Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 29-35 recibido 29-11-12 revisado 11-02-13 aceptado 19-02-13

WeFeelFine as Resource for Unsupervised Polarity Classification *

WeFeelFine *como recurso para clasificación de la polaridad no supervisada*

Arturo Montejo-Ráez
Departamento de Informática
Universidad de Jaén
Las Lagunillas s/n, Jaén - 23071
amontejo@ujaen.es

Resumen: En este trabajo se presenta una solución no supervisada al problema de la clasificación de la polaridad en micro-blogs. La propuesta no sólo no necesita de entrenamiento, sino que se construye a partir de las propias publicaciones de millones de usuarios en la web. Los resultados muestran la efectividad de esta propuesta, abriendo la puerta a una nueva forma de afrontar el análisis de sentimientos en micro-blogs.
Palabras clave: Análisis de emociones, clasificacin de la polaridad, Twitter, micro-blogging

Abstract: This papers shows the results obtained by a non supervised method in the task of sentiment polarity detection on micro-blogs. This method does not need of training, but it also is self-constructed from millions of publications on the web. The results show the effectiveness of the proposal, openining a new way of facing sentiment analysis in micro-blogs.
Keywords: Sentiment Analysis, polarity classification, Twitter, micro-blogging

1 Introduction

Twitter has become a key service in web-based communication. Its growth rate in terms of content and users has focused the attention of many other services, companies, communities and, of course, scientists. The amount of messages from Twitter users that floods the Internet turns this service into a very useful source of information about the topics on which people focus their interests. Nowadays, proper filtering, extraction and understanding of this overwhelming stream of text is the main subject of study for Natural Language Processing research. Besides, Sentiment Analysis on tweets is one of the most active topic of research taking place (Asiaee T. et al., 2012).

This work presents a novel unsupervised approach to tackle Sentiment Analysis on Twitter by associating to each tweet a list of "feelings" obtained by means of search over a corpus of micro-blogging publications gathered by the WeFeelFine project (Kamvar and Harris, 2011). In this way, tweets are caracterized by the most similar feelings associated by performing a retrieval over the sentences related to each tweet in WeFeelFine data. Then, a final measure of polarity is computed according to the list of feelings obtained. Our results show that this approach outperforms many state-of-the-art unsupervised solutions and that, due to its simplicity, may open a new way of understanding sentiment analysis of micro-blogs by using micro-blogs themselves.

The paper is organized as follows: first a brief introduction to the polarity classification problem is given. Then, the WeFeelFine project is described, with pointers to related research based on its data. Next, our approach is unveiled, describing the prepara-

* This work is partly funded by the European Comission, under the VII Framework Program (FP7 - 2007-2013), within the FIRST project (FP7-287607) and by the Spanish Government, within the TEXT-COOL project (TIN2009-13391-C04-02).

ISSN 1135-5948

tion of the system and its components. Experimental setup and results follows, to end with final conclusions and reflections on future lines to explore.

2 The polarity classification problem

Sentiment Analysis is one of the most active research areas in Natural Language Processing nowadays (Pang and Lee, 2008), with special interest in the classification of texts into positive, negative or neutral. This latter task is known as the Polarity Classification problem, and attracts the attention of the research community and also companies, politicians or personalities, due to the relevance in the study of reputation of products, people or any other item based on opinions of users in the web.

Polarity Classification is solved using both supervised and non-supervised approaches. Supervised strategies have reported the best results since early works (Pang, Lee, and Vaithyanathan, 2002) and it is still the choice for many solutions, from Information Theory based features (with SVM classifier) (Lin et al., 2012) to more complex learned rules (Tan et al., 2012). Unsupervised approaches have relied mainly on the use of lexicons where words are associated with polarity scores (Boldrini et al., 2010), although more avanced solutions using intensive lexical analysis are proposed (Chen et al., 2012). In any case, a value of 70% for F-score seems to be, still, far from these methods.

Turney (Turney, 2002), instead of manually generating a corpus of emotional words, used Altavista search engine to compute the *Semantic Orientation* (SO) of a phrase according to the proximity of well known emotional words (like *excellent* or *poor*) in millions of web pages. Unfortunately, the complexity of modern ranking algorithms used by main search engines has opened a gap between word statistics in the web and the actual results obtained, so the validity of such approach can be argued. Anyhow, the proposal of using the implicit knowledge in millions of texts has been an inspiration in our work.

Sentiment Analysis has been specially focused on Twitter, due to its relevance as social media (Martínez-Cámara et al., In press) and despite the inherent challenges of subjective micro-blogging, like irony (Reyes, Rosso, and Buscaldi, 2012). Our experiments are performed on tweets from this popular service, as explained later.

3 WeFeelFine

Since 2005, the website WeFeelFine[1] has been harvesting from social media millions of sentences containing "I feel" or "I am feeling" expressions, creating a huge database of sentences related to feelings or emotions (Kamvar and Harris, 2011). Although the main goal of the project is to serve as a monitor of human state at a global level, we found that the collected data could be useful in sentiment analysis. The authors of the website, indeed, perform this kind of analysis in order to produce semantic related data. Thanks to its API, it is possible to download a bunch of sentences (up to a limit of 1,500 imposed by the site) per each of the defined feelings. The current list of feelings stored contains 2,178 different feelings, although the 200 most frequent ones hold 70% from a total of almost 2 millions sentences. We can see the feelings with higher presence in the database along with the percentage over the total of sentences in Figure 1.

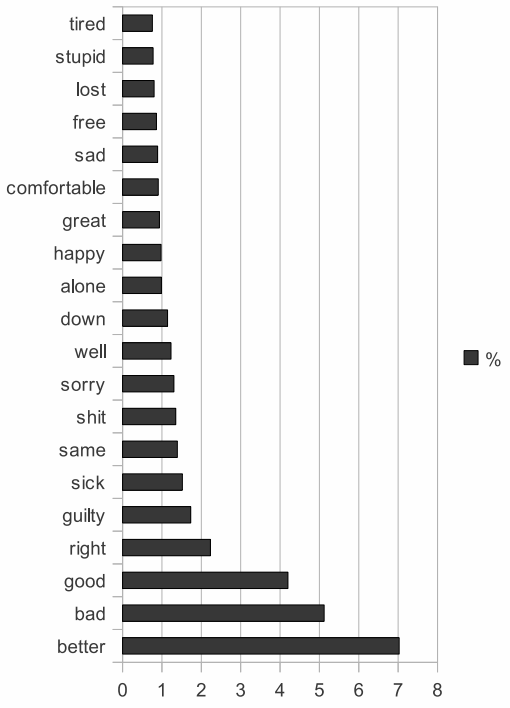

Figure 1: 20 most frequent feelings in WeFeelFine

[1] http://wefeelfine.org

WeFeelFine offers interactive tools to explore the data and relates the feelings with profile information like gender or age, as author details are also extracted from the web (see snapshots at Figure 2 and Figure 3. In our experiments, this information related to author profiles has been discarded, as such data had to be extracted when using other sources[2], although it could represent informative attributes (Schler et al., 2006). WeFeelFine is a very interesting project and its continuous crawling of data could represent a valuable resource in sentiment analysis, as considered by previous studies (Agarwal et al., 2008), where a bag of sentiment words is created using WeFeelFine list of feelings and augmented with synonyms and antonyms from Thesaurus[3].

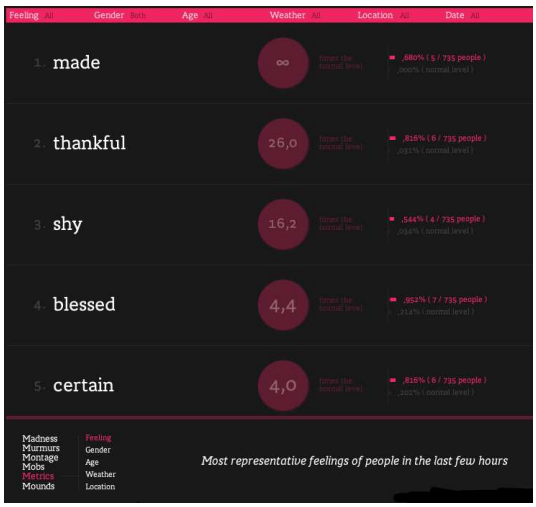

Figure 2: State monitor of feelings in We Feel Fine applet

4 *System architecture*

Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) is an interesting alternative to document modelling. Instead of generating a vector of word statistics, it produces a vector of related Wikipedia articles, being the similarity of the original document to each article the weight for that dimension. Therefore, the document is used as a query against a search engine over the whole Wikipedia, returning a list of articles ranked by their similarity. The novelty here is that the index is generated from data gathered from the blogosphere in a continous flow. So,

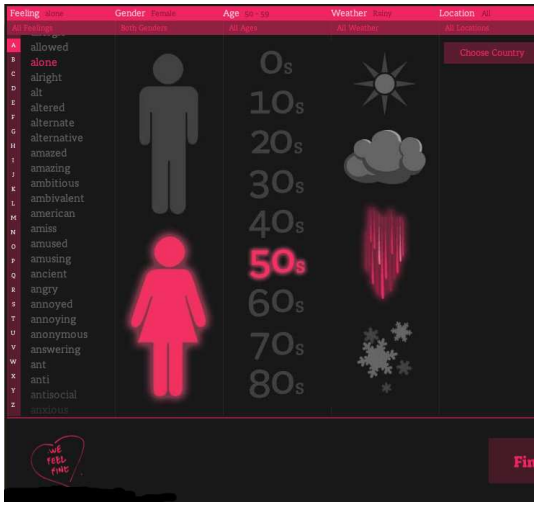

Figure 3: Exploring sentiments by gender, weather, age, country...

we can consider that the proposed approach is a mixed model of Gabrilovich and Turney (Turney, 2002) models.

Our approach is similar in that we represent each tweet by a vector of feelings. A manual labelling of the polarity (with +1 or -1 values) of those 200 feelings (which took just few minutes), is used to compute the final semantic orientation of the tweet by associating to it the list of most related feelings according to a search over the collection built from WeFeelFine data downloaded usign its API. Thus, the system can be splited into two different modules: the indexing module and the classification module.

4.1 Index generation

By means of the WeFeelFine API, we have generated a collection of 200 documents, corresponding to the most frequent feelings in this web, according to the state of its database on 10th October of 2012. Thus, for each feeling, there exists a document containing 1,500 sentences. These documents are indexed, as visualized in the whole process given in Figure 4. For indexing and retrieval the Lucene[4] engine has been used with default configuration (version 3.6.1). For both, sentences to feelings and testing tweets, hashtags and mentions ('#' and '@' strings) have been removed, along with URLs.

4.2 Search

Once the index is generated, we can take a tweet and ask the search engine with it

[2]PAN task on Author Profiling: `http://pan.webis.de`

[3]http://thesaurus.com

[4]http://lucene.apache.org/

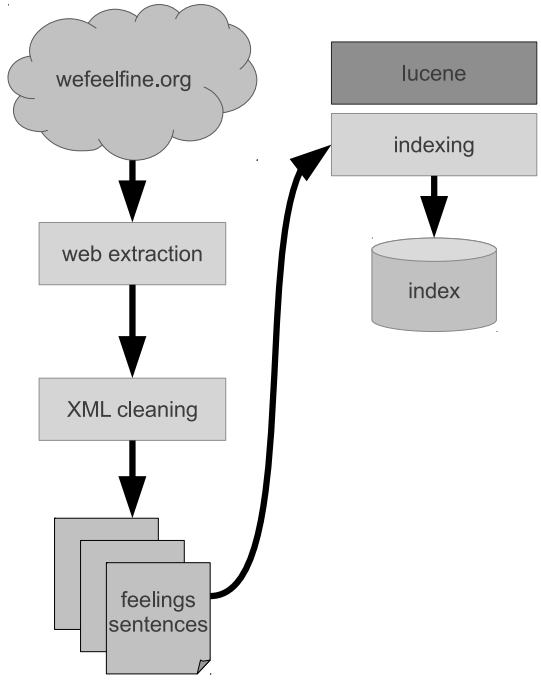

Figure 4: Indexing process

as a query, retrieving the closest feelings, as shown in Figure 5. Finally, from the ranked list of feelings, the final polarity of the tweet is computed based on the polarity value manually assigned to each feelilng. The only parameter that has to be specified is the number of results to be used before averaging, which determines the number of feelings to be taken into account when computing the polarity according to one of the two possible equations defined below.

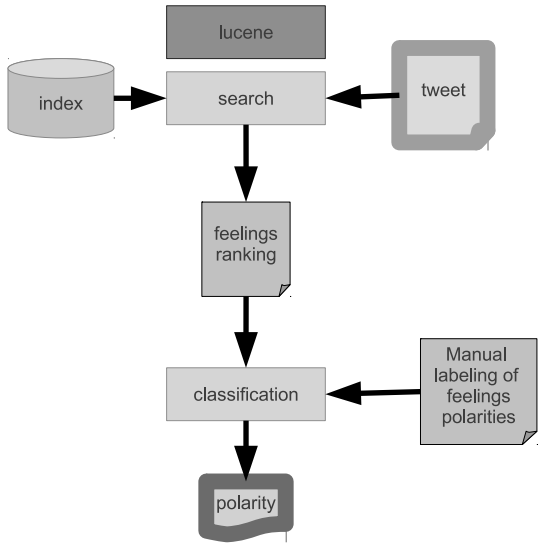

Figure 5: Classification process

As we will see later, the *Ranking Status Value* (RSV) computed by Lucene can also be useful when computing the final polarity. By using this ranking value (which reflects a distance between the tweet and a feeling), we can perform a weighted summatory. Therefore, two possible equations are proposed:

$$p(t) = \frac{1}{|R|} \sum_{r \in R} l_r \tag{1}$$

where
$p(t)$ is the polarity of tweet t
R is the list of retrieved feelings
l_r is the polarity label of feeling r

In the case of considering the RSV, the formula is very similar, but with RSV_r weighting the polarity.

$$p(t) = \frac{1}{|R|} \sum_{r \in R} RSV_r \cdot l_r \tag{2}$$

For example, the tweet "*The Nike Training Club beta iPhone app looks very interesting*" returns the top ten results with the given RSVs shown in Table 1. The table shows how feelings, according to the sentences that made up the representant document, are close to the tweet as query. The scoring formula of Lucene combines cosine and boolean similarities, but in summary is fully based on TF.IDF values, with other factors like document length. We have not changed the default *practical scoring formula* of the engine, altough an adjustment to Twitter nature is foreseen.

Rank	RSV	feeling	polarity
1	0.05166112	cool	+1
2	0.040141936	dumb	-1
3	0.03140159	lucky	+1
4	0.030341815	awesome	+1
5	0.029633064	fine	+1
6	0.029432593	used	-1
7	0.028811168	low	-1
8	0.027871676	missing	-1
9	0.027096074	complete	+1
10	0.026837287	proud	+1

Table 1: Resulting list from a Lucene search

5 *Experiments and results*

To prove this approach, we have taken the Emoticon data set from Stanford University (Go, Bhayani, and Huang, 2009). To enable the comparison of results with other approaches, only the test set is considered. It contains 177 negative tweets and 182 positive

tweets, manually labelled. Therefore, a total of 359 queries have been launched against Lucene, generating a list of results (feelings) for every tweet.

In order to explore the effect in the number of results considered, a range from 1 to 100 top results were taken into account, obtaining corresponding values of precision, recall, F-score and accuracy as performance scores.

5.1 Plain averaging

Results obtained applying Equation 1 are given in Table 2. As can be seen, an impressive F-score of 70.03% is reached when 55 top results are used as feelings to average the final polarity, although the performance of other values near 55 top results are small. Graphically represented in Figure 6, the effect of the number of feelings on the performance is clear. It is visible a constant increase in performance up to 20-30 results. We believe that this is due to the fact that semantic charge of a tweet (even if it is composed by few words) needs of a fined-grain representation under the shape of a list of feelings. Thus, when more feelings are considered, the tweet is modelled more properly. But also, a bit of performance drops beyond that number of 55 top feelings, this can be due to the integration of noise for larger list of results.

# results	accuracy	precision	recall	f-score
1	0.607242	0.677321	0.606584	0.640004
5	0.571031	0.643939	0.569551	0.604465
10	0.584958	0.680557	0.583132	0.628089
15	0.618384	0.700917	0.617030	0.656304
20	0.657382	0.717188	0.655724	0.685080
25	0.665738	**0.725959**	0.664509	0.693876
30	0.679666	0.720909	0.678323	0.698968
35	0.676880	0.723591	0.675421	0.698677
40	0.665738	0.715360	0.664199	0.688831
45	0.662953	0.711189	0.661296	0.685336
50	0.662953	0.714360	0.661374	0.686847
55	**0.682451**	0.721333	**0.680605**	**0.700377**
60	0.660167	0.695295	0.658471	0.676382
65	0.660167	0.699042	0.658704	0.678274
70	0.654596	0.699104	0.653287	0.675419
75	0.665738	0.703054	0.664432	0.683198
80	0.662953	0.698754	0.661995	0.679878
85	0.674095	0.699647	0.673061	0.686097
90	0.674095	0.704569	0.672984	0.688414
95	0.657382	0.697619	0.656267	0.676312
100	0.674095	0.708827	0.672984	0.690441

Table 2: Results obtained with plain averaging

5.2 RSV weighting

These configuration does not take into account the RSV when retrieving the feelings closest to the given tweet. The RSV is a useful measurement of the similarity between

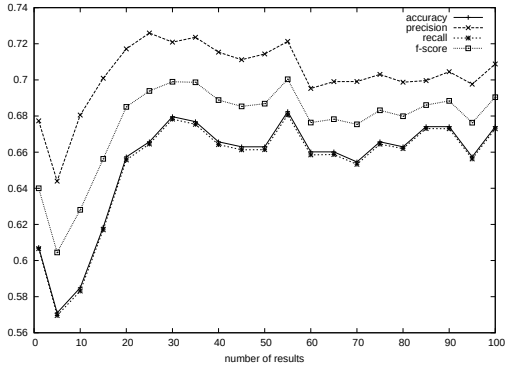

Figure 6: Effect of the number of results on performance for plain averaging

the tweet and the feeling. Thus, using it as weighting value in a linear combination to obtain the final polarity could lead to better performance values. Table 3 shows the results obtained when Equation 2 is applied. Our intuition is confirmed, with a 73% in F-score reached (again, with 55 results). Again this time, as can be observed in Figure 7, the usage of more results leads to better performance scores, with a constant increase up to 20-30 results. Both approaches are visually compared in Figure 8. From this graph we can conclude that applying the RSV as weight on the polarity of associated feelings leads to a better performances independently from the number of results considered.

# results	accuracy	precision	recall	f-score
1	0.607242	0.677321	0.606584	0.640004
5	0.635097	0.644961	0.633669	0.639265
10	0.654596	0.673526	0.652511	0.662852
15	0.662953	0.677771	0.661219	0.669392
20	0.676880	0.695655	0.674955	0.685149
25	0.710306	0.726939	0.708698	0.717702
30	0.701950	0.717462	0.700379	0.708817
35	0.713092	0.729327	0.711523	0.720315
40	0.699164	0.719044	0.697476	0.708096
45	0.713092	0.737645	0.711213	0.724187
50	0.715877	0.734548	0.714348	0.724307
55	**0.718663**	**0.743753**	**0.716785**	**0.730020**
60	0.701950	0.721432	0.700301	0.710710
65	0.693593	0.710775	0.692059	0.701292
70	0.685237	0.702374	0.683662	0.692892
75	0.685237	0.700415	0.683818	0.692017
80	0.696379	0.712209	0.694962	0.703480
85	0.690808	0.705399	0.689467	0.697342
90	0.693593	0.709749	0.692137	0.700832
95	0.690808	0.705399	0.689467	0.697342
100	0.704735	0.721573	0.703281	0.712310

Table 3: Results obtained with RSV weighting

6 Conclusions and further work

Being an unsupervised approach, the results obtained look very promising. Although supervised methods outperforms unsupervised ones, the need of a training corpus is a main drawback in the former approaches. Every

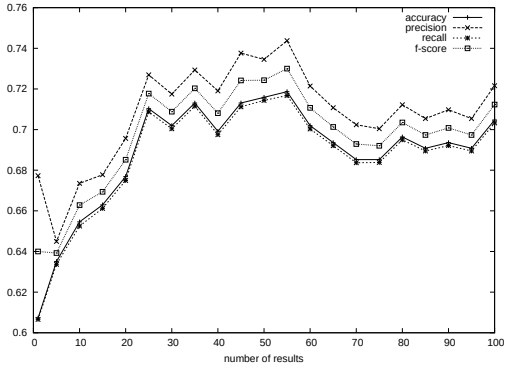

Figure 7: Effect of the number of results on performance for RSV weighting

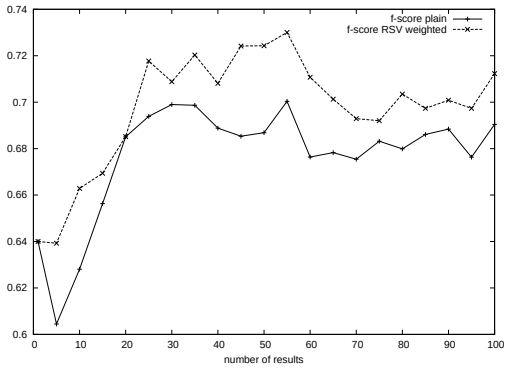

Figure 8: Plain averaging Vs. RSV weighting on F-score

day, millions of tweets flow from their authors to the web, tons of blogs are written and commented and, in many of them, feelings and emotions are expressed. The use of all the huge flow of data to semantically tag the emotions from the same flow of data represents an innovative solution to the attractive problem of sentiment polarity classification. Our experiments open a new way of tackling the problem.

Further experimentation is planned, including applying SentiWordNet (Baccianella, Esuli, and Sebastiani, 2010) as resource to determine the polarity of the crawled feelings without the need of manual intervention. Also, the method should be tested on additional data, like the i-Sieve corpus (Kouloumpis, Wilson, and Moore, 2011). Another open question is how to determine the optimal number of results from Lucene. The behaviour of Lucene RSVs could provide some clue on this issue. The normalization of the final RSVs values will be also studied. Besides, this method only performs a binary classification (positive/negative) and this is insufficient in many scenarios, where neutral or objective lables are also expected, along with a level of "intensity" in polarity values.

Despite the results derived from the experimentation to come, our approach can be easily moved to other languages. Current approaches on Multilingual Sentiment Analysis (Balahur and Turchi, 2012) rely on the traslation of lexicons or resources. In our case, a crawler of emotional publications by means of simple regular expression matching, as is done by WeFeelFine, would allow us to target any other language. This, also, is our intention in the case of Spanish, and the generation of a collection of tweets is undergoing.

References

Agarwal, N., H. Liu, J. Salerno, and S. Sundarajan. 2008. Understanding group interaction in blogosphere: a case study. In *Proc 2nd international conference on computational cultural dynamics (ICCCD), September*, pages 15–16.

Asiaee T., Amir, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1602–1606, New York, NY, USA. ACM.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Balahur, A. and M. Turchi. 2012. Multilingual sentiment analysis using machine translation? *WASSA 2012*, page 52.

Boldrini, Ester, Alexandra Balahur, Patricio Martínez-Barco, and Andrés Montoyo. 2010. Emotiblog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 1–10, Stroudsburg,

PA, USA. Association for Computational Linguistics.

Chen, L., W. Wang, M. Nagarajan, S. Wang, and A.P. Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 50–57.

Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611.

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

Kamvar, Sepandar D. and Jonathan Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 117–126, New York, NY, USA. ACM.

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

Lin, Y., J. Zhang, X. Wang, and A. Zhou. 2012. An information theoretic approach to sentiment polarity classification. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 35–40. ACM.

Martínez-Cámara, E., M.T. Martín-Valdivia, L.A. Ureña López, and A. Montejo-Ráez. In press. Sentiment analysis in twitter. *Natural Language Engineering*.

Pang, B. and L. Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers.

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.

Reyes, A., P. Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*, 74:1–12. cited By (since 1996) 0.

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, March.

Tan, L.K.W., J.C. Na, Y.L. Theng, and K. Chang. 2012. Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *Journal of Computer Science and Technology*, 27(3):650–666.

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, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.