ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm*

Delia Irazú Hernández Farías
Universitat Politècnica de València
Pattern Recognition and Human Language Technology
dhernandez1@dsic.upv.es

Emilio Sulis, Viviana Patti, Giancarlo Ruffo, Cristina Bosco
University of Turin
Dipartimento di Informatica
{sulis,patti,ruffo,bosco}@di.unito.it

Abstract

This paper describes the system used by the ValenTo team in the Task 11, Sentiment Analysis of Figurative Language in Twitter, at SemEval 2015. Our system used a regression model and additional external resources to assign polarity values. A distinctive feature of our approach is that we used not only word-sentiment lexicons providing polarity annotations, but also novel resources for dealing with emotions and psycholinguistic information. These are important aspects to tackle in figurative language such as irony and sarcasm, which were represented in the dataset. The system also exploited novel and standard structural features of tweets. Considering the different kinds of figurative language in the dataset our submission obtained good results in recognizing sentiment polarity in both ironic and sarcastic tweets.

1 Introduction

Figurative language, which is extensively exploited in social media texts, is very challenging for both traditional NLP techniques and sentiment analysis, which has been defined as "the computational study of opinions, sentiments and emotions expressed in text" (Liu, 2010). There is a considerable amount of works related to sentiment analysis and opinion mining (Pang and Lee, 2008; Liu, 2010; Cambria et al., 2013). In particular, the linguistic analysis of social media (microblogging like Twitter especially) has become a relevant topic of research in different languages (Rosenthal et al., 2014; Basile et al., 2014) and several frameworks for detecting sentiments and opinions in social media have been developed for different application purposes.

In a sentiment analysis setting, the presence in a text of figurative language devices, such as for instance irony, can work as an unexpected polarity reverser, by undermining the accuracy of the systems (Bosco et al., 2013). Therefore, several efforts have been recently devoted to detect and tackle figurative language phenomena in social media, following a variety of computational approaches, mostly focussing on irony detection and sarcasm recognition (Davidov et al., 2010; González-Ibáñez et al., 2011; Riloff et al., 2013) as classification tasks. Buschmeier et al. present an analysis of features, previously applied in irony detection, in a dataset from a product reviews corpus from Amazon (Buschmeier et al., 2014). Veale and Hao present a linguistic approach to separate ironic from non-ironic expressions in figurative comparisons over a corpus of web-harvested similes (Veale and Hao, 2010). Concerning Twitter, the problem of irony detection is addressed in (Reyes et al., 2013), where a set of textual features is used to recognize irony at a linguistic level. In (Riloff et al., 2013) the focus is on identifying sarcastic tweets that express a positive sentiment towards a negative situation. A model to classify sarcastic tweets using a set of lexical features is presented in (Barbieri et al., 2014). Moreover, a recent analysis on the interplay between sarcasm detection and sentiment analysis is in (May-

* The National Council for Science and Technology (CONACyT-Mexico) has funded the research work of the first author (218109/313683 grant).

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 694–698,
Denver, Colorado, June 4-5, 2015.

nard and Greenwood, 2014), where a set of rules has been proposed to improve the performance of the sentiment analysis in presence of sarcastic tweets.

In this paper we describe our participation to the *SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter* (Ghosh et al., 2015). The task concerned with classification of tweets containing different kinds of figurative language, in particular irony, sarcasm and metaphors. ValenTo system used a linear regression model, exploiting novel and standard structural and lexical features of tweets. Considering the different kinds of figurative language in the Semeval dataset - sarcasm, irony and metaphors - our submission had good results in recognizing sentiment polarity in both ironic and sarcastic tweets, than in the other cases.

2 Our System

We propose a supervised approach that consists in assigning a polarity value to tweets by using a linear regression model constructed from an annotated dataset. In order to catch characteristics that allow us to measure the polarity value in each tweet, we considered a set of features described below.

2.1 Feature Description

2.1.1 Structural Features

Among the several structural characteristics of tweets, in our study we consider: the length of tweets in amount of words (*lengthWords*); the length of a tweet as the number of characters that composes the textual message (*lengthChar*); the frequency of commas, semicolons, colons, exclamation and question marks (*punctuation marks*); the frequency of some Part of Speech categories as nouns, adverbs, verbs and adjectives (*POS*); the frequency of uppercase letters in each tweet *upperFreq*; the frequency or presence of URL *urlFreq*; and the amount of emoticons used in order to express some kind of emotion, we consider both positive (*emotPosFreq*) and negative ones (*emotNegFreq*).

We also consider some features that belongs to tweets, like: the presence or absence of hashtags (*hashtagBinary*) and mentions (*mentionsBinary*); the amount of hashtags (*hashtagFreq*) and mentions (*mentionsFreq*) in each tweet; and if the tweet is a retweet (*isRetweet*). Finally, we decide to take into account a feature (*polReversal*) in order to reverse the polarity (positive to negative, and vice versa) if a tweet includes the hashtag #sarcasm or #not.

2.1.2 Lexical Resources

In order to take into account sentiments, emotions and psycholinguistic features, and to count their frequency, we use the following lexical resources:

AFINN: it is a dictionary of $2,477$ English manually labeled words collected by Nielsen (Nielsen, 2011). Polarity values varies from -5 up to $+5$ [1].

ANEW: the Affective Norms for English Words provides a set of emotional ratings for a large number of English words (Bradley and Lang, 1999). Each word in is rated from 1 to 9 in terms of the three dimensions of Valence, Arousal and Dominance.

DAL: the Dictionary of Affective Language developed by Whissell (Whissell, 2009) contains $8,742$ English words rated in a three-point scale[2]. Each word is rated into the dimensions of Pleasantness, Activation and Imagery.

HL: Hu–Liu's lexicon (Hu and Liu, 2004) includes about $6,800$ positive and negative words[3].

GI: General Inquirer (Stone and Hunt, 1963) contains categories and subcategories for content analysis with dictionaries based on the Lasswell and Harvard IV-4 [4].

SWN: SentiWordNet (Baccianella et al., 2010) is a lexical resource for opinion mining and consists in three sentiment scores: positive, negative and objective[5]. We take into account the first two categories.

SN: SenticNet is a semantic resource for concept-level sentiment analysis (Cambria et al., 2012). We take into account the values of each one of the five dimensions (*senticnetDimensions*) provided by the lexical resource: Pleasantness (Pl), Attention (At), Sensitivity (Sn) and Aptitude (Ap) and Polarity (Pol); and also the polarity value p obtained by using the formula (*senticnetFormula*) below based

[1] https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

[2] ftp://perceptmx.com/wdalman.pdf

[3] http://www.cs.uic.edu/~liub/FBS/

[4] http://www.wjh.harvard.edu/~inquirer/homecat.htm.We are mostly interested in the positive and negative words.

[5] http://sentiwordnet.isti.cnr.it/download.php

on a combination of the first four dimensions:

$$p = \sum_{i=1}^{n} \frac{Pl(c_i) + |At(c_i)| - |Sn(c_i)| + Ap(c_i)}{3N}$$

where c_i is an input concept, N the total number of concepts which compose the tweet, and 3 a normalisation factor.

LIWC: Linguistic Inquiry and Word Counts dictionary[6] contains 127,149 words distributed in categories that can further be used to analyze psycholinguistic features in texts. We select two categories for positive and negative emotions: PosEmo (12,878) entries and NegEmo (15,115 entries).

NRC: in the NRC word-emotion association lexicon (Mohammad and Turney, 2013) each word is labeled according to the Plutchik's primary emotions.

3 Results

3.1 Task Description and Dataset

The goal of the Task 11 at SemEval 2015, is the following: *given a set of tweets rich w.r.t. the presence of such figurative devices, to determine for each message whether the user expressed positive, negative or neutral sentiment, and the sentiment degree.* To have a measure of the sentiment intensity expressed in the message, it was proposed a fine-grained 11-point sentiment polarity scale.

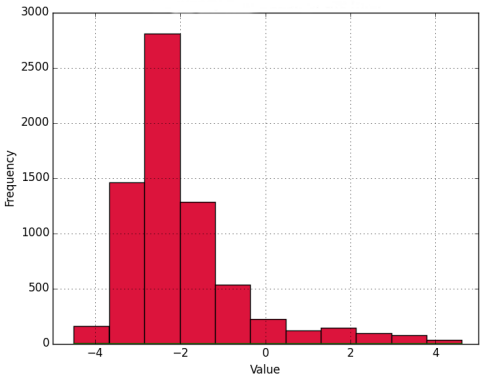

Figure 1: Frequency distribution of tweets by polarity intensity.

Two measures evaluated the similarity of the participant systems predictions to the manually annotated gold standard: Cosine Similarity (CS) and Mean Squared Error (MSE). The corpus available for training and trial consists of around $9,000$ figurative tweets with sentiment scores ranging from -5 to $+5$. Because of the perishability of Twitter data, some of them cannot be recovered by the published list of tweet identifiers; finally, we could rely on a corpus of $7,390$ messages considering both training and trial datasets. With respect to the polarity, the whole distribution is positively skewed (Fig. 1). The median value is very negative (-2.3) and the average of the tweets polarity is -2.

[6] http://www.liwc.net

Table 1: Criteria for assigning classes.

3c-approach		4c-approach	
Original	**New**	**Original**	**New**
pv>0	pos	*pv* > 0	pos
pv<0	neg	-2.5 > *pv* <= 0	nsn
pv=0	neu	-3.5 > *pv* <= -2.5	neg
		pv <= -3.5	vn

3.2 ValenTo System

As a first step, we decided to address the problem as a classification task. We experimented three approaches, each featured by a different amount of considered classes; in the first one (**3c-approach**) we used just three classes: *positive* (*pos*), *negative* (*neg*) and *neutral* (*neu*); in the second one (**4c-approach**) we used four classes: *positive*, *negative*, *not so negative (nsn)* and *very negative (vn)*; and in the third one (**11c-approach**) we used the original values included in the corpus, i.e. eleven classes from -5 to $+5$. For the first two approaches we changed the polarity values (*pv*) in each one of the tweets contained in the dataset according to the criteria summarized in Table 1. Based on polarity value distribution shown in Fig. 1, we separated the classes in different ranges that cover all the possible values. A small set of widely classification algorithms was used: Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM)[7]. We performed classification experiments using only the training set (i.e. 6,928 tweets); a ten fold-cross-validation criterium was applied. Table 2 presents results obtained in F-measure terms.

[7] We used Weka toolkit's version available at http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Table 2: Classification experiments: results.

Approach	NB	DT	SVM
3c-approach	0.829	0.804	0.790
4c-approach	0.458	0.440	0.462
11c-approach	0.324	0.311	0.302

As expected, from our classification results, the performance in terms of F-Measure drops while the number of classes increase. We decided to apply a different approach: Regression.

In order to build a regression model able to assign polarity values, we decided to merge both training and trial datasets (*fullTrainingSet* composed by 7,390 tweets). We used the Linear Regression Algorithm in Weka.

First, from the whole *fullTrainingSet* corpus we randomly extracted a set for training, containing the 70% of the tweets, and a set for the test, with the remaining 30%, obtaining *Subset-1*. We repeated the procedure two times more and we obtained *Subset-2* and *Subset-3*.Second, we made up 11 different combinations of features ***ft-conf[1-11]***. Each one contains a subset of the features described in Sec. 2.1. We built the features combination according to a preliminary analysis with respect to frequency distribution. Then, we applied our regression model for each *Subset* and ***ft-conf***. In order to evaluate the performance of our model, we used the script to obtain the cosine similarity measure provided by the organizers. Table 3 shows the results of these experiments for what concerns ***ft-conf2*** configuration, the one we selected for constructing the final model submitted to SemEval-Task 11 (due to lack of space, not have been included all results obtained). ***ft-conf2*** contains the following features:

lengthChar, *punctuation marks*, *POS*, *upperFreq*, *urlFreq*, *emotPosFreq*, *emotNegFreq*, *hashtagBinary*, *mentionsBinary*, *hashtagFreq*, *mentionsFreq*, *isRetweet*, *polReversal*, *AFINN*, *ANEW*, *DAL*, *HL*,*GI*, *SWN*, *senticnetDimensions*, *senticnetFormula*, *LIWC*, *NRC*

Table 3: Regression experiments: results.

Features	Subset-1	Subset-2	Subset-3
ft-conf2	0.8218	0.8161	0.8199

In order to measure the relevance of each feature used in our model, we applied the RELIEF algorithm[8]. The best ranked features are those related to emotional words (*NRC*) and polarity lexicons (*AFINN* and *HL*).

3.3 Official Results

We ranked 6th out of 15 teams in the SemEval-2015 Task 11 (Ghosh et al., 2015)[9]. ValenTo achieved the score of $\mathbf{0.634}$ using the CS measure, and a score of $\mathbf{2.999}$ using the MSE measure, while the best team achieved the score of $\mathbf{0.758}$ for CS, and a score of $\mathbf{2.117}$ for MSE.

Our results in terms of *irony* and *sarcasm* seem to be close to the best ones in each category (See Table 4).

Table 4: Official ValenTo and best results in each category of figurative type.

Category	CS		MSE	
	ValenTo	***Best***	ValenTo	***Best***
Overall	0.634	*0.758*	2.999	*2.117*
Sarcasm	**0.895**	***0.904***	**1.004**	***0.934***
Irony	**0.901**	***0.918***	**0.777**	***0.671***
Metaphor	0.393	*0.655*	4.730	*3.155*
Other	0.202	*0.612*	5.315	*3.411*

4 Conclusions

We described our participation at SemEval-2015 Task 11. A distinctive feature of our approach is that we used not only word-sentiment lexicons but also novel resources for dealing with emotions and psycholinguistic information. Based on both features analysis and evaluation results, we can draw a first insight about the importance of using such high-level information about affective value of the words in a tweet to tackle with figurative language such irony and sarcasm. As future work, the use of additional features for addressing figurative language under other perspectives (e.g. metaphor) will be explored.

[8] ReliefFAttributeEval version included in Weka (Robnik-Sikonja and Kononenko, 1997).

[9] http://alt.qcri.org/semeval2015/task11/index.php?id=task-results-and-initial-analysis-1, Table1.

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)*.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano, 2014. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, chapter Modelling Sarcasm in Twitter, a Novel Approach, pages 50–58.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 50–57.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Margaret Bradley and Peter Lang. 1999. Affective norms for english words (ANEW): Instruction manual and affective ratings. Technical report, Citeseer.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49.

Erik Cambria, Catherine Havasi, and Amir Hussain. 2012. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *AAAI FLAIRS Conference*, pages 202–207.

Erick Cambria, B. Schuller, Yunqing Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, March.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116.

A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, A. Reyes, and J. Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM*.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586.

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.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the EMNLP: Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

Marko Robnik-Sikonja and Igor Kononenko. 1997. An adaptation of relief for attribute estimation in regression. In *Fourteenth International Conference on Machine Learning*, pages 296–304.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, August.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963*, AFIPS '63 (Spring), pages 241–256.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, pages 765–770.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages. In *Psychological Reports*.