected that unpaid workers would label more honestly, compatible with [2], [58]. Also, we observed that there were no extreme judgements of Extreme False or True, and the workers always labelled away from the extreme judgements. This could be explained by their diminished confidence as they were not sure 100% about the truth of each statement.

C. CONSISTENCY SCORES

With respect to answering RQ1, we found that as expected, unpaid workers achieved very high consistency scores. From

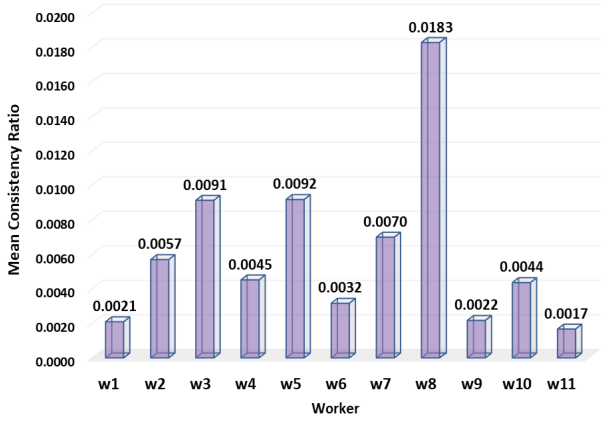

FIGURE 4. Mean consistency ratio (score) for each worker.

Fig. 4, we observed that all workers attained very low consistency ratios compared to the supposed realistic limit of 0.10. Even the worker with the worst score, worker 8, had a score of around 0.0183 that was still far less than the limit. This means that all workers achieved high consistency.

About Inter-annotator consistency, i.e. the consistency between workers, since we have more than two workers, we used Fleiss' Kappa which is a measure of agreement between multi-workers.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{17}$$

where this measure divides the degree of agreement that is attainable above chance, by the degree of agreement actually achieved above chance. Our result $\kappa = 0.1$. This is slight agreement. This is could be interpreted by the problem of the underestimation of agreement of Fleiss' kappa statistic in assessing high levels of inter-raters agreement as [59] argued. The high levels of agreement in our results are shown in Figure 3 where average consensus score is near gold truth in most of statements. We moreover, reduced the scale to binary [True, False] by merging the categories of the scale and rising κ to 0.27, which is Fair agreement.

D. EFFECTS OF THE FACTORS

In this section we present the results related to RQ2. The effects of the three factors on consistency were explored by comparing the overall mean consistency ratio against the means when each factor was absent. As the consistency ratio for all workers were nearly zero skewed, we normalized the data for all the ratios. We found the mean consistency ratio for all duplicates from all the workers to be 0.241.

The mean consistency ratio of the pairwise matrix without the placement duplicates was 0.246. This is very slightly larger than the mean overall consistency ratio. This indicates that the absence of placement duplicates did not have any noticeable negative effect on the consistency of the workers. The consistency across the duplicated statements highlights the honest labeling by the unpaid workers.

Regarding the difficulty, the mean consistency ratio for all workers without the difficult statements was 0.224, which was less than the mean overall ratio. This indicates that the absence of difficult duplicates increased the consistency of the workers. This was expected because the difficulty of the task could have led to differing impressions about the truth of the statement, and consequently the labeling.

Regarding the rephrasing, the mean for all workers was 0.207 in the consistency ratio matrix in the absence of the rephrased statements. This indicates that the effect of rephrasing was the same as the effect of difficulty. This was unexpected. We suspect that the rephrasing of the statements led to a distribution in the judgement, and hence, labeling, of a worker. All of these results are shown in Fig. 5.

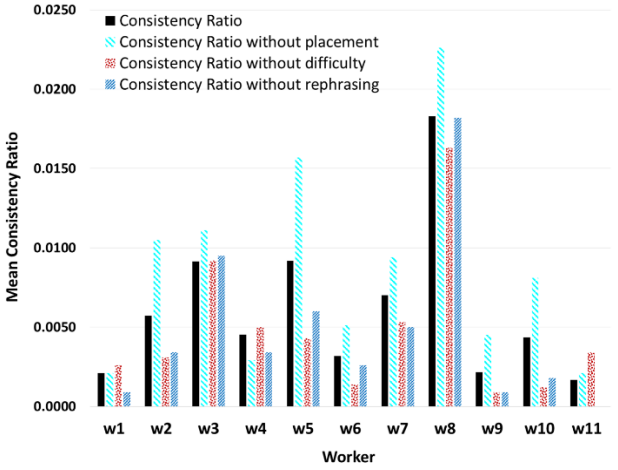

FIGURE 5. Comparison of mean consistency ratio for each for all cases.

E. RELATIONSHIP BETWEEN CONSISTENCY AND ACCURACY

To investigate the relationship between the workers' consistency and accuracy, we used the Pearson Correlation Coefficient r. We tested the correlation between the accuracy measure (mean MAE) (10) and the consistency measure (the mean of W_{cons} (8)) of the workers. We found a correlation coefficient of $r = 0.57$, with $p < 0.07$. This indicates that there is a marginally significant positive relationship between the accuracy and consistency. The positive correlation was expected in our experiment from the high accuracy and consistency score achieved by the unpaid workers. Moreover, we statistically tested the r between the mean MAE (10) and the consistency differences score as a consistency baseline Cb measure(16). We found that $r = 0.54$ with $p = 0.088$. This is similar to the previous result that a worker with larger differences in his rating (i.e., less consistency) was likely to have larger errors compared to the ground truth.

We also studied the statistical relationship between the mean consensus score (12) and mean MAE (10) of the

workers by using the r coefficient to estimate this relationship. We found $r = -0.67$ with $p < 0.05$. This negative correlation indicates that, as expected, workers with high consensus scores will have less errors with respect to the ground truth.

VI. PERFORMANCE ANALYSIS

A. RELIABILITY ANALYSIS

We measured the reliability of our experiment with respect to the selected scale through the internal consistency of our scale. We used the Cronbach alpha [60]:

$$\alpha = \frac{K}{1-K}\left(1 - \frac{\sum_1^k var(Y_i)}{var(X)}\right) \tag{18}$$

where K is the number of core statements, which is 18, $var(Y_i)$ is the variance of workers' labels of the statements, and $var(X)$ is the variance of the total labeling. The α in our experiment was 0.76, which indicates that it has good internal consistency.

B. ACCURACY ANALYSIS

To evaluate the performance of our model, due to the unavailability of consistency benchmark datasets, and a lack of works studying the consistency, we used the methods [26], [33] as baselines of comparison using our dataset. Williams *et al.* [26] introduced a method to calculate the consistency of 402 crowdsourcing workers. They created a dataset of 30 images and the task was asking to count the number of objects in each of them. a worker in each task counted objects in 10 images (two of them used as consistency probe. the same image with modification).

Naderi *et al.* [33] presented a survey contains 74 items, which was conducted with a total of 256 participants. They measured the consistency using some randomly selected items, which are asked twice in the questionnaire. We compared our method, which uses pair-wised differences, against [26]. and [33]. Williams *et al.* calculated the absolute difference between a worker's outputs for the original task and its duplicate where output is the counting number of object in an image. Slightly similar, Naderi *et al.* calculated the differences between the worker's answers of same questionnaire item. They used the weighted Euclidian distance. Their weights were calculated using responses of all workers, which is the consensus, score (12) in our methodology.

For the accuracy comparison, first, for each worker across all statements we calculated the difference between the original statement and the duplicated one. We used pairwise difference in our case and absolute difference in case of Williams *et al.* and weighted Euclidian difference in case of Naderi *et al.* Then, we calculated the average differences in each case and used it as threshold. After that, the accuracy of each method for each statement is determined based on the threshold. Finally, the average accuracy of each worker for each method is calculated using his/her accuracy of all statements. Fig. 6 shows the accuracy of all workers for each method. It illustrates that our method archives higher accuracy than the contemporary for almost all the workers. In average, our model achieved 73% average accuracy, which surpassed the 61% of Williams *et al.* [26] and the 67% of Naderi *et al.* [33].

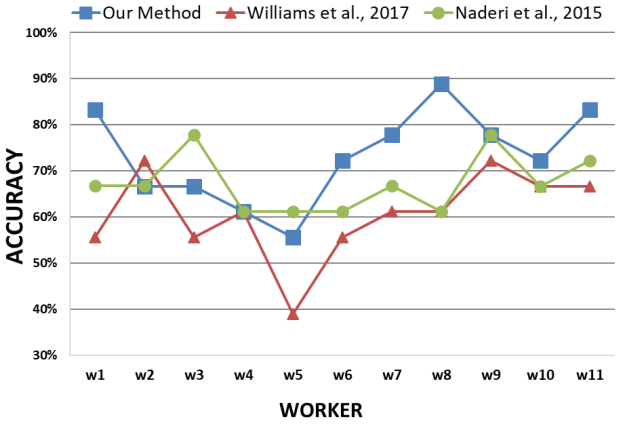

FIGURE 6. Accuracy comparison, our method vs. [26], [33].

VII. DISCUSSION AND FUTURE WORK

Measuring workers' reliability in crowdsourcing is a major challenge. Studying the level of consistency in their answers sheds light on their performance, and consequently their reliability. In this work, we studied the consistency of unpaid workers using a pair-wise comparison method to measure their internal consistency in rating the truthfulness of textual political statements. The effects of three different characteristics were examined in our experiment, namely the placement of the statements, the difficulty of the task, and the rephrasing of the statements.

Generally, unpaid workers perform repeated tasks in a consistent manner. This is expected because workers who are intrinsically motivated do well in the crowdsourcing [2], [58], such as in citizen sciences. An important result in this study is the consistency score of the workers. More accurate results were obtained from our model/method compared to the baselines [26], [33]. This can be attributed to the mathematical robustness of the pairwise comparison method compared to the limited approach of calculating the absolute differences of errors or weighted distances.

We compared our results for the effects of each characteristic with corresponding results from previous works, which differ from this work in terms of the measurements used, the pool of workers (unpaid vs paid ones), and the types of task (texts rating vs image objects counting). In our experiment, placement did not affect the consistency ratio. This could be because the completion time of the tasks included long breaks, as reflected in the post-questionnaire responses. Hence, placement-related effects like fatigue [61] would not be of impact. The results for the effects of difficulty are similar to those in prior work. The task difficulty affected the workers' consistency negatively. This is expected because

inconsistent results are expected even in the absence of difficulty. Our results are consistent with the relationship between the task difficulty and reliability found by [62].

Finally, rephrasing had same effect as the difficulty in our experiment. This is different from previous works. An explanation may be the confusion resulting from the modified texts which was absent in comparisons between image transformation duplicates in previous works.

An additional observation is that because our task was about political statements, the workers' reliability could be vulnerable to the bias effect [44], [52]. This is true to a large extent for the workers' accuracy but not their consistency. We mitigated the bias effects by omitting the name of the politician who made the statement. Furthermore, the workers did not have a major interest in political affairs of the US. This matched our expectation, and was further confirmed by their answers on the post-questionnaire.

We expected other effects such as recognition. This was clarified by the answers given by the workers in the questionnaire which indicated that they suspected that some statements were duplicated. The tasks could therefore be susceptible to recognition which might result in workers changing their earlier answers. We mitigated these effects by disallowing the workers from going back to earlier tasks to ensure that the workers moved forward in the tasks, even when they were suspicious of the similarity of the statements.

Regarding the limitations in this study, we plan to extend our work to more crowdsourcing settings. The implementation of our consistency measurement for paid workers will be an interesting future work. Crowdsourcing platforms such as AMT have an abundance of paid workers. Extending our work by recording different performance characteristics such as workers' time per task, hover time, out of focus time, scrolling, and answer switching is another future work that will open promising future research. Such extensions will be the cornerstone for modeling and developing machine learning algorithms for predictions of worker consistency. The correlation between accuracy and consistency can also be investigated because workers can be consistent but not accurate. Other effects such as learning and fatigue can be studied. This, together with studying paid workers, will enrich the research on crowdsourcing and facilitate consistency measurement for more types of workers like spammers and Sybils.

VIII. CONCLUSION

In this study, we propose a new model for measuring the consistency of unpaid workers in crowdsourcing. Our experiment studied how workers labeled the truthfulness of duplicate political claims. We assessed their consistency and studied the effects of different characteristics. Our results show that the volunteer workers achieved high consistency scores. The accuracy of our model outperformed the state-of-the-art methods. Future work includes implementing our model for paid workers in a featured crowdsourcing platform. Another future work is to extend this consistency study to include worker features. This will help in the development of models for machine learning techniques and for predicting worker consistency and reliability.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research, King Saud University, for funding and supporting this research through the initiative of DSR. They would also like to thank the Graduate Students Research Support (GSR) and the Researchers Support and Services Unit (RSSU) at King Saud University for their technical support.

REFERENCES

[1] N. Q. V. Hung, D. C. Thang, M. Weidlich, and K. Aberer, "Minimizing efforts in validating crowd answers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 999–1014.

[2] R. M. Borromeo and M. Toyama, "An investigation of unpaid crowdsourcing," *Hum.-Centric Comput. Inf. Sci.*, vol. 6, no. 1, p. 11, Dec. 2016.

[3] M. Poesio, J. Chamberlain, and U. Kruschwitz, "Crowdsourcing," in *Handbook of Linguistic Annotation*, N. Ide and J. Pustejovsky, Eds. Dordrecht, The Netherlands: Springer, 2017, pp. 277–295.

[4] S. Zong, A. Baheti, W. Xu, and A. Ritter, "Extracting COVID-19 events from Twitter," Jun. 2020, *arXiv:2006.02567*. [Online]. Available: http://arxiv.org/abs/2006.02567

[5] D. Zhang, Y. Zhang, Q. Li, T. Plummer, and D. Wang, "CrowdLearn: A crowd-AI hybrid system for deep learning-based damage assessment applications," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 1221–1232.

[6] A. Olivieri, S. Shabani, M. Sokhn, and P. Cudré-Mauroux, "Creating task-generic features for fake news detection," in *Proc. 52nd Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 5196–5205.

[7] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1313–1321, May 2016.

[8] X. Chen, Y. Zhang, H. Xu, Y. Cao, Z. Qin, and H. Zha, "Visually explainable recommendation," Jan. 2018, *arXiv:1801.10288*. [Online]. Available: http://arxiv.org/abs/1801.10288

[9] X. Yin, W. Liu, Y. Wang, C. Yang, and L. Lu, "What? How? Where? A survey of crowdsourcing," in *Frontier and Future Development of Information Technology in Medicine and Education*, vol. 269, S. Li, Q. Jin, X. Jiang, and J. J. Jong, and H. Park, Eds. Dordrecht, The Netherlands: Springer, 2014, pp. 221–232.

[10] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 76–81, Mar. 2013.

[11] Q. Hu, S. Wang, P. Ma, X. Cheng, W. Lv, and R. Bie, "Quality control in crowdsourcing using sequential zero-determinant strategies," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 5, pp. 998–1009, May 2020.

[12] J. Zhang, V. S. Sheng, and J. Wu, "Crowdsourced label aggregation using bilayer collaborative clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3172–3185, Oct. 2019.

[13] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 901–911, Apr. 2016.

[14] J. Le, J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution," in *Proc. Workshop Crowdsourcing Search Eval. (CSE)*, 2010, pp. 17–20.

[15] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing," in *Proc. Workshops 25th AAAI Conf. Artif. Intell.*, 2011, pp. 43–48.

[16] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Oct. 2016.

[17] J. Vuurens, A. de Vries, and C. Eickhoff, "How much spam can you take? An analysis of crowdsourcing results to increase accuracy," in *Proc. ACM SIGIR Workshop Crowdsourcing Inf. Retr. (CIR)*, 2011, pp. 21–26.

[18] V. S. Sheng, J. Zhang, B. Gu, and X. Wu, "Majority voting and pairing with multiple noisy labeling," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1355–1368, Jul. 2019.
[19] M. Nazariani and A. A. Barforoush, "Dynamic weighted majority approach for detecting malicious crowd workers," *Can. J. Electr. Comput. Eng.*, vol. 42, no. 2, pp. 108–113, 2019.
[20] F. Tao, L. Jiang, and C. Li, "Label similarity-based weighted soft majority voting and pairing for crowdsourcing," *Knowl. Inf. Syst.*, vol. 62, no. 7, pp. 2521–2538, Jul. 2020.
[21] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: Using implicit behavioral measures to predict task performance," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 13–22.
[22] J. Rzeszotarski and A. Kittur, "CrowdScape: Interactively visualizing user behavior and output," in *Proc. 25th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2012, pp. 55–62.
[23] S. Yuasa, T. Nakai, T. Maruichi, M. Landsmann, K. Kise, M. Matsubara, and A. Morishima, "Towards quality assessment of crowdworker output based on behavioral data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 4659–4661.
[24] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms," in *Proc. CrowdSearch Workshop*, 2012, pp. 26–30.
[25] A. Moayedikia, K.-L. Ong, Y. L. Boo, and W. Yeoh, "Bee colony based worker reliability estimation algorithm in microtask crowdsourcing," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 713–717.
[26] A. C. Williams, J. Goh, C. G. Willis, A. M. Ellison, J. H. Brusuelas, C. C. Davis, and E. Law, "Deja Vu: Characterizing worker reliability using task consistency," in *Proc. 5th AAAI Conf. Hum. Comput. Crowdsourcing (HCOMP)*, 2017, pp. 197–205.
[27] S.-W. Huang and W.-T. Fu, "Enhancing reliability using peer consistency evaluation in human computation," in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*, 2013, pp. 639–647.
[28] J. Cheng, J. Teevan, and M. S. Bernstein, "Measuring crowdsourcing effort with error-time curves," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI)*, 2015, pp. 1365–1374.
[29] K. Hata, R. Krishna, F.-F. Li, and M. S. Bernstein, "A glimpse far into the future?: Understanding long-term crowd worker quality," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput. (CSCW)*, 2017, pp. 889–901.
[30] L. Shamir, D. Diamond, and J. Wallin, "Leveraging pattern recognition consistency estimation for crowdsourcing data analysis," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 3, pp. 474–480, Jun. 2016.
[31] F. Scholer, A. Turpin, and M. Sanderson, "Quantifying test collection quality based on the consistency of relevance judgements," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. (SIGIR)*, 2011, pp. 1063–1072.
[32] P. Sun and K. T. Stolee, "Exploring crowd consistency in a mechanical Turk survey," in *Proc. 3rd Int. Workshop CrowdSourcing Softw. Eng. (CSI-SE)*, 2016, pp. 8–14.
[33] B. Naderi, I. Wechsung, and S. Moller, "Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms," in *Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–2.
[34] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. Conf. Hum. Factors Comput. Syst. (CHI)*, 2004, pp. 319–326.
[35] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Math. Comput. Model.*, vol. 57, nos. 11–12, pp. 2918–2932, Jun. 2013.
[36] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 614–622.
[37] A. Sheshadri and M. Lease, "SQUARE: A benchmark for research on computing crowd consensus," in *Proc. 1st AAAI Conf. Human Comput. Crowdsourcing (HCOMP)*, 2013, pp. 156–164.
[38] W. Lee, C. H. Huang, C. W. Chang, M. K. D. Wu, K. T. Chuang, P. A. Yang, and C. C. Hsieh, "Effective quality assurance for data labels through crowdsourcing and domain expert collaboration," in *Proc. Adv. Database Technol. (EDBT)*, 2018, pp. 646–649.
[39] M. Hirth, S. Scheuring, T. Hossfeld, C. Schwartz, and P. Tran-Gia, "Predicting result quality in crowdsourcing using application layer monitoring," in *Proc. IEEE 5th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2014, pp. 510–515.
[40] D. Zhu and B. Carterette, "An analysis of assessor behavior in crowdsourced preference judgments," in *Proc. SIGIR Workshop Crowdsourcing Search Eval.*, 2010, pp. 17–20.
[41] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 1941–1944.
[42] M. Brunelli, "A survey of inconsistency indices for pairwise comparisons," *Int. J. Gen. Syst.*, vol. 47, no. 8, pp. 751–771, Nov. 2018.
[43] M. R. Pinto, Y. O. de Lima, C. E. Barbosa, and J. M. de Souza, "Towards fact-checking through crowdsourcing," in *Proc. IEEE 23rd Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2019, pp. 494–499.
[44] M. Sameki, T. Zhang, L. Ding, M. Betke, and D. Gurari, "Crowd-O-meter?: Predicting if a person is vulnerable to believe political claims," in *Proc. 5th AAAI Conf. Human Comput. Crowdsourcing (HCOMP)*, 2017, pp. 157–166.
[45] L. Wang, Y. Wang, G. de Melo, and G. Weikum, "Five shades of untruth: Finer-grained classification of fake news," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 593–594.
[46] W. Y. Wang, "'liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 422–426.
[47] K. Shu, A. H. Awadallah, S. Dumais, and H. Liu, "Detecting fake news with weak social supervision," *IEEE Intell. Syst.*, early access, May 28, 2020, doi: 10.1109/MIS.2020.2997781.
[48] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.
[49] W. Willett, J. Heer, and M. Agrawala, "Strategies for crowdsourcing social data analysis," in *Proc. ACM Annu. Conf. Hum. Factors Comput. Syst. (CHI)*, 2012, pp. 227–236.
[50] J. Fülöp, W. W. Koczkodaj, and S. J. Szarek, "A different perspective on a scale for pairwise comparisons," in *Transactions on Computational Collective Intelligence I* (Lecture Notes in Computer Science), vol. 6220, no. 6. Berlin, Germany: Springer, 2010, pp. 71–84.
[51] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing—A study on mechanical Turk," in *Proc. 17th Americas Conf. Inform. Syst. (AMCIS)*, 2011, pp. 1–11.
[52] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, and G. Demartini, "Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background," in *Proc. 43rd Int. ACM Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2020, pp. 1–10.
[53] *The Text Labeling Interface*. Accessed: Feb. 13, 2020. [Online]. Available: http://dataplume.pythonanywhere.com
[54] J. A. ALONSO and M. T. LAMATA, "Consistency in the analytic hierarchy process: A new approach," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 14, no. 4, pp. 445–459, Aug. 2006.
[55] T. L. Saaty, "Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process," *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, vol. 102, no. 2, pp. 251–318, Sep. 2008.
[56] J. B. P. Vuurens and A. P. de Vries, "Obtaining high-quality relevance judgments using crowdsourcing," *IEEE Internet Comput.*, vol. 16, no. 5, pp. 20–27, Sep. 2012.
[57] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2416–2419.
[58] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan, "Quality through flow and immersion: Gamifying crowdsourced relevance assessments," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2012, pp. 871–880.
[59] R. Falotico and P. Quatto, "Fleiss' kappa statistic without paradoxes," *Qual. Quantity*, vol. 49, no. 2, pp. 463–470, Mar. 2015.
[60] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.
[61] Y. Zhang, X. Ding, and N. Gu, "Understanding fatigue and its impact in crowdsourcing," in *Proc. IEEE 22nd Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2018, pp. 57–62.
[62] S. Räbiger, Y. Saygın, and M. Spiliopoulou, "How does tweet difficulty affect labeling performance of annotators?" Aug. 2018, *arXiv:1808.00388*. [Online]. Available: http://arxiv.org/abs/1808.00388

FATTOH ALQERSHI received the B.S. degree in information technology from the Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt, in 2005, and the master's degree in information systems from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2012, where he is currently pursuing the Ph.D. degree. From 2007 to 2009, he was an Assistant Lecturer with the Department of Software Engineering, Faculty of Engineering and Information Technology, Taiz University, Yemen. His research interests include online social networks, social media analysis, and crowdsourcing.

MUHAMMAD AL-QURISHI (Member, IEEE) received the Ph.D. degree from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2017. He is currently a Postdoctoral Researcher with the Chair of Pervasive and Mobile Computing (CPMC), CCIS, KSU, and is one of the founding members of CPMC. He has published several articles in refereed journals (IEEE, ACM, Springer, and Wiley). His research interests include data science, big data analysis and mining, pervasive computing, and machine learning. He received the Innovation Award for a mobile cloud serious game from KSU, in 2013, the Best Ph.D. Thesis Award from CCIS, KSU, in 2018, and the IBM Data Science Professional Certificate and Deep Learning Certification from deeplearning.ai.

MEHMET SABIH AKSOY received the master's degree from Yildiz Technical University, Istanbul, Turkey, in 1985, and the Ph.D. degree in artificial intelligence from Cardiff University, U.K., in 1994. He graduated from Istanbul Technical University, in 1982. He worked with Istanbul Technical University, Sakarya University, and Istanbul University, Turkey, from 1984 to 2002. He joined the College of Computer and Information Sciences, King Saud University, in 2002, where he is currently working as a Full Professor. His research interests include machine learning, inductive learning, expert systems, artificial neural networks, and data mining.

MAJED ALRUBAIAN (Member, IEEE) received the Ph.D. degree from the Department of Information Systems, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2015. He has authored several papers in the refereed IEEE/ACM/Springer journals and conferences. His research interests include social media analysis, data analytics and mining, social computing, information credibility, and cyber security. He is a Student Member of ACM.

MUHAMMAD IMRAN (Member, IEEE) received the Ph.D. degree in information technology from the Universiti Teknologi PETRONAS, Malaysia, in 2011. He is currently an Associate Professor with the College of Applied Computer Science, King Saud University, Saudi Arabia. His research was financially supported by several grants. He has completed a number of international collaborative research projects with reputable universities. He has published more than 250 research papers in peer-reviewed and well-recognized international conferences and journals. Many of his research articles are among the highly cited and most downloaded. His research interests include the Internet of Things, mobile and wireless networks, big data analytics, cloud computing, and information security. He has been consecutively awarded with the Outstanding Associate Editor of IEEE Access, in 2018 and 2019, besides many others. He served as the Editor-in-Chief for the *EAI Endorsed Transactions on Pervasive Health and Technology*. He also serves as an Associate Editor for top ranked international journals, such as the *IEEE Communications Magazine*, the IEEE Network, *Future Generation Computing Systems*, and IEEE Access. He served/serving as a Guest Editor for about two dozen special issues in journals, such as the *IEEE Communications Magazine*, the *IEEE Wireless Communications Magazine*, *Future Generation Computing Systems*, IEEE Access, and *Computer Networks*. He has been involved in about 100 peer-reviewed international conferences and workshops in various capacities, such as the Chair, the Co-Chair, and the Technical Program Committee Member.