TASS: Detecting Sentiments in Spanish Tweets

Xabier Saralegi and Iñaki San Vicente
Elhuyar Fundazioa

hizkuntza ELHUYAR zerbitzuak

Introduction

- Knowledge discovery useful for decision making and market analysis.
- Explosion of Web 2.0, very rich source of user-generated information.
  - Social media like twitter a very valuable source for seeking opinions.
- *TASS*: Opinion mining or sentiment analysis over Spanish tweets.

State of the Art

- Main approaches:
  - Knowledge/Lexicon/Rules based approach (Turney, 2002; Kim and Hovy, 2004).
  - Supervised approach (Pang et al., 2002).

- Dealing with tweets:
  - POS and lemmas (Barbosa and Feng, 2010).
  - Emoticons (O'Connor et al., 2010).
  - Discourse (Somasundaran et al., 2009).
  - Follower graph (Speriosu et al., 2011).

- Approaches for tweets:
  - Supervised combined with lexicons (Barbosa and Feng, 2010; Kouloumpis, Wilson, and Moore, 2011).
  - Semi-supervised (label propagation) combined with lexicons (Sindhwani and Melville, 2008).

Experiments

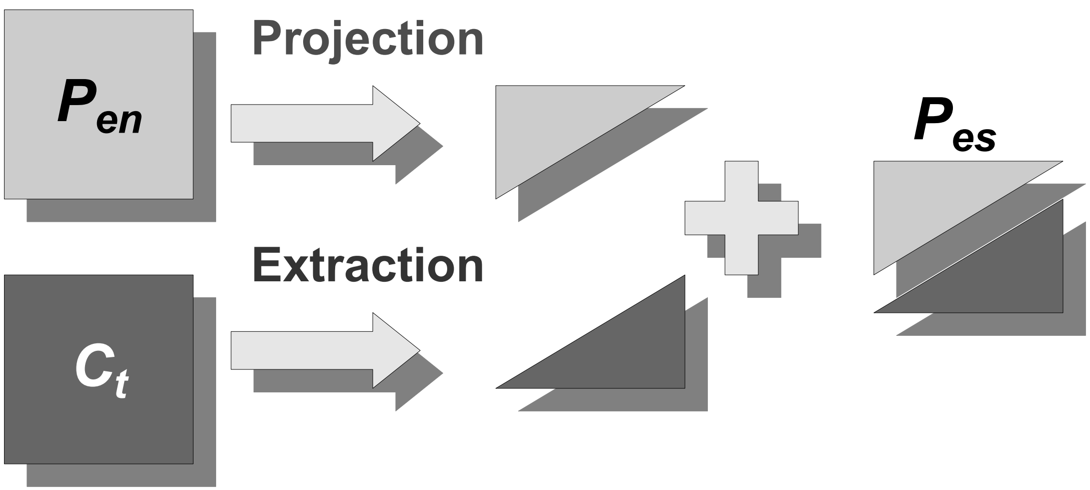
Training Data

- Training data C_t consists of 7,219 tweets:

Polarity	# of tweets	% of tweets
P+	1,764	22,44%
P	1,019	14,12%
NEU	610	8,45%
N	1,221	16,91%
N+	903	12,51%
NONE	1,702	23,58%
Total	7,219	100%

Polarity Lexicon

- A new polarity lexicon for Spanish $\boldsymbol{P_{es}}$ created from two different sources:
  a) An existing English polarity lexicon $\boldsymbol{P_{en}}$ (Projection).
  b) Training corpus $\boldsymbol{C_t}$ (Extraction).



- An English polarity lexicon (Wilson et al., 2005) $\boldsymbol{P_{en}}$ automatically translated into Spanish:
  - Translation by a English-Spanish dictionary $\boldsymbol{D_{en \rightarrow es}}$

- An English polarity lexicon (Wilson et al., 2005) $\boldsymbol{P_{en}}$ automatically translated into Spanish:
  - Translation by a English-Spanish dictionary $\boldsymbol{D_{en \rightarrow es}}$

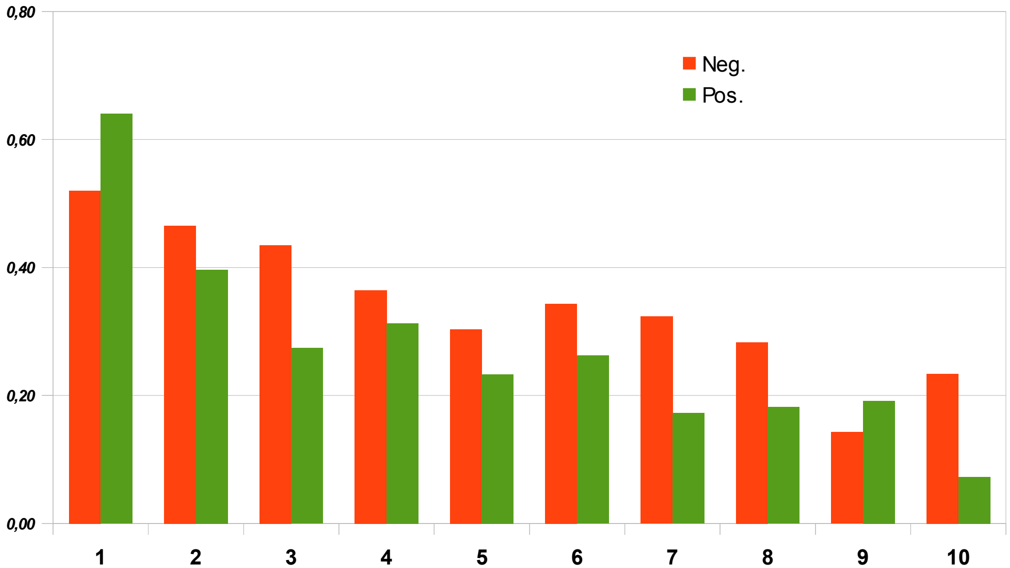
	# of headwords	# of pairs	Avg # of trans.
$\boldsymbol{D_{en \rightarrow es}}$	15,134	31,884	2.11

  - Ambiguous translations solved manually:
    - Polarity was also revised.

- Translated dictionary:

Polarity	English words in P_{en}	Words translated by $D_{en \rightarrow es}$	Translation candidates	Selected candidates
N	4,144	2,416	3,480	**2,164**
P	2,304	2,057	2,271	**1,180**
Total	6,878	4,473	5,751	**3,344**

- Polarity words automatically extracted from the training corpus $\boldsymbol{C_t}$:
  - Extraction of the words most associated with a certain polarity by using *Loglikelihood ratio* (*LLR*).
  - Top 1,000 negative and top 1,000 positive words manually checked:
    - 338 negative and 271 positive words selected.



- Merging projection and extraction based dics.:

	***Projection* based lexicon**	***Extraction* based lexicon**	***Final* lexicon P_{es}**
N	2,164	338	2,435
P	1,180	271	1,518
Total	3,344	609	3,953

Supervised system

- SMO implementation of the Support Vector Machine algorithm (*Weka*).
- All the classifiers built over the training data.
- All the classifiers evaluated by the 10-fold cross validation.

- Pre-process: Some heuristics for dealing with normalization
  - Replication of characters (e.g., "*Sueñooo*"):
    - Removed according to Freeling's dictionary.
  - Abbreviations (e.g., "*q*", "*dl*", …):
    - Extended by using a equivalents list.
  - Overuse of upper case (e.g., "*MIRA QUE BUENO*"):
    - If a sequence of two common words change to lower case.
  - Normalization of urls:
    - complete url replaced by "URL".

Baseline::Supervised System

- Unigram representation using all lemmas (Freeling) as features (15,069).
- Frequency of the lemmas as values.

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578

Selection of Polarity Words::Supervised System

- Only lemmas included in the polarity lexicon $\boldsymbol{P_{es}}$:
  - More precise features and less computational cost (From 15,069 to 3,730 features).

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578
SP	**0.484**	**0.594**	**0.254**	**0.098**	**0.397**	**0.422**	**0.598**

Emoticons and Interjections::Supervised System

- Two new features: # of positive emoticons, # of negative emoticons:
  - A list of 23 positive and 34 negative emoticons.
- Two new features: # of positive interjections, # of negative interjections:
  - A list of 28 positive and 54 negative interjections.

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578
SP	0.484	0.594	0.254	0.098	0.397	0.422	0.598
SP+EM	**0.49**	**0.612**	**0.253**	**0.097**	**0.402**	**0.428**	**0.6**

POS Information::Supervised System

- POS tags as features.
- Useful for distinguishing between subjective and objective texts.

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578
SP	0.484	0.594	0.254	0.098	0.397	0.422	0.598
SP+EM	0.49	0.612	0.253	0.097	0.402	0.428	0.6
SP+POS	**0.496**	**0.596**	**0.245**	**0.093**	**0.414**	**0.438**	**0.634**

Frequency of Polarity Words::Supervised System

- Two new features: a score of the positivity and a score of the negativity of a tweet:

$$spos = \sum_{w_i \in tweet} positive(P_{es}, w_i)$$

$$sneg = \sum_{w_i \in tweet} negative(P_{es}, w_i)$$

Frequency of Polarity Words::Supervised System

- Treatment of negations and adverbs:
  - Change the polarity of a word it is included in a negative clause.
  - Increase (e.g., "*mucho*", "*absolutamente*") or decrease (e.g., "*poco*") the polarity of a word depending on the adverb.
- Weight polarity of words depending on Syntactic Nesting Level:
  - Importance of each word *w* by the relative syntactic nesting level 1/ln(*w*):

$$spos = \sum_{w_i \in tweet} \left(positive(P_{es}, w_i) + \frac{1}{\ln(w_i)} \right)$$

$$sneg = \sum_{w_i \in tweet} \left(negative(P_{es}, w_i) + \frac{1}{\ln(w_i)} \right)$$

Frequency of Polarity Words::Supervised System

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578
SP	0.484	0.594	0.254	0.098	0.397	0.422	0.598
SP+EM	0.49	0.612	0.253	0.097	0.402	0.428	0.6
SP+POS	0.496	0.596	0.245	0.093	0.414	0.438	0.634
SP+FP	**0.514**	**0.633**	**0.261**	**0.115**	**0.455**	**0.438**	**0.613**

All features combined::Supervised System

Features/ Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.45	0.574	0.267	0.137	0.368	0.385	0.578
SP	0.484	0.594	0.254	0.098	0.397	0.422	0.598
SP+EM	0.49	0.612	0.253	0.097	0.402	0.428	0.6
SP+POS	0.496	0.596	0.245	0.093	0.414	0.438	0.634
SP+FP	0.514	0.633	0.261	0.115	0.455	0.438	0.613
All	**0.523**	**0.648**	**0.246**	**0.111**	**0.463**	**0.452**	**0.657**

Using Additional Corpora::Supervised System

- Additional training data C_{tw} was retrieved using the attitude feature of the twitter search:
  - Search is based on emoticons as in (Go et al., 2009).
- Retrieved tweets were classified according to their attitude (P or N):

Corpora/ Tweets	P	N	Total
C_{tw}	11,363	9,865	21,228

- Compiled corpus used in two ways:

  A)Find new polarity words for polarity lexicon P_{es} (AC1).

  B)Adding C_{tw} to the training data (AC2).

Using Additional Corpora::Supervised System

A)Extraction of polarity words from $\boldsymbol{C_{tw}}$ (AC1)

- Same methodology as used for building $\boldsymbol{P_{es}}$:
  - *LLR* for extracting positive and negative candidates.
  - First 500 positive and first 500 negative candidates manually revised (110 positive and 95 negative selected).

Features/ Metric	**Acc. (6 cat.)**	**P+**	**P**	**NEU**	**N**	**N+**	**NONE**
All	0.523	0.648	0.246	0.111	0.463	0.452	0.657
All+AC1	**0.523**	**0.647**	**0.248**	**0.116**	**0.46**	**0.451**	**0.655**

Using Additional Corpora::Supervised System

B)Adding examples from $\boldsymbol{C_{tw}}$ to the training data (AC2):

- Original Training data $\boldsymbol{C_t}$ divided into two parts:
  - $\boldsymbol{C_{t\text{-}test}}$ (15%) and $\boldsymbol{C_{t\text{-}train}}$ (85%).
- Adding examples from $\boldsymbol{C_{tw}}$ to $\boldsymbol{C_{t\text{-}train}}$:
  - All of examples for training (All+AC2).
  - Only examples containing OOV words ($w \in P_{es} \wedge freq(w, C_{t-train}) = 0$): (All+AC2/OOV)

Features/ Metric	# of training examples	Accuracy
All	6,137	0.573
All+AC2	**27,365**	**0.507**
All+AC2/OOV	**7,807**	**0.569**

Evaluation & Results

- Test data $\boldsymbol{C_t}$ consists of 60,798 tweets:

Polarity	# of tweets	% of tweets
P+	20,745	34.12%
P	1,488	2.45%
NEU	1,305	2.15%
N	11,287	18.56%
N+	4,557	7.5%
NONE	21,416	35.22%
Total	60,798	100%

Results

Features/ Metric	Acc. (4 cat.)	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Baseline	0.616	0.527	0.638	0.214	0.139	0.483	0.471	0.587
All	0.702	0.641	0.752	0.323	0.166	0.563	0.564	0.683
All+AC1 (**submitted run**)	0.711	0.653	0.753	0.32	0.167	0.566	0.566	0.685

- AC1 provides improvement.
- Best performance over P+ and NONE.
- Worst performance over NEU and P.
- Better results than those achieved over the training data:
  - The best system (ALL+AC1): 0.653 vs. 0.523.

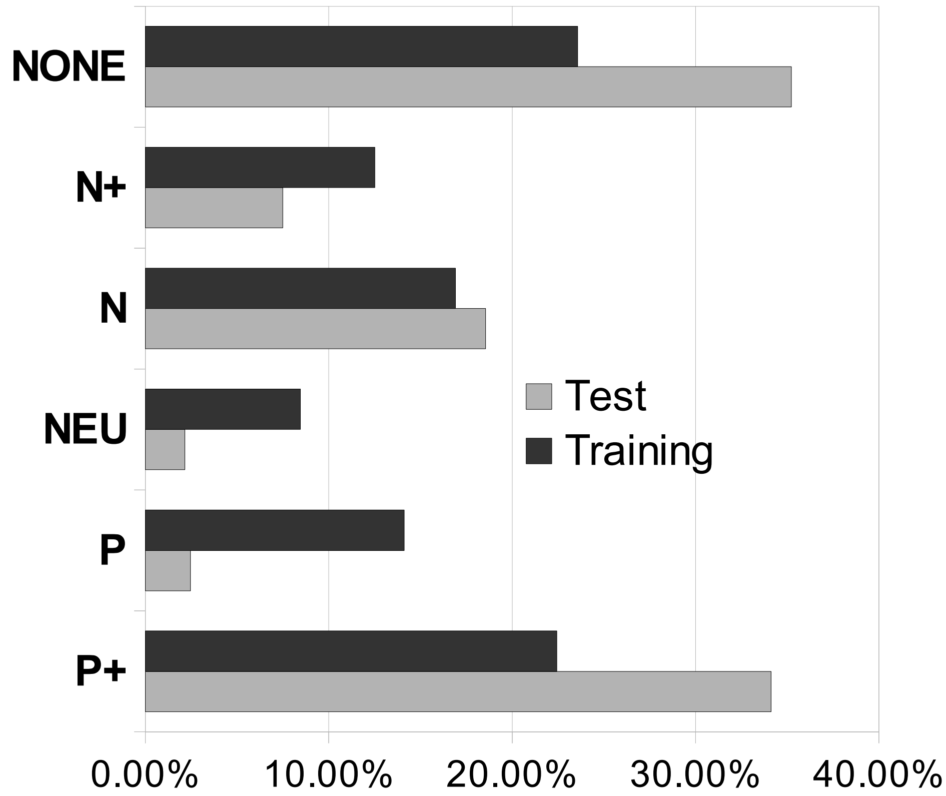
Results

Features/ Metric	**Acc. (4 cat.)**	**Acc. (6 cat.)**	**P+**	**P**	**NEU**	**N**	**N+**	**NONE**
Baseline	0.616	**0.527**	0.638	0.214	0.139	0.483	0.471	0.587
All	0.702	**0.641**	0.752	0.323	0.166	0.563	0.564	0.683
All+AC1	0.711	**0.653**	0.753	0.32	0.167	0.566	0.566	0.685

Features/ Metric	**Acc. (6 cat.)**	**P+**	**P**	**NEU**	**N**	**N+**	**NONE**
Baseline	**0.45**	0.574	0.267	0.137	0.368	0.385	0.578
All	**0.523**	0.648	0.246	0.111	0.463	0.452	0.657
All+AC1	**0.523**	0.647	0.248	0.116	0.46	0.451	0.655

Results

- The distribution difference between training and test data:



Conclusions

Conclusions

- Our system effectively combines several features based on linguistic knowledge:
  - Lemmas, POS tags, polarity words...
- Good contribution of semi-automatically built polarity dictionary.
- Robust performance of the system.

References

- Barbosa, Luciano and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. COLING '10.
- Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classication using distant supervision. CS224N Project Report.
- Han, Bo and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. ACL-HLT.
- Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. COLING.
- Kouloumpis, E., T. Wilson, and J. Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! AAAI.
- Liu, Fei, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. ACL.
- O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. AAAI.
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classication using machine learning techniques. EMNLP.
- Sindhwani, Vikas and Prem Melville. 2008. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. ICDM-
- Somasundaran, Swapna, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classication. EMNLP.
- Speriosu, Michael, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classication with label propagation over lexical links and the follower graph. EMNLP.
- Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classication of reviews. ACL.
- Wilson, Theresa, Paul Homann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Rilo, and Siddharth Patwardhan. 2005. OpinionFinder. HLT/EMNLP.

Thank You