Opinion Mining of Spanish Customer Comments with Non-Expert Annotations 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 data. 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 report high inter-annotator agreement between non-experts and expert annotators. We evaluate the advantages/drawbacks of each HIT design and show that, in our case, the use of non-expert annotations is a viable and cost-effective alternative to expert annotations.

1 Introduction

The recent advent of crowdsourcing platforms such as Amazon's Mechanical Turk[1], Crowdflower[2] and others has attracted a lot of attention both from companies as from academia. Crowdsourcing enables requesters to tap from a global pool of non-experts to obtain rapid and affordable answers to simple Human Intelligence Tasks (HITs), which can be subsequently used to train data-driven AI systems.

A number of recent papers on this subject point out that non-expert annotations, if produced in a sufficient quantity, can rival and even surpass the quality of expert annotations, often at a much lesser cost (Snow et al., 2008), (Su et al., 2007). However, this possible increase in quality depends on the task at hand and on an adequate HIT design (Kittur et al., 2008).

In this paper, we evaluate the usefulness of AMT annotations to train an Opinion Mining system to detect opinionated contents (Polarity Detection) in Spanish customer comments on car brands. Currently, a large majority of AMT tasks is designed for English speakers. One of the reasons for participating in this shared task was to find out how easy it is to obtain annotated data for Spanish. In addition, we want to find out how useful these data are, by comparing them to expert annotations and using them to train an Opinion Mining system for polarity detection.

This paper is structured as follows. Section 2 contains an explanation of the task outline and our goals. Section 3 contains an analysis of three different HIT designs that we used in this task. In Section 4, we provide a detailed analysis of the retrieved HITs and focus on geographical information of the workers, the correlation between the different HIT designs, the quality of the retrieved answers and the total cost of the experiment. In Section 5, we evaluate the incidence of AMT-generated annotations on a polarity classification task using two different experimental settings. Finally, we conclude in Section 6.

2 Task Outline and Goals

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk

[1] https://www.mturk.com

[2] http://crowdflower.com/

(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 because it contains a large and varied pool of Spanish customer comments suitable to train an Opinion Mining System and because opinions include simultaneously global numeric and specific ratings over particular attributes of the subject matter. Section 5.1.1 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. Each HIT design contains a single sentence to be evaluated. HIT1 is a simple categorization scheme in

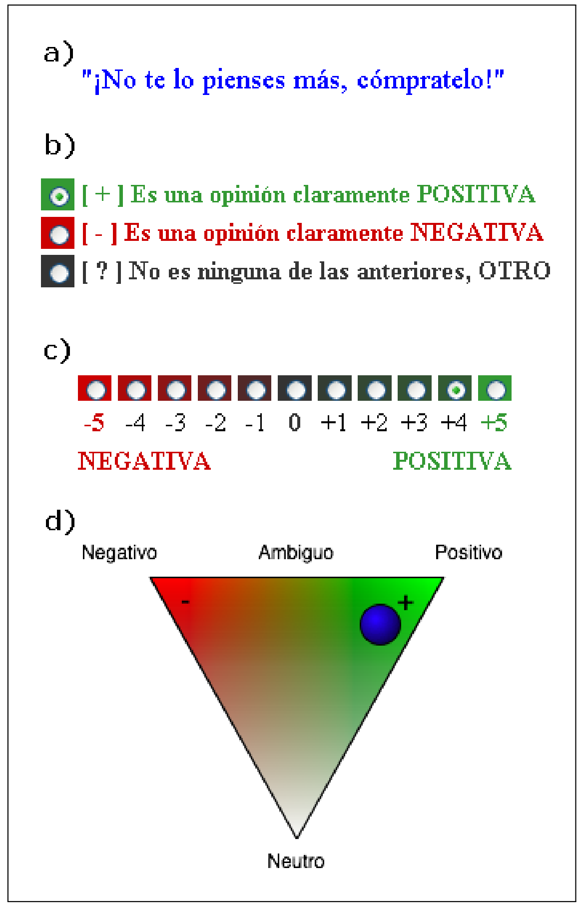

Figure 1: An example sentence (a) and the three HIT designs used in the experiments: (b) HIT1: a simple categorization scheme, (c) HIT2: a graded categorization scheme, and (d) HIT3: a continuous triangular scoring scheme containing both a horizontal positive-negative axis and a vertical subjective-objective axis.

which workers are asked to classify the sentence as being either *positive*, *negative* or *neutral*, as is shown in Figure 1, section b. HIT2 is a graded categorization template in which workers had to assign a score between -5 (negative) and +5 (positive) to the example sentence, as is shown in Figure 1, section c. Finally, HIT3 is a continuous triangular scoring template that allows workers to use both a horizontal positive-negative axis and a vertical subjective-objective axis by placing the example sentence anywhere inside the triangle. The subjective-objective 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 the subjective end, while the sentence *'The car has six cilinders'* should be located towards the objective end. Figure 1, section d contains an example of HIT3. In order not to burden the workers with overly complex instructions, we did not mention this subjective-objective axis but asked them instead to place ambiguous sentences towards the center of the horizontal positive-negative axis and more objective, non-opinionated sentences towards the lower *neutral* tip of the triangle.

For each of the three HIT designs, we specified the requirement of three different unique assignments per HIT, which led to a total amount of $3 \times 3 \times 1000 = 9000$ HITs being uploaded on AMT. Mind that setting the requirement of unique assigments ensures a number of unique workers *per individual HIT*, but does not ensure a consistency of workers over a single batch of 1000 HITs. This is in the line with the philosophy of crowdsourcing, which allows many different people to participate in the same task.

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. In order to discourage the use of automatic translation tools, a time limit of two minutes was imposed and test sentences contain idiomatic constructions notoriously difficult for Machine Translation Systems.

4.1 HIT Statistics

Table 1 contains statistics on the workers who completed our HITs. A total of 19 workers passed the competence test and submitted at least one HIT. Of those, four workers completed HITs belonging to two different designs and six submitted HITs in all three designs. Twelve workers are located in the US (64%), three in Spain (16%), and one in Mexico (5%), Equador (5%), The Netherlands (5%) and an unknown location (5%).

As to a comparison of completion times, it took a worker on average 11 seconds to complete an instance of HIT1, and 9 seconds to complete an instance of HIT2 and HIT3. At first sight, this result might seem surprising, since conceptually there is an increase in complexity when moving from HIT1 to HIT2 and from HIT2 to HIT3. These results might suggest that users find it easier to classify items on a graded or continuous scale such as HIT2 and HIT3, which allows for a certain degree of flexibility, than on a stricter categorical template such as HIT1, where there is no room for error.

4.2 Annotation Distributions

In order to get an overview of distribution of the results of each HIT, a histogram was plotted for each different task. Figure 2a shows a uniform distribution of the three categories used in the simple categorization scheme of HIT1, as could be expected from a balanced dataset.

Figure 2b shows the distribution of the graded categorization template of HIT2. Compared to the distribution in 2a, two observations can be made: (i) the proportion of the zero values is almost identical to the proportion of the neutral category in Figure 2a, and (ii) the proportion of the sum of the positive values [+1,+5] and the proportion of the sum of the negative values [-5,-1] are equally similar to the proportion of the positive and negative categories in 2a. This suggests that in order to map the graded annota-

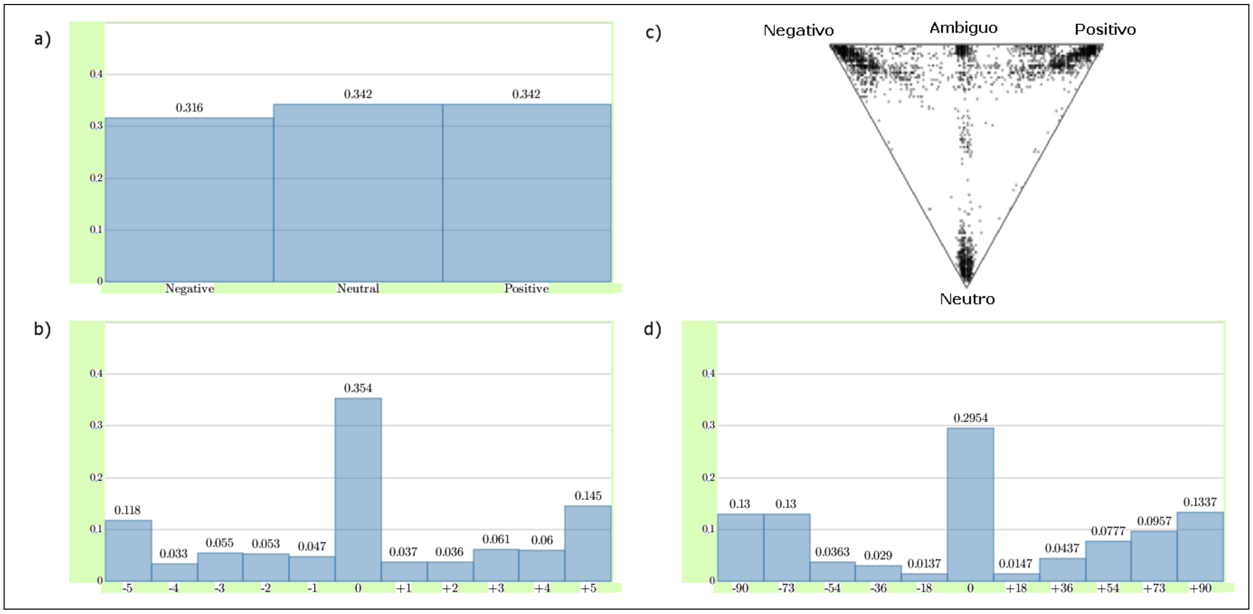

Figure 2: Overview of HIT results: a) distribution of the three categories used in HIT1, b) distribution of results in the scaled format of HIT2, c) heat map of the distribution of results in the HIT3 triangle, d) distribution of projection of triangle data points onto the X-axis (positive/negative).

Overall			HIT1		HIT2		HIT3	
ID	C	%	#	sec.	#	sec.	#	sec.
1	mx	29.9	794	11.0	967	8.6	930	11.6
2	us	27.6	980	8.3	507	7.8	994	7.4
3	nl	11.0	85	8.3	573	10.9	333	11.4
4	us	9.5	853	16.8	-	-	-	-
5	es	9.4	-	-	579	9.1	265	8.0
6	ec	4.1	151	9.4	14	16.7	200	13.0
7	us	3.6	3	15.7	139	8.5	133	11.6
8	us	2.2	77	8.2	106	7.3	11	10.5
9	us	0.6	-	-	-	-	50	11.2
10	us	0.5	43	5.3	1	5	-	-
11	us	0.4	-	-	38	25.2	-	-
12	us	0.4	-	-	10	9.5	27	10.8
13	es	0.4	-	-	-	-	35	15.1
14	es	0.3	-	-	30	13.5	-	-
15	us	0.3	8	24.7	18	21.5	-	-
16	us	0.2	-	-	-	-	22	8.9
17	us	0.2	-	-	17	16.5	-	-
18	?	0.1	6	20	-	-	-	-
19	us	0.1	-	-	1	33	-	-

Table 1: Statistics on AMT workers for all three HIT designs: (fictional) worker ID, country code in ISO 3166-1, % of total number of HITs completed, number of HITs completed per design and average completion time.

tions of HIT2 to the categories of HIT1, an intuitive partitioning of the graded scale into three equal parts should be avoided. Instead, a more adequate alternative would consist of mapping [-5,-1] to *negative*, 0 to *neutral* and [+1,+5] to *positive*. This means that even slightly positive/negative grades correspond to positive/negative categories.

Figure 2c shows a heat map that plots the distribution of the annotations in the triangle of HIT3. It appears that worker annotations show a spontaneous tendency of clustering, despite the continuous nature of the design. This suggests that this HIT design, originally conceived as continuous, was transformed by the workers as a simpler categorization task using five labels: *negative*, *ambiguous* and *positive* at the top, *neutral* at the bottom, and *other* in the center.

Figure 2d shows the distribution of all datapoints in the triangle of Figure 2c, projected onto the X-axis (positive/negative). Although similar to the graded scale in HIT2, the distribution shows a slightly higher polarization.

These results suggest that, out of all three HIT designs, HIT2 is the one that contains the best balance between the amount of information that can be obtained and the simplicity of a one-dimensional annotation.

4.3 Annotation Quality

The annotation quality of AMT workers can be measured by comparing them to expert annotations. This is usually done by calculating inter-annotator agreement (ITA) scores. Note that, since a single HIT can contain more than one assignment and each assignment is typically performed by more than one annotator, we can only calculate ITA scores between batches of assignments, rather than between individual workers. Therefore, we describe the ITA scores in terms of batches. In Table 4.4, we present a comparison of standard kappa[3] calculations (**?**) between batches of assignments in HIT1 and expert annotations.

We found an inter-batch ITA score of 0.598, which indicates a moderate agreement due to fairly consistent annotations between workers. When comparing individual batches with expert annotations, we found similar ITA scores, in the range between 0.628 and 0.649. This increase with respect to the inter-batch score suggests a higher variability among AMT workers than between workers and experts. In order to filter out noise in worker annotations, we applied a simple majority voting procedure in which we selected, for each sentence in HIT1, the most voted category. This results in an additional batch of annotations. This batch, refered in Table 4.4 as *Majority*, produced a considerably higher ITA score of 0.716, which confirms the validity of the majority voting scheme to obtain better annotations.

In addition, we calculated ITA scores between three expert annotators on a separate, 500-sentence dataset, randomly selected from the same corpus as described at the start of Section 3. This collection was later used as test set in the experiments described in Section 5. The inter-expert ITA scores on this separate dataset contains values of 0.725 for κ_1 and 0.729 for κ_2, only marginally higher than the *Majority* ITA scores. Although we are comparing results on different data sets, these results seem to indicate that multiple AMT annotations are able to produce a similar quality to expert annotations. This might suggest that a further increase in the number of HIT assignments would outperform expert ITA scores, as was previously reported in (Snow et al., 2008).

[3] In reality, we found that fixed and free margin Kappa values were almost identical, which reflects the balanced distribution of the dataset.

4.4 Annotation Cost

`<This is another useful reference: (Mason and Watts, 2009)>`

At the specified cost of 0.02$ per HIT, a total sum of 180$ (minus Amazon fees) was spent on the task.

In order to decide to use the Amazon's "mechanical turk" it's important to measure the quality of the results, but, of course, the cost of performing the annotation using turkers instead of in-house experts and the time needed to perform the annotation must also be taken into account. To compute the in-House costs, the expert or engineer is paid at a rate of 70$ an hour, including salary and all other structural costs. The in-house annotation tasks of 1000 sentences would last for three hours, giving a cost of 210$ and a marginal cost of preparing the data. By means of Amazon we spent 75$ to annotate the same sentences three times, to get the same quality of results. But we must add the costs of designing the Hits and the qualifying test and the task of upload the data into Amazon's servers. These tasks highly depend on the experience of the engineers developing it, and can range from a couple of hours to a couple of days (for the first times). As the volume of data to annotate increases, the economical benefit of the AMT is more evident, because the design time becomes a small portion of the budget. For annotating 30,000 sentences the differences rises up to 1350$ that are more than enough to design the hit. But the main important economic impact may not be the cost but the time to perform the task. In-house massive annotation may take a lot of time, and can become a hard task, as a few users need to annotate a lot of sentences. As this task can become tiresome and hard it needs to be elapsed in time. In AMT the a hit may be distributed into many volunteers that 24 hours a day are ready to work in parallel to produce the results in a few hours.

5 Incidence of annotations on supervised polarity classification

This section intends to evaluate the incidence of AMT-generated annotations on a polarity classification task. We present two different evaluations. In

	κ_1	κ_2
Inter-batch	0.598	0.598
Batch_1 vs. Expert	0.628	0.628
Batch_2 vs. Expert	0.649	0.649
Batch_3 vs. Expert	0.626	0.626
Majority vs. Expert	0.716	0.716
Experts[4]	0.725	0.729

Table 2: Interannotation Agreement as a measure of quality of the annotations in HIT1. κ_1 = Fixed Margin Kappa. κ_2 = Free Margin Kappa.

the evaluation of Section 5.1, we compare the results of training a polarity classification system with noisy available metadata and with AMT generated annotations of HIT1. In the evaluation of Section 5.2, we compare the results of training several polarity classifiers on six different training sets, each of them generated from the AMT annotations of HIT1.

5.1 Experiment one: AMT annotations vs. original Ciao annotations

In this section, 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.

5.1.1 Description of datasets

As was mentioned in Section 3, all sentences were extracted from a corpus of user opinions on cars from the automotive section of `www.ciao.es` (Spanish). 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[5] equal to 5 were assigned to category ‘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 in HIT1. This dataset is used for training the contrastive classifiers which are to be compared with baseline system. The three independent annotations generated by AMT workers for each sentence within this dataset were consolidated into one unique annotation by majority voting: if the three provided annotations happened to be different[6], the sentence was assigned to category ‘neutral’; otherwise, the sentence was assigned to the category with at least two annotation agreements. This dataset contains a total of 1000 sentences, with a vocabulary coverage of 3022 words.

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[7]. 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 user comments containing more than 1000000 sentences in total. The sampling was conducted with the following constraints in mind: (i) the three resulting datasets should not overlap, (ii) only sentences containing more than 3 tokens are considered, and (iii) each resulting dataset must be balanced, as much as possible, in terms of the amount of sentences per category. Table 3 presents the distribution of sentences per category for each of the three considered datasets.

[5]The corpus at `www.ciao.es` contains consumer opinions marked with a score between 1 (negative) and 5 (positive). `<Rafael, please correct.>`

[6]This kind of total disagreement among annotators occurred only in 13 sentences out of 1000.

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

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

Table 3: Sentence-per-category distributions for baseline, annotated and evaluation datasets.

5.1.2 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 were evaluated by computing precision and recall for each considered category, 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’.

class	precision	recall
positive	50.10 (3.79) 60.21 (2.07)	62.00 (7.47) 71.00 (2.18)
not_positive	69.64 (2.70) 77.95 (1.32)	58.05 (7.54) 68.54 (2.75)
negative	35.25 (2.63) 39.07 (1.78)	53.46 (10.55) 55.52 (3.26)
not_negative	78.04 (2.19) 79.73 (1.10)	62.62 (6.76) 66.87 (2.31)
neutral	32.51 (3.02) 44.72 (2.00)	48.03 (7.33) 67.12 (2.96)
not_neutral	68.17 (2.65) 79.41 (1.58)	52.81 (3.84) 60.40 (2.96)

Table 4: Mean precision and recall over 20 independent simulations (with standard deviations provided in parenthesis) for each considered class in classifiers trained with either the baseline dataset (upper values) or the annotated dataset (lower values).

5.1.3 Results and discussion

Table 4 presents the resulting mean values of precision and recall for each considered category in classifiers trained with either the baseline or the annotated dataset. As observed in the table, with the exception of recall for category ‘negative’ and precision for category ‘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.

Table 5 presents the resulting mean 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.

After considering all evaluation metrics, the benefit provided by human-annotated data availability for categories ‘neutral’ and ‘positive’ is evident. However, in the case of category ‘negative’, although some gain is also observed, the benefit of human-annotated data does not seem to be as much as for

classifier	baseline	annotated
positive/not_positive	59.63 (3.04)	69.53 (1.70)
negative/not_negative	60.09 (2.90)	63.73 (1.60)
neutral/not_neutral	51.27 (2.49)	62.57 (2.08)

Table 5: Mean accuracy over 20 independent simulations (with standard deviations provided in parenthesis) for each classification subtasks trained with either the baseline or the annotated dataset.

System	Experts	Batch1	Batch2	Batch3	Majority	All
Winnow	44.2	43.6	40.4	47.6	46.2	50.6
SVM	57.6	53.0	55.4	54.0	57.2	52.8
C45	42.2	33.6	42.0	41.2	41.6	45.0
Maxent	59.2	55.8	57.6	54.0	57.6	58.6

Table 6: Accuracy figures of four different classifiers (Maxent, C45, Winnow and SVM) trained on six different datasets: Experts=training set annotated by experts, Batch1=first batch of HIT1, Batch2=second batch of HIT1, Batch3=third batch of HIT1, Majority= batch obtained by majority voting between Batch1, Batch2 and Batch3, All=batch obtained by aggregating Batch1, Batch2 and Batch3.

the two other categories. 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 category 'negative' than medium and high rating comments do with respect to classes 'neutral' and 'positive'.

In any case, this experimental work only 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.

5.2 Experiment two: AMT annotations vs. expert annotations

In this section, we compare the results of training several polarity classifiers on six different training sets, each of them generated from the AMT annotations of HIT1. The different training sets are: (i) the original dataset of 1000 sentences annotated by experts (*Experts*), (ii) the first set of 1000 AMT results (*Batch1*), (iii) the second set of 1000 AMT results (*Batch2*), (iv) the third set of 1000 AMT results (*Batch3*), (v) the batch obtained by majority voting between Batch1, Batch2 and Batch3 (*Majority*), and (vi) a batch of 3000 training instances obtained by aggregating Batch1, Batch2 and Batch3 (*All*). In (**?**),

Table 6 contains results of four different classifiers (Maxent, C45, Winnow and SVM), trained on these six different datasets and evaluated () on the same 500-sentence test set as explained in Section 5.1.1. The Winnow classifier improves

The Maxent classifier ... The C45 decision tree ... . Finally, the SVM ...

6 Conclusions

In this paper we have examined the usefulness of non-expert annotations on Amazon's Mechanical Turk to annotate the polarity of Spanish consumer comments. We discussed the advantages/drawbacks of three different HIT designs, ranging from a simple categorization scheme to a continous scoring template. We report high inter-annotator agreement scores between non-experts and expert annotators and show that training an Opinion Mining System with non-expert AMT annotations outperforms original noisy annotations and obtains competitive results when compared to expert annotations using a variety of classifiers. In conclusion, we found that, in our case, the use of non-expert annotations through crowdsourcing is a viable and cost-effective alternative to the use of expert annotations.

References

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.

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.