TASS 2015, septiembre 2015, pp 81-86 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Spanish Twitter Messages Polarized through the Lens of an English System

Mensajes de Twitter en español polarizados desde la perspectiva de un sistema para inglés

Marlies Santos Deas
Columbia University
New York, NY USA
ms5072@columbia.edu

Or Biran
Columbia University
New York, NY USA
orb@cs.columbia.edu

Kathleen McKeown
Columbia University
New York, NY USA
kathy@cs.columbia.edu

Sara Rosenthal
Columbia University
New York, NY USA
sara@cs.columbia.edu

Resumen: En este artículo describimos la adaptación al español de un sistema basado en aprendizaje automático con clasificación supervisada que fue desarrollado originalmente para el idioma inglés. El equipo de la Universidad de Columbia adaptó este sistema para participar en la Tarea 1 propuesta en TASS 2015, que consiste en determinar la polaridad a nivel global de un grupo de mensajes escritos en español en la red social Twitter.
Palabras clave: análisis de sentimiento, clasificación de polaridad

Abstract: In this paper we describe the adaptation of a supervised classification system that was originally developed to detect sentiment on Twitter texts written in English. The Columbia University team adapted this system to participate in Task 1 of the 4th edition of the experimental evaluation workshop for sentiment analysis focused on the Spanish language (TASS 2015). The task consists of determining the global polarity of a group of messages written in Spanish using the social media platform Twitter.
Keywords: sentiment analysis, polarity classification

1 Introduction

Sentiment analysis is the field concerned with analyzing the sentimental content of text. Most centrally, it involves the task of deciding whether an utterance contains *subjectivity*, as opposed to only objective statements, and determining the *polarity* of such subjective statements (e.g., whether the sentiment is positive or negative). Automatic sentiment analysis has important applications in advertizing, social media, finance and other fields. One variant that has become popular in recent years is sentiment analysis in microblogs, notably *Twitter*, which introduces difficulties common in that genre such as very short utterances, non-standard language and frequent out-of-vocabulary words.

The vast majority of work on sentiment analysis has been on English texts. Since methods for determining sentiment often rely on language-specific resources such as sentiment-tagged thesauri, they are often difficult to adapt to languages beyond English, as other language often have scarcer computational resources.

This paper describes the efforts of the Columbia University team at *Task 1* of TASS[1] 2015. TASS is an annual workshop focusing on sentiment analysis in Spanish, especially of short social media texts such as tweets. Each year, TASS proposes a number of tasks and collects the results of different participating systems.

In 2015, Task 1 is a combined subjectivity-polarity task: for each tweet, the competing system is expected to provide a label. There are two variants - the fine-grained variant, where there are six labels: {P+, P, Neu, N, N+, NONE}, and the coarse variant, where there are four labels: {P, Neu, N, NONE}. TASS distributes a standard data set of over 68,000 Spanish tweets for participants in this task (Villena-Román et al., 2015).

Instead of creating a new Spanish-specific system, we have adapted our existing English system to the Spanish language. We show that with relatively small engineering efforts and the proper resources, but without

[1] Taller de Análisis de Sentimientos

Publicado en http://ceur-ws.org/Vol-1397/. CEUR-WS.org es una publicación en serie con ISSN reconocido ISSN 1613-0073

any language-specific feature engineering, our system can be adapted to a new language and achieve performance that is competitive with other systems at TASS. As a side effect, we formalized the process of adapting our system to any new language.

2 Related Work

Sentiment analysis in Twitter is a recent but popular task. In English, the SemEval Task of Sentiment Analysis in Twitter was the most popular task in both 2013 and 2014 (Rosenthal et al., 2014). In Spanish, TASS has organized a Twitter sentiment analysis task every year since 2012.

Multiple papers focusing on this task have been recently published. Most focus on supervised classification, using lexical and syntactic features (Go, Bhayani, and Huang, 2009; Pak and Paroubek, 2010; Bermingham and Smeaton, 2010). The latter, in particular, compare polarity detection in twitter to the same task in blogs, and find that despite the short and linguistically challenging nature of tweets, it is easier to detect polarity in Twitter than it is in blogs using lexical features, presumably because of more sentimental language in that medium.

Other work focused on more specialized features. Barbosa and Feng (2010) use a polarity dictionary that includes non-standard (slang) vocabulary words as well as Twitter-specific social media features. Agarwal et al. (2011) use the Dictionary of Affect in Language (DAL) (Whissell, 1989) and social media features such as slang and hashtags. Rosenthal, McKeown, and Agarwal (2014) use similar features, as well as features derived from Wiktionary, WordNet and emoticon dictionaries.

In Spanish, most work on Twitter sentiment analysis has been in the context of TASS. Many of the top-performing systems utilize a combination of lexical features, POS and specialized lexicons: the Elhuyar system relies on the Elhuyar Polar lexicon (Roncal and Urizar, 2014), while the LyS system (Vilares, Doval, and Gómez-Rodrıguez, 2014) and the CITIUS-CILENIS system (Gamallo and Garcia, 2013) each evaluate several Spanish-language lexicons. Other systems rely on distributional semantics (Montejo-Raez, Garcia-Cumbreras, and Diaz-Galiano, 2014) and on social media features (Zafra et al., 2014; Fernández et al., 2013).

3 Method

The main effort consisted of adapting an English sentiment analysis system for Spanish tweets, particularly for Task 1 of TASS 2015. The English system has been successfully applied to two editions of the SemEval Task 9 ("Sentiment Analysis in Twitter") - 2013 and 2014 (Rosenthal et al., 2014). The system consists of a Logistic Regression classifier that utilizes a variety of lexical, syntactic and specialized features (detailed in Section 3.2). It has two modes that can be run independently or in conjunction:

1. Subjectivity detection (distinguish between subjective and objective tweets)
2. Polarity detection (classify subjective tweets into positive, negative, or neutral).

The system is described in detail in Rosenthal and McKeown (2013). For the TASS task, four new modes were added:

1. Four-way classification, where the possible classes are P, N, NEU, and NONE
2. Four-way composite classification, where tweets are run through a two-step process: a binary classification (subjective, objective) followed by a three-way classification (P,N,NEU) of subjective tweets. Objective tweets in turn are given the label "NONE". Consequently, this two-step classification process yields to a four-way classifier. To train the subjectivity classifier, we grouped all labels other than "NONE" into one subjective label.
3. Six-way classification, where the possible classes are P, P+, N, N+, NEU, and NONE
4. Six-way composite classification (similar to four-way composite, and including two more labels: P+ and N+)

3.1 Preprocessing of tweets

Special tokens such as emoticons are replaced by a related word (e.g. "smiley") and supplemented with its affect values as represented in the DAL (Whissell, 1989). URLs and Twitter handles are converted to fixed tags that are not analyzed further to determine whether they are carriers of polarity. This process is unchanged from the English system.

We use the Stanford NLP library [2] for tag-

[2] http://nlp.stanford.edu/software/index.shtml

ging and parsing the tweet. Using the parse tree labels, we chunk the tweet into its shallow syntactic constituents (e.g. grup.nom). As in the English system, the chunker outputs one of three labels per token to indicate the position of the latter within a chunk: 'B' for beginning, 'I' for in (or intermediate, a continuation of the current chunk), and 'O' for out-of-vocabulary.

3.2 Features

The set of features used for Spanish is the same as that of the English system; we did not incorporate any Spanish-specific features for this task. The features currently utilized are essentially those described in Rosenthal, McKeown, and Agarwal (2014), and have evolved over time from the original system detailed in Agarwal, Biadsy, and Mckeown (2009).

In addition to lexical features (n-grams and POS), the system utilizes a variety of specialized features for social media text: emoticons; expanded web acronyms (LOL → laugh out loud) and contractions (xq → porque); punctuation and repeated punctuation; lengthened words in the tweet (e.g., largooooooo); all-caps words; and slang. We also use statistics of the DAL values for the words in the tweet (e.g., the mean activation, the max imagery, etc.).

3.3 Adaptation to Spanish

Adapting the English system to Spanish included two parts. First, we had to find Spanish equivalents to the English lexical resources (dictionaries, word lists etc.) that our system relies on. Second, we had to find equivalent Spanish NLP tools (a tokenizer, POS tagger and chunker).

3.3.1 Lexical Resources

The major challenge we faced was the lack of readily available resources in Spanish. In some cases, Spanish resources could be found and incorporated without a major effort - for example, the Spanish version of the DAL (Dell Amerlina Ríos and Gravano, 2013) was simple to integrate. In other cases, we had to put in more significant work - especially for the social media resources (e.g. the lists of contractions and emoticons). Table 1 details the English lexical resources used by our system and the Spanish equivalents, in addition to the location in which we found them or the method we used to create or adapt them.

We integrated the Standard Spanish dictionary distributed with Freeling[3] as our non-slang dictionary. For the DAL, we use the Spanish version created by Dell Amerlina Ríos and Gravano (2013). We leveraged the Google Translate service to create a Spanish version of our list of emoticons, and manually created a list of Spanish contractions.

The resulting Spanish resources are not identical to the original English ones. For example, the DAL scores for the word "abandon" and its Spanish translation "abandonar" are close but not exactly the same. Furthermore, the number of entries in the English DAL is more than three times that of the Spanish one, which results in a significant difference in coverage. In the standard dictionary, due to the highly inflected nature of the Spanish language, the number of entries more than quintuples when compared to the English version. Table 2 shows the percentage of the vocabulary (unique tokens) found in the training corpus for each resource. The standard dictionary has the highest coverage, followed by the DAL.

The English system utilizes a few additional resources, namely Wiktionary, WordNet and SentiWordNet. We have not yet integrated a Spanish version of these into the system, and consider that our first priority in future work. While Spanish equivalents of Wiktionary and WordNet do exist (Wikcionario and EuroWordNet), SentiWordNet does not have non-English counterparts. Our planned solution is to use MultiWordNet, a resource in which the English WordNet is aligned with other languages, to translate the English Synsets included in SentiWordNet into Spanish.

Resource	# Found	Percentage
Standard dictionary	10801	45.3 %
DAL	1845	7.7 %
NNP	8293	34.8 %
Punctuation and Numbers	259	1.1 %
Emoticons	11	∼ 0 %

Tabla 2: Coverage of the training set vocabulary by various resources

3.3.2 NLP Tools

We use the Spanish version of the Stanford Maxent Tagger (Toutanova and Manning,

[3]http://nlp.lsi.upc.edu/freeling

Resource	English	Spanish	Location or Method
Dictionary of Affect in Language (DAL)	abandon pleasantness mean: 1.0 activation mean: 2.38 imagery mean: 2.4 ...8,742 entries	abandonar 1.2 2.8 2.0 ...2,669 entries	http://habla.dc.uba.ar/gravano/sdal.php?lang=esp
Contractions	aren't → are not can't → cannot ...52 entries	pal → para el pallá → para allá ...21 entries	Manually created. Included variations with and without accent marks, apostrophes, and slang spelling.
Emoticons	:) happy :D laughter :-(sad ...99 entries	:) feliz :D risa :-(triste ...99 entries	Translated using Google translate
Standard dictionary (i.e. not slang)	98,569 entries, including proper nouns	556,647 entries, including inflected forms	Concatenated files contained in the Freeling installation and removed duplicates

Tabla 1: Parallel Lexical Resources

2000) for tagging. For chunking, we use the Spanish version of the Stanford Parser, and derive chunks from the lowermost syntactic constituents (or the POS, if the token is not a part of an immediate larger constituent).

For example, the Spanish phrase "Buen viernes" is chunked as follows:

Parse tree:
(ROOT (sentence (sn (grup.nom (s.a (grup.a (aq0000 Buen))) (w viernes)))))

Chunked phrase:
Buen/aq0000/B-grup.a viernes/w/B-w

4 Experiments and Results

We submitted two experiments (one simple and one composite; see Section 3) for each combination of classification task (four-way, six-way) and test corpus (full, 1k), for a total of eight experiments. The results are shown in Table 3. For each combination we show the accuracy and the macro-averaged precision, recall and F1 score.

We trained five models with the training data provided by TASS:

1. Four-way (P, N, Neu, NONE)
2. Six-way (P+, P, N+, N, Neu, NONE)
3. Subjectivity model (subjective, objective)
4. Three-way polarity (P, N, Neu)
5. Five-way polarity (P+, P, N+, N, Neu)

The last two were used in conjunction with the subjectivity model to form the composite classifier, as explained in Section 3.

4.1 Discussion

The task in which our system performs the best is the three-label classification using the joint four-way classifier described in Section 3. Both joint models (four-way and six-way) outperform their composite counterparts on the full test corpus. However, there is an improvement when using the composite model on the balanced 1k corpus, for both the three-label and five-label classification.

In terms of labels, our system consistently has the most difficulty classifying neutral (Neu) tweets across all experiments. In comparison, it did well in classifying strongly positive (P+) and objective (NONE) tweets, as well as positive (P) in the three-label subtask. Negative (N, N+) tweets were in between. Table 4 shows the performance of each system (for each task) on individual labels.

To assess the usefulness of our features in discriminating among the different classes, we looked at the odds ratios of the features for each class. Table 5 shows a few of the most discrimative features from each category: n-grams, POS and social media (SM). We found that social media features dominate across all classes, which is not a surprising outcome given the popular use of such features in Twitter communication. As shown in Table 5, emoticons such as a smiley face can be highly discriminative between positive and negative tweets, with a significantly stronger association with the former. Polar N-grams such as "felices" (happy) also constitute a relevant group and tend to be discriminative for the polar classes N and P. In the POS group, interrogative pronouns (pt) mar-

Task	TestSet	Variant	Acc.	Prec.	Rec.	F1	System Rank	Group Rank
5 Labels	Full	Six-way	**0.495**	0.393	0.441	**0.416**	28 / 37	11 / 16
		Composite	0.362	0.313	0.334	0.323	34 / 37	
	1k	Six-way	0.397	0.345	0.372	0.358	23 / 32	
		Composite	**0.419**	0.365	0.372	**0.369**	14 / 32	7 / 16
3 Labels	Full	Four-way	**0.597**	0.492	0.503	**0.497**	26 / 39	13 / 15
		Composite	0.481	0.404	0.403	0.404	36 / 39	
	1k	Four-way	0.578	0.450	0.493	0.470	31 / 39	
		Composite	**0.600**	0.461	0.481	**0.471**	25 / 39	13 / 16

Tabla 3: Sentiment Analysis results at global level (all measures are macro-averaged)

Task	Test Corpus	Variant	P	P+	N	N+	NEU	NONE
5 Labels	Full	Six-way	0.160	0.577	0.434	0.413	0.123	0.599
		Composite	0.096	0.490	0.370	0.278	0.088	0.376
	1k	Six-way	0.194	0.566	0.399	0.332	0.081	0.456
		Composite	0.260	0.583	0.438	0.335	0.078	0.493
3 Labels	Full	Four-way	0.676	N/A	0.603	N/A	0.108	0.544
		Composite	0.576	N/A	0.493	N/A	0.074	0.376
	1k	Four-way	0.667	N/A	0.584	N/A	0.079	0.483
		Composite	0.695	N/A	0.595	N/A	0.088	0.493

Tabla 4: F-measure of each class

king words such as "qué" (what) and "dónde" (where) are most important across all categories, followed by various types of verbs including semiauxiliary gerunds (vsg) and past indicative auxiliary (vais).

Group	P	N	NEU
Social Media	u	lol	:\(
	:D	u	\$
	\$	:D	k
n-grams	esfuerzo	esfuerzo	en @telediariointer 20:30
	gracias	petición de	))))
	felices	pide a rajoy	petición de
Part of Speech	pt000000	pt000000	pt000000
	vsg0000	vais000	vaif000
	vssp000	vsg0000	vsg0000

Tabla 5: Features with high Odds Ratios per class in four-way classification joint model

While it is difficult to compare our system's Spanish results with the results on English - the TASS dataset is quite different from the SemEval dataset - it is evident that the Spanish task is harder. This is not surprising, since we have fewer resources, and the ones which were adapted are in some cases not as comprehensive. However, the fact that we can get competitive results in Spanish using a system that was originally designed for English sentiment analysis shows that relatively quick and painless adaptation to other languages is possible.

5 Conclusion

We have adapted a sentiment analysis system, the original target language of which was English, to classifying the subjectivity and polarity of tweets written in Spanish for participation in Task 1 of TASS 2015. The English system provided significant leverage, allowing for direct reuse of most of its components, from the processing pipeline down to the features used by the classifier. The experimental results are encouraging, showing our system to be competitive with others submitted to TASS despite being adapted into Spanish from another language. From here on we will pursue further enhancements.

The main challenge we encountered was the need to substitute several English lexical resources that the system extensively employs with analogous Spanish variants that were not always easily attainable. In future work, we will incorporate the final missing pieces - Spanish versions of Wiktionary, WordNet and SentiWordNet - so that our Spanish system uses equivalents of all resources used by the English system.

While adapting our system to Spanish, we have compiled a list of necessary resources and presented some automated methods of quickly attaining such reasources in other languages (e.g., using Google Translate to quickly convert a list of emoticons). These along with resources and tools that we expect to be able to find for most languages (e.g.,

a standard dictionary and a list of contractions; a POS tagger and a constituent parser) comprise the bulk of the list. Some resources, such as the DAL, will potentially present a bigger challenge in other languages, but can possibly be automated through token translation as well. In future work, we will experiment with our system in additional languages and further refine our adaptation process.

Acknowledgements

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Bibliography

Agarwal, A., F. Biadsy, and K. R. McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *EACL*.

Agarwal, A., X. Boyi, I. Vovsha, O. Rambow, and R. Passonneau. 2011. Sentiment analysis of Twitter data. In *Workshop on Languages in Social Media*.

Barbosa, L. and J. Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *COLING*.

Bermingham, A. and A. F. Smeaton. 2010. Classifying sentiment in microblogs: Is brevity an advantage? In *International Conference on Information and Knowledge Management*.

Dell Amerlina Ríos, M. and A. Gravano. 2013. Spanish DAL: A Spanish dictionary of affect in language.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martinez-Barco, A. Montoyo, and R. Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams.

Gamallo, P. and M. García. 2013. TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. In *TASS 2013*.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision.

Montejo-Raez, A., M.A. Garcia-Cumbreras, and M.C. Diaz-Galiano. 2014. Participación de Sinai Word2vec en TASS 2014. In *TASS 2014*.

Pak, E. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Conference on International Lang. Resources and Evaluation*.

Roncal, I. S. V. and X. S. Urizar. 2014. Looking for features for supervised tweet polarity classification. In *TASS 2014*.

Rosenthal, S. and K. McKeown. 2013. Columbia NLP: Sentiment detection of subjective phrases in social media.

Rosenthal, S., K. McKeown, and A. Agarwal. 2014. Columbia NLP: Sentiment detection of sentences and subjective phrases in social media. In *SemEval*.

Rosenthal, S., P. Nakov, A. Ritter, and V. Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in Twitter.

Toutanova, K. and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Joint SIGDAT Conference on Empirical Methods in NLP*.

Vilares, D., Y. Doval, and C. Gómez-Rodríguez. 2014. LyS at TASS 2014: A prototype for extracting and analysing aspects from Spanish tweets. In *TASS 2014*.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Urena-Lopez. 2015. Overview of TASS 2015.

Whissell, C. 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4.

Jiménez Zafra, S. M., E. Martínez Cámara, M.T. Martín Valdivia, and L.A. Ureña López. 2014. Sinai-esma: An unsupervised approach for sentiment analysis in Twitter. In *TASS 2014*.