

ELiRF: A Support Vector Machine Approach for Sentiment Analysis Tasks in Twitter at SemEval-2015

Mayte Giménez Ferran Pla Lluís-F. Hurtado

Universitat Politècnica de València

Camí de Vera s/n, 46022 València

{mgimenez, fpla, lhurtado}@dsic.upv.es

Abstract

This paper describes our participation at tasks 10 (sub-task B, Message Polarity Classification) and 11 task (Sentiment Analysis of Figurative Language in Twitter) of Semeval2015. We describe the Support Vector Machine system we used in this competition. We also present the relevant feature set that we take into account in our models. Finally, we show the results we obtained in this competition and some conclusions.

1 Introduction

Nowadays social media, such as Twitter, produce a vast amount of information that lead us to new challenges in Machine Learning (ML) and in Natural Language Processing (NLP) fields.

Twitter¹ is a micro-blogging service, which according to latest statistics, has 284 million active users, 77 % outside the US that generate 500 million tweets a day in 35 different languages. That means 5,700 tweets per second and they had peaks of activity of 43,000 per second. This numbers justify the great interest in the automatic processing of this information.

The study (Analytics, 2009) estimates that 50.9% of tweets have some useful information that are capable of mobilize opinions in Internet and also in the real world. Therefore, social media users opinions have great strategic value for different organizations.

Our work is focused on automatically identify the prevailing sentiment in a tweet using ML and NLP

¹About twitter,inc. <https://about.twitter.com/company>. Accessed: 30-12-2014.

techniques. We developed a system for determining the tweets polarity for 10B and 11 tasks at the SemEval-2015 competition.

The aim of task 10 (subtask B) (Rosenthal et al., 2015) is to classify tweets among positive, negative, and neutral polarity. In task 11 (Ghosh et al., 2015) we had to deal with figurative language, and we should assign a polarity to each tweet with a score that vary in the range [-5..5], this score represents the degree of the sentiment. Due to this last requirement, we formalized this task as a regression problem.

Our approach shared some points for solving both tasks. Preprocessing and feature extraction processes from the corpora were similar. We considered some common problems when we are dealing with text from social media and in particular from Twitter: short texts, slang, peculiarities of the language (*hashtags, retweets, user mentions*, etc.). We represented features extracted using a bag of n-grams. We used Support Vector Machine (SVM) formalism due to the fact to its ability to handle large feature space and to determine the relevant features.

Task 10B has been considered as a classification problem and it has been modeled by means of SVM classifiers. For Task 11 we used regression SVM, due to the granularity of the scores.

Both tasks were solved using a supervised technique. Our systems learned from the training set supplied by the Semeval organization. We also used external resources such as polarity dictionaries.

The rest of this paper is organized as follows. In section 2, we briefly present some relevant works related to these tasks. In section 3, we describe

the main features of the used corpora. In section 4, we present the system we developed to solve these tasks. Section 5 is dedicated to show the results of our experimental work and the results we obtained for the SemEval tasks. Finally, in section 6, we will share some conclusions from our work and possible future directions.

2 Related Work

Sentiment Analysis has been widely studied in the last decade in multiple domains. Most work focuses on classifying the polarity of the texts as positive, negative, mixed, or neutral. The pioneering works in this field used supervised (Pang et al., 2002) or unsupervised (knowledge-based) (Turney, 2002) approaches. In (Pang et al., 2002), the performance of different classifiers on movie reviews was evaluated. In (Turney, 2002), some patterns containing POS information were used to identify subjective sentences in reviews to then estimate their semantic orientation.

In (Pang and Lee, 2008) we can find a comprehensive study of the different techniques used to identify the polarity of a text.

Many efforts have been made to transfer this knowledge to language extracted from social media. In the literature we can find recent attempts to solve this problem using different machine learning approaches such as, SVM, Maximum Entropy, Naive Bayes, etc, (Barbosa and Feng, 2010; O'Connor et al., 2010a; Zhu et al., 2014). At best, these works achieve F1-score close to 70%, therefore we still could improve these proposed systems.

The construction of polarity lexicons is another widely explored field of research. Opinion lexicons have been obtained for English (Liu et al., 2005; Wilson et al., 2005) and also for Spanish (Perez-Rosas et al., 2012). A good presentation of the SA problem and a description of the state-of-the-art of the more relevant approaches to SA can be found in (Liu, 2012).

Research works about SA on Twitter are much more recent. Twitter appeared in the year 2006 and the early works in this field are from 2009 when Twitter started to achieve popularity. Some of the most significant works are (Barbosa and Feng, 2010), (Jansen et al., 2009), and (O'Connor et al.,

2010b). A survey of the most relevant approaches to SA on Twitter can be seen in (Vinodhini and Chandrasekaran, 2012). The SemEval competition has also dedicated specific tasks for SA on Twitter (Wilson et al., 2013; Rosenthal et al., 2014a,b) which shows the great interest of the scientific community in this field.

TASS workshop has proposed different tasks for SA focused on the Spanish language (Villena-Román and García-Morera, 2013) and (Villena-Román et al., 2014). In this paper, we have included some ideas that we have used in previous works in the context of some SA tasks at TASS competition for Spanish (Pla and Hurtado, 2013, 2014b,a)

3 Corpus Description

In the following section, we describe the main features of SemEval2015 corpora used in 10B and 11 tasks, respectively.

3.1 Task 10 B

The corpora supplied by the Semeval2015 organization is composed by 7,236 tweets for training, 1,242 tweets for tuning (development set) and 2,880 tweets for test-time development composed by part of the Semeval2013 corpora used in that edition (Nakov et al., 2013). The test corpora has an official test with 2,390 tweets and a progress test with 8,987 tweets.

Figure 1 plots the polarity distribution over these train, tuning and test-time development corpora. On average, 16.53% of the tweets are negatives, 45.75% are neutrals and 37.72% are positives. Vocabulary from training corpus has 25,973 words, development corpus has 6,700 words and test-time development corpus has 13,672 words after we deleted the *stop-words*. We found that 57.57% of the words from test-time development were never seen in training.

We studied the Zipf's distribution of the words from train, tune and test-time development corpora and we find out that words with less number of synsets, less ambiguity, appear with more frequency. We used this information in the normalization of the SentiWordNet Lexicon.

Since we used lexicons as a features for training our systems, it is important to know the percentage of words from corpus which appear in these lexicons.

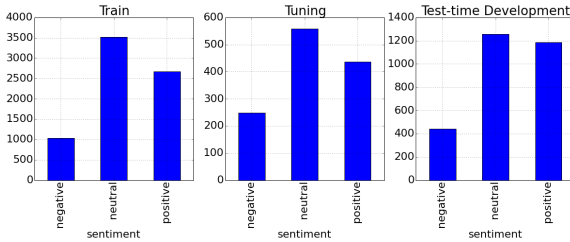


Figure 1: Polarity distribution studied over train, tune (dev) and test-time development corpora in Task 10.

Table 1 highlights how less than 10 % of the vocabulary from the corpus can be found in the lexicons; with the exception of the lexicons *NRC* and *SentiWordNet* (Baccianella et al., 2010) but in this lexicon we have to deal with the semantic ambiguity of the words.

Lexicon	Train	Test
Afinn	3.14 %	3.85 %
Pattern	4.28 %	5.21 %
SentiWordNet	45.21 %	51.26 %
Jeffrey	4.01 %	4.56 %
NRC	29.42 %	33.26 %

Table 1: Percentage of words from task 10’s corpora with polarity using different lexicons in Task 10.

It is noteworthy that only 19.98% of training tweets and 20.31% of tweets from the test-time development set have *hashtags*. Users tag the content of their tweets with *hashtags*, consequently its meaning may be relevant when we try to classify a tweet. However *hashtags* often have multiple words together and segmentation of these words it is a problem in itself.

3.2 Task 11

The Task 11 corpus is similar to previous one, but its main feature is that it contains figurative language such as irony and affective metaphor. This kind of language will increase the complexity of the task. Also this task requires a much more fine grained polarity identification. Two corpora were provided to address this task.

- A trial corpus with 1,000 figurative tweets an-

notated. We were able to retrieve 925 tweets –86.6 % from total–.

- A train corpus with 8,000 tweets, of these we recover 6,928 tweets – 92.5 % from total–.

Trial and a train corpus share some tweets. We had 7,135 unique tweets to train and tune our systems. The corpus has 22,227 words without *stop words*.

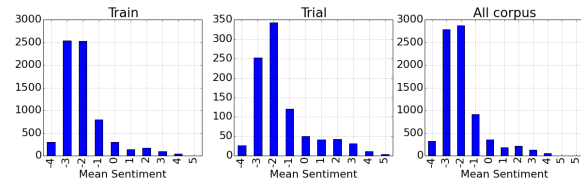


Figure 2: Polarity distribution in the development corpora in Task 11.

Table 2 shows the percentage of words from task 11 corpus we could find in the lexicons. Just like vocabulary from task 10, a small percentage of the vocabulary will have a polarity score.

Lexicon	Corpus
Afinn	5.75 %
Pattern	5.69 %
SentiWordNet	43.23 %
Jeffrey	5.64 %
NRC	38.09 %

Table 2: Percentage of words from task 11’s corpora with polarity using different lexicons in Task 11.

As expected, the corpora for this task has a lot of figurative language. If we assume that Twitter’s users tag semantically its tweets using *hashtags* and tags as #irony or #sarcasm indicates the presence of figurative text then at least 46.22 % of the corpus has figurative text. This was the only knowledge we add to deal with task 11 differently from the knowledge used in task 10. Finally, a remarkable 85.58% of tweets have at least one *hashtag*. Therefore these features will be relevant in our classification system.

4 Our System

In this section we describe the main features of the system developed for SemEval tasks We determined

the baseline for both tasks by selecting the most probable class in the training set. In task 10B we got a 26.49% of F1-score, a 43.61% of precision, and a 43.61 % of recall. In task 11 we got a 19.53% of F1-score, a 36.51 % of precision and a 36.51% of recall.

After studying the corpus, we train and tune different classifiers using features extracted from the text and from the lexicons. We did a 10-cross validation to tune the SVM models.

4.1 Feature Extraction

We selected the best set of features in order to solve each task. The best features considered were:

N-grams We used a *bag-of-words* approach to represent each tweet as a feature vector that contains the tf-idf factors of the selected features of the training set. After tokenizing the tweet and deleting its *stop words* we extract n-grams of characters. We have two approaches: we got all n-grams joining words or just n-grams within words. In task 10 we used 1-grams to 6-grams and we vectorized them using tf-idf coefficients. In task 11 we used the same approach but we used 3-grams to 9-grams.

Negation We need to deal with negation to predict polarity correctly. Thus, we label every word in a negation context. We assume that a *negation context* begins with a negation word as: “never”, “no”, “nothing”, “none”, ..., and ends with a punctuation mark, following the approach of (Pang et al., 2002). We used this strategy only in task 10. After labeling negation context, our system extracted the n-grams from labeled tweets.

Lexicons In order to use lexicons, tweets are tokenized, cleaned the *stop words* and all the tokens are converted to lowercase. We applied five lexicons.

1. *Pattern* (De Smedt and Daelemans, 2012): Given a tweet this lexicon will return a score with the polarity and another one with the objectivity.
2. *Afinn-111* (Hansen et al., 2011): This lexicon has a set of words tagged with a score. We sum the polarity of every word in a tweet to get a score for the whole tweet. $\sum_{w \in W} Afinn(w)$

3. *Jeffrey* (Hu and Liu, 2004): This lexicon has two sets of words: a positive and a negative word set. We got two scores from this lexicon. First score is the count of positive words and the second one is the count of negative words. $\sum_{w \in W} Jeffrey(w)$

4. *NRC* (Mohammad et al., 2013): Likewise, we obtain a score for each tweet adding the polarity of each word from this lexicon. Also we return a score normalized by the length of the tweet. $\frac{1}{|W|} \sum_{w \in W} NRC(w)$

5. *SentiWordNet* (Baccianella et al., 2010): In this lexicon each word could belong to multiple sets of meaning (*Synsets S*), therefore we normalize the score of a word by its number of meanings. This lexicon provides three scores for: positive, negative and objective words, and we used these three scores. $\sum_{w \in W} \frac{1}{|S|} \sum_{s \in S} SentiWordNet(w, s)$

Features from Twitter: We count the number of *hashtags*, *retweets*, mentions and URLs for each tweet.

Some *hashtags* like: #irony, #sarcasm o #not,... are useful in order to identify the presence of figurative text in a tweet. We count the number of these *hashtags* as a feature.

Encoding We consider number of capitalized words and the number of words with elongated characters.

Obviously we tried different set of features like: POS tags, word n-grams, binary bag of words, ... also we tried different combinations of features in order to optimize the system.

4.2 Clasification

We classified tweets using a SVM approach. In task 10B we used a linear kernel for classification and in task 11 we also used a linear kernel for regression. Feature selection process was performed in task 10 using the development corpus and in task 11 using a cross-validation technique (10-fold cross validation) on training set. We selected the set of features that optimized the accuracy of the system

on the development set.

We used *scikit-learn toolkit* (Pedregosa et al., 2011), and we developed a framework to define functional classification models. These models included: preprocess, mining, vectorization features, and classification functions. This framework receive 1 to N models. A tweet is classified using the most voted category or using the mean of predictions if we are doing regression.

5 Experiments

We tested a set of configurations in order to obtain a competitive classifier. In this section, we present only the systems which achieved best performance in development time. We submitted only the best system to the SemEval 2015 competition.

5.1 Task 10B

1. **Model 1:** We used a linear SVM. The set of features considered were:
 - 1-gram to 6-grams of characters from tweet.
 - 1-gram to 6-grams of characters from negation labelled tweet.
 - Lexicons 1, 2, and 5.
 - Features extracted from Twitter
2. **Model 2:** A linear SVM trained using these features:
 - 1-gram to 6-grams of characters from negation labelled tweet.
 - All lexicons described in section 4.1.
 - Features extracted from Twitter
3. **Model 3:** A linear SVM trained using these set of features:
 - 1-gram to 6-grams of characters from tweet.
 - Lexicons 1, 2, and 5.
 - Features extracted from Twitter
4. **Model 4:** We created three SVMs classifiers. Each one of them were trained with this set of features:

- 1-gram to 6-grams of characters from tweet.
- A lexicon. Each SVM has its own lexicon. We used lexicons 1, 2, and 5.

Then we used a majority voting system to combine these classifiers.

Table 3 shows the best systems in development phase. The accuracy is computed globally. Precision and recall are the average of these metrics for each class.

	<i>accuracy</i>	<i>precision</i>	<i>recall</i>
Model 1	0.6899	0.7035	0.6942
Model 2	0.7073	0.7201	0.7024
Model 3	0.6989	0.7146	0.7026
Model 4	0.6920	0.7074	0.6190

	F1	$F1_{neg}$	$F1_{neu}$	$F1_{pos}$
Model 1	0.6826	0.5014	0.7303	0.6994
Model 2	0.7013	0.5365	0.7407	0.7209
Model 3	0.6901	0.4802	0.7391	0.7162
Model 4	0.6816	0.4759	0.7307	0.7060

Table 3: Performance in development phase from our best systems in Task 10B.

For the competition we submitted the model 2 which achieved the best performance in the development phase. Table 4 shows evaluation performance. Forty teams participated in this task. In the official rank our system achieved the 24th position and the 35th position in the progress test.

5.2 Task 11

Our best model for this task was trained using these features:

- 3-grams to 9-grams of characters from tweet.
- Lexicons 1, 2, and 5.
- Features extracted from Twitter including the number of figurative *hashtags*.

We selected this set of features by cross validation. We tuned our system using the official measure, the cosine distance.

		F1	Rank	Best	Worst
Official Test	Twitter 2015	58.58	24	64.84	24.80
Progress Test	LiveJournal 2014	68.33	28	75.34	34.06
	SMS 2013	60.20	28	68.49	26.14
	Twitter 2013	57.05	32	93.62	32.14
	Twitter 2014	61.17	35	74.42	32.2
	Twitter 2014 sarcasm	45.98	24	59.11	35.58

Table 4: Evaluation results in Task 10B.

	Cosine	Rank	Best	Worst
Overall	0.6579	5	0.758	0.059
Sarcasm	0.904	1	0.904	0.412
Irony	0.905	4	0.918	-0.209
Metaphor	0.411	5	0.655	-0.023
Other	0.247	8	0.584	-0.025

Table 5: Official evaluation results in Task 11.

	MSE	Rank	Best	Worst
Overall	3.096	8	2.117	6.785
Sarcasm	1.349	9	0.934	4.375
Irony	1.034	8	0.671	7.609
Metaphor	4.565	4	3.155	9.219
Other	5.235	5	3.411	12.16

Table 6: MSE evaluation results in Task 11.

Table 5 shows the official results of our system in task 11. We achieved the 5th position in the rank. Our system obtained the first position in detecting sarcasm. We achieved a 0.918 of cosine similarity measure. For non figurative language, our system performed worse, obtaining the 8th position in the rank. We think this is due to the fact that training corpus lacks of non-figurative tweets, therefore our system was not able to learn this class properly.

Mean square error metric (MSE) was also considered by Task 11 organizers. Table 6 shows the results achieved using this metric. We obtained worse results because we didn't tune the system for this metric.

6 Conclusions

We have presented a system for 10B and 11 tasks at SemEval 2015. We used a machine learning

approach based on SVM formalism for both tasks. We handled both tasks uniformly with regard to the preprocessing, feature extraction and feature representation. We have not included any knowledge about the tasks, except from resources used, that is, corpora and dictionaries. In this respect, our system will be easy to adapt to other SA tasks and other languages with this kinds of resources.

Even we did not include any external knowledge we plan to study the impact of including external resources to improve our system. Moreover, we also find interesting to extend existing corpora based on Twitter in order to increase the accuracy of the machine learning system.

Acknowledgments

This work has been partially funded by the projects, DIANA: DIScourse ANALYSIS for knowledge understanding (MEC TIN2012-38603-C02-01) and Tímpano: Technology for complex Human-Machine conversational interaction with dynamic learning (MEC TIN2011-28169-C05-01).

References

- Pear Analytics. Twitter study–august 2009. 2009.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC*, 2010.
- Luciano Barbosa and Junlan Feng. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

- Tom De Smedt and Walter Daelemans. "vreselijik mooii!"(terribly beautiful): A subjectivity lexicon for dutch adjectives. In *LREC*, pages 3568–3572, 2012.
- A. Ghosh, G. Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015)*, June 2015.
- Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news-affect and virality in Twitter. In *Future information technology*, pages 34–43. 2011.
- Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
- Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.
- Bing Liu. *Sentiment Analysis and Opinion Mining. A Comprehensive Introduction and Survey*. 2012.
- Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web, WWW '05*, pages 342–351, New York, NY, USA, 2005. ISBN 1-59593-046-9.
- Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.
- Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 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 2013.
- Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010a.
- Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for Twitter. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010b.
- Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 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 2012. ISBN 978-2-9517408-7-7.
- Ferran Pla and Lluís-F Hurtado. ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter. In *XXIX Congreso de la Sociedad Española para*

- el Procesamiento del Lenguaje Natural (SEPLN 2013)*. TASS, pages 220–227, 2013.
- Ferran Pla and Lluís-F. Hurtado. Political tendency identification in Twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 183–192, Dublin, Ireland, August 2014a.
- Ferran Pla and Lluís-F. Hurtado. Sentiment analysis in Twitter for spanish. In *Natural Language Processing and Information Systems*, volume 8455 of *Lecture Notes in Computer Science*, pages 208–213. 2014b. ISBN 978-3-319-07982-0.
- Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in Twitter. *Proc. SemEval*, 2014a.
- Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August 2014b.
- Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '2015*, Denver, Colorado, June 2015.
- Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424, 2002.
- Julio Villena-Román and Janine García-Morera. Workshop on sentiment analysis at sepln 2013: An over view. In *Proceedings of the TASS workshop at SEPLN 2013*. IV Congreso Español de Informática, 2013.
- Julio Villena-Román, Miguel Angel Garcia Cumberas, Janine García-Morera, Eugenio Martínez Cámara, César de Pablo Sánchez, Alfonso Ureña López, and Maria Teresa Martín Valdivia. Tass2014-workshop on sentiment analysis at sepln-overview. In *Proceedings of the TASS workshop at SEPLN 2014*. IV Congreso Español de Informática, 2014.
- G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: A survey. *International Journal*, 2(6), 2012.
- Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.
- Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. Semeval-2013 task 2: Sentiment analysis in Twitter. *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, 13, 2013.
- Xiaodan Zhu, Svetlana Kiritchenko, and Saif M Mohammad. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. *SemEval 2014*, page 443, 2014.