Evaluating Crowdsourced Relevance Assessments Using Self-Reported Traits and Task Speed

Christopher Chow
Research School of Computer Science
The Australian National University
Australia
christopher.chow@anu.edu.au

Tom Gedeon
Research School of Computer Science
The Australian National University
Australia
tom.gedeon@anu.edu.au

ABSTRACT

Relevance is the strength of the relationship between a user's perceived information need and an information object. Systems designed to help users identify relevant information can often rely on high quality labelled datasets. However, the subjective and personal nature of relevance means that establishing ground truth labels is difficult. In this work, we conduct a user study on text documents to crowdsource relevance assessments against four topics. Workers' self-reported measures and task completion speed are used to calculate a range of ground truth measures against which classification performance can be assessed. Our results indicate that average subjective relevance and confidence-weighted measures are on par with the annotations from an expert panel. Further work is planned to expand these findings.

CCS CONCEPTS

• **Human-centered computing → User studies** • Human-centered computing → Empirical studies in HCI • *Information systems → Relevance assessment*

KEYWORDS

Relevance assessments, ground truth, subjective relevance, crowdsourcing

ACM Reference format:
C. Chow and T. Gedeon, 2017, Evaluating Crowdsourced Relevance Assessments Using Self-Reported Traits and Task Speed. In *Proceedings of the 29th Australian Conference on Human-Computer Interaction, Brisbane, QLD, Australia, November 2017 (OzCHI 2017)*, 4 pages.
https://doi.org/10.1145/3152771.3156146

1 INTRODUCTION AND PREVIOUS WORK

Relevance assessments – acts of deciding whether or not an information object (e.g. documents, images, or emails) is of current value – are commonplace in everyday life. This might include during tasks such as identifying interesting items in a news feed, browsing search results, or filtering spam emails. In order to help prevent information overload in users, decision aids powered by artificial intelligence and machine learning are becoming increasingly pervasive. However, supervised machine learning performance can be limited by the quality of the initial training dataset, and in particular the dataset's labels (often known as the 'ground truth' or 'gold standard' [18]).

In many cases, expert panels are employed to generate the relevance labels, however this can be expensive [2, 25], and introduce further complications. Relevance is complex and personal; it is the strength of the match between an information object and a user's own perceived information need [18]. These needs can be influenced by many factors including the users' cognitive state, environment and circumstances, or the object under consideration. As such, it is not something that can be accurately described in a single label. As a result, relying on one person's subjective labelling for a dataset – even if they are considered an expert – could bias model training away from the social norm [10].

Crowdsourcing is one potential method for reducing the reliance on expert labelling, with previous work demonstrating its usefulness across several disciplines [12, 14, 25]. Work by [19] illustrates the benefits of crowdsourcing for relevance assessments in particular, which is made possible through the many different crowdsourcing marketplaces now available [32]. However, the use of crowdsourcing is not without it's own barriers, as contribution quality (or reputation [28]) is often a concern. In cases such as these, previous work has investigated measuring quality via expert review [31], peer assessments [28], or intrinsic behavioural properties such as mouse clicks [20]. However, in a subjective domain like relevance, establishing worker quality is perhaps even more difficult, as there may not be a clear point of reference from which response deviations can be measured. Therefore, finding ways to generate higher quality machine learning labels from the responses of workers is of considerable interest.

Recent studies have explored ways to address these challenges. For example, [2] demonstrated the importance of collaborative approaches to labelling machine learning datasets, introducing a system to help workers label data akin to an expert annotation workflow. Similarly, in their entity annotation task [10] devised a method for *crowd parting* based on shared patterns of responses to uncover interesting trends, and amend labels as

OzCHI '17, November 28-December 1, 2017, Brisbane, QLD, Australia
© 2017 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-5379-3/17/11…$15.00
https://doi.org/10.1145/3152771.3156146

appropriate. Separately, work by [25] and [29] showed that calibrating worker bias by using expert annotations yielded improved results, whilst others have demonstrated that traits of the workers themselves are useful indicators. For example, a study by [11] demonstrated that gender, location and personality type can be indicative of label quality.

We extend this work by proposing an approach to generating ground truth label sets based on self-reported or behavioural traits. We focus on crowdsourcing topical relevance assessments toward text documents, with a goal to create a unified ground truth label useful for machine learning applications, similar to the work in [2, 3, 10]. The relevance assessments we collected represent personal truths to each user, and are thus valid despite their divergence. Therefore, our approach investigates methods to weight worker responses more likely to be of higher quality, without too heavily discounting the contributions of others. Our approach reveals that self-reported confidence measures are on par with expert level performances, and thus could increase the capacity for crowdsourcing to provide cheap and distributed relevance assessment labels for subjective datasets.

2 USER STUDY METHODOLOGY

We conduct a user study to crowdsource topical relevant assessments, and aggregate responses based on the weightings of two self reported measures, as well as task completion speed. The range of documents chosen and the task design were intentionally challenging so as to elicit a range of relevance judgments.

2.1 Document Corpus

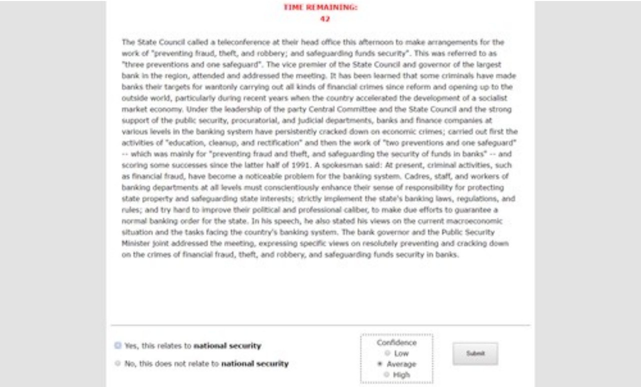
We expand a text document corpus introduced in [3]. A further 10 documents are combined with the existing 30 to include an additional topic, raising the total number of topics to four. New documents were sourced from media reports, academic material including textbooks, and other standard test collections including TREC and REUTERS. Document modifications were made to maintain conformity with the existing corpus, including length modulation, and readability balancing. Full details regarding source credit, topic labels, and control attributes are described within the dataset, which is available upon request.

Each document is assigned a relevance label from one, two or none of four topic classes: national security, natural disasters, computer science and psychology. For example, some documents pertain to more than one defined topic, whilst others are not relevant to any. Topic labels were derived through unanimous agreement of an expert panel as per the original method [3].

2.2 Task and Experimental Procedure

User studies took place in a controlled experimental environment. Participants were given a short brief, during which it was explained that they should quickly assess whether or not each document was relevant to their given topic (i.e., perform binary classification).

Each participant was asked a series of demographic questions, before being issued with one of the four corpus topics. A description of the topic was given as a rubric to help positively shape worker performance [31], as we expected some participants to have expertise in some of the areas (i.e., computer science), yet inexperience with the others. Participants rated their level of familiarity with the topic on a 5-point Likert scale as a measure of their expertise.



TIME REMAINING:
42

The State Council called a teleconference at their head office this afternoon to make arrangements for the work of "preventing fraud, theft, and robbery; and safeguarding funds security". This was referred to as "three preventions and one safeguard". The vice premier of the State Council and governor of the largest bank in the region, attended and addressed the meeting. It has been learned that some criminals have made banks their targets for wantonly carrying out all kinds of financial crimes since reform and opening up to the outside world, particularly during recent years when the country accelerated the development of a socialist market economy. Under the leadership of the party Central Committee and the State Council and the strong support of the public security, procuratorial, and judicial departments, banks and finance companies at various levels in the banking system have persistently cracked down on economic crimes; carried out first the activities of "education, cleanup, and rectification" and then the work of "two preventions and one safeguard" -- which was mainly for "preventing fraud and theft, and safeguarding the security of funds in banks" -- and scoring some successes since the latter half of 1991. A spokesman said: At present, criminal activities, such as financial fraud, have become a noticeable problem for the banking system. Cadres, staff, and workers of banking departments at all levels must conscientiously enhance their sense of responsibility for protecting state property and safeguarding state interests; strictly implement the state's banking laws, regulations, and rules; and try hard to improve their political and professional caliber, to make due efforts to guarantee a normal banking order for the state. In his speech, he also stated his views on the current macroeconomic situation and the tasks facing the country's banking system. The bank governor and the Public Security Minister joint addressed the meeting, expressing specific views on resolutely preventing and cracking down on the crimes of financial fraud, theft, and robbery, and safeguarding funds security in banks.

○ Yes, this relates to **national security**
○ No, this does not relate to **national security**

Confidence
○ Low
● Average
○ High

Submit

Figure 1: A single document displayed to a participant. A 45 second countdown timer is presented at the top, followed by the document content and a relevance assessment form.

Each document within the corpus was presented in turn to the participant using a custom interface shown in Fig. 1. Participants could select their binary relevance choice, as well as rate their self-confidence using a 3-point Likert scale. This task was intentionally a binary classification as work by [16] demonstrated that novice users benefited from a binary approach, with little to no adverse effect on experts. A red countdown timer at the top of the screen indicated their time remaining on each individual sub-task. A maximum of 45 seconds was allowed before automatic progression to the next item, controlling the users' ability to exhaustively analyse each document, and to improve judgment quality [17].

To help negate any potential order effects during the study, a Latin Square was used. Due to the high number of texts, eight stratified folds were used for the Latin Square, with each fold being randomised internally.

2.3 Participants

We recruited 58 university students (35 male; 23 female) aged between 18 and 39. Participants took part in exchange for course credit, as approved by our Human Research Ethics Committee. Participants were randomly allocated to one of four conditions which determined the topic they would perform relevance assessments against. Of the participants, 46 (79%) were from computer science backgrounds, whilst 33 (57%) were native English speakers.

3 GROUND TRUTH MEASURES

The expert panel annotations derived as part of the document corpus curation serves as a baseline upon which each of the workers' personal truths can be compared. Expert opinions are a common method for serving as a reference point upon which the performance of classifiers or other groups are measured (e.g. [2, 10, 15, 21]). However, using expert panels as the primary measure of evaluation is problematic. It is difficult to delineate how much experience or knowledge is necessary for one to be considered an expert [5, 24], plus the chances of reaching unanimity amongst the group is low [9]. This highlights the benefits of crowdsourcing for such tasks.

We consider several different methods for computing a unified ground truth measure based on the crowdsourced labels. After each workers' labels have been weighted as per the following methods, a double majority vote was used to aggregate labels across all participants. For example, in order for a document to be assigned a label, it must have received more 'Relevant' than 'Non-relevant' votes within the condition, and more 'Relevant' votes than any other condition. In the event of a tie, multiple labels were assigned.

3.1 Average Subjective Relevance (ASR)

For each document, the average subjective relevance [30] is calculated by taking the mean of all relevance evaluations by the 58 participants. We use a simple majority decision rule so that topicality is denoted if more than half of participants decided in favour of relevance. As the participants were performing a binary classification, our average subjective relevance is not indicative of *degree* of relevance, despite each document having a degree of topicality.

3.2 Weighted Average Subjective Relevance

We propose *weighted average subjective relevance* assessments. User behaviours or self-reported measures can provide additional insight into how they perceive relevance. This information can add weight to a user's explicit assessment, rewarding user opinions that contribute the most benefit or accuracy, and penalizing the impact of potentially harmful decisions (e.g. guesses).

3.2.1 Confidence-weighted Average Subjective Relevance (CASR). We postulate that users who report higher confidence in a decision are more likely to have a better understanding of the topicality, and thus are more likely to be reliable.

During the task, participants self-rated their confidence on a 3-point Likert scale. To calculate *confidence-weighted* average subjective relevance, we used these confidence ratings to alter how much impact their decision has overall. A rating of 3 (*high* confidence) meant that the participant's full decision was incorporated—we accept their choice as truth regardless of how it compared to the expert panels choice. A rating of 2 (*average* confidence) altered the participants' choice to have only half an impact. A rating of 1 (*low* confidence) we took as meaning the participant guessed, and therefore the opposite of their choice was used (i.e., if they answered relevant, it was changed to irrelevant and vice versa).

3.2.2 Time-weighted Average Subjective Relevance (TASR). When user performance against the expert panel's ground truths is compared with their time spent making the decision, there is a significant negative correlation (r = -0.262, p < 0.05), shown in Fig. 2. This apparent inconsistency with the speed-accuracy tradeoff phenomenon [7] could be attributed to users answering faster if they are sure of their decision, whilst those who are unsure, and thus likely to be wrong, maximise the time available. These timing behaviours form the basis for computing *time-weighted* average subjective relevance measures. This result confirms previous findings in [17].

If a user is quicker to answer (within the 0.25 quantile for that document), we accept their decision as truth. Meanwhile if the user took more than 44 seconds to answer, we assume they did not know and therefore the inverse of their decision was taken. All other decisions in between were treated with half impact.

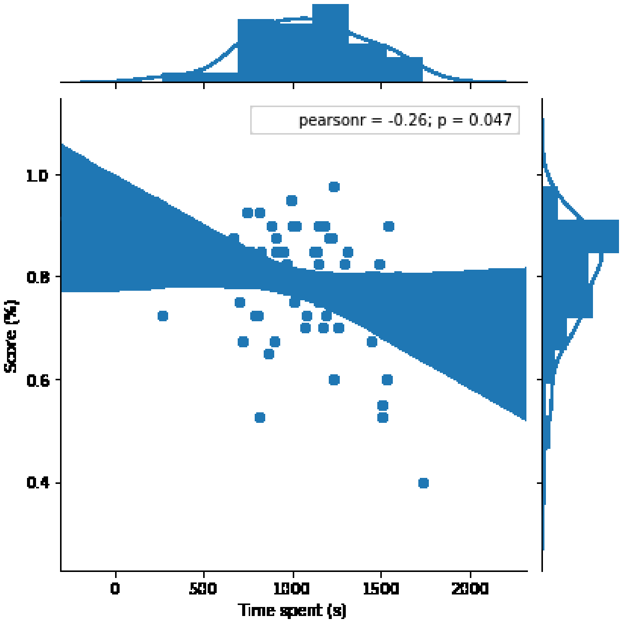


Figure 2: Relationship between time spent and user performance against the expert panels' ground truths.

3.2.3 Familiarity-weighted Average Subjective Relevance (FASR). Although during our user study we defined the necessary topic to each participant to establish a baseline understanding, the nature of the population meant that each participant had a varying degree of experience, expertise, exposure to their assigned topic. This might particularly be the case for the computer science topic, as our participants were predominantly from computer science backgrounds. As such, we hypothesised that a participant's self-rated familiarity with the topic prior to the study could be used a method for calculating *familiarity-weighted* average subjective relevance.

Familiarity was self-reported via a 5-point Likert scale. Therefore, each level contributed impacted on their decision's contribution by 25% (e.g. a familiarity score of 5 meant that the decision was taken as truth, whereas a familiarity of 1 meant that the decision was taken as likely to be a guess, and therefore the inverse was used).

4 RESULTS

To compare each of the annotation sets, we compute inter-annotator agreement using Fleiss' kappa (K) [6], as shown in Table 1. There is substantial agreement [13, 26] between the expert panel baseline, and the average subjective relevance and confidence-weighted measures. Surprisingly, there was very poor agreement between the expert label set and the time and familiarity weighted measures.

We next use each of these unified ground truth label sets as a measure for assessing user performance. This provides a broad indication of the practical implications of using the label sets as ground truths. For each of our measures, the ability for the

Table 1: K scores comparing all computed measures

	ASR	CASR	TASR	FASR
Expert	0.71	0.65	-0.34	-0.45
ASR		0.77	-0.46	-0.57
CASR			-0.41	-0.53
TASR				0.57

participants to correctly classify the documents against their prescribed topic is shown in Table 2.

Performance of users was best when expert panel, average subjective relevance, and confidence-weighted measures were set as the ground truths. The average subjective relevance has the highest recall of 0.82, meaning that when workers were judged against the simple average of their peers, they were reasonably good at finding all the documents that were relevant to their topic. However, user performance against the confidence-weighted measure showed an increase in ability to identify only the relevant documents, with a precision of 0.76. These trade-offs lead to them having comparable F1 scores of 0.74 and 0.72, respectively. Time and familiarity weighted measures scored poorly across all performance measures.

5 DISCUSSION AND FUTURE WORK

From the results, no single measures allowed participants to complete the task without error. This is an unfortunate side effect of the subjectivity of relevance, as task ambiguity and differences in opinion or method will lead to inconsistent labels [2]. Users would therefore inevitably find false positives (Type I errors; characterized by lower precision) or false negatives (Type II errors; characterized by lower recall) when compared to the responses of others. Therefore, the F1 scores suggest that choosing between the most appropriate method of calculating ground truth measures is dependent on the situation in which it will be deployed. We discuss the practical implications of this within the scope of the topics in our document corpus.

Making a larger number of Type II errors can be detrimental in safety critical situations. As one of the topics of the dataset, national security applications would be an area where scrutinising false positives would be preferred over discarding important information, even if it was wasteful of time. As such, a ground truth measure with a low Type II error rate such as the average subjective relevance would be most beneficial for these circumstances. Relevance assessments could be crowdsourced in parallel by multiple users, with an intelligent system computing average subjective relevance to produce predictions with lower Type II errors compared to other methods.

Meanwhile, circumstances exist where making fewer Type I errors over Type II is preferred. This includes in cyber security, where perceiving phishing emails as relevant when in fact they are not could have severe consequences. Spam or phishing filtering could therefore be trained on datasets labelled using confidence-weighted average subjective relevance assessments so that fewer Type I errors are made, even if this means some legitimate and relevant emails are relegated to the spam folder.

By limiting time in this study, we emulate real-world crowdsourcing situations where workers are demotivated or uninterested [32]. The results suggest that even if that were the case, aggregated relevance assessments are still fit for purpose as machine learning dataset labels. In fact, our results suggest that ASR (simple majority voting) provides the best performance and hence the best intermediary 'truth'. As the results proved promising when compared to the judgements of an expert panel, it could suggest a reduced reliance on seeking expensive expert opinion in future applications. Other methods for crowdsourcing labels in real-world applications have already proven to be successful, such as image classification problems [2, 4], language translation [27], and affective text analysis [25]. Therefore, the results support the generation of crowdsourced dataset labels in circumstances where there is greater subjectivity or ambiguity.

Table 2: Participant performance against each of the ground truth measures. Bold values indicate highest value for each performance score.

	Expert	ASR	CASR	TASR	FASR
Accuracy	0.81	**0.85**	0.81	0.21	0.19
Precision	0.59	0.67	**0.76**	0.38	0.23
Recall	0.75	**0.82**	0.69	0.17	0.11
F1 Score	0.66	**0.74**	0.72	0.23	0.15

Finally, the complicated nature of relevance means that there are many more interactions yet to be explored. Future work could investigate different exclusion or weighting criteria during worker vote aggregation, as well as additional decision rules such as using unanimity or varying thresholds instead of simple majority. Combining several user behaviours into a single weighted measure may also prove to be beneficial. More work is needed to investigate the application of worker quality measures with implicit measures of relevance, particularly in cases where explicit user judgments are not available. Examples include looking at detecting quality from research into variations in eye gaze or pupil dilation [23], behavioural data such as mouse-click rates [8, 20], or facial data and other biometrics [1, 22]. The exploration of these approaches would be useful for the human-computer interaction (HCI) community.

6 CONCLUSION

In this work, we conduct a user study on text documents to crowdsource relevance assessments against four topics. Worker judgements are used to calculate a range of ground truth measures against which classification performance is measured. Our results indicate that average subjective relevance and confidence-weighted average subjective relevance measures are on par with the annotations from an expert panel, which supports other lines of work in this area. In our study, measures based upon self-reported familiarity or task completion speed were not useful.

The results contribute toward developing higher quality label sets from crowdsourced relevance tasks. We explore user behaviours and attributes, and how these can be applied as indicators of quality in crowdsourcing. Further research into this area is useful to the HCI community, as it will result in ensuring high quality outcomes from machine learning systems, from which end-users ultimately benefit. This work only begins to explore the possibilities of weighted average subjective relevance assessments, and further work is planned to expand on these findings.

ACKNOWLEDGMENTS

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

REFERENCES

[1] Arapakis, I., Konstas, I., & Jose, J. M. (2009, October). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM international conference on Multimedia* (pp. 461-470). ACM

[2] Chang, J. C., Amershi, S., & Kamar, E. (2017, May). Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2334-2346). ACM.

[3] Chow, C., & Gedeon, T. (2015, October). Classifying document categories based on physiological measures of analyst responses. In *Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on* (pp. 421-425). IEEE.

[4] Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM, 54*(4), 86-96.

[5] Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). The Cambridge handbook of expertise and expert performance. Cambridge University Press.

[6] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

[7] Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E. J., ... & Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences, 107*(36), 15916-15920.

[8] Guo, Q., & Agichtein, E. (2012, April). Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web* (pp. 569-578). ACM.

[9] Herrera-Viedma, E., Herrera, F., & Chiclana, F. (2002). A consensus model for multiperson decision making with different preference structures. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 32*(3), 394-402.

[10] Kairam, S., & Heer, J. (2016, February). Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1637-1648). ACM.

[11] Kazai, G., Kamps, J., & Milic-Frayling, N. (2012, October). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2583-2586). ACM.

[12] Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456). ACM.

[13] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

[14] Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., & Williams, A. C. (2017, February). Crowdsourcing as a Tool for Research: Implications of Uncertainty. In *CSCW* (pp. 1544-1561).

[15] Loh, C. S., & Sheng, Y. (2015). Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*(1), 5-19.

[16] Liu, D. R., & Wu, I. C. (2008). Collaborative relevance assessment for task-based knowledge support. *Decision Support Systems, 44*(2), 524-543.

[17] Maddalena, E., Basaldella, M., De Nart, D., Degl'Innocenti, D., Mizzaro, S., & Demartini, G. (2016, September). Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing.*

[18] Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

[19] Moshfeghi, Y., Rosero, A. F. H., & Jose, J. M. (2016). A game-theory approach for effective crowdsource-based relevance assessment. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(4), 55.

[20] Radlinski, F., & Joachims, T. (2005, August). Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 239-248). ACM.

[21] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. arXiv preprint arXiv:1707.01836.

[22] Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing & Management, 28*(4), 457-466.

[23] Salojärvi, J., Kojo, I., Simola, J., & Kaski, S. (2003, September). Can relevance be inferred from eye movements in information retrieval. In *Proceedings of WSOM* (Vol. 3, pp. 261-266).

[24] Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research, 136*(2), 253-263.

[25] Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Association for Computational Linguistics.

[26] Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med, 37*(5), 360-363.

[27] von Ahn, L. (2013, March). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 1-2). ACM.

[28] Whiting, M. E., Gamage, D., Gaikwad, S. N. S., Gilbee, A., Goyal, S., Ballav, A., ... & Sarma, T. S. (2017, February). Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1902-1913). ACM.

[29] Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999, June). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246-253). Association for Computational Linguistics.

[30] Xu, Y., & Wang, D. (2008). Order effect in relevance judgment. *Journal of the Association for Information Science and Technology, 59*(8), 1264-1275

[31] Yuan, A., Luther, K., Krause, M., Vennix, S. I., Dow, S. P., & Hartmann, B. (2016, February). Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1005-1017). ACM.

[32] Yuen, M. C., King, I., & Leung, K. S. (2011, October). A survey of crowdsourcing systems. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 766-773). IEEE.