Opinion Mining of Spanish Customer Comments with Non-Expert Annotations on Mechanical Turk

Bart Mellebeek, Francesc Benavent, Jens Grivolla,
Joan Codina, Marta R. Costa-jussà and Rafael E. Banchs

Barcelona Media Centre d'Innovació

Creating Speech and Language Data
With Amazon's Mechanical Turk

NAACL, June 2010, LA

Outline

Why participate in the Shared Task?

HIT Design

Annotation Results

Experimental Results

Conclusions

Outline

Why participate in the Shared Task?

HIT Design

Annotation Results

Experimental Results

Conclusions

Barcelona Media

- An R&D center in Barcelona, Spain.
- Associated to Pompeu Fabra University.
- Applied research, commercial projects.
- Currently: focus on Opinion Mining.

Goals

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk (MechTurk) annotations to train an Opinion Mining System on Spanish consumer data.

Research questions:

1. Annotation quality: MechTurk workers vs. experts.
2. Annotation applicability: training of batch of classifiers.
3. Language barriers: is there a large enough worker pool for Spanish?
4. Return on Investment: MechTurk workers vs. experts.

Outline

Why participate in the Shared Task?

HIT Design

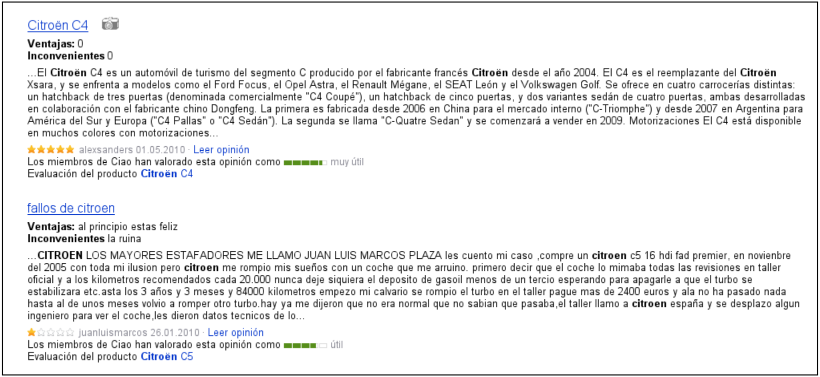
Annotation Results

Experimental Results

Conclusions

Source of our data: `www.ciao.es`

- `ciao` = a European based online-shopping portal.
- Forum containing user reviews and opinions on a wide variety of products.
- All sentences were extracted from user reviews on the automotive section of `www.ciao.es`.



Citroën C4

Ventajas: 0
Inconvenientes 0

...El **Citroën** C4 es un automóvil de turismo del segmento C producido por el fabricante francés **Citroën** desde el año 2004. El C4 es el reemplazante del **Citroën** Xsara, y se enfrenta a modelos como el Ford Focus, el Opel Astra, el Renault Mégane, el SEAT León y el Volkswagen Golf. Se ofrece en cuatro carrocerías distintas: un hatchback de tres puertas (denominada comercialmente "C4 Coupé"), un hatchback de cinco puertas, y dos variantes sedán de cuatro puertas, ambas desarrolladas en colaboración con el fabricante chino Dongfeng. La primera es fabricada desde 2006 en China para el mercado interno ("C-Triomphe") y desde 2007 en Argentina para América del Sur y Europa ("C4 Pallas" o "C4 Sedán"). La segunda se llama "C-Quatre Sedan" y se comenzará a vender en 2009. Motorizaciones El C4 está disponible en muchos colores con motorizaciones...

★★★★★ alexsanders 01.05.2010 · Leer opinión
Los miembros de Ciao han valorado esta opinión como muy útil
Evaluación del producto **Citroën** C4

fallos de citroen

Ventajas: al principio estas feliz
Inconvenientes la ruina

...**CITROEN** LOS MAYORES ESTAFADORES ME LLAMO JUAN LUIS MARCOS PLAZA les cuento mi caso ,compre un **citroen** c5 16 hdi fad premier, en novienbre del 2005 con toda mi ilusion pero **citroen** me rompio mis sueños con un coche que me arruino. primero decir que el coche lo mimaba todas las revisiones en taller oficial y a los kilometros recomendados cada 20.000 nunca deje siquiera el deposito de gasoil menos de un tercio esperando para apagarle a que el turbo se estabilizara etc.asta los 3 años y 3 meses y 84000 kilometros empezo mi calvario se rompio el turbo en el taller pague mas de 2400 euros y ala no ha pasado nada hasta al de unos meses volvio a romper otro turbo.hay ya me dijeron que no era normal que no sabian que pasaba,el taller llamo a **citroen** españa y se desplazo algun ingeniero para ver el coche,les dieron datos tecnicos de lo...

★☆☆☆☆ juanluismarcos 26.01.2010 · Leer opinión
Los miembros de Ciao han valorado esta opinión como útil
Evaluación del producto **Citroën** C5

www.ciao.es: some example sentences

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

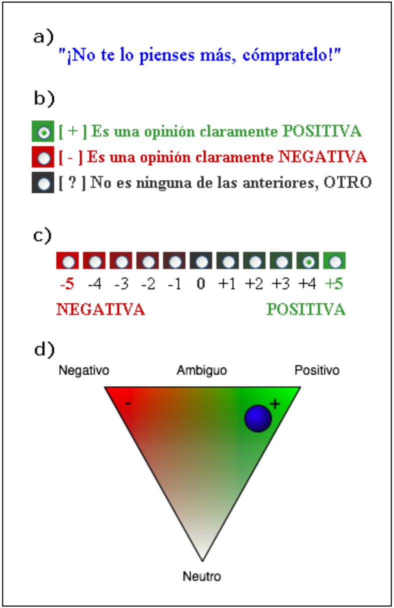
Tiene muchas piezas defectuosas.
(It contains many defective parts.)

La conducción es genial pero el diseño una porquería.
(The car drives great but the design is crap.)

El Volvo es mejor que el Fiat.
(Volvo is better than Fiat.)

Este coche tiene 6 cilindros.
(This car has 6 cylinders.)

HIT Design



a)

"¡No te lo pienses más, cómpratelo!"

b)

● [+] Es una opinión claramente POSITIVA

○ [-] Es una opinión claramente NEGATIVA

○ [?] No es ninguna de las anteriores, OTRO

c)

○	○	○	○	○	○	○	○	○	●	○
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5

NEGATIVA POSITIVA

d)

Competence Test

- Uploaded test batch of 30 HITs.
- Performed in matter of minutes but . . .
- most answers were random clicks.

Outline

Why participate in the Shared Task?

HIT Design

Annotation Results

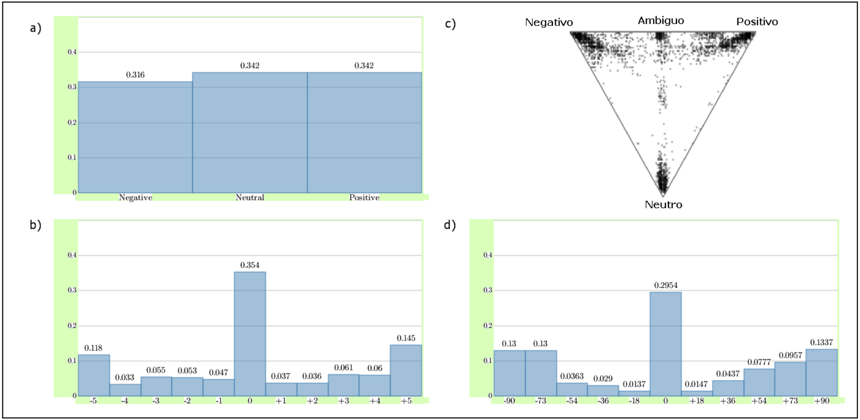
Experimental Results

Conclusions

HIT Statistics

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	-	-

Annotation Distributions



Annotation Quality

	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[1]	0.725	0.729

[1] Results on a different 500-sentence random sample of the same corpus.

Annotation Costs

- Three MechTurk annotators per HIT.
- 1000 sentences $\times 3 \times 3$ HIT designs = total of 9000 assignments.
- Reward of .02$ per assignment → total sum spent = 225$ (180$ + 45$ Amazon fees).
- In-house expert: 1000 assignments in three hours at 70$ per hour.
- We saved 210$ ($= 70 \times 3$) - 75$ ($= \frac{225}{3}$) = 135$, which constitutes almost 65% of the cost of an expert annotator.
- A lower reward (.01$) would most probably have worked equally well.

Outline

Why participate in the Shared Task?

HIT Design

Annotation Results

Experimental Results

Conclusions

Two Experiments

- Experiment 1: practical utility. Train a single polarity classifier on different data sets. Compare system with noisy available metadata and with MechTurk generated annotations of HIT1.
- Experiment 2: theoretical utility. Train several polarity classifiers using the same data set, annotated by experts and MechTurk workers.

Description of Data Sets

All sentences were extracted from a corpus of user opinions on cars from the automotive section of `www.ciao.es`.

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

Experiment 1

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)

Experiment 2

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

Outline

Why participate in the Shared Task?

HIT Design

Annotation Results

Experimental Results

Conclusions

Conclusions: 1. Annotation Quality

Question?

How do MechTurk worker annotations compare to expert annotations?

Result

High inter-annotator agreement between MechTurk workers and experts.

Conclusions: 2. Annotation Applicability

Question?

How useful are MechTurk worker annotations to train polarity classifiers?

Result

MechTurk worker annotations outperform initial noisy data and produce results on par with expert annotations.

Conclusions: 3. Language Barriers

Question?

Is there a large enough worker pool for Spanish?

Result

After introducing a competence test, getting results for Spanish was easy and rapid.

Conclusions: 4. Return on Investment

Question?

What is the Return on Investment of using MechTurk annotations instead of expert annotations.

Result

For the proposed task, we saved 65% of an expert annotator at a relatively high MechTurk reward.

Questions?