Measuring the Quality of Annotations for a Subjective Crowdsourcing Task

Raquel Justo(✉), M. Inés Torres, and José M. Alcaide

Universidad del País Vasco UPV/EHU, Sarriena s/n, 48940 Leioa, Spain
raquel.justo@ehu.es

Abstract. In this work an algorithm devoted to the detection of low quality annotations is proposed. It is mainly focused on subjective annotation tasks carried out by means of crowdsourcing platforms. In this kind of task, where a good response is not necessarily prefixed, several measures should be considered in order to pick the different behaviours of annotators associated to bad quality results: time, inter-annotator agreement and repeated patterns in responses. The proposed algorithm considers all these measures and provide a set of workers whose annotations should be removed. The experiments carried out, over a sarcasm annotation task, show that once the low quality annotations were removed and acquired again a better labeled set was achieved.

Keywords: Supervised learning · Annotation · Crowdsourcing · Subjective language

1 Introduction

Within the Pattern Recognition framework supervised learning methods make use of great amounts of annotated data in order to build robust models. However, the annotation process is a challenging task that usually requires a lot of effort, time and/or money. Traditionally, expert annotators, trained for the specific task, were involved in this process. But finding this kind of annotators is difficult and the whole process results expensive and tedious. Moreover, there are subjective tasks, such as emotion or sentiment analysis, for which the training of the annotators is not desirable and it is more interesting to pick up the diversity of people's opinion as a reflection of reality. In the last years, crowdsourcing have emerged as an alternative and more efficient method to carry out annotations and it has been used in different areas related to information retrieval, speech recognition, natural language processing, etc. Amazon's Mechanical Turk[1] or CrowdFlower[2] are good examples of crowdsourcing platforms that have been extensively used for completing annotation tasks. Within this framework task requesters can reach a large number of freelance employees to solve the annotation of microtasks. Dividing the work in microtasks makes possible to have

[1] www.mturk.com.
[2] www.crowdflower.com.

L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 58–68, 2017.
DOI: 10.1007/978-3-319-58838-4_7

an annotation task completed by a wide variety of different annotators, in cases where the diversity means a plus.

In this work, we dealt with a very subjective task focused on the identification of text excerpts as sarcastic or not sarcastic. Since the sarcasm cannot be unambiguously defined and often depends on diverse factors like the perception of the reader, the cultural environment, etc. [8,20], the diversity of people's opinion is very valuable in this case. That is, we wanted to use what people understand as sarcasm more than what the dictionary says. Thus, the most appropriate way of carrying out the annotation process is to use crowdsourcing.

Although crowdsourcing is now widely accepted and represents the basis for data collection and resource annotation or validation, the quality of the provided results are still in question. In fact, measuring the reliability of data is not straightforward and there are different works dealing with it [9,13–15, 17,21]. According to [10] there are several dysfunctional worker types, *Incapable workers* that do not fulfil the needed requirements. *Malicious workers* that try to invalidate experiments by submitting wrong answers on purpose and *Distracted workers* that do not pay full attention to the task. All of these workers provide poor quality results. The origin of this behaviour can be diverse but the incentives associated to the completion of microtasks can explain it to some extent.

The most common way of detecting unreliable workers is to introduce gold-standard work units alongside normal work units. They consist of very simple questions with a known answer that have to be well answered by annotators not to be rejected. Although it is an effective mechanism for some tasks it is not applicable in all the cases (tasks with open questions, etc.). Additionally, malicious workers can find innovative ways to circumvent gold-standard work units such as learning the answers to test questions and then reusing those answers under different accounts [22]. Moreover, in subjective tasks, the gold-standard does not seem to be an effective mechanism since the selection of questions without ambiguous cases is subjective itself. Let us note that our annotation task is drastically different from an audio transcription task, or a medical cancer diagnosis, where there is a real transcription or a biopsy result that can be objectively assumed as a ground truth.

In fact, if we focus on those subjective tasks, there are additional problems related to the data quality not gathered in the aforementioned cases. In our specific task, for instance, we have detected workers whose perception is highly different from the rest of annotators. It does not compulsorily mean that they are *Incapable workers*, *Malicious workers*, or *Distracted workers*, the problem seems to be more related to a misunderstanding of what sarcasm means that might be due to a different cultural background. In these cases, although gold-standard could help, additional information is needed to evaluate the reliability of data.

In a similar task, where irony and sarcasm were annotated using crowdsourcing [11], majority voting was used along with the quality control algorithm designed specifically for quality management on Amazon Mechanical Turk [15]. However, a high multiplicity is needed to use majority voting as a good quality control measure and this has associated a high cost. On the other hand,

the algorithm described in [15] was based on the inter-annotator agreement. It was inspired in [7] which proposed an expectation maximisation algorithm that iterates until convergence. However, the agreement information among annotators on its own might fail to detect malicious workers that uses bots and can learn the estimated distribution of the provided answers. Moreover, the proposed algorithm was based on the error rate estimation of workers, thus, a pool of data, that should be big enough to achieve good estimations, is needed to start working.

Up to our knowledge, it does not exist a reliable way of measuring the quality of data for highly subjective annotation tasks. Moreover, even when considering measures related to inter-annotator agreement there is a great confusion and the selected measures are not well motivated. The main contribution of this work is (1) to select one of the most general inter-annotator agreement coefficient and use its difference to provide a score for each annotator (2) to define a new agreement coefficient that is capable of picking a different kind of information and (3) to propose an algorithm that can detect low quality annotations, due to different behaviors, in a subjective task where ambiguity is very significant. It can be applied in any time during the annotation process and takes into consideration the main issues related to all kind of annotators that provide bad quality results.

This paper is organized as follows, Sect. 2 provides a definition of the annotations task and Sect. 3 describes the measures and the methodology employed to detect unreliable workers. In Sect. 4 a description of the experiments carried out is provided and finally, the conclusions and future work are detailed in Sect. 5.

2 Definition of the Task and the Annotation Procedure

The aim of this work is to detect bad quality results in a subjective crowdsourcing annotation task. To this end Spanish SOFOCO corpus, which draws on the website Menéame[3], was considered. Menéame is a social news aggregation site, modeled after Digg. Registered users can submit content in the form of stories: links to news published on other web sites accompanied by a short comment. Besides, they can post their own comments on each story, resulting in a comment thread. The term "comment" is the word used in Menéame for the different interactions of the users. However, from now on we will refer to those "comments" with the term "post". A procedure for finding, retrieving and processing stories and posts related to some specific controversial categories *Terrorism*, *Independence of Catalonia*, *Abortion*, *Gay Marriage* and *Creationism* was carried out.

Once the selected set of posts was retrieved from the Menéme website we wanted to had the posts labeled as *sarcastic* or *not-sarcastic* to develop an automatic sarcasm detection system. We used a custom crowdsourcing platform, CrowdScience[4] [16] to carry out the annotation task. The motivations of choosing this option are (1) Amazon's Mechanical Turk is not available in European

[3] www.meneame.net.

[4] Available for the scientific community under specific constraints. http://cz.efaber.net.

Countries only in USA. (2) In other available platforms like Crowdflower only around 19% of contributors who speaks Spanish are from Spain. Let us note that in the proposed task it is very important to have as much annotators from Spain as possible, because the differences with regard to american Spanish are very noticeable and (3) We wanted to have a controlled set of annotators at first to drive a pilot annotation task.

The platform shows the registered user a post and asks him to *Indicate whether a sarcastic tone is present in the post* and the annotator has to respond *yes* or *no*. In order to take some control of the annotators involved in the task, they have to make a registration at the website to start annotating. When the worker wanted to receive any incentive, identifying information (complete name and identity card number) was needed. Additionally, they had to indicate whether they had some skills related to the specific task, like linguistic knowledge or fluency in Spanish language, etc. Depending on the skills they indicated, the annotators were able to access different active tasks. However, as we could see looking at randoms examples, this was not enough to avoid bad quality annotations. Thus, a methodology was designed to detect annotations that should be removed and acquired again.

3 Measuring the Annotations Quality

In this section an algorithm that takes into account different measures employed to evaluate the quality of annotations is proposed.

3.1 Measures Related to the Time Employed in the Annotation

The time employed in carrying out annotations might provide information about the quality of them. It may initially seem that shorter average annotation times are indicators of bad quality annotators. However, after a bit deeper analysis, it can be concluded that average annotation times are very dependent on the skills of the annotators. Thus, there are very fast annotators, may be people used to the annotation procedure, or people that are able to read and understand very fast, that carry out high quality annotations. But there also slower annotators, that need more time, and also provide annotations with a high agreement percentage. Therefore, the time related measures were only used to detect very extreme behaviours associated to programmed bots or similar fraudulent annotators. In this work two measures were collected for each user:

Maximum period of time working continuously (without a stop interval above a threshold). It is assumed that a real annotator cannot work, for instance, more than 6 h without pauses longer than $10'$. Detecting such kind of behaviors would indicate an automatic way annotating, i.e. a bot.

Average annotation time and the corresponding standard deviation. The items to be annotated in the proposed tasks are very different between each others (in terms of length, annotation difficulty, etc.) and they usually require very different annotation times even for the same annotator. Thus, very low standard

deviations would indicate that the annotator is using the same annotation time for all the items, which is considered a suspicious behavior.

3.2 Measures Related to the Interannotator Agreement

According to [1,18], researchers who wish to use hand-coded data, that is, data in which items are labeled with categories, need to show that such data are reliable. Data are reliable if coders agree on the categories assigned to units to an extent determined by the purposes of the study. Different coefficients which measure agreement have been discussed in the literature, such us percentage agreement, χ^2, κ-like measures, etc. The simplest measure of agreement between two coders is the percentage of agreement or observed agreement in Eq. (1).

$$A_o = \frac{1}{n_I} \sum_{i \in I} agr_i \tag{1}$$

where n_I is the number of labeled items, or posts in this case, in set I and agr_i is defined as follows:

$$agr_i = \begin{cases} 1 \text{ if the two coders assign i to the same category} \\ 0 \text{ if the two coders assign i to different categories} \end{cases} \tag{2}$$

Although observed agreement is considered in the computation of all the measures of agreement presented in the literature [1], since some agreement is due to chance, it does not yield values that can be compared across studies. Thus, the observed agreement has to be adjusted for chance agreement like in S [2], π [23] or κ [5] coefficients as shown in Eq. (3), where A_e represents how much agreement is expected by chance. Expected agreement is the probability of two coders c_1 and c_2 agreeing on any class and is obtained according to Eq. (4), where T is the set of all items that the two coders have classified.

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{3}$$

$$A_e = \sum_{t \in T} P(t|c_1) \cdot P(t|c_2) \tag{4}$$

The difference between S, π, and κ lies in the assumptions leading to the calculation of $P(t|c_i)$. S is based on the assumption that if coders were operating by chance alone, we would get a uniform distribution: that is, for any two coders c_m and c_n and any two categories k_j, k_l, $P(t_j|c_m) = P(t_l|c_n)$. π assumes that if coders were operating by chance alone, we would get the same distribution for each coder: for any two coders c_m, c_n and any category k, $P(t|c_m) = P(t|c_n)$. Finally, κ considers that if coders were operating by chance alone, we would get a separate distribution for each coder.

Generalizations of π and κ coefficients, *Fleiss' Multi-*π [12] and *Multi-*κ [6], were proposed for the cases in which more than two coders were involved in the

classification of an item. However, a serious limitation of the aforementioned coefficients is that all disagreements are treated equally, although it is not relevant for this specific task it would be a disadvantage in similar subjective tasks like annotation of nastiness, disgust, etc., where annotators should provide a number in a scale. Thus, in this work we will focus on the more general and versatile Krippendorff's α [18] that is appropriate for use with multiple coders, different magnitudes of disagreement, and missing values, and is based on assumptions similar to those of π; and weighted kappa κ_w [4] (defined only for two coders). It provides a measure of the overall agreement found in the task that can be compared to other studies. It is given by Eq. (5), where D_o^α and D_e^α are the observed and expected disagreements respectively.

For the specific task proposed in this work where nominal data (Sarc Yes (1)/Sarc No(0)) and more than two annotators were involved the specific way of computing this coefficient is given by Eq. (5) according to [19].

$$\alpha = 1 - \frac{D_o^\alpha}{D_e^\alpha} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \tag{5}$$

where $o_{ck} = \sum_u \frac{\text{Number of c-k pairs in reliability matrix}}{m_u - 1}$, m_u is the number of values of the item u in the reliability matrix (in our case m_u is always 5), $n_c = \sum_k o_{ck}$ and $n = \sum_c n_c$. For further details see [1].

The terms in (5) are computed considering a reliability $p \cdot q$ matrix, where p is the number of annotators and q the number of items that would be labeled. Since in this work we wanted to detect low quality annotations, which is to say low quality annotators, we propose to measure the variations of Krippendorff's α when an annotator k was removed from the set as follows: $\Delta\alpha_k = \frac{\alpha - \alpha_k}{n_k}$ where n_k is the number of annotations carried out by annotator k. Given that Krippendorff's α is an agreement measure, a negative $\Delta\alpha_k$ means that when removing the k annotator a better agreement is obtained. The lower the $\Delta\alpha_k$ value the higher the disagreement introduced by the k annotator.

$\Delta\alpha_k$ is measured to evaluate an specific annotator's influence. However, other items and agreement percentages included in the overall task influence the obtained results. Thus, in this work we define a new coefficient, β_k, defined as an observed agreement of an annotator and the rest of annotators that labeled the same items. In this coefficient only the items classified by an specific k annotator are involved.

Let K be a set of annotators and J a set of items (posts in this case) labeled by the different annotators $k \in K$. Thus, K_j will represent the set of annotators that labeled a specific item $j \in J$ and J_k will represent the set of items labeled by an specific annotator $k \in K$. β_k is defined as shown in Eq. (6):

$$\beta_k = \frac{1}{|J_k|} \sum_{j \in J_k} \sum_{m \in K_j - k} \frac{c_{mk}}{|K_j| - 1} \tag{6}$$

where $|J_k|$ is the number of posts labeled by the annotator k, $|K_j|$ is the number of different annotations given to the j item and c_{mk} is defined as follows:

$$c_{mk} = \begin{cases} 1 \text{ if } m \text{ and } k \text{ coders assign j to the same category} \\ 0 \text{ if } m \text{ and } k \text{ coders assign j to different categories} \end{cases} \quad (7)$$

3.3 Measures Related to the Given Responses

It is also very valuable to detect when an annotator is providing random answers to the proposed questions, or always the same response. When binary questions are proposed (with a yes/no response for instance) it is even easier to say always "yes", or always "no". Thus, measures that detect such kind of response patterns would also be useful to detect fraudulent annotators. Moreover, the ratio of a kind of response, depending on the nature of the task, might indicate a low quality in annotations.

In the proposed task, where annotators are asked about the presence of a sarcastic tone, the *percentage of sarcastic responses* was measured for each annotator. In this specific case, since the sarcasm is less frequent that the absence of it, a high percentage of sarcastic responses would indicate a fraudulent annotator, an annotator that do not make the annotations carefully, an annotator that did not properly understand the task, etc.

3.4 Algorithm for the Detecion of Low Quality Data

Once the different measures presented in previous sections were collected for each annotator we proceed as shown in Algorithm 1 to select the annotators and the corresponding annotations that should be removed. The annotators in the resulting set have to be considered fraudulent, so their annotations are not reliable enough and should be removed and picked up again.

4 Experiments and Results

Using the platform described in Sect. 2, 13.717 annotations were carried out by 207 annotators fluent in European Spanish. Each post was labeled by 5 different annotators. The inter-annotator agreement was measured in terms of Kripendorff's α to evaluate the reliability of the task and a value of 0.24 was achieved. Although this is a low value, it is similar to the one obtained with IAC (0.22) [24], its counterpart corpus in English, in which SOFOCO is inspired. Furthermore, similar values were also achieved when considering different tasks for emotionally annotating synthesised speech [3] where subjectivity is also present.

However, a manual analysis of the annotations reveals that there were bad quality annotations that would lead to a poor training set for the sarcasm detection system. Thus, the quality measures and the methodology described in Sect. 3 were employed to detect annotators that should be removed.

Considering that annotators that wanted to obtain incentives for their work had to label at least 50 posts, M = 50 was selected to build the first set of

Algorithm 1. Detection of low quality data

Input: K_M: A set of annotators that carried out at least a minimum number of annotations (M). The annotators that made very few annotations were not considered to be influential enough. N : selected threshold for the number of worst elements in a set. T_{min} : Minimum required stop interval for continuous work. TP_{max} : Maximum time working continuously without stops longer than T_{min}. SD_{min} : Minimum standard deviation time for the different annotations of a worker.

Output: K_R: A subset of annotators from K_M whose annotations are not reliable.

$K_1, K_2, K_3 = \emptyset$
$K_{M_{Sorted\alpha}} \leftarrow$ Sort K_M in terms of descendent $\Delta\alpha_k$ values.
while $|K_1| <= N$ **do** ▷ K_1: set of the N annotators with the highest $\Delta\alpha_k$ values.
    $K_1 \leftarrow k \in K_{M_{Sorted\alpha}}$
$K_{M_{Sorted\beta}} \leftarrow$ Sort K_M in terms of ascendent β_k values.
while $|K_2| <= N$ **do** ▷ K_2: set of the N annotators with the lowest β_k values.
    $K_2 \leftarrow k \in K_{M_{Sorted\beta}}$
$K_{M_{Sorted\%sarc}} \leftarrow$ Sort K_M in terms of descendent percentage of sarcastic responses.
while $|K_3| <= N$ **do** ▷ K_3: set of the N annotators with the highest % sarc. resp.
    $K_3 \leftarrow k \in K_{M_{Sorted\%sarc}}$
$K_R = \{K_1 \cap K_2 \cap K_3\}$
for all $k \in K_M$ **do**
    Compute $TP_{max}(t_k)$ ▷ Compute maximum period of time working continuously without a minimum stop T_{min}.
    if $TP_{max}(t_k) < TP_{max}$ **then**
        $K_R \leftarrow k$
for all $k \in K_M$ **do**
    Compute $SD(t_k)$ ▷ Compute the time standard deviation of annotator k.
    if $SD(t_k) < SD_{min}$ **then**
        $K_R \leftarrow k$

annotators. The achieved set was comprised by 116 annotators. Then, N = 20 was selected to build the different sets with the highest percentage of sarcastic responses and worse $\Delta\alpha_k$ and β_k values. Besides, annotators with a maximum period of 6 h of continuous work without stops longer than 10′ were included in the suspicious set (TP_{max} = 6 h, T_{min} = 10′) along with those with standard deviation from average time lower than 10″ (SD_{min} = 10″). Finally, there were 5 annotators in the resulting suspicious user set (see Stage 1 column in Table 1).

Focusing on agreement coefficients, all the annotators shown in Stage 1 have significantly lower β_k values than the highest β_k value achieved by the best annotator that was 0.802 and also than the average β_k achieved with all 116 annotators that is 0.566. In fact, the β_k-s shown in the Table are the 5 worst values among the 116 annotators. Moreover, all the annotators have also negative values of $\Delta\alpha_k$, as expected, meaning that Kripendorff's α value increased when those annotators were removed. Besides, $\Delta\alpha_k$ values of 296, 288, 282 and 40 annotators are the worst values among the 116 annotators that were considered and the value of 279 annotator is among the 10 worst ones. This means that although the information gathered in both β_k and $\Delta\alpha_k$ is related to inter-annotator agreement, it is not exactly the same and can be complementary. Additionally, the percentage of sarcastic responses for all the annotators in the Table is very high, note 91% achieved by 282 that is the annotator with the highest percentage of sarcastic responses in the set. The percentage of the sarcastic responses given in Table 1 are among the 11 highest values in the set and all of them are significantly higher than the average value (42.77%). Regarding maximum working

Table 1. Values of the different parameters (number of posts, β_k, $\Delta\alpha_k$, ratio of sarcastic responses, maximum working time without stops loner than $10'$, average time and standard deviation) for the annotators of the K_R set given by the proposed algorithm in the first annotation stage. Additionally, values for the parameters obtained in a second annotation stage are also given.

	Stage 1					Stage 2	
	Ann. ID					Ann. ID	
	282	40	296	288	279	308	325
No. posts	500	50	150	303	497	67	65
β_k	**0.311**	0.373	0.377	0.430	0.439	0.522	0.523
$\Delta\alpha_k$	−0.211	−0.127	**−0.215**	−0.122	−0.076	−0.077	−0.114
Resp. sarc. ratio (%)	**91.0**	78.0	65.3	74.6	82.5	59.7	67.7
Max. work. time (sec.)	3964	2013	1752	2825	**4669**	1254	777
t_avg. (sec.)	33.42	27.20	**8.42**	37.07	17.03	35.14	12.55
t_dev. (sec.)	51.20	39.45	**2.90**	42.88	20.88	61.01	14.65

times, the highest one is 4669 meaning that the annotator has been continuously working around 1.3 h without stops higher than $10'$. According to our judgement it is not a suspicious working time and it seems that there are not automatic bots carrying out the annotations in our set. In the same way, the average times and the corresponding standard deviation do not indicate suspicious behaviours, except when regarding 296 annotator, which standard deviation is below $10''$.

Finally, the annotations carried out by the annotators given in Stage 1 were removed and a new annotation stage was carried out to get the labels again. When incorporating the new annotations, a higher Kripendorff's α is obtained (it changed from 0.22407 to 0.23193). Applying the same criterion described above a new set of suspicious annotators was achieved with the quality measures gathered in Stage 2 column of Table 1. All the quality measures in this set are better than those in the previous column. The values of β_k and $\Delta\alpha_k$ are much higher in this case and the percentage of sarcastic responses much lower. The same happens with the values associated to annotation times and there are not standard deviations under $10''$. Thus, we keep the annotations in the new set as the correct ones to build the training sets for the sarcasm detection system.

5 Conclusions

This work is devoted to the design of an algorithm capable of detect low quality annotations in a very subjective crowdsourcing task. It considers different measures to take into account different behaviours of workers that provide poor performing annotations: measures related to time, inter-annotator agreement and repeated patterns in responses. Regarding inter-annotator agreement measures a deep study of the most appropriate existing coefficients was carried out and the increment of one of them was employed. Besides, a new coefficient was

proposed to obtain complementary information. An extended comparison of the two coefficients would be carried out in future work. The experimental results show that the algorithm performs well in the proposed task and we propose to use it in other subjective tasks, like the annotation of nasty or disgusting posts.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
2. Bennet, E.M., Alpert, R., Goldstein, A.C.: Communications through limited response questioning. Public Opin. Q. **18**, 303–308 (1954)
3. Buchholz, S., Latorre, J., Yanagisawa, K.: Crowdsourced Assessment of Speech Synthesis. Wiley, Chichester (2013)
4. Cohen, J.: Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213–220 (1968)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
6. Davies, M., Fleiss, J.L.: Measuring agreement for multinomial data. Biometrics **38**(4), 1047–1051 (1982)
7. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Appl. Stat. **28**(1), 20–28 (1979)
8. Dress, M.L., Kreuz, R.J., Link, K.E., Caucci, G.M.: Regional variation in the use of sarcasm. J. Lang. Soc. Psychol. **27**(1), 71–85 (2008)
9. Eickhoff, C., de Vries, A.P.: How crowdsourcable is your task? In: Workshop on Crowdsourcing for Search and Data Mining (CSDM), Hong Kong, China (2011)
10. Eickhoff, C., de Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Inf. Retrieval **16**(2), 121–137 (2013)
11. Filatova, E.: Irony and sarcasm: corpus generation and analysis using crowdsourcing. In: Proceedings of LREC 2012, Istanbul, Turkey, pp. 392–398, 23–25 May 2012
12. Fleiss, J., et al.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
13. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In: Proceedings of the ACM CHI 2015, Seoul, Republic of Korea, pp. 1631–1640 (2015)
14. Gennaro, R., Gentry, C., Parno, B.: Non-interactive verifiable computing: outsourcing computation to untrusted workers. In: Rabin, T. (ed.) CRYPTO 2010. LNCS, vol. 6223, pp. 465–482. Springer, Heidelberg (2010). doi:10.1007/978-3-642-14623-7_25
15. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on Amazon mechanical turk. In: Proceedings of the ACM SIGKDD, pp. 64–67. New York, USA (2010)
16. Justo, R., Alcaide, J.M., Torres, M.I.: Crowdscience: crowdsourcing for research and development. In: Proceedings of IberSpeech 2016, Portugal, pp. 403–410 (2016)
17. Kou, Z., Stanton, D., Peng, F., Beaufays, F., Strohman, T.: Fix it where it fails: pronunciation learning by mining error corrections from speech logs. In: Proceedings of ICASSP 2015, South Brisbane, Australia, pp. 4619–4623, 19–24 April 2015
18. Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage, Thousand Oaks (2004)

19. Krippendorff, K.: Computing Krippendorff's Alpha Reliability. Technical report, University of Pennsylvania, Annenberg School for Communication, June 2007
20. Nunberg, G.: The Way we Talk Now: Commentaries on Language and Culture from NPR's "Fresh Air". Houghton Mifflin, Boston (2001)
21. Rodrigues, F., Pereira, F.C., Ribeiro, B.: Learning from multiple annotators: distinguishing good from random labelers. Pattern Recogn. Lett. **34**(12), 1428–1436 (2013)
22. Rothwell, S., Elshenawy, A., Carter, S., Iraga, D., Romani, F., Kennewick, M., Kennewick, B.: Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In: Proceedings of Interspeech 2015, Dresden, Germany, pp. 2784–2788. ISCA, 6–10 September 2015
23. Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Q. **19**(3), 321–325 (1955)
24. Swanson, R., Lukin, S.M., Eisenberg, L., Corcoran, T., Walker, M.A.: Getting reliable annotations for sarcasm in online dialogues. In: Proceedings of LREC 2014, Reykjavik, Iceland, pp. 4250–4257, 26–31 May 2014