Looking for Features for Supervised Tweet Polarity Classification

Buscando Características para Clasificación Supervisada de Polaridad de Tuits

Iñaki San Vicente Roncal
Elhuyar Fundazioa
Zelai Haundi 3, 20170 Usurbil
i.sanvicente@elhuyar.com

Xabier Saralegi Urizar
Elhuyar Fundazioa
Zelai Haundi 3, 20170 Usurbil
x.saralegi@elhuyar.com

Resumen: Este artículo describe el sistema presentado por Elhuyar para la tarea 1 de análisis de sentimiento enmarcada en la campaña de evaluación TASS 2014. Nuestro sistema trabaja sobre la base de un algoritmo Máquinas de Soporte vectorial (SVM). El sistema combina la información extraida a partir de lexicos de polaridad con características linguísticas. La incorporación de colocaciones basadas en estrucuras sintácticas así como el enriquecimiento de los léxicos de polaridad son los elementos más determinantes en la mejora de los resultados con respecto a TASS 2013. El sistema obtiene una precisión del 61% para la detección de polaridad de alta granularidad y un 69% para baja granularidad.
Palabras clave: TASS, Análisis de sentimiento, Minería de opiniones, Detección de polaridad

Abstract: This article describes the system presented by Elhuyar for the task 1 of the TASS 2014 sentiment analysis evaluation campaign. Our system implements a Support Vector Machine (SVM) algorithm. The system combines the information extracted from polarity lexicons with linguistic features. Incorporating syntax based ngrams and enriching the polarity lexicons prove to be the most influential factors in the improvement of the system with respect to our TASS 2013 participation. The system achieves an 61% accuracy fine granularity and an 69% accuracy for coarse granularity polarity detection.
Keywords: TASS, Sentiment Analysis, Opinion-mining, Polarity detection

1 Introduction

Knowledge management is an emerging research field that is very useful for improving productivity in different activities. Knowledge discovery, for example, is proving very useful for tasks such as decision making and market analysis. With the explosion of Web 2.0, the Internet has become a very rich source of user-generated information, and research areas such as opinion mining or sentiment analysis have attracted many researchers. Prove of that is that in the last years a growing number of Sentiment Analysis related shared tasks have been organized, such as TASS workshops (Villena-Román et al., 2012; Villena-Román et al., 2014), SemEval shared tasks (Nakov et al., 2013; Pontiki et al., 2014; Rosenthal et al., 2014) or the Concept-Level Sentiment Analysis Challenge at ESWC2014[1].

Being able to identify and extract the opinions of users about topics, events, or products is becoming an essential part of market analysis and reputation management systems, and social media is the main source for such information. Because of its special nature (limited length, non standard language), extracting such information from Twitter presents a challenge for Natural Language Processing systems. The TASS evaluation workshop aims "to provide a benchmark forum for comparing the latest approaches in this field". Our team only took part in the first task, which involved predicting the polarity of a number of tweets, with respect to 6-category classification, indicat-

[1]http://challenges.2014.eswc-conferences.org/index.php/SemSA

ing whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. It must be noted that most works in the literature only classify sentiments as positive or negative, and only in a few papers are neutral and/or objective categories included. We developed a supervised system based on a polarity lexicon and a series of additional linguistic features.

The rest of the paper is organized as follows. Section 2 reviews the state of the art in the social media polarity detection field, placing special interest on Twitter and its special characteristics. The third section describes the system we developed, the features we included in our supervised system and the experiments we carried out over the training data. The next section presents the results we obtained over the test data-sets. The last section draws some conclusions and future directions.

2 State of the Art

Much work has been done on the sentiment analysis field, from polarity lexicon induction to sentiment labeling and opinion extraction. There are extensive surveys on the field (Pang and Lee, 2008; Liu, 2012). In the last years microblogging sites such as Twitter have attracted the attention of many researchers with diverse objectives: stock market prediction (Bollen, Mao, and Zeng, 2010), polling estimation (O'Connor et al., 2010) or crisis situations analysis (Nagy and Stamberger, 2012).

The special characteristics of the language of Twitter require a special treatment when analyzing the messages. A special syntax (RT, @user, #tag,...), emoticons, ungrammatical sentences, vocabulary variations and other phenomena lead to a drop in the performance of traditional NLP tools (Foster et al., 2011; Liu et al., 2011). In order to solve this problem, a normalization of the text has been proposed (Brody and Diakopoulos, 2011; Han and Baldwin, 2011), as a preprocess of any analysis.

Once the normalization has been performed, traditional NLP tools may be used to analyse the tweets and extract features such as lemmas or POS tags (Barbosa and Feng, 2010). Emoticons are also good indicators of polarity (O'Connor et al., 2010). Other features analyzed in sentiment analysis such as discourse information (Somasundaran et al., 2009) can also be helpful. Speriosu et al. (2011) explore the possibility of exploiting the Twitter follower graph to improve polarity classification, under the assumption that people influence one another or have shared affinities about topics. Sindhwani and Melville (2008) adopt a semi-supervised approach using a polarity lexicon combined with label propagation. (Barbosa and Feng, 2010; Kouloumpis, Wilson, and Moore, 2011) combined polarity lexicons with machine learning for labelling sentiment of tweets. We adopt this strategy too, which has proven a successful approach in previous shared tasks (Saralegi and San Vicente, 2012; Mohammad, Kiritchenko, and Zhu, 2013).

3 Experiments

3.1 Training Data

The same as in previous editions, the training data C_t consists of 7,219 Twitter messages. Each tweet is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 6 levels have been defined: two positive (P and P+), two negative (N and N+), neutral (NEU) and no sentiment (NONE). The corpus is skewed towards positive polarity (see category distribution in the second column of Table 4), having nearly the 40% of the tweets P or P+ category.

3.2 Polarity Lexicon

3.2.1 Elhuyar Polar

Our main resource is the Elhuyar Polar (ElhPolar) polarity lexicon which was created for previous editions of the TASS workshop. The lexicon was semiautomatically built, on the one hand, by translating an existing English lexicon, and on the other by extracting positive and negative words from the training corpus C_t relying on association measures. All polarities in the lexicon were manually corrected by two annotators, in order to ensure their correctness to the greatest extent. A detailed explanation of building process is included in (Saralegi and San Vicente, 2013a). In addition, for TASS 2014 edition, ElhPolar was enriched with a manually compiled list of locutions, mainly verbals ("*agachar las orejas*", "*mantener el tipo*"), and some set phrases ("*ir a por lana y salir trasquilado*").

3.2.2 Additional lexicons

Experiments were conducted in order to include other polarity lexicons. Combining polarity lexicons will allow us to increase the coverage of the lexicon. We want to stress that even if we are trying to improve the coverage of our lexicon, it is important for us to minimize the noise other lexicons may introduce. That is why we gave preference to manually corrected resources and took some measures to discard entries which may have ambiguous (e.g., "infantil") or weak polarities (e.g., "desechable"). Table 1 provides statistics of the lexicons used. Following we describe briefly the lexicons used in our experiments:

- *Mihalcea's Lexicon (Perez-Rosas, Banea, and Mihalcea, 2012) (Mih)*: Perez Rosa's paper describes two lexicons. We only use here the one regarded as *"full strength"* lexicon, because it integrates manual annotations from OpinionFinder (Wilson et al., 2005).
- *Spanish Emotion Lexicon (SEL) (Sidorov et al., 2013)*: the lexicon provides a Probability Factor of Affective use (PFA) for each of its entries, with respect to at least one of six basic emotions: joy, anger, fear, sadness, surprise and disgust. We map emotions to a binary polarity scale, considering positive words most related to joy, and negative all the others except those related to surprise. We consider surprise an ambiguous sentiment and thus discard those words.
- *SO-CAL lexicon (Taboada et al., 2011)* has the polarities of the words graded in a $[-5, 5]$ scale, from most negative to most positive. The less polar levels $[-3, 3]$ presented some conflicts with respect to other lexicons. Experiments were carried out in order to determine the most suitable words to be included included in our lexicon.

3.3 Supervised System

We used the SMO implementation of the SVM algorithm included in the Weka (Hall et al., 2009) data mining software. All the classifiers built over the training data were evaluated by means of the 10-fold cross validation strategy. Complexity parameter was optimized ($C = 0.666667$).

Lexicon \Polarity	negative	positive	Total
ElhPolar	2,857	1,654	4,511
Mih (full)	476	871	1,347
SO-CAL	2,572	2,119	4,691
SEL	1,193	668	1,861 (+175 discarded)

Table 1: Statistics of the polarity lexicons used by our system.

3.3.1 Preprocessing

As mentioned in section 2, microblogging in general and Twitter, in particular, suffers from a high presence of spelling errors. This hampers any knowledge-based processing as well as supervised methods. Thus prior to any other process, we apply a microtext normalization step. We apply a two step normalization algorithm (Saralegi and San Vicente, 2013b). First, candidates for each unknown word are generated by means of various methods dealing with different error-sources: extension of usual abbreviations, correction of colloquial forms, correction of replication of characters, normalization of interjections, and correction of spelling errors by means of edit-distance metrics. Then, the correct candidates are selected using a language model trained on correct Spanish text corpora.

In addition, all URLs are replaced by the *"URL"* string, and text is converted to lower case (upper case information is saved for later use).

3.3.2 Baseline

The SVM system presented to last year's task 1 was used (Saralegi and San Vicente, 2013a) as baseline. Following we give a brief overview of the features the system uses:

- *ElhPolar*: Frequency of lemmas in Elhuyar Polar polarity lexicon.
- *POS information*: the frequency of the POS tags in a message.
- *Frequency of Polarity Words (FP)*: Two features including the polarity information of the lexicon. Positivity and negativity scores of a tweet are computed based on the polarities in ElhPolar. Various phenomena, such as negation or intensity modifiers are taken into account.

- *Emoticons and Interjections*: Emoticon and interjection lists were compiled from various sources. Emoticons are grouped in 3 positive and 5 negative categories. Interjections are grouped into two classes: positive and negative interjections. Frequency of each category is included as a feature of the classifier.

- *Upper case*: Overuse of upper case (e.g., *"MIRA QUE BUENO"*) is often used to give more intensity to the tweet. The proportion of upper-cased characters in a tweet is stored as a feature.

The features described in the next sections were added on top of this initial configuration. Experiments carried out with various lexicons (section 3.3.6) influence the FP values described above.

Features / Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
All features (Elh2014)	**51.54**	64.6	29.0	13.4	48.8	43.6	65.9
- Ngrams	**-0.51**	-0.9	-0.2	-0.2	-0.1	-0.9	-0.7
- Neg	**-0.13**	-0.1	-0.7	-0.1	0.3	-0.9	0.1
- Punct	-0.06	-0.2	-0.1	0.1	-0.2	0.4	0.1

Table 2: Ablation experiments on C_t corpus. Only the information of ElhPolar lexicon is used in these experiments. Columns 3rd to 8th show F-scores for each of the class values.

3.3.3 Syntax based ngrams (Ngrams)

Frequent ngram combinations can help to better identify the polarity of texts. For example, *"merecer la pena"* (to be worth), is a positive expression, but *"pena"* (pity) is negative. Detecting such structures would be helpful for identifying prior polarities more accurately. So, we extract ngrams from the training corpus based on certain syntactic patterns. Specifically, [N+Adj] and [Verb+Noun] patterns were used to extract locutions (e.g., "perro faldero"). A minimum frequency of 3 occurrences was required for a locution to be accepted. Following this methodology, a total amount of 192 ngrams were extracted. Each of them is included as a new feature in the classifier, storing their occurrence frequency.

Experiments carried out on C_t training data-set ("- Ngrams" row in Table 2, indicate that those locutions are indeed helpful, specially for detecting extreme polarities (P+ and N+).

3.3.4 Punctuation marks (Punct)

Some authors (Proisl et al., 2013; Barbosa and Feng, 2010) suggest that punctuation marks may be good hints for detecting polarity. It is difficult to discern a specific polarity based solely on the information provided by punctuation marks, but they may be a good hint to determine intensity of the sentiment, specially when appearing at the end of a sentence. Following this intuition, we added four new features: the number of exclamation and interrogation marks in a tweet, and whether a tweet ends with and interrogation or exclamation marks.

Results on C_t show that such features do provide some improvement. Looking at the results of the training set, a single feature was included in the final configuration: whether the tweet ends with an interrogation mark or not. "- Punct" row in Table 2 represent the ablation study for this configuration.

3.3.5 Treatment of Negations (Neg)

The polarity of a word changes if it is included in a negative clause[2]. Our baseline system only takes into account negation phenomena when computing FP values. Instead, we include this information explicitly to our learning model. For each lexicon and ngram feature f, another feature NOT_f is created. This nearly duplicates the feature number used by the classifier (from 8k to 14k features).

Experiments on training data (see "- Neg" row in Table 2) showed that the classifier obtains a slight improvement by using those features.

Lexicons \Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elh2014 (All features)	51.54	64.6	29.0	13.4	48.8	43.6	65.9
Elh2014+SEL **(Run1)**	**51.74**	65.0	29.2	12.9	48.9	43.7	66.3
Elh2014+Mih	51.50	64.9	28.8	14.3	48.5	43.3	66.2
Elh2014+SO-CAL3	51.18	64.8	28.2	13.8	47.7	43.3	66.0
Elh2014+SEL+ Mih+SO-CAL3 **(Run2)**	51.63	64.8	28.9	14.2	48.4	43.7	66.9
Elh2014+SEL+ Mih+SO-CAL4 **(Run3)**	51.59	64.7	29.3	14.2	48.2	43.8	66.8

Table 3: Lexicon combination experiments on training data. Columns 3rd to 8th show F-scores for each of the class values.

[2]Syntactic information provided by FreeLing (Padró and Stanilovsky, 2012) is used for detecting those cases.

3.3.6 Lexicon Combination

As we have already mentioned in section 3.3.2, FP features are the solution we have to explicitly provide the classifier with the polarity information stored in the polarity lexicons. This allows the system to take into account those polarity words not appearing in the training data. Rather than adding new influential features to the model, we expect combining lexicons will help to more accurately compute polarity score values.

Since we have combined several lexicons, conflicts arise due to words having several polarities. In order to solve those conflicts, we established a preference order. ElhPolar lexicon is first in this order, followed by SEL, SO-CAL, and Mih. We made this decision because ElhPolar is the most adapted lexicon to the corpus we are working with and it includes information extracted from the training data.

Table 3 presents the results of combining the various lexicons. Results are computed using all the features described in the previous sections. According to those results, neither SO-CAL nor Mih lexicons would be useful. However it is difficult to measure the real impact of such lexicons against the training data, due to the fact that most frequent polarity words in C_t are already included in the ElhPolar lexicon. That could also explain the little improvement achieved overall (0.2%). Hence, we decided to send runs for those configurations with results over the system using only ElhPolar.

Note that there are several configurations using the SO-CAL lexicon. The SO-CAL3 notation refers to using those entries in the lexicon with a polarity score > 3 or < -3. Similarly, SO-CAL4 refers to those entries with scores > 4 or < -4. Including the complete SO-CAL led to a drop in performance for us, so we conducted experiments in order to determine if using only its most polar words could still be helpful. We only include here the configurations which achieved the best results on C_t.

4 Evaluation and Results

The organization provided two evaluation test-sets. On the one hand, for comparison purposes, TASS 2013's test-set C_{e2013} was used (Villena-Román et al., 2014). On the other hand, a 1,000 tweet subset was also prepared C_{e1k}, containing a more similar category distribution compared with the training corpus. Then again, it must be noted that C_{e1k} is yet more skewed towards positive polarity (50% of the whole corpus, as show in the last column of Table 4) and NONE tweets have been reduced considerably.

Polarity	**tweets in C_t**	**tweets in C_{e2013}**	**tweets in C_{e1k}**
P+	22.88% (1,652)	34.12% (20,745)	29.1% (291)
P	17.07% (1,232)	2.45% (1,488)	21.6% (216)
NEU	9.28% (670)	2.15% (1,305)	6.3% (63)
N	18.49% (1,335)	18.56% (11,287)	20.7% (207)
N+	11.73% (847)	7.5% (4,557)	10% (100)
NONE	20.54% (1,483)	35.22% (21,416)	12.3% (123)
Total	100% (7,219)	100% (60,798)	100% (1000)

Table 4: Polarity classes distribution in train and test corpora

Each participant was allowed to send up to three runs per task where 6-category classification (5 polarities + NONE) and 4-category classification (3 polarities + NONE) were considered different tasks. For the 4-category results, all tweets regarded as positive are grouped into a single category, and the same is done for negative tweets. Table 5 presents the results for both evaluations against the C_{e2013} corpus, using the best scored classifiers in the training process. Table 6 presents the results for the evaluation against the C_{e1k} data-set. In addition to the accuracy results, both tables show F-scores for each class for the 6-category classification. For the sake of readability, we will refer to our submitted systems as follows:

- **Run1**: Elh2014+SEL.
- **Run2**: Elh2014+SEL+Mih+SO-CAL3.
- **Run3**: Elh2014+SEL+Mih+SO-CAL4.

Results over the C_{e2013} data-set, show the tendency of improving the results obtained over the training set. Overall, a 1% improvement improvement is achieved over last year's system. Although the system ranked second with this corpus, it is 3% and 1% beyond the best results achieved by ELiRF-UPV team, for 6 and 4 category classifications, respectively.

Results over the C_{e1k} data-set (see Table 6) are overall lower than those obtained with the C_{e2013} corpus. Accuracy is below training corpus results in all cases. However, the improvement our new features obtain over last

Metric/ System	Acc. (4 cat.)	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elhuyar Tass2013	68.6	60.1	72.5	22.8	14.4	54.5	46.5	66.9
Run 1	**69.9**	**61.0**	73.6	22.8	14.0	55.9	52.7	66.7
Run 2	69.7	60.6	73.1	22.7	14.6	55.8	53.0	66.1
Run 3	69.8	60.6	73.0	22.9	14.6	55.9	53.0	66.0
Best Results (ELiRF-UPV)	**70.9**	**64.3**	-	-	-	-	-	-

Table 5: Results obtained on the evaluation of the C_{e2013} data.

Metric/ System	Acc. (4 cat.)	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elhuyar Tass2013	61.0	44.8	66.5	24.7	14.0	40.8	39.8	45.9
Run 1	62.3	46.7	67.6	19.2	19.1	46.0	46.9	46.5
Run 2	63.2	**47.4**	67.1	23.1	19.1	46.4	45.9	48.0
Run 3	**63.5**	47.3	66.5	21.5	19.6	47.4	47.1	47.6
Best Results (ELiRF-UPV)	**65.9**	**48.0**	-	-	-	-	-	-

Table 6: Results obtained on the evaluation of the C_{e1k} data.

year's system is more notable over this corpus. Also the gap between our system and the best results narrows, specially in the 6 category classification task.

It is worth mentioning that lexicon combinations' performance has a boost compared to the training data. Results on C_{e2013} (Table 5) behave similarly as on C_t, with run1 above the other two, although the differences are minimal, specially in the 4 category classification. In turn, Table 6 shows that, Mih and SO-CAL lexicons which had even a negative contribution on the training data (see Table 3, runs 2 and 3), provide the best results on C_{e1k} improving the results more than 1% and 2% over run 1 and baseline systems, respectively. These results remark the importance of the FP values, because many of the polarity words added by those lexicons only influence the classifier model through the FP values, because they had no occurrence in the training data.

If we take a look at the individual category results, first thing we notice is that neutral tweets are very difficult to classify. Such tweets do contain polarity words, but often they have mixed polarities. We should try to find features that better characterize such messages. The performance of negative categories drops significantly from C_{e2013} to C_{e1k}, but the results on C_{e1k} are in concordance with C_t results. NONE tweets have that same behavior on test data, but in that case, results on C_t agree with those on C_{e2013}. In any case it is difficult for us to draw conclusions, because 43% of the tweets in C_{e1k} are annotated differently in C_{e2013}.

5 Conclusions

We have presented a SVM classifier for detecting the polarity of Spanish tweets. Our classifiers ranked second among 7 participant groups. Starting on the system developed for the TASS 2013 challenge, we have successfully incorporated a series of new features, such as syntax based ngrams or negated elements. The combination of various polarity lexicons has also contributed to the performance improvement. It must be noted that such improvement is not reflected on the experiments carried out on the training corpus. The limited size of the training data and the fact that most influential polarity words were already included in our initial lexicon, make difficult to determine to what extent the additions may help.

There is still room for improvement. We would like to further explore the use of ngram-based locutions on the one hand, for example by collecting polarity annotated locutions. On the other hand, neutral polarity is the hardest one to determine. A future line of work is to direct our efforts towards researching on how to characterize such messages.

Acknowledgments

This work has been partially founded by the Industry Department of the Basque Government under grant IE12-333 (Ber2tek project) and Spanish Government MICINN project *Skater* (Grant No. TIN2012-38584-C06-01).

References

Barbosa, Luciano and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA.

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *1010.3003*, October.

Brody, Samuel and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!:

using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570.

Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. #hardtoparse: POS tagging and parsing the twitterverse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, August.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, november.

Han, Bo and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June.

Kouloumpis, E., T. Wilson, and J. Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, X., S. Zhang, F. Wei, and M. Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), Portland, Oregon*.

Mohammad, Saif, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, page 321–327. Association for Computational Linguistics.

Nagy, Ahmed and Jeannie Stamberger. 2012. Crowd sentiment detection during disasters and crises. In *Proceedings of the 9th International ISCRAM Conference*, pages 1–9.

Nakov, Preslav, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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. In *Fourth International AAAI Conference on Weblogs and Social Media*, May.

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Perez-Rosas, Veronica, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Pontiki, Maria, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Proisl, Thomas, Paul Greiner, Stefan Evert, and Besim Kabashi. 2013. KLUE: Simple and robust methods for polarity classification. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, page 395–401, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Rosenthal, Sara, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval*, volume 14.

Saralegi, Xabier and Iñaki San Vicente. 2012. Tass: Detecting sentiments in spanish tweets. In *Proceedings of the TASS Workshop at SEPLN.*

Saralegi, Xabier and Iñaki San Vicente. 2013a. Elhuyar at tass 2013. In *Proceedings of the TASS 2013 Workshop at SEPLN.*

Saralegi, Xabier and Iñaki San Vicente. 2013b. Elhuyar at tweetnorm 2013. In *Proceedings of the TweetNorm Workshop at SEPLN.*

Sidorov, Grigori, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence*. Springer, pages 1–14.

Sindhwani, V. and P. Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, pages 1025 –1030, December.

Somasundaran, Swapna, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 -*, EMNLP '09, pages 170–179, Stroudsburg, PA, USA.

Speriosu, Michael, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Villena-Román, Julio, Janine García-Morera, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2014. Tass 2013 - a second step in reputation analysis in spanish. *Procesamiento del Lenguaje Natural*, 52(0).

Villena-Román, Julio, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. 2012. Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50:37–44.

Wilson, Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations -*, pages 34–35, Vancouver, British Columbia, Canada.