Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing

Eddy Maddalena
University of Southampton
Southampton
United Kingdom
E.Maddalena@soton.ac.uk

Davide Ceolin
Centrum Wiskunde & Informatica
Amsterdam
The Netherlands
davide.ceolin@cwi.nl

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

Abstract

In the age of fake news and of filter bubbles, assessing the quality of information is a compelling issue: it is important for users to understand the quality of the information they consume online. We report on our experiment aimed at understanding if workers from the crowd can be a suitable alternative to experts for information quality assessment. Results show that the data collected by crowdsourcing seem reliable. The agreement with the experts is not full, but in a task that is so complex and related to the assessor's background, this is expected and, to some extent, positive.

1 Introduction and Background

Online information is used by a variety of stakeholders as a basis for decision making, knowledge discovery, studies, and many more activities. However, as a consequence of the democratic nature of the Web, such information shows an extremely diverse level of quality. Making explicit this level of quality for each information item is crucial to allow the stakeholders an overall adequate information perusal. Given their pervasiveness and influence on the public opinion, online news are a kind of information whose quality assessment becomes a particularly critical task to contrast the spread of misinformation and disinformation.

Assessing the quality of online news and information in general is a challenging task, because of its intrinsic complexity. Information quality can be assessed by considering diverse points of views; how they can be assessed, and how the assessment results should be combined, depends on the assessors and on their requirements. This calls for a combined approach, where automated computation is required to handle the huge amount of information available on the Web, while human computation is required to understand how the quality dimensions are assessed and combined. An important aspect of human computation in this context is its regularity: when human assessments are consistent enough, automated computation can leverage them to scale the computation up.

In a previous work by Ceolin, Noordegraaf, and Aroyo [CNA16], two user studies are performed to collect quality assessments regarding Web documents on the vaccination debate. Assessments were collected by means of a Web application, in a scenario similar to crowdsourcing with the only difference that the assessments were expressed by a few experts (media scholars and journalism students) rather than a large crowd of anonymous workers. This approach has been named *nichesourcing* [Boe+12]. Ceolin, Noordegraaf, and Aroyo noted that, when the task at hand is constrained, experts who show a similar background tend to significantly agree with each other. However, they also noted that the task of deeply assessing online information is rather demanding, and expert availability is limited. Crowdsourcing could be a solution to the limited availability of human assessors.

In this paper, we repeat that study [CNA16] though crowdsourcing to analyse similarities and differences among the two ways of collecting human assessments. Our ultimate goal is to determine if and how crowdsourcing is a suitable alternative to nichesourcing for information quality assessment. Section 2 briefly surveys related work, Section 3 describes the experimental setup we adopted, Section 4 presents the results, and

Section 5 concludes the paper.

2 Related Work

In the age of fake news [Laz+18; VRA18] and of the filter bubble [Par11], assessing the quality of information is a compelling issue: it is important for users to understand the quality of the information they consume online. Two important initiatives that are worth being mentioned in this field are the W3C Credible Web Community Group (https://credweb.org/) and the Credibility Coalition (http://credibilitycoalition.org). While the first is meant to establish standards to model and share data about the credibility of information online, the second aims at identifying markers and strategies for establishing the credibility of the same information. To this extent, the work we present in this paper is complementary to these initiatives, as it aims at providing gold standards to reason on the credibility (and, more broadly, quality) of online information.

3 Experimental Setup

3.1 Dataset Description

We ran our experiment on a sample from the vaccination debate dataset provided by the QuPiD project (http://qupid-project.net) and used by Ceolin, Noordegraaf, and Aroyo [CNA16]. In 2015, a measles outbreak took place at Disneyland, California. Such outbreak triggered a fierce debate that fleshed out the already hot discussions regarding vaccinations, where pro and anti vaccination individuals blamed each other for the responsibility of the event. The vaccination debate dataset collects a number of documents regarding that specific debate. While the dataset is limited in size (about 50 documents), it is rather diverse in terms of types of documents represented (newspaper articles, activist blog posts, etc.) and stances (pro, anti, neutral).

3.2 The Crowdsourcing Task

The crowdsourcing task we ran aimed at collecting laymen judgments concerning the quality of a subset of 20 articles assessed by the experts (media scholars and journalism students). We asked each worker to assess one document along eight different quality dimensions derived from Ceolin, Noordegraaf, and Aroyo [CNA16] (we slightly reformulated some of them to have a shorter description, more adequate for crowd workers):

1. Accuracy - How accurate is the information in this article?
2. Neutrality - Is the document neutral with respect to the topic addressed, or does it clear stance (e.g., pro, against)?
3. Readability - Does the document read well?
4. Precision - How precise is the information in this document (as opposed to vague)?
5. Completeness - How complete is the information in this document?
6. Trustworthiness - How trustworthy is the source? Is the source trustworthy or does it exhibit malicious intentions?
7. Relevance - How relevant is the article to the task?
8. Overall quality - Which is your general opinion about the quality of the article?

We also asked two further questions requiring workers personal opinion, to understand how personal belief affects quality judgment:

9. Your personal opinion - Do you agree with the document content?
10. Your confidence - How knowledgeable/expert are you about the topic?

All the 10 assessments were collected on a 5-stars Likert scale, as in the original experiment [CNA16]. For each quality dimension, we also asked the users to motivate their judgment by some free text.

The task ran on the Figure Eight (https://www.figure-eight.com/) crowdsourcing platform by selecting level-three workers who are highest accuracy contributors. Each worker was paid 0.2 USD and could not judge more than three articles. Besides redundancy (each article was judged by 10 workers), we also adopted some standard quality checks: each worker was shown a pair of articles of clearly low and high quality, and the work was rejected if the collected values were ranked in the wrong way; there was also a time threshold (the worker needed to spend at least 120 seconds on the task), and some syntactic checks on the free text motivations.

3.3 Research Questions

This experiment allows us to address three research questions:

Q1. Relationships between quality dimensions: what are the correlations between the quality dimensions? Do some of the quality dimensions correlate in a way that makes one derivable from another? What is the difference between experts and workers?

Q2. Internal agreement (between individual workers): can different workers agree to a reasonable extent when assessing quality dimensions? Are there differences among the dimensions?

Q3. External agreement (between individual workers and experts): what is the individual external agreement, i.e., the agreement between the individual workers and the experts, on all dimensions? What is the aggregate external agreement,

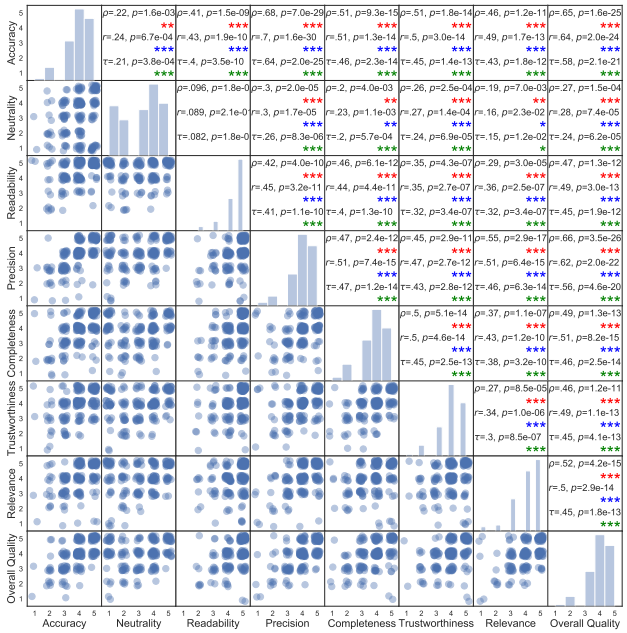

Figure 1: Scatterplots and correlations between the dimensions pairs, for raw worker values

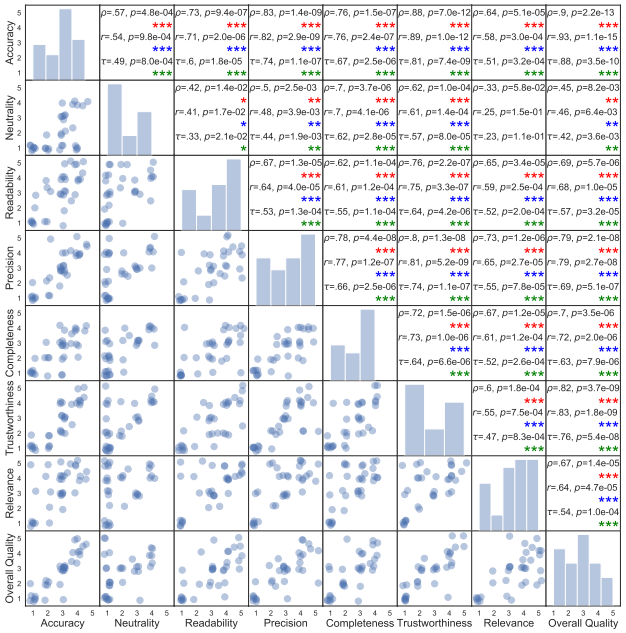

Figure 2: Scatterplots and correlations between the dimensions pairs, for the experts

i.e., the agreement between the aggregated assessments by the workers and the experts, on all dimensions?

4 Results

The main results are grouped on the basis of the research questions.

4.1 Q1: Quality Dimensions Relationships

A first result is presented in Figure 1, that shows a scatterplot matrix. For each pair of dimensions (indicated on the diagonal), a scatterplot is shown (in the bottom triangular matrix, with some random jitter to avoid some overlap). Each dot in a scatterplot represents one individual worker/article pair, and its coordinates are the values expressed by the worker on the corresponding two dimensions. In the upper triangular part, the correlation values are shown with their p-values to measure statistical significance.

Figure 2 allows to compare the data to experts. Comparing correlation values, it is clear that experts are more consistent across dimensions; p-values are roughly similar in the two cases.

As it is common practice in crowdsourcing, in place of using raw values by individual workers, we compute aggregated values. We select a simple (if not the simplest) aggregation function: the arithmetic mean. Figure 3 shows the correlations obtained when aggregating with the mean the 10 values expressed by 10 workers on the same article. When comparing to Figure 1, one can see that correlations increase, although they are less statistically significant. When comparing to Figure 2 one can see that usually the correlation between dimensions are higher for the experts than for the aggregate workers, but values are definitely more comparable than the individual raw values, and indeed the aggregate workers have higher correlations than the experts in three cases (the correlations between Accuracy and Relevance those between Overall Quality and both Neutrality and Precision). We also tried aggregating with the median, obtaining worse results.

Another remark that can be made by observing the histograms on the diagonals of Figures 1 and 2 is that the values provided by the experts tend to follow a more Bimodal distributions (they use more the extremes of the scale) than the workers. This is even clearer when looking at the aggregated values since the mean of the values will pull them even more towards the middle of the scale, as it can be seen in Figure 3. The distributions also show that the workers tend to express higher values than the experts.

4.2 Q2: Internal Agreement among Workers

Table 1 shows the agreement among the workers, overall and on each quality dimension, measured by both Krippendorff's α [Kri07] and Φ [Che+17]. Both measures assume values in $[-1, +1]$ (with -1 corresponding to complete disagreement, 0 to random agreement, and $+1$ to complete agreement). For Φ the table also shows, besides the most likely Φ value, the Highest Posterior Density (HPD) interval, i.e., the interval that contains the actual Φ value with a 95% probability: these are quite small intervals, so we can be confi-

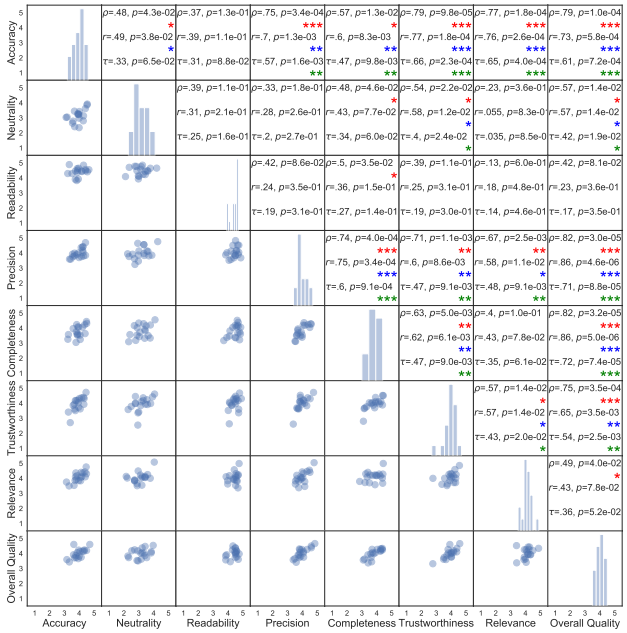

Figure 3: Scatterplots and correlations between the dimensions pairs, for aggregated (mean) worker values.

Dimension	α	Φ	HPD [2.5, 97.5]
All	0.132	0.084	[0.014, 0.146]
Accuracy	0.057	0.800	[0.747, 0.836]
Neutrality	0.016	0.703	[0.609, 0.778]
Readability	0.012	0.687	[0.500, 0.831]
Precision	0.026	0.807	[0.773, 0.868]
Completeness	0.065	0.876	[0.816, 0.903]
Trustworthiness	0.108	0.904	[0.827, 0.954]
Relevance	0.022	0.739	[0.716, 0.783]
Overall Quality	0.011	0.833	[0.805, 0.852]

Table 1: Agreement among the workers

dent that the most likely Φ value is correct. α values are quite low, but Φ ones are much higher. Most likely, as we have discussed above, assessment values have a quite low variability. In such a case, α exhibits a pathological behavior, which is of the issues with α that is solved by Φ as discussed by Checco et al. [Che+17]. The much higher Φ values, together with the narrow HPD intervals, show that the agreement among the workers is consistent even if not complete.

The results presented so far hint that the data collected by our crowdsourcing experiment are reliable. It is also important to remark that although the workers in some cases fail to exactly replicate the assessments by the experts (as we discuss shortly), the task is quite complex and assessor background might have a critical role. In this respect, a full agreement might even be a problem rather than a feature. If this is the case, it might be necessary to treat in a different way different worker groups, and/or decrease the granularity and ask to evaluate passages of an article instead of a full article. In this light, we observe a low correlation (between 0 and 0.20) between the workers confidence, i.e., question number 10, and all the quality dimensions and a moderate correlation (about 0.6) between the workers agreement, i.e., question 9, with the article assessed and Precision, Accuracy, and Overall Quality scores. While this correlation is not complete, it still hints at the possibility that a subgroup of the workers shows a confirmation bias, meaning that these tend to judge positively the articles they agree with, and vice-versa. In this short paper we do not have the space to discuss these issues in full, and we leave them for future work.

4.3 Q3: External Agreement with the Experts

Turning to the agreement between workers and experts, the scatterplots and correlations values in Figure 4 (top row) show that the agreement of the individual workers with the experts is rather low, as correlation values are positive but quite small, and often not significant. Figure 4 (center row) shows the agreement with the experts that is obtained when aggregating the worker values with the mean. Correlation values are systematically higher than individual workers, although almost never greater than 0.5 and often not statistically significant. As previously observed, the aggregation reduces the range of the values: whereas the experts usually use the full spectrum, the aggregated workers score is more limited. In all these plots, the eight dimensions show quite similar correlation values with the exception of Neutrality: workers particularly disagree with the experts about it.

Figure 4 (bottom row) demonstrates the previous claim that in general the median is a worse aggregation function: lower correlation values are obtained for Completeness, Trustworthiness, Relevance, and, especially, Overall Quality (which has not correlation with the experts when using the median). However, Readability and Precision are similar, and Neutrality and, especially, Accuracy are higher. This suggests that different and more sophisticate aggregation functions might lead to a higher agreement with the experts, an issue that for space limits we leave for future work.

5 Conclusions and Future Work

In this paper we present an experiment that aims at comparing crowd and nichesourcing as methods for assessing the quality of online information from a multidimensional standpoint. We collect 10 assessments about 20 articles from a dataset on the vaccination debate, and we analyze them internally and in comparison to previously published expert assessments. We

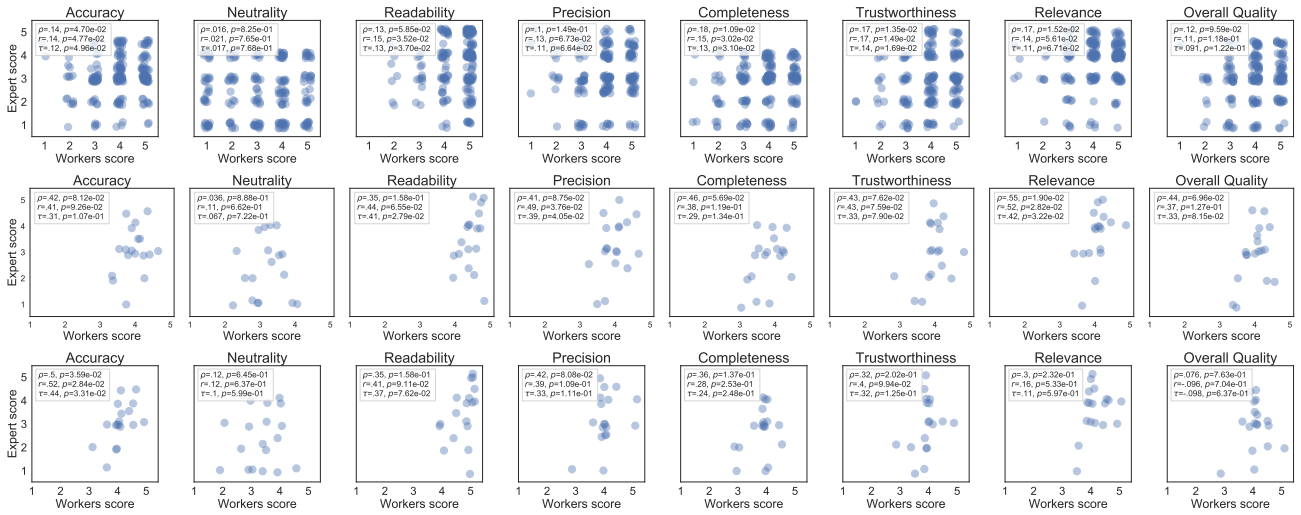

Figure 4: Scatterplots and correlations between experts and: (i) individual workers (top row); (ii) aggregated workers, with mean as aggregation function (center row); and (iii) aggregated workers, with median as aggregation function (bottom row).

observe that workers tend to use higher values than experts, and that aggregate workers values show a higher correlation in three cases (between Accuracy and Relevance, and between Overall Quality and Neutrality and Precision). When looking at the internal agreement among workers, we note that this is high, but not complete. This might be due to the fact that, at least some workers, show a confirmation bias, *i.e.*, tend to rate higher documents they agree with, and vice-versa. Lastly, when looking at the agreement between workers and experts, we can see that this is generally high, except for the Neutrality dimension.

In the future, we plan to extend our dataset to increase the number of assessments, of articles analysed, and of topics covered to help us generalise our findings. We plan to extend the depth of our analyses, for example to identify an assessability measure for documents (hinting at how easy it is to assess them), and to identify similar groups of workers with higher internal agreement.

Acknowledgements This study was partially supported by the *H2020* project *QROWD* (grant agreement ID: 732194).

References

[Boe+12] Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. "Nichesourcing: Harnessing the Power of Crowds of Experts". In: *Knowledge Engineering and Knowledge Management*. Springer Berlin Heidelberg, 2012, pp. 16–20.

[Che+17] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. "Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing". In: *The 5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2017)*. 2017.

[CNA16] Davide Ceolin, Julia Noordegraaf, and Lora Aroyo. "Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments". In: *Knowledge Engineering and Knowledge Management*. Springer International Publishing, 2016, pp. 83–97.

[Kri07] Klaus Krippendorff. "Computing Krippendorff's alpha reliability". In: *Departmental papers (ASC)* (2007), p. 43.

[Laz+18] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. "The science of fake news". In: *Science* 359.6380 (2018), pp. 1094–1096.

[Par11] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, 2011.

[VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151.