Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias

David La Barbera[†], Kevin Roitero[†], Gianluca Demartini[‡], Stefano Mizzaro[†], and Damiano Spina[◇]

[†] University of Udine, Udine, Italy, {labarbera.david,roitero.kevin}@spes.uniud.it, mizzaro@uniud.it

[‡] University of Queensland, Brisbane, Australia, g.demartini@uq.edu.au

[◇] RMIT University, Melbourne, Australia, damiano.spina@rmit.edu.au

Abstract. News content can sometimes be misleading and influence users' decision making processes (e.g., voting decisions). Quantitatively assessing the truthfulness of content becomes key, but it is often challenging and thus done by experts. In this work we look at how experts and non-expert assess truthfulness of content by focusing on the effect of the adopted judgment scale and of assessors' own bias on the judgments they perform. Our results indicate a clear effect of the assessors' political background on their judgments where they tend to trust content which is aligned to their own belief, even if experts have marked it as false. Crowd assessors also seem to have a preference towards coarse-grained scales, as they tend to use a few extreme values rather than the full breadth of fine-grained scales.

1 Introduction

The credibility of information available online may vary and the presence of untrustworthy information has big implications on our safety online [5,12,15]. The recent increase of misinformation online is to be blamed on technologies that have enabled the next level of strategic politic propaganda. Social media platforms and their data allow for extreme personalization of content which makes it possible to individually customise information. Given that the majority of people access news from social media platforms [13] such strategies can be used towards the goal of influencing decision making processes [1,14].

In this constantly evolving scenario, it is key to understand how people perceive the *truthfulness of information* presented to them. To this end, in this paper we collect data from US-based crowd workers and compare it with expert annotation data generated by fact-checkers such as PolitiFact. Our dataset contains multiple judgments of truthfulness of information collected from several non-expert assessors to measure agreement levels and to identify controversial content. We also collect judgments over two different judgment scales and collect information about assessors' background that allows us to analyse assessment bias. The dataset we created is publicly available at https://github.com/KevinRoitero/crowdsourcingTruthfulness.

The results of our analysis indicate that: 1) crowd judgments can be aggregated to approximate expert judgments, 2) there is a political bias in crowd-generated truthfulness labels where crowd assessors tend to believe more to statements coming from speakers off the same political party they have voted for in the last election; and 3) there seems to be a preference for coarse-grained scales where crowd assessors tend to use the extreme values in the scale more often than other values.

2 Related Work

Crowdsourcing has been previously used as a methodology in the context of information credibility research. For example, Zubiaga and Heng [17] looked at how tweet credibility can be assessed by means of Amazon MTurk workers in the context of disaster management. Their results show that it is difficult for crowd workers to properly assess the truthfulness of tweets in this context, but that the reliability of the source is a good indicator for trusted information. Kriplean et al. [6] analyse how volunteer crowdsourcing can be used for fact-checking by simulating the democratic process. The Fact-checking Lab at CLEF [9,3] looks at this problem by defining the task of ranking sentences according to their need to be fact-checked. Maddalena et al. [7] focus on the ability of the crowd to assess news quality along eight different quality dimensions. Roitero et al. [10] use crowdsourcing to study user perception of fake news statements. As compared to previous studies looking at crowdsourcing for information credibility tasks, we look at bias in the data due to the assessor and the rating scale used to collected labels in the context of the truthfulness of statements by US politicians.

3 Methodology

3.1 Dataset Description

In our study[1] we use the PolitiFact dataset constructed by Wang [16]. This dataset contains 12 800 statements by politicians with truth labels produced by expert fact-checkers on a 6-level scale: i.e., `True`, `Mostly True`, `Half True`, `Barely True`, `False`, and `Lie`.[2] For this work, we selected a subset of 120 statements randomly sampled from the PolitiFact dataset to make sure that a balanced number of statements per class and per political party was included in the sample.

3.2 Crowdsourcing Setup

We crowdsourced 120 statements each judged by 10 distinct crowd workers across 400 HITs on Amazon MTurk asking US-based workers to label the truthfulness

[1] The setup was reviewed and approved by the Human Research Ethics Committee at the University of Queensland.

[2] In the original dataset, `Pants on Fire` is used; we preferred `Lie` to facilitate workers.

Table 1. Most frequently used support URLs over both scales, with and without gold questions.

	PolitiFact	**Wikipedia**	**WashingtonPost**	**Google**	**Youtube**	**CNN**	**NYTimes**
S6 + Gold	0.55	0.16	0.01	0.02	0.01	0.00	0.1
S6 − Gold	0.73	0.06	0.01	0.01	0.00	0.01	0.01
S100 + Gold	0.54	0.13	0.02	0.02	0.03	0.00	0.1
S100 − Gold	0.72	0.03	0.01	0.01	0.00	0.01	0.1

of statements from the dataset. Each HIT, rewarded $1.20 (i.e., $0.15 for each statement), consisted of 8 statements for which we asked an assessment either using the original 6-level scale (S6) or a 100-level scale (from 0 to 100) using a slider set by default at 50 (S100). The 8 statements contained 2 gold questions used to quality check the workers' responses by means of providing judgments consistent with the expert ground truth. Other than gold questions, each HIT contained 3 statements by Republican party speakers and 3 by Democratic party speakers. More than the judgments, crowd workers where also asked to provide a justification for each of their judgments, and a URL pointing to the source of information supporting their judgment. At the beginning of the HIT each worker was asked to complete a demographics questionnaire; it also included questions about their political orientation, used to classify crowd assessors as aligned to the US Democratic party (Dem) or the US Republican party (Rep).

4 Results

Participants (400 in total) were well balanced across the political spectrum (108 Dem and 92 Rep for S6; 109 Dem and 91 Rep for S100) and most of them have a college degree. Many crowd assessors used the PolitiFact website as a source of evidence for their judgments (55.4% of judgments done with S6 and 54.9% of judgments done with S100) as shown in Table 1. The list of URLs in this table also shows how the majority of crowd assessors performed the task correctly, as they tried to refer to reliable sources such as PolitiFact, Wikipedia, Washington Post, CNN, and New York Times.

4.1 Judgment Distributions

Figure 1 shows the raw assessment score distributions given by crowd workers both for the S6 scale and the S100 scale. These results hint that workers tend to use fewer values than those available: the extremes of the scales are more used; for S100, the middle value (50) is also frequently used and some smaller peaks can be seen in correspondence of the multiples of 10. These outcomes are much less manifest in the aggregated scores (i.e., the arithmetic mean of the scores provided by the ten crowd workers judging the same statement), shown in Figure 2: as usual, these are more evenly distributed. Also, values at the lower end of the scale are much less frequent. These outcomes suggest that perhaps a two- or three-level scale would be more appropriate for this task, although

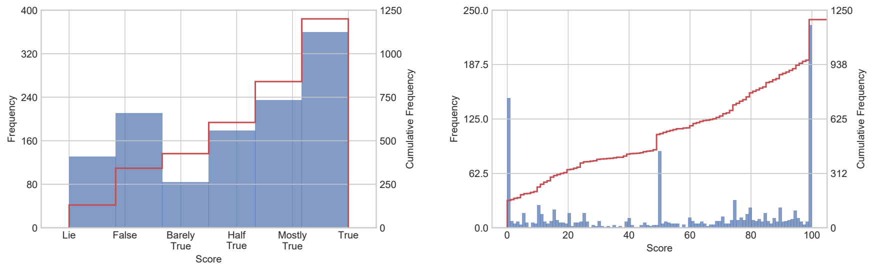

Fig. 1. Individual score distributions: S_6 (left, raw judgments), S_{100} (right, raw judgments). The red line shows the cumulative distribution of judgments.

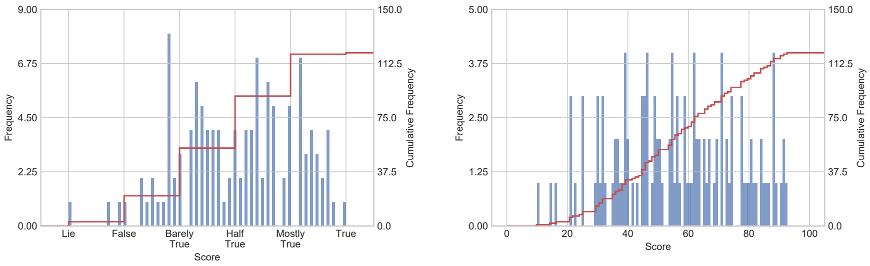

Fig. 2. Distribution of scores aggregated by mean: S_6 (left), S_{100} (right). The red line shows the cumulative distribution of judgments.

fine-grained scales have been successfully used in a crowdsourcing setting for relevance assessment [8,11]. We intend to further address this issue in our future work also using the scale transformation techniques proposed by Han et al. [4].

4.2 Crowd vs. Experts

Figure 3 shows the crowd assessor labels as compared to expert judgment of truthfulness over both judgments scales. Crowd assessors seem able to distinguish among the different levels of the S6 scale as the median values are increasing following the expert assessments over the levels of the scale. A t-test comparing crowd assessor scores across expert judgment levels shows that crowd scores are significantly different ($p < 0.01$) across all levels except for the class combinations `Lie`–`False`, `False`–`Barely True`, and `Mostly True`–`True`. The crowd appears to be more lenient than experts in assessing the truthfulness of statements as scores for the lowest categories tend not to reach the bottom end of the scale in both S6 and S100. This suggests the need to align scales when used by crowd assessors and experts in order to identify misleading content using crowdsourcing. Overall, S6 seems more adequate than S100, not only because (as noted above) workers tend to use coarse-grained scales but also because the agreement with experts

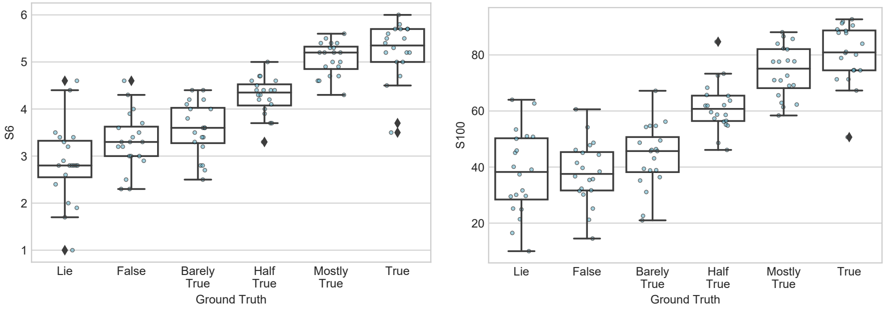

Fig. 3. Comparison of crowd labels with expert ground truth: S_6 (left), S_{100} (right).

seems higher for S6: S100 presents wider boxplots and less separable categories, especially for the first three classes (`Lie`, `False`, `Barely True`).

4.3 Crowd Assessor Bias

Figure 4 shows how crowd assessors labelled statements as compared to ground truth expert labels based on their political background. We can see that crowd assessors who voted for the Rep party tend to assign higher truthfulness scores, especially for the `Lie` and `False` ground truth labels, showing how, on average, they believe to content more than crowd assessors who voted for the Dem party.

When comparing how crowd workers assess statements differently based on who the speaker is, we can observe that `True` statements obtain higher scores from crowd assessors who voted for the speaker's party. That is, Dem workers assigned an average score of 84.54 on S100 and 5.48 on S6 to `True` statements by Dem speakers and only 81.83 on S100 and 5.00 on S6 to `True` statements by Rep speakers. Rep workers assigned an average score of 81.89 on S100 and 5.35 on S6 to `True` statements by Rep speakers and only 73.24 on S100 and 4.73 on S6 to `True` statements by Dem speakers. While this is an expected behaviour, we also notice that Dem crowd assessors appear to be more skeptical than Rep crowd assessors by showing a lower average judgment score for untrue statements (e.g., Fig. 4, top row).

5 Conclusions

In this paper we presented a dataset of crowdsourced truthfulness judgments for political statements and compared the collected judgments across different crowd assessors, judgments scales, and with expert judgments.

Our results show that 1) crowd judgments, if properly aggregated, are comparable to expert ones 2) crowd assessors political background has an impact on how they label political statements: they show a tendency to be more lenient towards statements by politicians of the same political orientation as their

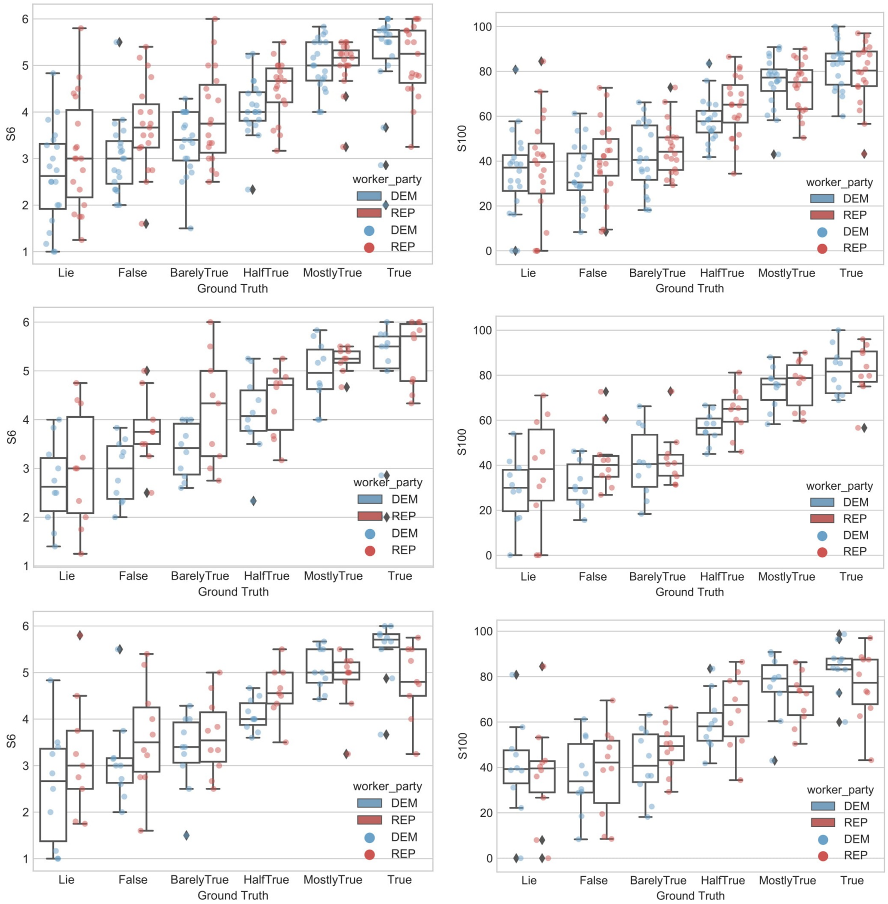

Fig. 4. Comparison with ground truth for Dem workers (blue) and Rep workers (red): S_6 (left), S_{100} (right). All statements (first row), Rep speaker statements (second row) and Dem speaker statements (third row).

own; and 3) crowd assessors seem to have a preference towards coarse-grained judgment scales for truthfulness judgements.

Acknowledgements. This work is partially supported by an Australian Research Council Discovery Project (DP190102141) and a Facebook Research award.

References

1. Bittman, L., Godson, R.: The KGB and Soviet disinformation: an insider's view. Pergamon-Brassey's (1985)
2. Cuzzocrea, A., Bonchi, F., Gunopulos, D. (eds.): Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information

and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018, CEUR Workshop Proceedings, vol. 2482. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2482

3. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims. In: Proceedings of CLEF. pp. 301–321 (2019)
4. Han, L., Roitero, K., Maddalena, E., Mizzaro, S., Demartini, G.: On transforming relevance scales. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM) (2019)
5. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision support systems **43**(2), 618–644 (2007)
6. Kriplean, T., Bonnar, C., Borning, A., Kinney, B., Gill, B.: Integrating on-demand fact-checking with public dialogue. In: Proceedings of CSCW. pp. 1188–1199 (2014)
7. Maddalena, E., Ceolin, D., Mizzaro, S.: Multidimensional news quality: A comparison of crowdsourcing and nichesourcing. In: Cuzzocrea et al. [2], http://ceur-ws.org/Vol-2482/paper17.pdf
8. Maddalena, E., Mizzaro, S., Scholer, F., Turpin, A.: On crowdsourcing relevance magnitudes for information retrieval evaluation. ACM Transactions on Information Systems **35**(3), 19:1–19:32 (Jan 2017). https://doi.org/10.1145/3002172, http://doi.acm.org/10.1145/3002172
9. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In: Proceedings of CLEF (2018)
10. Roitero, K., Demartini, G., Mizzaro, S., Spina, D.: How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. In: Cuzzocrea et al. [2], http://ceur-ws.org/Vol-2482/paper38.pdf
11. Roitero, K., Maddalena, E., Demartini, G., Mizzaro, S.: On fine-grained relevance scales. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 675–684. SIGIR '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3210052, http://doi.acm.org/10.1145/3209978.3210052
12. Self, C.C.: Credibility. In: An integrated approach to communication theory and research, pp. 449–470. Routledge (2014)
13. Shearer, E., Gottfried, J.: News use across social media platforms 2017. Pew Research Center **7** (2017)
14. Starbird, K.: Disinformations spread: bots, trolls and all of us. Nature p. 449 (2019)
15. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health informationa survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(5), e1209 (2017)
16. Wang, W.Y.: "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426 (2017)
17. Zubiaga, A., Ji, H.: Tweet, but verify: epistemic study of information verification on twitter. Social Network Analysis and Mining **4**(1), 163 (2014)