How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing

Kevin Roitero[+], Gianluca Demartini[*], Stefano Mizzaro[‡], and Damiano Spina[◇]

[+] University of Udine, Udine, Italy, roitero.kevin@spes.uniud.it
[*] University of Queensland, Brisbane, Australia, g.demartini@uq.edu.au
[‡] University of Udine, Udine, Italy, mizzaro@uniud.it
[◇] RMIT University, Melbourne, Australia, damiano.spina@rmit.edu.au

Abstract

We report on collecting truthfulness values (i) by means of crowdsourcing and (ii) using fine-grained scales. In our experiment we collect truthfulness values using a bounded and discrete scale with 100 levels as well as a magnitude estimation scale, which is unbounded, continuous and has infinite amount of levels. We compare the two scales and discuss the agreement with a ground truth provided by experts on a six-level scale.

1 Introduction

Checking the validity of statements is an important task to support the detection of rumors and *fake news* in social media. One of the challenges is the ability to scale the collection of validity labels for a large number of statements.

Fact-checking has been shown as a task difficult to be performed in crowdsourcing platforms.[1] However, crowdworkers are often asked to annotate truthfulness of statements using a few discrete values (e.g., true/-false labels).

Recent work in information retrieval [Roi+18; Mad+17] has shown that using more fine-grained scales (e.g., a scale with 100 levels) presents some advantages with respect to classical few levels scales. Inspired by these works, we look at different truthfulness scales and experimentally compare them in a crowdsourcing setting. In particular, we compare two novel scales: a discrete scale on 100 levels, and a continuous Magnitude Estimation scale [Mos77]. Thus our specific research question is: *What is the impact of the scale to be adopted when annotating statement truthfulness via crowdsourcing?*

[1] https://fullfact.org/blog/2018/may/crowdsourced-factchecking/

2 Background

Recent work looked at the methods to automatically detect fake news and fact-check. Kriplean et al. [Kri+14] look at the use of volunteer crowdsourcing to fact-check embedded into a socio-technical system similar to the democratic process. As compared to them, we look at the more systematic involvement of humans in the loop to quantitatively assess the truthfulness of statements.

Our work looks at experimentally comparing different schemes to collect labelled data for truthful facts. Related to this, Medo and Wakeling [MW10] investigate how the discretization of ratings affects the co-determination procedure, i.e., where estimates of user and object reputation are refined iteratively together.

Zubiaga et al. [Zub+18] and Zubiaga and Ji [ZJ14] look at how humans assess credibility of information and, by means of a human study, identify key credibility perception features to be used for automatic detection of credible tweets. As compared to them, we also look at the human dimension of credibility checking but rather focus on which is the most appropriate scale for human assessors to make such assessment.

Kochkina, Liakata, and Zubiaga [KLZ18b] and Kochkina, Liakata, and Zubiaga [KLZ18a] look at ru-

mour verification by proposing a supervised machine learning model to automatically perform such a task. As compared to them, we focus on understanding the most effective scale used to collect training data to then build such models.

Besides the dataset we used for our experiments in this paper, other datasets related to fact checking and the truthfulness assessment of statements have been created. The Fake News Challenge[2] addresses the the task of stance detection: estimate the stance of a body text from a news article relative to a headline. Specifically, the body text may agree, disagree, discuss or be unrelated to the headline. Fact-checking Lab at CLEF 2018 [Nak+18] addresses a ranking task, i.e., to rank sentences in a political debate according to their worthiness for fact-checking, and a classification task, i.e., given a sentence that is worth checking, to decide whether the claim is true, false or unsure of its factuality. In our work we use the dataset first proposed by Wang [Wan17] as it has been created using six-level labels which is in-line with our research question about how many levels are most appropriate for such labelling task.

3 Experimental Setup

3.1 Dataset

We use a sample of statements from the dataset detailed by Wang [Wan17]. The dataset consists of a collection of 12,836 labelled statements; each statement is accompanied by some meta-data specifying its "speaker", "speaker's job", and "context" (i.e., the context in which the statement has been said) information, as well as the the truth label made by experts on a six-level scale: pants-fire (i.e., lie), false, barely-true, half-true, mostly-true, and true.

For our re-assessment, we perform a stratified random sampling to select 10 statements for each of the six categories, obtaining a total of 60 statements. The screenshot in Figure 1 shows one of the statements included in our sample.

3.2 The Crowdsourcing Task

We obtain for each statement a crowdsourced truth label by 10 different workers. Each worker judges six statements (one for each category) plus two additional "gold" statements used for quality checks. We also ask each worker to provide a justification for the truth value he/she provide.

We pay the workers 0.2$ for each set of 8 judgments (i.e., one Human Intelligent Task, or HIT). Workers are allowed to do one HIT for each scale only, but they are allowed to provide judgments for both scales.

[2] http://www.fakenewschallenge.org/

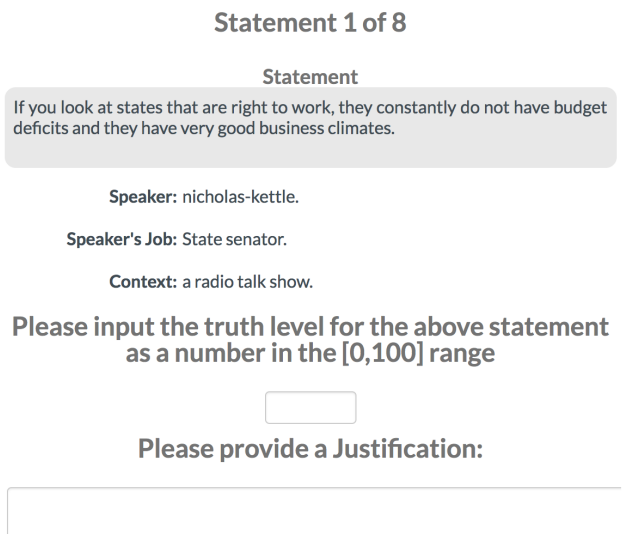

Figure 1: Example of a statement included in a crowdsourcing HIT.

We use randomized statement ordering to avoid any possible document-ordering effect/bias.

To ensure a good quality dataset, we use the following quality checks in the crowdsourcing phase:

- the truth value of the two gold statements (one patently false and the other one patently true) has to be consistent;
- the time spent to judge each statement has to be greater than 8 seconds;
- each worker has two attempts to complete the task; at the third unsuccessful attempt of submitting the task the user is prevented to continue further.

We collected the data using the Figure-Eight platform.[3]

3.3 Labeling Scales

We consider two different truth scales, keeping the same experimental setting (i.e., quality checks, HITs, etc.):

1. a scale in the $[0, 100]$ range, denoted as S_{100};
2. the Magnitude Estimation [Mos77] scale in the $(0, \infty)$ range, denoted as ME_{∞}.

The effects and benefits of using the two scales in the setting of assessing document relevance for information retrieval evaluation has been explored by Maddalena et al. [Mad+17] and Roitero et al. [Roi+18].

Overall, we collect 800 truth labels for each scale, so 1,600 in total, for a total cost of 48$ including fees.

[3] https://www.figure-eight.com/.

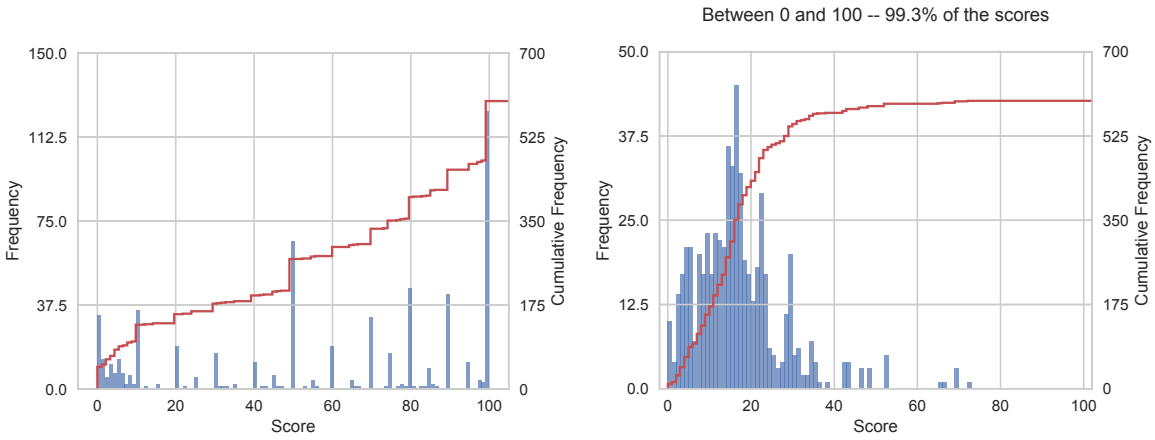

Figure 2: Individual score distributions: S_{100} (left, raw), and ME_{∞} (right, normalized). The red line is the cumulative distribution.

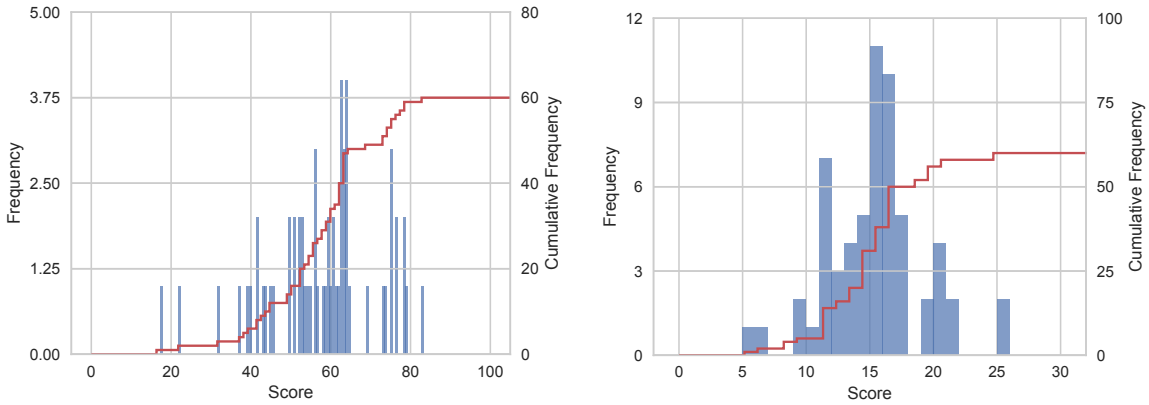

Figure 3: Aggregated score distributions: S_{100} (left), and ME_{∞} (right).

4 Results

4.1 Individual Scores

While the raw scores obtained with the S_{100} scale are ready to use, the scores from ME_{∞} need a normalization phase (since each worker will use a personal, and potentially different, "inner scale factor" due to the absence of scale boundaries); we computed the normalized scores for the ME_{∞} scale following the standard normalization approach for such a scale, namely geometric averaging [Ges97; McG03; Mos77]:

$$s^* = \exp\big(\log s - \mu_H(\log s) + \mu(\log s)\big),$$

where s is is the raw score, $\mu_H(\log s)$ is the mean value of the $\log s$ within a HIT, and $\mu(\log s)$ is the mean of the logarithm of all ME_{∞} scores.

Figure 2 shows the individual scores distributions: for S_{100} (left) the raw scores are reported and for ME_{∞} (right) the normalized scores. The x-axis represents the score, while the y-axis its absolute frequency; the cumulative distribution is denoted by the red line. As we can see, for S_{100} the distribution is skewed towards higher values, i.e., the right of the plot, and there is a clear tendency of giving scores which are multiple of ten (an effect that is consistent with the findings by Roitero et al. [Roi+18]).

For the ME_{∞} scale, we see that the normalized scores are almost normally-distributed (which is consistent with the property that scores collected on a ratio scale like ME_{∞} should be log-normal), although the distribution is slightly skewed towards lower values (i.e., left part of the plot).

4.2 Aggregated Scores

Next, we compute the aggregated scores for both scales: we aggregate the scores of the ten workers judging the same statement. Following the standard practices, we aggregate the S_{100} values using the arithmetic mean, as done by Roitero et al. [Roi+18], and the ME_{∞} values using the median, as done by Maddalena et al. [Mad+17] and Roitero et al. [Roi+18]. Figure 3 shows the aggregated scores; comparing with Figure 2, we notice that for S_{100} the distribution is

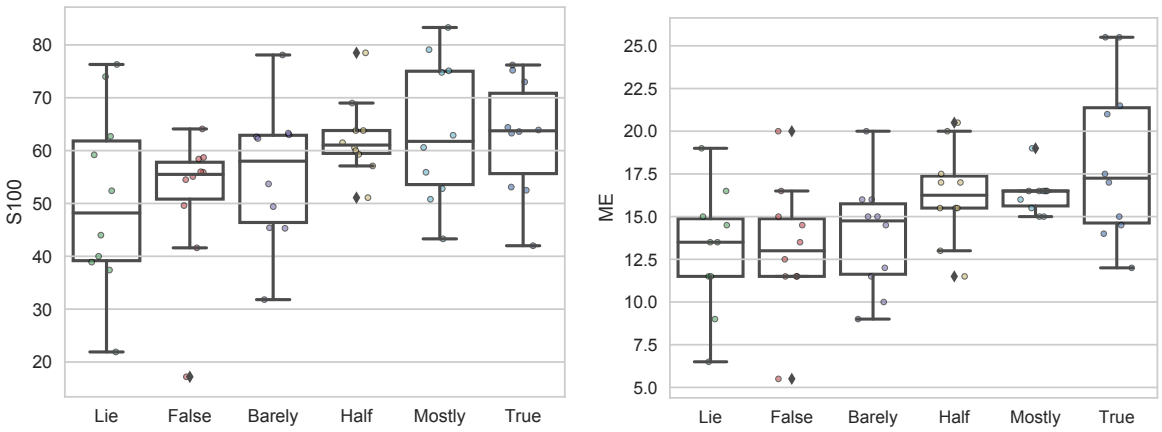

Figure 4: Comparison with ground truth: S_{100} (top), and ME_∞ (bottom).

more balanced, although it can not be said to be bell-shaped, and the decimal tendency effect disappears; furthermore, the most common value is not 100 (i.e., the limit of the scale) anymore. Concerning ME_∞, we see that the scores are still roughly normally distributed.[4] However, the x-range is more limited; this is an effect of the aggregation function, which tends to remove the outlier scores.

4.3 Comparison with Experts

We now turn to compare with the ground truth our truth levels obtained by crowdsourcing. Figure 4 shows the comparison between the S_{100} and ME_∞ (normalized and) aggregated scores with the six-level ground truth. In each of the two charts, each box-plot represents the corresponding scores distribution. We also report the individual (normalized and) aggregated scores as colored dots with some random horizontal jitter. We can see that, even with a small number of documents (i.e., ten for each category), the median values of the box-plots are increasing; this is always the case for S_{100}, and true for most of the cases for ME_∞ (where there is only one case in which this is untrue, for the two adjacent categories "Lie" and "False"). This behavior suggests that both the S_{100} and ME_∞ scales allow to collect truth levels that are overall consistent with the ground truth, and that the S_{100} scale leads to a slightly higher level of agreement with the expert judges than the ME_∞ scale. We analyze agreement in more detail in the following.

[4]Running the omnibus test of normality implemented in `scipy.stats.normaltest` [DP73], we cannot reject the null hypothesis, i.e., $p > .001$ for both the aggregated and raw normalized scores. Although not rejecting the null hypothesis does not necessary tell us that they follow a normal distribution, we can say we are pretty confident they came from a normal distribution.

4.4 Inter-Assessor Agreement

Figure 5 shows the inter-assessor agreement of the workers, namely the agreement among all the ten workers judging the same statement. Agreement is computed using Krippendorff's α [Kri07] and Φ Common Agreement [Che+17] measures; as already pointed out by Checco et al. [Che+17], Φ and α measure substantially different notions of agreement. As we can see, while the two agreement measures show some degree of similarity for S_{100}, for ME_∞ the agreement computed is substantially different: while α has values close to zero (i.e., no agreement), Φ shows a high agreement level, on average around 0.8. Checco et al. [Che+17] show that α can have an agreement value of zero even when the agreement is actually present in the data. Although agreement values seem higher for ME_∞, especially when using Φ, it is difficult to clearly prefer one of the two scales from these results.

4.5 Pairwise Agreement

We also measure the agreement within one unit. We use the definition of pairwise agreement by Roitero et al. [Roi+18, Section 4.2.1] that allows to compare (S_{100} and ME_∞) scores with a ground truth on different scales (six levels). Figure 6 shows that the pairwise agreement with the experts of the scores collected using the two scales is similar.

4.6 Differences between the two Scales

As a last result, we note that the two scales measure something different, as shown by the scatter-plot in Figure 7. Each dot is one statement and the two coordinates are its aggregated scores on the two scales. Although Pearson's correlation between the two scales is positive and significant, it is clear that there are some differences, that we plan to study in future work.

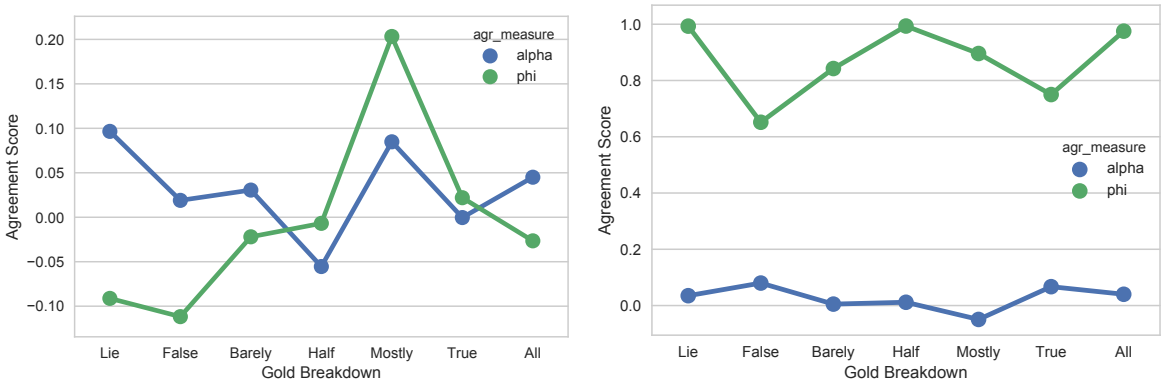

Figure 5: Assessor agreement: S_{100} (left), and ME_{∞} (right).

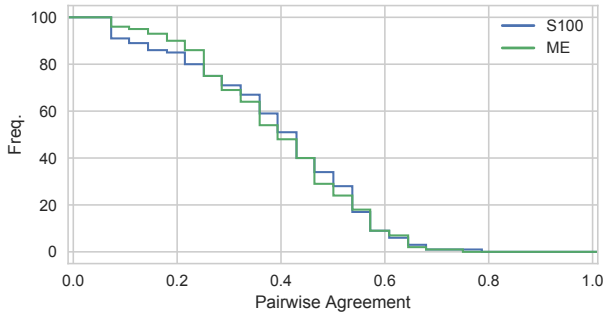

Figure 6: Complementary cumulative distribution function of assessor agreement for S_{100} and ME_{∞}.

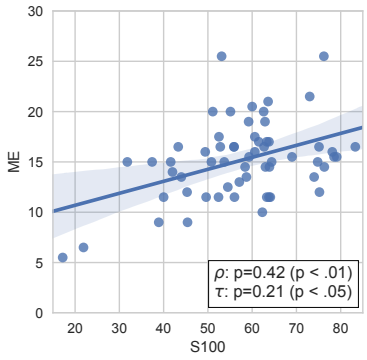

Figure 7: Agreement of the aggregated scores between S_{100} and ME_{∞}.

5 Conclusions and Future Work

We performed a crowdsourcing experiment to analyze the impact of using different fine-grained labeling scales when asking crowdworkers to annotate truthfulness of statements. In particular, we tested two labeling scales: S_{100} [RMM17] and ME_{∞} [Mad+17]. Our preliminary results with a small sample of statements from Wang's dataset [Wan17] suggest that:

- Crowdworkers annotate truthfulness of statements in a way that is overall consistent with the experts ground truth collected on a six-levels scale (see Figure 4), thus it seems viable to crowdsource truthfulness of statements.
- Also due to the limited size of our sample (10 statements), we cannot quantify which is the best scale to be used in this scenario: we plan to further address this issue in future work. In this respect, we remark that whereas the reliability of the S_{100} scale is perhaps expected, it is worth noticing that the ME_{∞} scale, for sure less familiar, leads anyway to truthfulness values that are of comparable quality to the ones collected by means of the S_{100} scale.
- The scale used has anyway some effect, as it is shown by the differences in Figure 4, the different agreement values in Figure 5, and the rather low agreement between S_{100} and ME_{∞} in Figure 7.
- S_{100} and ME_{∞} scales seems to lead to similar agreement with expert judges (Figure 6).

For space limits, we do not report on other data like, for example, the justifications provided by the workers or the time taken to complete the job. We plan to do so in future work.

Our preliminary experiment is an enabling step to further explore the impact of different fine-grained labeling scales for fact-checking in crowdsourcing scenarios. We plan to extend the experiment with more and more diverse statements, also from other datasets, which will allow us to perform further analyses. We plan in particular to understand in more detail the differences between the two scales highlighted in Figure 7.

References

[Che+17] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. "Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing". In: *Proc. HCOMP*. 2017, pp. 11–20.

[DP73] Ralph D'Agostino and Egon S. Pearson. "Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$". In: *Biometrika* 60.3 (1973), pp. 613–622.

[Ges97] George Gescheider. *Psychophysics: The Fundamentals*. 3rd. Lawrence Erlbaum Associates, 1997.

[KLZ18a] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. "All-in-one: Multi-task Learning for Rumour Verification". In: *arXiv preprint arXiv:1806.03713* (2018).

[KLZ18b] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. "PHEME dataset for Rumour Detection and Veracity Classification". In: (2018).

[Kri+14] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. "Integrating On-demand Fact-checking with Public Dialogue". In: *Proc. CSCW*. 2014, pp. 1188–1199.

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

[Mad+17] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. "On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation". In: *ACM TOIS* 35.3 (2017), p. 19.

[McG03] Mick McGee. "Usability magnitude estimation". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47.4 (2003), pp. 691–695.

[Mos77] Howard R Moskowitz. "Magnitude estimation: notes on what, how, when, and why to use it". In: *Journal of Food Quality* 1.3 (1977), pp. 195–227.

[MW10] Matúš Medo and Joseph Rushton Wakeling. "The effect of discrete vs. continuous-valued ratings on reputation and ranking systems". In: *EPL (Europhysics Letters)* 91.4 (2010).

[Nak+18] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. "Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims". In: *Proc. CLEF*. 2018.

[RMM17] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. "Do Easy Topics Predict Effectiveness Better Than Difficult Topics?" In: *Proc. ECIR*. 2017, pp. 605–611.

[Roi+18] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. "On Fine-Grained Relevance Scales". In: *Proc. SIGIR*. 2018, pp. 675–684.

[Wan17] William Yang Wang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection". In: *Proc. ACL*. 2017, pp. 422–426.

[ZJ14] Arkaitz Zubiaga and Heng Ji. "Tweet, but verify: epistemic study of information verification on Twitter". In: *Soc. Net. An. and Min.* 4.1 (2014), p. 163. ISSN: 1869-5469.

[Zub+18] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. "Detection and resolution of rumours in social media: A survey". In: *ACM Computing Surveys (CSUR)* 51.2 (2018), p. 32.