Elhuyar at TASS 2013

Elhuyar en TASS 2013

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

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

Resumen: Este artículo describe el sistema presentado por nuestro grupo para la tarea de análisis de sentimiento enmarcada en la campaña de evaluación TASS 2013. Adoptamos una aproximación supervisada que hace uso de conocimiento lingüístico. Este conocimiento lingüístico comprende lematización, etiquetado POS, etiquetado de palabras de polaridad, tratamiento de emoticonos y tratamiento de negación. También se lleva a cabo un preprocesado para el tratamiento de errores ortográficos. La detección de las palabras de polaridad se hace de acuerdo a un léxico de polaridad para el castellano creado en base a dos estrategias: Proyección o traducción de un léxico de polaridad de inglés al castellano, y extracción de palabras divergentes entre los tuits positivos y negativos correspondientes al corpus de entrenamiento. El sistema obtiene una precisión del 60% para la detección de polaridad de alta granularidad y un 68% para baja granularidad.
Palabras clave: TASS, Análisis de sentimiento, Minería de opiniones, Detección de polaridad

Abstract: This article describes the system presented for the task of sentiment analysis in the TASS 2012 evaluation campaign. We adopted a supervised approach that includes some linguistic knowledge-based processing for preparing the features. The processing comprises lemmatisation, POS tagging, tagging of polarity words, treatment of emoticons and treatment of negation. A pre-processing for treatment of spell-errors is also performed. Detection of polarity words is done according to a polarity lexicon built in two ways: projection to Spanish of an English lexicon, and extraction of divergent words of positive and negative tweets of training corpus. The system achieves an 60% accuracy fine granularity and an 68% 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. Being able to identify and extract the opinions of users about topics or products would enable many organizations to obtain global feedback on their activities. Some studies (O'Connor et al., 2010) have pointed out that such systems could perform as well as traditional polling systems, but at a much lower cost. In this context, social media like Twitter constitute a very valuable source when seeking opinions and sentiments.

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, indicating 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 polarity detection field, placing special interest on sentence level detection, and on Twitter messages, in particular. 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 with our system first in the training-set and later in the test data-set. The last section draws some conclusions and future directions.

2 State of the Art

Much work has been done in the last decade in the field of sentiment labelling. Most of these words are limited to polarity detection. Determining the polarity of a text unit (e.g., a sentence or a document) usually includes using a lexicon composed of words and expressions annotated with prior polarities (Turney, 2002; Kim and Hovy, 2004; Riloff, Wiebe, and Phillips, 2005; Godbole, Srinivasaiah, and Skiena, 2007). Much research has been done on the automatic or semi-automatic construction of such polarity lexicons (Riloff and Wiebe, 2003; Esuli and Sebastiani, 2006; Rao and Ravichandran, 2009; Velikovich et al., 2010).

Regarding the algorithms used in sentiment classification, although there are approaches based on averaging the polarity of the words appearing in the text (Turney, 2002; Kim and Hovy, 2004; Hu and Liu, 2004; Choi and Cardie, 2009), machine learning methods have become the more widely used approach. Pang et al. (2002) proposed a unigram model using Support Vector Machines which does not need any prior lexicon to classify movie reviews. Read (2005) confirmed the necessity to adapt the models to the application domain, and (Choi and Cardie, 2009) address the same problem for polarity lexicons.

In the last few years many researchers have turned their efforts to microblogging sites such as Twitter. As an example, Bollen, Mao and Zeng (2010) have studied the possibility of predicting stock market results by measuring the sentiments expressed in Twitter about it. 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, many authors have proposed a normalization of the text, as a preprocess of any analysis, reporting an improvement in the results. Brody (2011) deals with the word lengthening phenomenon, which is especially important for sentiment analysis because it usually expresses emphasis of the message. Han and Baldwin (2011) use morphophonemic similarity to match variations with their standard vocabulary words, although only 1:1 equivalences are treated, e.g., *'imo = in my opinion'* would not be identified. Instead, they use an Internet slang dictionary to translate some of those expressions and acronyms. Liu et al. (2012) propose combining three strategies, including letter transformation, "priming" effect, and misspelling corrections.

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. (Barbosa and Feng, 2010; Kouloumpis, Wilson, and Moore, 2011) combined polarity lexicons with machine learning for labelling sentiment of tweets. Sindhwani and Melville (2008) adopt a semi-supervised approach using a polarity lexicon combined with label propagation.

3 *Experiments*

3.1 Training Data

The training data C_t provided by the organization consists of 7,219 Twitter messages (see Table 1). 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: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and no sentiment (NONE). The numbers of tweets corresponding to P+ and NONE are higher than the rest. NEU is the class including the least tweets. In addition, each message includes its Twitter ID, the creation date and the Twitter user ID.

Polarity	**#tweets**	**% of #tweets**
P+	1652	22.88%
P	1232	17.07%
NEU	670	9.28%
N	1335	18.49%
N+	847	11.73%
NONE	1483	20.54%
Total	7,219	100%

Table 1: Polarity classes distribution in corpus C_t.

3.2 Polarity Lexicon

We created a new polarity lexicon for Spanish P_{es} from two different sources:

a) An existing English polarity lexicon (Wilson et al., 2005) P_{en} was automatically translated into Spanish by using an English-Spanish bilingual dictionary D_{en-es} (see Table 2). Despite P_{en} including neutral words, only positive and negative ones were selected and translated. Ambiguous translations were solved manually by two annotators. We adopt a semi-automatic process in order to maximize the accuracy of the final lexicon. Altogether, 5,751 translations were checked. Polarity was also checked and corrected during this manual annotation. It must be noted that as all translation candidates were checked, many variants of the same source word were selected in many cases. Finally, 2,361 negative words and 1,289 positive words were included in the polarity lexicon (see fifth column of Table 3). We detected a significant number of Out Of Vocabulary (OOV) words (43%) in this translation process (see second and third columns of Table 3). Most of these words were inflected forms: pasts (e.g., *"terrified"*), plurals (e.g., *"winners"*), adverbs (e.g., *"vibrantly"*), etc. Plurals and participles were automatically lemmatized and translated. In the case of derivational adverbs, lemmas and their suffixes were translated separately, and then the corresponding translation was constructed and manually revised (e.g., *"alarmingly"* = alarming+ly → alarmante+mente= *"alarmantemente"*). By means of this process we reduced the rate of OOV words down to 31%.

	#headwords	**#pairs**	**avg. #translations**
D_{en-es}	15,134	31,884	2.11

Table 2: Characteristics of the D_{en-es} bilingual dictionary.

b) As a second source for our polarity lexicon, words were automatically extracted from the training corpus C_t. In order to extract the words most associated with a certain polarity; let us say positive, we divided the corpus into two parts: positive tweets and the rest of the corpus. Using the Log-likelihood ratio (LLR) we obtained the ranking of the most salient words in the positive part with respect to the rest of the corpus. The same process was conducted to obtain negative candidates. The top 1,000 negative and top 1,000 positive words were manually checked. Among them, 338 negative and 271 positive words were selected for the polarity lexicon (see sixth column in Table 3).

polarity	**English words in P_{en}**	**Words translated by D_{en-es}**	**Translation candidates**	**Manually selected candidates**	**Manually selected from C_t**	**Colloquial words P_{es}**	**Final lexicon P_{es}**
negative	4,144	2,765	3,481	2,361	271	225	2,857
positive	2,304	1,659	2,270	1,289	338	27	1,654
Total	6,448	4,424	5,751	3,344	609	252	4,511

Table 3: Statistics of the polarity lexicons.

Additionally, we created a list of colloquial polarity vocabulary (e.g., *'chupóptero'*, *'dabuten'*) by collecting words from two sources: *"Diccionario de jerga y expresiones coloquiales"*[1] dictionary and *www.diccionariojerga.com*, a crowdsourcing

[1]http://www.ual.es/EQUAL-ARENA/Documentos/coloquio.pdf

web including colloquial vocabulary edited by users.

3.3 Supervised System

Although some preliminary experiments were conducted using an unsupervised approach, we chose to build a supervised classifier, because it allowed us to combine the various features more effectively. We used the SMO implementation of the Support Vector Machine algorithm included in the Weka (Hall et al., 2009) data mining software. Default configuration was used. All the classifiers built over the training data were evaluated by means of the 10-fold cross validation strategy.

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 the normalization system presented in the TweetNorm 2013 evaluation campaign (Saralegi and San Vicente, 2013). The system follows a two step strategy: 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, we also apply some heuristics in order to look for elements we can influence the polarity of a tweet:

- Overuse of upper case (e.g., *"MIRA QUE BUENO"*). Upper case is used to give more intensity to the tweet. If we detect a sequence of two words all the characters of which are upper case and which are included in Freeling's dictionary as common, we change them to lower case.
- Normalization of urls. The complete url is replaced by the *"URL"* string.

3.3.1 Baseline

As baseline we implemented a unigram representation using all lemmas in the training corpus as features (14,760 altogether). Lemmatisation was done by using Freeling. We stored the frequency of the lemmas in a tweet.

3.3.2 Selection of Polarity Words (SP)

Only lemmas corresponding to words included in the polarity lexicon P_{es} (see section 3.2) were selected as features. This allows the system to focus on features that express the polarity, without further noise. Another effect is that the number of features decreases significantly (from 14,760 to 4,511), thus reducing the computational costs of the model. In our experiments relying on the polarity lexicon (see Table 4, first and second rows) clearly outperforms the unigram-based baseline. The rest of the features were tested on top of this configuration.

3.3.3 Emoticons and Interjections (EM)

Emoticons and interjections are very strong expressions of sentiments. A list of emoticons is collected from a Wikipedia article about emoticons and all of them are classified as positive (e.g., ":)", ":D" ...) or negative (e.g., ":(" , "u_u" ...). 23 emoticons were classified as positive and 35 as negative. A list of 54 negative (e.g., *"mecachis"*, *"sniff"*, ...) and 28 positive (e.g., *"hurra"*, *"jeje"*, ...) emotive interjections including variants modelled by regular expressions were also collected from different webs as well as from the training corpora. The frequency of each emoticon and interjection type (positive or negative) is included as a feature of the classifier.

The number of upper-case letters in the tweet was also used as an orthographical clue. In Twitter where it is not possible to use letter styling, people often use the upper case to emphasize their sentiments (e.g., *GRACIAS*), and hence, a large number of uppercase letters would denote subjectivity. So, the relative number of upper-case letters in a tweet is also included as a feature.

According to the results (see Table 4, 4th row), these clues did not provide a significant improvement. Nevertheless, they did show a slight improvement. Moreover, other literature shows that such features indeed help to detect the polarity (Kouloumpis, Wilson, and Moore, 2011). The low impact of these features could be explained by the low density of such elements in our data-set: only 622 out of 7,219 tweets in the training data (8.6%) include emoticons or interjections. Emoticon, interjection and capitalization features were included in our final model.

3.3.4 POS Information (PO)

Results obtained among the literature are not clear as to whether POS information helps to determine the polarity of the texts (Kouloumpis, Wilson, and Moore, 2011), but POS tags are useful for distinguishing between subjective and objective texts. Our hypothesis is that certain POS tags are more frequent in opinion messages, e.g., adjectives. In our experiments POS tags provided by Freeling were used. We used as a feature the frequency of the POS tags in a message.

Results in Table 4 show that this feature provides a notable improvement and it is especially helpful for detecting objective messages (view difference in F-score between SP and SP+PO for the NONE class).

3.3.5 Frequency of Polarity Words (FP)

The SP classifier does not interpret the polarity information included on the lexicon. We explicitly provide that information as a feature to the classifier. Furthermore, without the polarity information, the classifier will be built taking into account only those polarity words appearing in the training data. Including the polarity frequency information explicitly, the polarity words included in the P_{es} but not in the training corpus will be used by the classifier. By dealing with those OOV polarity words, our intention is to make our system more robust.

Two new features are created to be included in the polarity information: a score of the positivity and a score of the negativity of a tweet. In principle, positive words in P_{es} add 1 to the positivity score and negative words add 1 to the negativity score. However, depending on various phenomena, the score of a word can be altered. These phenomena are explained below.

Treatment of Negations and Adverbs

The polarity of a word changes if it is included in a negative clause. Syntactic information provided by Freeling is used for detecting those cases. The polarity of a word increases or decreases depending on the adverb which modifies it. We created a list of increasing (e.g., *"mucho"*, *"absolutamente"*, ...) and decreasing (e.g., *"apenas"*, *"poco"*, ...) adverbs. If an increasing adverb modifying a polarity word is detected, the polarity is increased (+1). If it is a decreasing adverb, the polarity of the words is decreased (−1). Syntactic information provided by Freeling is used for detecting these cases.

Intensity of polarity

Some words denote polarity more intensely than others; e.g., *'aborrecer'* is clearly negative, while *'abundancia'* can be negative in some contexts, although it is generally considered positive. We manually analyzed those words in P_{es} that occurred in the training corpus C_t, and we annotated strongly polar words. We consider those words better polarity words and thus, we give them a higher weight (1.6 instead of 1).

Features/ Metric	**Acc. (6 cat.)**	**P+**	**P**	**NEU**	**N**	**N+**	**NONE**
Baseline	0.436	0.566	0.278	0.174	0.371	0.369	0.59
SP	0.463	0.587	0.269	0.098	0.142	0.413	0.581
SP+PO	0.471	0.587	0.27	0.127	0.403	0.422	0.618
SP+EM	0.475	0.615	0.261	0.133	0.411	0.41	0.598
SP+FP	0.495	0.627	0.279	0.161	0.457	0.429	0.624
All	0.506	0.642	0.287	0.144	0.47	0.427	0.655

Table 4: Accuracy results obtained on the evaluation of the training data. Columns 3rd to 8th show F-scores for each of the class values.

4 Evaluation and Results

The evaluation test-set C_e provided by the organization consists of 60,798 Twitter messages (see Table 5). Each participant was allowed to send an unlimited number of runs. Although the results include classification into 6 categories (5 polarities + NONE), the results were also given on a 4-category basis (3 polarities + NONE). 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 6 presents the results for both evaluations using the best scored classifiers in the training process. In addition to the accuracy results, Table 6 shows F-scores for each class for the 6-category classification.

The first thing we notice is that the results obtained with the test data are better than those achieved with the training data for all configurations. The best system (ALL) achieves 0.601 of accuracy while the same system scored 0.506 of accuracy in training. Even the baseline shows the same tendency. Regarding the differences between configurations, tendencies observed in the cross validation evaluation of the train-

Polarity	**#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%

Table 5: Polarity classes distribution in test corpus C_e.

ing data are confirmed in the evaluation of the test data. Then again, improvement of ALL over Baseline is also higher in test data-based evaluation than in the training cross-validation evaluation: a 16.06% improvement in the accuracy over the baseline was obtained in training cross-validation, while in the test data evaluation, the improvement rose to 18.54%. P+ and NONE classes are those our classifier identifies best, being NEU and P the classes with the worst performance (tables 4 and 6). If we look at the distribution of the polarity classes (tables 1 and 5), we can see that the proportion of the P+ and NONE classes increases significantly in the test data with respect to the training data. By contrast, the NEU and P classes decreased dramatically. These distribution differences between development and test datasets lead us to the conclusion that both datasets have been annotated following different criteria and/or methodologies. The distribution differences together with the performance of the system regarding specific classes could explain the gap in accuracy between test and training evaluations.

Metric/ System	**Acc. (4 cat.)**	**Acc. (6 cat.)**	**P+**	**P**	**NEU**	**N**	**N+**	**NONE**
Baseline	0.595	0.507	0.61	0.175	0.125	0.462	0.406	0.593
ALL	**0.686**	**0.601**	0.725	0.228	0.144	0.545	0.465	0.669

Table 6: Results obtained on the evaluation of the test data.

5 Conclusions

We have presented a SVM classifier for detecting the polarity of Spanish tweets. Our system effectively combines several features based on linguistic knowledge. In our case, using a semi-automatically built polarity lexicon improves the system performance significantly over a unigram model. Other features such as POS tags, and especially word polarity statistics were also found to be helpful. We have improved the tweet normalization step over last year's algorithm. Overall, the system shows robust performance when it is evaluated against test data different from the training data.

There is still much room for improvement. Some authors (Pang and Lee, 2004; Barbosa and Feng, 2010) have obtained positive results by including a subjectivity analysis phase before the polarity detection step. We would like to explore that line of work. Lastly, it would be worthwhile conducting in-depth research into the creation of polarity lexicons including domain adaption and treatment of word senses.

Acknowledgments

This work has been partially founded by the Industry Department of the Basque Government under grant IE11-305 (knowTOUR project).

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.

Choi, Yejin and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 590–598, Stroudsburg, PA, USA.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

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.

Godbole, N., M. Srinivasaiah, and S. Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 219–222.

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.

Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA.

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, Fei, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1035–1044, Jeju Island, Korea, July.

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*.

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.

Pang, Bo and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA.

Rao, Delip and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682, Stroudsburg, PA, USA.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA.

Riloff, E., J. Wiebe, and W. Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceeding of the national conference on Artificial Intelligence*, volume 20, page 1106.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* -, pages 105–112.

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

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.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 417, Philadelphia, Pennsylvania.

Velikovich, Leonid, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 777–785, Stroudsburg, PA, USA.

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.