SINAI: Voting System for Twitter Sentiment Analysis

Eugenio Martínez-Cámara, Salud María Jiménez-Zafra,
M. Teresa Martín-Valdivia, L. Alfonso Ureña-López
SINAI Research Group
University of Jaén
E-23071, Jaén (Spain)
{emcamara, sjzafra, maite, laurena}@ujaen.es

Abstract

This article presents the participation of the SINAI research group in the task Sentiment Analysis in Twitter of the SemEval Workshop. Our proposal consists of a voting system of three polarity classifiers which follow a lexicon-based approach.

1 Introduction

Opinion Mining (OM) or Sentiment Analysis (SA) is the task focuses on the computational treatment of opinion, sentiment and subjectivity in texts (Pang and Lee, 2008). Currently, OM is a trendy task in the field of Natural Language Processing due mainly to the fact of the growing interest in the knowledge of the opinion of people from different sectors of the society.

The interest in the research community for the extraction of the sentiment in Twitter posts is reflected in the organization of several workshops with the aim of promoting the research in this task. Two are the most relevant, the first is the task Sentiment Analysis in Twitter celebrated within the SemEval workshop whose first edition was in 2013 (Nakov et al., 2013). The second is the workshop TASS[1] , which is a workshop for promoting the research in sentiment analysis in Spanish in Twitter. The first edition of the workshop took place in 2012 (Villena-Román et al., 2013).

The 2014 edition of the task Sentiment Analysis in Twitter proposes a first subtask, which has as challenge the sentiment classification at entity level, and a second subtask that consists of the polarity classification at document or tweet level. The training corpus is the same than the former edition, but this year the test corpus is considerably bigger than the prior one. A wider description of the task and the corpus can be read in (Rosenthal et al., 2014).

We present an unsupervised polarity classification system for the subtask B of the task Sentiment Analysis in Twitter. The system is based on a voting strategy of three lexicon-based sentiment classifiers. The sentiment analysis research community broadly knows the lexicons selected. They are, SentiWordNet (Baccianella et al., 2010), the lexicon developed by Hu and Liu (Hu and Liu, 2004) and the MPQA lexicon (Wilson et al., 2005).

The rest of the paper is organized as follows. The following section focuses on the description of the different sentiment resources used for developing the sentiment classifiers. The subsequent section outlines the system proposed for the 2014 edition of the task. The last section exposes the analysis of the results reached this year.

2 Sentiment lexical resources

Sentiment lexicons are lexical resources composed of opinion-bearing words and some of them also of sentiment phrases of idioms. Most of the sentiment lexicons are formed by a list of words without any additional information.

A sentiment classifier based on list of opinion-bearing words usually consists of finding out the words of the list in a given document. This method can be considered very simple for the complexity of OM, but it has reached acceptable results in different domains and also is applied in real systems like Trayt.com[2].

Our experience in the field of SA allows us to assert that sentiment lexicons can be divided depending on the information linked to each word,

[1] http://www.daedalus.es/TASS

[2] Trayt.com is a search engine of reviews of restaurants. The polarity classifier of Tryt.com is a lexicon-based system which uses the opinion list compiled by Bing Liu.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 572–577,
Dublin, Ireland, August 23-24, 2014.

so three groups can be found:

- List of opinion-bearing words: These lexicons are usually two lists of polar words, one of them of positive words and another one of negative terms. Some examples of this kind of sentiment lexicons are for English the one compiled by (Hu and Liu, 2004), and for Spanish, the iSOL lexicon (Molina-González et al., 2013).

- List of opinion-bearing words with syntactic information: As it is wider known, OM is a domain-dependent task and can be also said that a context-dependent task. Thus, some lexicons add syntactic information with the aim of offering some information for disambiguating the term, and also provide a different orientation of the word depending on its POS-tags. One example of this kind of lexicon is MPQA subjectivity lexicon (Wilson et al., 2005).

- Knowledge base sentiment lexicons: These lexicons usually indicate the semantic orientation of the different senses of each word, whereas the previous lexicons only indicate the polarity of each word. Also, it is very common that in the knowledge base sentiment lexicons each sense is linked to the likelihood of being positive, negative and neutral. One example of this kind of polar lexicon is SentiWordNet (Baccianella et al., 2010).

In the polarity classifier developed for the workshop a lexicon of each type has been utilised. The sentiment linguistic resources used has been:

- Sentiment lexicon compiled by Bing Liu: The lexicon was used the first time in (Hu and Liu, 2004). Since then, the authors have been updating the list, and currently the list is formed by 2006 positive words and 4783 negative words. Also, the lexicon includes some misspellings with the aim of better representing the language used in the Internet.

- MPQA Subjectivity lexicon (Wilson et al., 2005): The lexicon is formed by over 8000 subjectivity clues. Subjectivity clues are words and phrases that have subjective usages. The lexicon was developed joining words compiled by the authors and with words taken from General Inquirer. Each word is linked with its grade of subjectivity, with its part of speech tag and with its semantic orientation. Due to the fact that each word has its POS-tag there are some words that depending on its POS have a different semantic orientation.

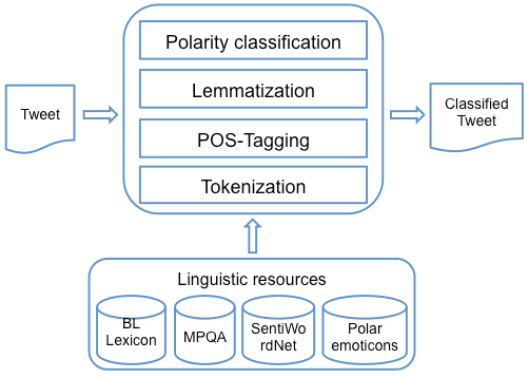

Figure 1: Architecture of the system.

- SentiWordNet 3.0 (Baccianella et al., 2010): is a lexical resource which assigns three sentiment scores to each synset of WordNet: positivity, negativity and objectivity.

3 Polarity classification

We wanted to take advantage from our experience in meta-classification in OM for the 2014 edition of the task, Sentiment Analysis in Twitter. We have reached good results in OM using meta-classifiers in different domains (Perea-Ortega et al., 2013) and (Martín-Valdivia et al., 2013). Therefore, we propose a voting system that combines three polarity classifiers. The general architecture of the system is shown in Figure 1.

Tokenization is a common step of the three classifiers. Due to the specific characteristics of the language used in Twitter, a specific tokenizer for Twitter was preferred to use. The tokenizer published by Christopher Potts[3] was selected and updated, with the aim of recognizing a wider range of tokens.

When the tweet is tokenized, the following step is discover its polarity. Each of the three polarity classifiers follows the same strategy for the classification, but they perform different operations on each tweet. The classifier based on the lexicon compiled by Bing Liu (C_BingL) consists of seeking each token in the opinion-bearing words

[3] http://sentiment.christopherpotts.net/tokenizing.html

list. Therefore, after the tokenization, any linguistic operation has to be performed on the tweet. This classifier classifies a tweet as positive if the number of positive tokens is greater or equal than the number of negative tokens. If there are not polar tokens, the polarity of the tweet is neutral.

The second polarity classifier is the based on MPQA lexicon (C_MPQA). Some of the words that are in the MPQA lexicon are lemmatized, and also the sentiment depends on their POS-tag. Thus, to take advantage of all the information offered by MPQA is needed to perform a morphological analysis to each tweet. The morphological analysis firstly identifies the POS-tag of each token of the tweet, and then the lemmatizer extracts the lemma of the token.

Recently, some linguistic tools have been published to carry out linguistic analysis in tweets. Currently, two POS-taggers for Twitter are available. One of them, is the described in (Gimpel et al., 2011) and the second one in (Derczynski et al., 2013). Although the authors of the two systems are competing for which of the two taggers are better, our selection was based on the usability of the two systems. To use the tagger developed by Gimpel et al. is needed to download their software, meanwhile the one developed by Derczynski et al. can be integrated in other taggers. On our point of view, the tagger of Derczynski et al. has the advantage of offering the training model of the tagger[4], which allows us to integrate it in other POS-tagging tools. The training model of the tagger was integrated in the Stanford Part-of-Speech Tagger[5]. When each token of the tweet is associated with its corresponding POS-tag, the lemmatizer is run over the tweet. The lemmatizer used is the offered by the toolkit for Natural Language Processing, NLTK (Bird et al., 2009). When each token is accompanied by its corresponding POS-tag and lemma, the polarity classifier can seek each token in the MPQA subjective lexicon.

Besides the label of the polar class (positive or negative), each entry in the MPQA corpus has a field called type, which indicates whether the term is considered strongly subjective or the term is considered weakly subjective. Thus, in the calculation of the polarity score these two levels of subjectivity are considered, so when the term is strong subjective it is considered to have a score of 1, and when the term is weak subjective the system considers the term as less important and its score is 0.75.

The polarity classifier based on the use of SentiWordNet (C_SWN) needs that each word of the tweet is linked with its POS-tag and its lemma, so the same pipeline that the classifier based on MPQA follows is also followed by the classifier based on SentiWordNet.

In the bibliography about OM can be found different ways to calculate the polarity class when SentiWordNet is used as a sentiment knowledge base. Some works perform a disambiguation method with the aim of selecting only the synset that corresponds with the sense of the word in the context of the given document. But there are other works that do not perform any disambiguation method, and also reach good results. Denecke in (Denecke, 2008) describes a very simple method to calculate the polarity of each of the words of a document without the need of a disambiguation process. The method consists of calculating per each word in the document, which is in SentiWordNet, the arithmetic mean of the positive, negative and neutral score of each of the synsets that the word has in SentiWordNet. When the scores of each word are calculated, the score of the document is determined as the arithmetic mean of each score of the words. The class of the document is corresponded with the greatest polar score (positive, negative, neutral). Due to the acceptable results that the Denecke formula reaches, we have introduced a soft disambiguation process with the aim of improving the classification accuracy. This soft disambiguation process consist of only taking those synsets corresponding with POS-tag of the word whose polarity are being calculated. For example, the word "good" can do the function of an adverb, a noun or an adjective. In SentiWordNet, there are two synsets of "good" as an adverb, four synset of "good" as a noun, and twenty-one synsets as an adjective. If the polarity score is calculated with the Denecke formula, the twenty-seven synsets are used. Meanwhile, if it used our proposal, and the word "good" in the given sentence is acting as an adverb, then only the two synsets of the word "good" when it is adverb are considered to calculate the polarity score.

During the development of the system, we noticed that synsets have a lower probability to be positive or negative, and most of them in Senti-

[4]https://gate.ac.uk/wiki/twitter-postagger.html
[5]http://nlp.stanford.edu/software/tagger.shtml

WordNet are neutral. With the aim of boosting the likelihood to be positive or negative, the polarity classifier does not consider the neutral score of the synset. If the positive score is greater than the negative score and greater than 0.15 then the term is positive. If the negative score is greater than the positive score and greater than 0.15 then the word is negative, in other case the word is neutral.

Each of the polarity classifiers take into consideration the presence of emoticons, the expressions of laughing and negation. The emoticons are processed as words, so for determining their polarity a sentiment lexicon of emoticons was built. The polar lexicon of emoticons consists of fifty-eight positive emoticons and forty-four negative ones. Laughing expressions usually express a positive sentiment, so when a laughing expression is detected the counter of positive words is increased by one. The strategy for negation identification is a bit straightforward but effective. Due to the specific linguistic characteristics of tweets, a strategy based on windows of words has been implemented. When a polar word is identified, it is sought in the previous three words whether there is a negative particle. In those cases that a negative particle is found, the polarity of the sentiment word is reversed, that it to say if a positive (negative) word is negated the system considers it as negative (positive).

The last step of the polarity classifier is the running of a voting system among the three polarity classifiers. Three are the possible output values of the three base classifiers {negative, neutral, positive}. When the majority class is positive, the tweet is classified as positive, when the majority class is negative then negative is the class assigned to the tweet and when majority class is neutral or there is not a majority class then the tweet is classified as neutral.

4 Analysis of the results

Before showing the results reached in the evaluation of the task, the results accomplished in the development phase of the system will be shown. Three main systems were assessed during the development phase:

- Baseline (BL): The three base classifiers compose the baseline system, but the three polarity scores of SentiWordNet are considered and negation is not taken into account.
- Neutral scores are not considered (NN): It is the same than the Baseline system but the neutral scores of SentiWordNet are not considered.
- Negation identification (NI): The neutral scores of SentiWordNet are not taken into account and the negation is identified.

The results are show in Table 1.

	Precision	Recall	F1	Accuracy	Improve (Acc.)
BL	55.85%	52.02%	53.87%	60.32%	-
NN	56.03%	52.27%	54.09%	60.46%	0.23%
NI	57.22%	53.41%	55.25%	61.12%	1.33%

Table 1: Results achieved during the developing phase.

As can be seen in Table 1 the systems (NN) and (NI) reach better results than the baseline, so all the modifications to the baseline are good for the polarity classification process. The results confirm our hypothesis that the neutral score of the synsets in SentiWordNet are not contributing positively to the sentiment classification. Also, a straightforward strategy for identifying the scope of the negation improves the accuracy of the classification. The results help us to choose the final configuration of the system. As is described in the former section the final polarity classification system follows a voting scheme of three base lexicon-based polarity classifiers. The three base classifiers take into consideration the presence of emoticons, laughing expressions, identifies the scope of negation, and the classifier based on SentiWordNet does not take into consideration the neutral score of the synsets.

The edition 2014 of the task Sentiment Analysis in Twitter has assessed the systems with five different corpus tests: LiveJournal2014, SMS2013, Twitter2013, Twitter2014, Twitter2014Sarcasm. The results reached with each of the test corpus are shown in Table 2.

Some of the results shown in Table 2 are much closed to the results reached during the development phase, because all of the F1 scores are closed to 55%. The lower results have been reached with the corpus Twitter2014 and Twitter2014Sarcasm. The poor results in Twitter2014Sarcasm are due to the lack of a module in the system for the detection of sarcasm. A sarcastic sentence is usually a sentence with a sentiment that expresses the opposite

		Precision	Recall	F1
LiveJournal2014	Positive	60.19%	76.95%	67.54%
	Negative	36,51%	75,00%	49.12%
	Neutral	82.48%	51.36%	63.31%
	Overall	–	–	58.33%
SMS2013	Positive	63.01%	60.19%	61.57%
	Negative	42.13%	71.86%	53.12%
	Neutral	82.27%	73.72%	77,76%
	Overall	–	–	57.34%
Twitter2013	Positive	60.56%	70.15%	65.01%
	Negative	28.29%	50.15%	36.17%
	Neutral	73.66%	57.06%	64.31%
	Overall	–	–	50.59%
Twitter2014	Positive	57.13%	77.49%	65.77%
	Negative	27.23%	42.64%	33.23%
	Neutral	73.54%	49.20%	58.96%
	Overall	–	–	49.50%
Twitter2014Sarcasm	Positive	57.58%	48.72%	52.78%
	Negative	5.00%	100.00%	9.52%
	Neutral	84.62%	24.44%	37.93%
	Overall	–	–	31.15%

Table 2: Results reached with the test corpus.

		F1		
		C_BingL	C_SWN	C_MPQA
LiveJournal2014	Positive	68.11%	42.62%	65.20%
	Negative	55.43%	39.81%	49.60%
	Neutral	64.03%	58.07%	58.43%
	Overall	61.77%	41.21%	57.40%
SMS2013	Positive	61.67%	43.53%	53.56%
	Negative	54.19%	28.79%	52.678%
	Neutral	76.00%	75.85%	68.38%
	Overall	57.93%	36.16%	53.12%
Twitter2013	Positive	68.30%	23.40%	62.37%
	Negative	46.20%	11.60%	37.75%
	Neutral	61.17%	62.11%	57.39%
	Overall	57.25%	17.50%	50.06%
Twitter2014	Positive	69.33%	22.17%	66.74%
	Negative	41.55%	9.79%	33.00%
	Neutral	53.25%	55.63%	52.76%
	Overall	55.44%	15.98%	49.87%
Twitter2014Sarcasm	Positive	56.10%	27.27%	52.06%
	Negative	17.78%	9.52%	8.51%
	Neutral	44.44%	30.24%	30.77%
	Overall	36.94%	18.40%	30.28%

Table 3: Results reached by each base classifier with the test corpus.

sentiment, so a polarity classifier without a specific module to treat this linguistic phenomenon will be probably misclassified the sarcastic sentences. The results for Twitter2014Sarcasm for the negative class indicate this problem. The low value of the precision and the high value of the recall in the negative class mean that a high number of negative sentences have been classified as positive.

The analysis of the results is completed with the assessment of our method. We proceed from the hypothesis that a combination of several classifiers will improve the final classification. Our hypothesis is based on own previous publications, (Perea-Ortega et al., 2013) and (Martín-Valdivia et al., 2013). We have classified the test corpus with each of the three base classifiers, with the aim of knowing the performance of each one. The results are shown in Table 3.

Table 3 shows that the classifier C_BingL reaches better results than the combination of the three classifiers. The first conclusion we draw from this fact is that the good performance of meta-classifiers with large opinions is not achieved with the short texts of Twitter. But, this conclusion is preliminary, because the lower results of the voting system may be due to a not good combination of the three classifiers. So we have to continue working in the analysis on how to build a meta-classifier for OM in Twitter. The rest of the classifiers reached lower results than the voting system. Another reason that the voting system achieved lower results than C_BingL may be because the three classifiers are not heterogeneous, in other words, when one of the systems misclassified a document the other ones classify it correctly, so the base classifiers help each other, and the combination of systems reaches better results than the individual systems. But, in our case may be that the systems are not heterogeneous, so our ongoing work is the study of the heterogeneity between the three classifiers.

If we focus only in the results achieved by C_BingL, it is remarkable that the higher difference is in the negative class. C_BingL reaches greater results than the voting system in negative class, and it has the same negation treatment module that the voting system. This fact allow us to say that the low results in the negative class reached by the voting system is not due to the negation treatment module, and may because by the own combination method.

To sum up, after analysing the results, we have noticed that the same meta-classifier methodology that we usually apply to large reviews cannot be directly apply to tweets. Therefore, our ongoing work is focused firstly on conducting a deep analysis of the results presented in this work, and secondly in the study on how to improve of polarity classification in Twitter following a unsupervised methodology, and thirdly on how to build a good meta-classifier for OM in Twitter.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional

(FEDER), ATTOS project (TIN2012-38536-C03-0) from the Spanish Government, AORESCU project (P11-TIC-7684 MO) from the regional government of Junta de Andalucía and CEATIC-2013-01 project from the University of Jaén.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Kerstin Denecke. 2008. Using SentiWordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512, April.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. ACL.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

María-Teresa Martín-Valdivia, Eugenio Martínez-Cámara, Jose-M. Perea-Ortega, and L. Alfonso Ureña López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Syst. Appl.*, 40(10):3934–3942, August.

M. Dolores Molina-González, Eugenio Martínez-Cámara, Maria Teresa Martín-Valdivia, and José M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Syst. Appl.*, 40(18):7250–7257.

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

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

José M. Perea-Ortega, M. Teresa Martín-Valdivia, L. Alfonso Ureña López, and Eugenio Martínez-Cámara. 2013. Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. *Journal of the American Society for Information Science and Technology*, 64(9):1864–1877.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. 2013. TASS - Workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. ACL.