Opinion Mining with Non-Expert Annotations: A Comparison of HIT Designs on Mechanical Turk.

Abstract

One of the major bottlenecks in the development of data-driven AI Systems is the lack of reliable human annotations. The recent advent of several crowdsourcing platforms such as Amazon's Mechanical Turk, allowing requesters the access to affordable and rapid results of a global workforce, greatly facilitates the creation of massive training sets. Most of the available studies on the effectiveness of crowdsourcing report on English data. We use Mechanical Turk annotations to train an Opinion Mining System to classify Spanish consumer comments. We design three different Human Intelligence Task (HIT) strategies and show high inter-annotator agreement between non-experts and expert annotators. We evaluate the advantages/drawbacks of each HIT design and report `<results of the classifier>`.

1 Introduction

`<Intro.>` `<Some possible citations for remainder paper.>` In (Snow et al., 2008), it is shown that . . .
In (Sheng et al., 2008), it is shown that . . .
In (Kittur et al., 2008), it is shown that . . .
In (Su et al., 2007), it is shown that . . .

2 Task Outline and Goals

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk (AMT) annotations to train an Opinion Mining System on Spanish consumer data. More specifically, we address the following research questions:

(i) Annotation quality: how do the different AMT annotations compare to expert annotations? We compare the inter-annotator agreement in expert annotations with the inter-annotator agreement in the different AMT annotations.

(ii) Annotation applicability: how does the performance of an Opinion Mining classifier vary after training on different (sub)sets of AMT and expert annotations? Given a simple classification technique, we evaluate the system performance by using AMT annotations, expert annotations and the combination of both as training data. The idea is not to evaluate the classification technique *per se*, but to measure the influence of the training material.

(iii) Return on Investment (ROI): how does the use of AMT annotations compare economically against the use of expert annotations? AMT offers the possibility of obtaining inexpensive annotations the quality of which tends to be worse than expert annotations `<include ref>`. We show that, for the task at hand, the ROI is positive.

(iv) Language barriers: `x%` of all AMT tasks are designed for English speakers `<include ref>`. How easy is it to get reliable AMT results for Spanish?

3 HIT Design

We selected a dataset of 1000 sentences containing user opinions on cars from the automotive section of `www.ciao.es` (Spanish). This website was chosen since it contains a large and varied amount of opinions in Spanish and because opinions include simultaneously global numeric and specific ratings

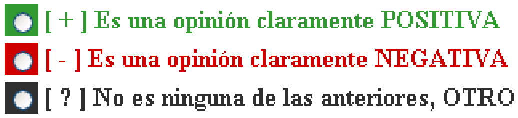

Figure 1: Figure 1.

Figure 2: Figure 2.

over particular attributes of the subject matter. Section 5.1 contains more detailed information about the selection of the dataset. An example of a sentence from the data set can be found in (1):

(1) 'No te lo pienses más, cómpratelo!'
(= 'Don't think twice, buy it!')

The sentences in the dataset were presented to the AMT workers in three different HIT designs. HIT1 is a simple categorization scheme in which workers are asked to classify each sentence as being either 'positive', 'negative' or 'neutral', as is shown in Figure 1. HIT2 is a graded categorization template in which workers had to assign a score between -5 (negative) and +5 (positive) to each example sentence, as is shown in Figure 2. Finally, HIT3 is a continuous triangular categorization template that allows workers to place <improve> examples both on a horizontal positive-negative axis and on a vertical subjective-objective axis. This latter axis expresses the degree to which the sentence contains opinionated content and was earlier used by (Esuli and Sebastiani, 2006). For example, the sentence 'I think this is a wonderful car' clearly marks an opinion and should be positioned towards subjective scale, while the sentence 'The car has six cilinders' should be placed high on the objective scale. An example of HIT is is provided in Figure 2. In order not to burden workers with ..., we kept things simple and did this and that.

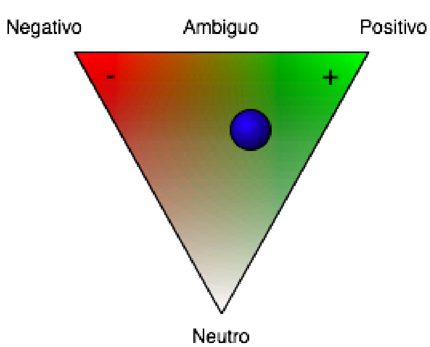

Figure 3: Figure 3.

ID	Country	HIT1	HIT2	HIT3	Acc
A16MC82ITK70QZ	US	77/8.2			x%
A19835WFUL4B52	US	43/5.3			x%
A198YDDSSOBP8A	Mexico	794/11.0			x%
A1COK1GRYUJA1M	US	3/15.7			x%
A1F70TQGR00PTQ	US	980/			x%

Table 1: Statistics on AMT workers: (fictional) ID, Country, per HIT type nr. hits/average completion time, Accuracy.

4 Annotation Task Results and Analysis

After designing the HITs, we uploaded 30 random samples for testing purposes. These HITs were completed in a matter of seconds, mostly by workers in India. After a brief inspection of the results, it was obvious that most answers corresponded to random clicks. Therefore, we decided to include a small competence test to ensure that future workers would possess the necessary linguistic skills to perform the task. The test consists of six simple categorisation questions of the type of HIT1 that a skilled worker would be able to perform in under a minute.

4.1 Annotation Statistics

These are the statistics:
<among non-experts and experts vs. non-experts. (Snow et al., 2008) is good for this.>

4.2 Annotation Quality

<take into account annotator bias>
<This is a very useful reference: (Dawid and Skene, 1979)>
<This is another useful reference: (Mason and Watts, 2009)>

5 Incidence of annotations on suppervised polarity classification

This section intends to evaluate the incidence of AMT-generated annotations on the polarity classification task. According to this, a comparative evaluation between two polarity classification systems is conducted. More specifically, baseline or reference classifiers trained with noisy available metadata are compared with contrastive classifiers trained with AMT generated annotations. Although more sophisticated classification schemas can be conceived for this task, a simple SVM-based binary supervised classification approach is considered here.

5.1 Description of datasets

For conducting the experimental evaluation, three different datasets were considered:

1. Baseline: constitutes the dataset used for training the baseline or reference classifiers. Automatic annotation for this dataset was obtained by using the following naive approach: those sentences extracted from comments with ratings equal to 5 were assigned to class ”positive”, those extracted from comments with ratings equal to 3 were assigned to ”neutral”, and those extracted from comments with ratings equal to 1 were assigned to ”negative”. This dataset contains a total of 5570 sentences, with a vocabulary coverage of 11797 words.

2. Annotated: constitutes the dataset that was manually annotated by AMT workers. This dataset is used for training the contrastive classifiers which are to be compared with baseline systems. The three independent annotations generated by AMT workers for each sentence within this dataset were consolidated into one unique annotation by using the following criterion: if the three provided annotations happened to be different[1], the sentence was assigned to class ”neutral”; otherwise, the sentence was assigned to the class with at least two annotation agreements. This dataset contains a total of 1000 sentences, with a vocabulary coverage of 3022 words.

[1]Actually, this kind of total disagreement among annotators occurred only in 13 sentences out of 1000.

	Baseline	Annotated	Evaluation
Positive	1882	341	200
Negative	1876	323	137
Neutral	1812	336	161
Totals	5570	1000	500

Table 2: Sentence-per-class distributions for baseline, annotated and evaluation datasets.

3. Evaluation: constitutes the gold standard used for evaluating the performance of classifiers. This dataset was manually annotated by three experts in an independent manner. The gold standard annotation was consolidated by using the same criterion used in the case of the previous dataset[2]. This dataset contains a total of 500 sentences, with a vocabulary coverage of 2004 words.

These three datasets were constructed by randomly extracting sample sentences from an original corpus of over 25000 comments containing more than 1000000 sentences in total. The sampling was conducted with the following constraints in mind: the three resulting datasets should not overlap, only sentences containing more than 3 tokens could be extracted, each resulting dataset must be balanced, as much as possible, in terms of the amount of sentences per class. Table 2 presents the distribution of sentences per class for each of the three considered datasets.

5.2 Experimental settings

As mentioned above, a simple SVM-based supervised classification approach was considered for the polarity detection task under consideration. According to this, two different groups of classifiers were considered: a baseline or reference group, and a contrastive group. Classifiers within these two groups were trained with data samples extracted from the baseline and annotated datasets, respectively. Within each group of classifiers, three different binary classification subtasks were considered: positive/not_positive, negative/not_negative and neutral/not-neutral. All trained binary classifiers

[2]In this case, annotator inter-agreement was above 80%, and total disagreement among annotators occurred only in 1 sentence out of 500

were evaluated by computing precision and recall for each considered class, as well as overall classification accuracy, over the evaluation dataset.

A feature space model representation of the data was constructed by considering the standard bag-of-words approach. In this way, a sparse vector was obtained for each sentence in the datasets. Stop-word removal was not conducted before computing vector models, and standard normalization and TF-IDF weighting schemes were used.

Multiple-fold cross-validation was used in all conducted experiments to tackle with statistical variability of the data. In this sense, twenty independent realizations were actually conducted for each experiment presented and, instead of individual output results, mean values and standard deviations of evaluation metrics are reported.

Each binary classifier realization was trained with a random subsample set of 600 sentences extracted from the training dataset corresponding to the classifier group, i.e. baseline dataset for reference systems, and annotated dataset for contrastive systems. Training subsample sets were always balanced with respect to the original three categories: ”positive”, ”negative” and ”neutral”.

5.3 Results and discussion

Table 3 presents the resulting average values of precision and recall for each considered class in classifiers trained with either the baseline or the annotated dataset. As observed in the table, with the exception of recall for class ”negative” and precision for class ”not_negative”, both metrics are substantially improved when the annotated dataset is used for training the classifiers. The most impressive improvements are observed for ”neutral” precision and recall, and for ”positive” precision.

Table 4 presents the resulting average values of accuracy for each considered subtask in classifiers trained with either the baseline or the annotated dataset. As observed in the table, all subtasks benefit from using the annotated dataset for training the classifiers; however, it is important to mention that while similar absolute gains are observed for the ”positive/not_positive” and ”neutral/not_neutral” subtasks, this is not the case for the subtask ”negative/not_negative”, which actually gains much less than the other two subtasks.

	baseline	baseline	annotated	annota
class	precision	recall	precision	recall
positive	54.23 (3.52)	44.65 (3.68)	68.33 (3.09)	53.65
not_positive	66.88 (1.79)	74.75 (2.85)	72.88 (1.21)	83.28
negative	40.49 (3.22)	39.93 (4.18)	44.96 (2.08)	38.26
not_negative	77.16 (1.27)	77.53 (2.33)	77.69 (1.07)	82.02
neutral	34.37 (3.57)	31.43 (7.93)	49.69 (3.39)	50.43
not_neutral	68.75 (1.60)	71.72 (5.84)	76.26 (1.89)	75.65

Table 3: Average precision and average recall (with standard deviations provided in parenthesis) for each considered class in classifiers trained with either the baseline or the annotated dataset.

classifier	baseline	annotated
positive/not_positive	62.69 (2.35)	71.40 (1.64)
negative/not_negative	67.13 (1.90)	69.92 (1.19)
neutral/not_neutral	58.72 (2.55)	67.52 (2.10)

Table 4: Average accuracy (with standard deviations provided in parenthesis) for each classification subtasks trained with either the baseline or the annotated dataset.

After considering all evaluation metrics, it is evident the important benefit provided by human-annotated data availability for classes ”neutral” and ”positive”. However, in the case of class ”negative”, although some gain is also observed, the benefit of human-annotated data does not seem to be as much as for the other two classes. This, along with the fact that the ”negative/not_negative” subtask is actually the best performing one (in terms of accuracy) when baseline training data is used, might suggest that low rating comments contains a better representation of sentences belonging to class ”negative” than medium and high rating comments do with respect to classes ”neutral” and ”positive”.

In any case, this experimental work just verifies the feasibility of constructing training datasets for opinionated content analysis, as well as it provides an approximated idea of costs involved in the generation of this type of resources, by using AMT.

6 Conclusions

Future work: HIT Design: what is the optimal design of the annotation task for our purpose and what is the effect of a suboptimal design on the system

```
scores? We present AMT workers
with three different HIT designs
and evaluate the impact of each
of them on the overall system
performance.
```

Acknowledgments

We thank Amazon for generously sponsoring this Shared Task.

References

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6.

A. Kittur, E. H Chi, and B. Suh. 2008. Crowdsourcing user studies with mechanical turk.

W. Mason and D. J Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85.

V. S Sheng, F. Provost, and P. G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y Ng. 2008. Cheap and fastbut is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Q. Su, D. Pavlov, J. H Chow, and W. C Baker. 2007. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web*, pages 231–240.