TASS: Detecting Sentiments in Spanish Tweets

TASS: Detección de Sentimientos en Tuits en Español

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 2012. 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, tratamiento de negación, y ponderación de polaridad según el nivel de anidamiento sintáctico. 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. Los resultados de la evaluación final muestran un buen rendimiento del sistema así como una notable robustez tanto para la detección de polaridad a alta granularidad (65% de exactitud) como a baja granularidad (71% de exactitud).
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, treatment of negation, and weighting of polarity words depending on syntactic nesting level. 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. Evaluation results show a good performance and also good robustness of the system both for fine granularity (65% of accuracy) as well as for coarse granularity polarity detection (71% of accuracy).
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 challenge consisted of two tasks: predicting the sentiment of Spanish tweets, and identifying the topic of

the tweets. 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 pre-process 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.

A common problem of the supervised approaches is to gather labelled data for training. In the case of the TASS challenge, we would tackle this problem should we want to collect additional training data. In order to automatically build annotated corpora, (Go, Bhayani, and Huang, 2009) collect tweets containing the ":)" emoticon and regard them as positive, and likewise for the ":(" emoticon. Kouloumpis (2011) uses a similar approach based on most common positive and negative hashtags. Barbosa (Barbosa and Feng, 2010) rely on existing web services such as Twend or Tweetfeel to collect annotated emoticons. One major problem of the aforementioned strategies is that only positive and negative tweets can be collected.

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+	1,764	24.44%
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%

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. Altogether, 7,751 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, 2,164 negative words and 1,180 positive words were included in the polarity lexicon (see fifth column of Table 3). We detected a significant number of OOV words (35%) 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. So they were not dealt with.

	#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). We found a higher concentration of good candidates among the best ranked candidates (see Figure 1).

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

polarity	English words in P_{en}	Words translated by D_{en-es}	Translation candidates	Manually selected candidates	Manually selected from C_t	Final lexicon P_{es}
negative	4,144	2,416	3,480	2,164	271	2,435
positive	2,304	2,057	2,271	1,180	338	1,518
Total	6,878	4,473	5,751	3344	609	3,953

Table 3: Statistics of the polarity lexicons.

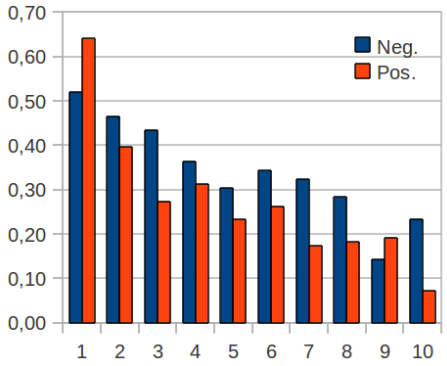

Figure 1: Precision of candidates from C_t depending on LLR ranking intervals (100 candidates per interval {1-100,101-200,...}).

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, except for the one including additional training data (see section 3.3.6 for details).

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. We rejected the use of spell-correctors such as Google spell-checker because they try to treat many correct words that they do not know. Therefore, we apply some heuristics in order to preprocess the tweets and solve the main problems we detected in the training corpus:

- Replication of characters (e.g., *"Sueñooo"*): Sequences of the same characters are replaced by a single character when the pre-edited word is not included in Freeling's[1] dictionary and the post-edited word appears in Freeling's dictionary.
- Abbreviations (e.g., *"q"*, *"dl"*, ...): A list of abbreviations is created from the training corpus. These abbreviations are extended before the lemmatisation process.
- 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.

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

3.3.1 Baseline

As baseline we implemented a unigram representation using all lemmas in the training corpus as features (15,069 altogether). Lemmatisation was done by using Freeling. Contrary to (Pang, Lee, and Vaithyanathan, 2002), we stored the frequency of the lemmas in a tweet. Although using presence performed slightly better in the baseline configuration (improvement was not significant), as other features were included, we achieved better results by using frequency. Thus, for the sake of simplicity, all the experiments shown make use of the frequency.

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 15,069 to 3,730), thus reducing the computational costs of the model. In our experiments relying on the polarity lexicon (see Table 4) 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"*, ...) 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 upper-case 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), 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 (Koulompis, 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 (Koulompis 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.

Syntactic Nesting Level

The importance of the word in the tweet determines the influence it can have on the polarity of the whole tweet. We measured the importance of each word w by calculating the relative syntactic nesting level $l_n(w)$. The lower the syntactic level, the less important it is. The relative syntactic nesting level is computed as the inverse of the syntactic nesting level $(1/l_n(w))$.

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+PO	0.496	0.596	0.245	0.093	0.414	0.438	0.634
SP+EM	0.49	0.612	0.253	0.097	0.402	0.428	0.6
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
ALL+AC1	0.523	0.647	0.248	0.116	0.46	0.451	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.

3.3.6 Using Additional Corpora (AC)

Additional training data were retrieved using the Perl Net::Twitter API. Different searches were conducted during June 2012 using the attitude feature of the twitter search. Using this feature, users can search for tweets expressing either positive or negative opinion. The search is based on emoticons as in (Go et al., 2009). Retrieved tweets were classified according to their attitude.

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

Table 5: Characteristics of the tweet corpus collected from Twitter.

The corpus C_{tw} including retrieved tweets (see Table 5.) was used in two ways: on the one hand, we used it to find new words for our polarity lexicon P_{es}, by using the automatic method described in section 3.2. The first 500 positive candidates and 500 negative candidates were manually checked. Altogether, 110 positive words and 95 negative ones (AC1) were included in the polarity lexicon P_{es}. According to the results (see ALL+AC1 in Table 4), these new polarity words do not provide any improvement. The reason is that most relevant polarity words included in the training corpus C_t are already included in P_es as explained in section 3.2. In order to measure the contribution of these words better, evaluation was carried out against the test corpus where more OOV polarity words would be likely to appear (see section 4).

On the other hand (AC2), we added C_{tw} to the training data, in the hypothesis that more training data would lead to a better model, although polarity strength was not distinguished. Thus, only P and N examples are obtained. In order to evaluate the effect of the new data, the original training data were divided into two parts: 85% (6,137 tweets) for training ($C_{t-train}$) and 15% (1,082) for testing (C_{t-test}). The test data were randomly selected and the proportions of the polarity classes were maintained equal in both parts. Our first classifier (ALL+AC2) was trained with all the retrieved tweets included in C_{tw} as well as the tweets in $C_{t-train}$. Results show (see Table 6) that accuracy decreased when using these data for training. A second experiment was carried out (ALL+AC2-OOV), adding to the training data $C_{t-train}$ only those tweets of C_{tw} containing at least one word w from P_{es} but not appearing in the training corpus ($w \in P_{es} \wedge freq(w, C_{t-train}) = 0$). Only 7.9% of the retrieved tweets were added. Results were still unsatisfactory, and so, additional training data were left out of the final model.

It must be noted that the tweet retrieval effort was very simple, due to the limited time we had to develop the system. We conclude that these additional training data were unhelpful due to the differences with the original data provided: C_{tw} contained many more ungrammatical structures and nonstandard tokens than the original data; the dates of the tweets were different which could even lead to topic and vocabulary differences; and especially, the fact that the additional data collected did not include neutral or objective tweets and neither did it include different degrees of polarity in the case of positive and negative tweets.

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

Table 6: Results obtained by including additional examples in the training data.

4 Evaluation and Results

The evaluation test-set C_e provided by the organization consists of 60,798 twitter messages (see Table 7) annotated as explained in section 3.1. Only one run of results was allowed for submission. 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 8 presents the results for both evaluations using the best scored classifiers in the training process. In addition to the accuracy results, Table 8 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+AC1) achieves 0.653 of accuracy

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 7: Polarity classes distribution in test corpus C_e.

while the same system scored 0.523 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 training data are confirmed in the evaluation of the test data. Then again, improvement of ALL+AC1 over Baseline is also higher in test data-based evaluation than in the training cross-validation evaluation: a 16.22% improvement in the accuracy over the baseline was obtained in training cross-validation, while in the test data evaluation, the improvement rose to 23.91%. P+ and NONE classes are those our classifier identifies best, being NEU and P the classes with the worst performance (tables 4 and 8). If we look at the distribution of the polarity classes (tables 1 and 7), 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. The distribution difference together with the performance of the system regarding specific classes could explain the difference in accuracy between test and training evaluations. It remains unclear to us why the F-scores for all the classes improved with respect to the training phase. We should analyse the characteristics of the training and test corpora, looking for differences in the samples and annotation.

As for the results of the individual classes, it is worth mentioning that neutral tweets are very difficult to classify because they do contain polarity words. We looked at its confusion matrix (both for training and test evaluations) and it shows that NEU tweets wrongly classified are evenly distributed between the other classes, except for the NONE class, with almost no NEU tweets classified as NONE. Most of the NEU tweets contain positive and negative sentences, which leads us to think that a discourse treatment could be useful in order to determine the importance of each sentence with respect to the whole tweet. In the case of positive tweets, P tweets, many of them are classified as P+.

In the experiment (AC1) described in section 3.3.6 we did not obtain any improvement by adding the words extracted from an additional corpus of tweets to the polarity lexicon P_{es}. If we take into account that the most significant words of the training corpus (C_t) were already included in P_{es}, it could be expected that the words in AC1 would have little effect on the training data. In the evaluation against the test data where the vocabulary is larger, the AC1 lexicon provides a slight improvement (see difference between ALL and All+AC1 in Table 8).

Metric/ System	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

Table 8: Results obtained on the evaluation of the test data.

5 Conclusions

We have presented an 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. In our experiments, including external training data was unsuccessful. However, our approach was very simple, and so, a more exhaustive experimentation should be carried out in order to obtain conclusive results. In any case, the system shows robust performance when it is evaluated against test data different from the training data.

There is still much room for improvement. Tweet normalization was naïvely implemented. 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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.