Applying Basic Features from Sentiment Analysis for Automatic Irony Detection

Irazú Hernández-Farías(✉), José-Miguel Benedí, and Paolo Rosso

Pattern Recognition and Human Language Technology, Universitat Politècnica de València, Valencia, Spain
{dhernandez1,prosso}@dsic.upv.es,jmbenedi@prhlt.upv.es
https://www.prhlt.upv.es/

Abstract. People use social media to express their opinions. Often linguistic devices such as irony are used. From the sentiment analysis perspective such utterances represent a challenge being a polarity reversor (usually from positive to negative). This paper presents an approach to address irony detection from a machine learning perspective. Our model considers structural features as well as, for the first time, sentiment analysis features such as the overall sentiment of a tweet and a score of its polarity. The approach has been evaluated over a set classifiers such as: Naïve Bayes, Decision Tree, Maximum Entropy, Support Vector Machine, and for the first time in irony detection task: Multilayer Perceptron. The results obtained showed the ability of our model to distinguish between potentially ironic and non-ironic sentences.

Keywords: Automatic irony detection · Figurative language processing · Sentiment analysis

1 Introduction

The ability to recognize ironic intent in utterances is performed by humans in a relatively easy way although not always. We develop this ability since childhood and, over years with social interaction we increase it. In many cases we are able both to understand and to produce such utterances without a strict definition of what is or may be considered an ironic expression. Irony is a sophisticated, complex and prized mode of communication; it is intemately connected with the expression of feelings, attitudes or evaluations [2]. Moreover, irony can be considered as a strategy, which is intented to criticise or to praise. Sometimes but not always, it means the opposite of the literal meanings; generally irony shows or express some kind of contradiction [1].

Recently interest for discover information in social media has been growing. Twitter, offers a face-saving ability that allows users to express themselves employing linguistic devices such as irony. User-generated content is difficult to analyse: Internet language is hard to analyze due to the lack of paralinguistic cues; in addition one needs to have a good understanding of the context of the

R. Paredes et al. (Eds.): IbPRIA 2015, LNCS 9117, pp. 337–344, 2015.
DOI: 10.1007/978-3-319-19390-8_38

situation, the culture in question, and the people involved [8]. For research areas such sentiment analysis (SA), irony detection is important to avoid misinterpreting ironic statement as literal [11].

For computational linguistic purposes, most of the time irony and sarcasm are often viewed as the same figurative language device. Irony is often considered as an umbrella term that covers also sarcasm [12]. Previous works are mainly based on the classification of tweets as ironic or sarcastic and rely solely on text analysis.

This paper presents an approach for irony detection using a set of features that combine both surface text properties and information exploited from sentiment analysis lexicons. The main contribution of this paper is to take advantage of the classification of utterances according to their polarity. We consider in order to detect irony it is important to take into account the sentiment expressed in a tweet. Our model improves state-of-the-art results. The rest of this article is organized as follows: previous works on automatic irony detection are introduced in Sect. 2. In Sect. 3 we describe the set of features used. In Sect. 4, dataset, classifiers, experimental setting and evaluation of our approach are presented. Finally, in Sect. 5 we draw some conclusions and discuss future work.

2 Related Work

Recently automatic irony detection has attracted the attention of researchers from both machine learning and natural language processing [11]. A shared task on figurative language processing has been organized at SemEval 2015[6][1].

A survey that includes both philosophical and literary works investigating ironic communication and some computational efforts to operationalize irony detection is presented by Wallace in [11]. Reyes et al. [10] address the problem of irony detection as a classification task; the authors proposed a model employing to four types of conceptual features: signatures, unexpectedness, style and emotional scenarios. Bosco et al. in [4] present a study that investigates sentiment and irony in online political discussion social media in Italian. Buschmeier et al. [5] present an analysis of 29 features (such as punctuation marks, emoticons, interjections and bag-of-words); the authors' main goal is to investigate the impact of features removal on the performance of their approach. Barbieri and Saggion [3] used six groups of lexical features (frequency, written-spoken, intensity, structure, sentiments, synonyms, ambiguity), in order to classify ironic tweets (the same dataset of [10] was used).

3 Proposed Features

We address irony detection as a classification problem, considering different types of features. In our model, we consider some features previously applied

[1] Given a set of tweets the task consist in determining whether the user has expressed a positive, negative or neutral sentiment; more information is available at: http://alt.qcri.org/semeval2015/task11/.

in irony detection. Moreover, we propose two sentiment analisys features (*Sentiment Score* and *Polarity Value*) in order to take advantage of resources that allow to measure the overall sentiment expressed in each tweet. We can distinguish the set of features into *Statistical-based* and *Lexical-based*. *Statistical-based* are surface patterns that can be obtained taking into account the frequency of some words or characters in the tweet. *Lexical-based* are obtained by using information beyond the textual content of the tweet, i.e. applying external resources.

The first set, ***Statistical-based*** features is composed of four dimensions: **a)** *Textual Markers (TM)*, features widely used in this task, which include frequency of visual cues as: lenght of tweet, capitalization, punctuation marks, and emoticons[2]; **b)** *Counter-Factuality (CF)*[3], the frequency of discursive terms that hint at opposition or contradiction in a text such as "nevertheless"[4]; **c)** *Temporal Compression (TC)*[3], the frequency of terms that identify elements related to opposition in time, i.e. terms that indicate an abrupt change in a narrative; and **d)** *POS-based features (POS)*, where each tweet has been processed using a POS-tagger developed for this kind of texts called ARK[5]; we take into account frequency of verbs, nouns, adjectives and adverbs.

Our second set of features, ***Lexicon-based***, exploits different knowledge bases to represent each tweet: **a)** *Semantic Similarity (SIM)*[3], consists in obtaining the degree of inconsistency measuring the relationship between the concepts contained in each tweet using the WordNet::Similarity[6] module; **b)** *Emotional Value (EV)*[3], where the emotional value is calculated taking into account the categories described by Whissel [13], in her Dictionary of Affect in Language (DAL)[7]. **c)** *Sentiment Score(SS)*, in order to catch the overall sentiment (positive, negative or neutral) expressed in a tweet. We applied a lexicon developed by Hu-Lui in [7][8]; and **d)** *Polarity Value(PV)*, this feature allows to identify the rate of evaluation, either to criticize (negative) or to praise (positive). We use

[2] Using emoticons, with few characters is possible to display one's true feeling; sometimes they are virtually required under certain circumstances in text-based communication, where the absence of some kind of cues can hide what was originally intended to be humorous, sarcastic, ironic, and often negative [14].

[3] Feature previously applied by Reyes et al. [10].

[4] The complete list of words can be downloaded from http://users.dsic.upv.es/grupos/nle.

[5] http://www.ark.cs.cmu.edu/TweetNLP/.

[6] https://codegoogle.com/p/ws4j/. This module allows to calculate a set of seven different similarity measures.

[7] DAL is composed by 8,000 English words, distributed in three categories: *Activation*, refers to the degree of response, either passive or active, that humans exhibit in an emotional state; *Imagery*, quantifies how easy or difficult is to form a mental picture for a given word; and *Pleasantness*, quantifies the degree of pleasure suggested by a word.

[8] http://ww.cs.uic.edu/~liub/FBS/.

AFINN[9] lexicon, which contains a list of words labelled with a polarity valence value between minus five (negative) and plus five (positive) for each word.

The last two features in this set (*Sentiment Score(SS)* and *Polarity Value(PV)*) have not been previously used in irony detection. Our main motivation to use sentiment analysis features is that an ironic utterance is subjective, hence contains a positive or negative opinion. On the other hand, we taking into account a feature that allows us obtaining a polarity value from each tweet, so we have both the "overall"sentiment and a score of the polarity. In sentiment analysis, there are several resources that could help to improve the detection of ironic tweets.

4 Experiments and Results

The dataset used in this work was compiled by Reyes et al. [10] and consists of a total of 40,000 tweets written in English, distributed in four different classes: Irony, Education, Humor and Politics. The corpus was built retrieving 10,000 tweets that contain one of the following hashtags: #irony, #education, #humor and #politics. These hashtags allow to have tweets in which users explicitly declare their ironic attempt, and a large sample of non-ironic tweets. In order to perform classification process, we apply a set of classifiers widely used in text classification tasks. Some of them has been used in irony identification. The set of classifiers[10] is composed by: Decision Tree (*DT*), Maximum Entropy (*ME*), Naïve Bayes (*NB*), Random Forest (*RF*) and Support Vector Machine (SVM, with a RBF kernel)[11] and Multilayer Perceptron (*MLP*, we used a backpropagation based multilayer perceptron, with sigmoid functions, a learning rate of 0.3 and 500 epochs in each run; we did not perform any parameter tuning.). In this paper we propose to apply *MLP*, that has never been used for irony detection.

As in [3] and [10], we perform a set of binary classifications between Irony and Education/Humor/Politics. Each experiment has been performed in a 10-fold-cross-validation setting. We run experiments for one baseline: *Bag Of Words (BOW)*. We exploit only most frequent unigrams per class (1,000) in order to represent each tweet. This baseline relies on standard text classification features. According to [11], words counts alone offer an insufficient representation for verbal irony detection.

We apply two different vector representation approaches for experimental purposes. Each tweet was converted to a vector composed by 16 features. No feature selection technique was performed. In the first approach the features belonging to *Statistical-based* were taking into account the frequency of each one; while *Lexicon-based* are represented in different ways: the semantic similarity is the value obtained using the above-mentioned module; emotional value is

[9] http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt.

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

[11] Default parameters for each algorithm were used.

calculated taking into account values in DAL over words that compose each tweet; the sentiment score can be *positive* (more positive than negative terms), *negative* (more negative than positive terms) or neutral (same amount of positive and negative terms); finally, the polarity value is assigned by calculating the difference between the positive and the negative polarity of each tweet according to AFINN lexicon.

In the second approach we applied the representativeness criterion presented by Reyes et al. [10] in order to assign a value for *Statistical-based* features; the representativeness of a given document d_k (e.g. a tweet) is computed according to:

$$\delta_{i,j}(d_k) = \frac{f_{ij}}{|d_k|} \tag{1}$$

where i is the i-th feature; j is the j-th dimension; f is the feature dimension frequency; and $|d_k|$ is the lenght of the k-th document d_k. If $\delta_{i,j}(d_k)$ is $\geq$0.5, a value of 1 is assigned; otherwise, a representativeness value of 0 (not representative at all) is assigned; and the *Lexicon-based* features were represented as the same way above described for the first approach.

Three experiments were carried out using the classification algorithms mentioned above. Each experiment are constructed under different criteria. Two of them (***Lesk*** and ***Wu-Palmer***) are based in the first representation approach while the third (***Rep, Representativeness***) takes into account the second approach. The difference between *Lesk* and *Wu-Palmer* is the semantic similarity[12], that take into account, using Lesk and Wu-Palmer measures respectively.

In Table 1, we report F-measure results of our classification experiments. It can be observed that all results overcome the baseline. The bold values are used to highlight those F-measures greater than state-of-the-art (See Table 3). The best result is achieved by *SVM* in the three sub-tasks (binary classification Irony vs. Education, Irony vs. Humor and Irony vs. Politics). As reported by [3] and [10], higher results in F-measure are achieved by *ironic-vs-politics* classification, while lower F-measure lie in *ironic-vs-humor*. We carried out the t-test (with a 95 % confidence level) in order to see if the best results are statistically significant.

Moreover, we calculated the *Classification Error Rate (CER)*. In Table 2 CER values for each binary classification (*Iro-Edu*, *Iro-Hum* and *Iro-Pol*) are presented. As can be seen, our model obtains satisfactory CER rates. The best results(bold values in Table 2) are obtained by: SVM, MLP and RF.

As mentioned above, the dataset has been used before ([3] and [10]). The results reported by their authors are shown in Table 3. In both works a Decision Tree classifier was used. The last two rows in the table correspond to our results using the Decision Tree classifier.

As Table 2 shows, our approach improves the F-measure obtained previously by state-of-the-art approches. In order to determine which features are more rel-

[12] We performed experiments using each similarity measure of the WordNet::Similarity module. Due to lack of space, we report only the results with highest classification rates. The similarity measures are described in detail in [9].

Table 1. Results in F-measure for the baseline and each representation approach corresponding to binary classification. The underlined values are statistically significant.

	Irony-Education				*Irony-Humor*				*Irony-Politics*			
	BOW	Lesk	Wu-Palmer	Rep	BOW	Lesk	Wu-Palmer	Rep	BOW	Lesk	Wu-Palmer	Rep
DT	0.34	**0.78**	**0.78**	0.68	0.34	0.75	0.74	0.70	0.34	**0.79**	**0.79**	0.63
ME	0.37	**0.75**	**0.75**	0.66	0.37	0.74	0.74	0.69	0.36	**0.76**	**0.76**	0.59
MLP	0.50	**0.78**	**0.78**	0.67	0.50	0.75	0.76	0.70	0.50	**0.79**	**0.79**	0.61
NB	0.44	0.70	0.70	0.66	0.46	0.69	0.70	0.65	0.45	0.70	0.71	0.57
RF	0.16	**0.79**	**0.79**	0.68	0.16	0.76	0.76	0.70	0.16	**0.81**	**0.81**	0.63
SVM	0.63	**0.80**	**0.80**	0.68	0.59	**0.77**	**0.78**	0.69	0.64	**0.81**	**0.80**	0.63

Table 2. Results in terms of CER

	Irony-Education				***Irony-Humor***				***Irony-Politics***			
	BOW	Lesk	Wu-Palmer	Rep	BOW	Lesk	Wu-Palmer	Rep	BOW	Lesk	Wu-Palmer	Rep
DT	66.01	43.65	43.75	63.12	65.1	49.67	51.72	59.58	65.41	41.05	41.36	72.22
ME	62.58	49.88	49.93	67.46	62.48	50.51	50.64	60.17	63.13	46.82	46.59	79.24
MLP	50	43.8	42.87	64.76	50	48.25	**42.87**	60.07	50	40.82	40.96	76.53
NB	55.18	59.62	59.31	66.31	53.68	60.43	59.27	68.82	54.91	53.36	57.46	77.55
RF	84.11	40.5	40.71	63.22	84.31	46.19	45.97	59.71	84.2	**37.09**	**36.88**	72.59
SVM	55.18	**40.1**	**40.15**	63.65	55.08	**44.17**	44.07	60.7	54.46	37.93	37.88	73.82

evant in our model, Information Gain[11] was calculated. There are some features that seem to contribute more than others in our model to discriminate between classes (see Fig. 1). As can be seen, the textual markers (TM) features are a good indicator of this kind of utterances. Moreover, also the sentiment analysis features (SS and PV) showed to have an important impact on irony detection. This strenght the idea that irony detection is strongly related to sentiment analysis.

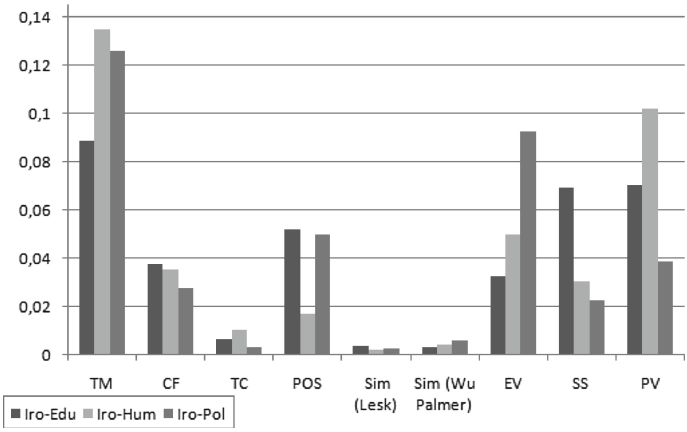

Fig. 1. Information gain for our set of features

Table 3. Results in F-measure of our model against state-of-the-art

	Irony vs.		
	Education	*Humor*	*Politics*
Reyes et al.	0.70	0.76	0.73
Barbieri and Saggion	0.73	0.75	0.75
Our approach *Lesk*	**0.78**	0.75	**0.79**
Our approach *Wu-Palmer*	**0.78**	**0.79**	**0.79**

According to Fig. 1, features related to SA seem to be quite important to identify ironic from non-ironic tweets. From this we may say that using features and resources for SA could improve performance of models for irony detection.

5 Conclusions

Given the growing interest in exploiting knowledge generated in social media, irony detection has attracted the attention of different research areas. Different approaches have been proposed to tackle this task. In this paper we proposed a model for ironic tweets classification, taking advantage for the first time of sentiment analysis features. The proposed model obtained higher values in terms of f-measure than those reported in the state-of-the-art using the same dataset. One of the best results was obtained by *MLP*, a method has not been previously used for irony detection. Also in terms of CER, our model showed good performance in classification rates of ironic tweets in the experiments we carried out. As future work an in-depth analysis of the impact of the proposed features is needed. We plan to exploit further features and resources from sentiment analysis.

Acknowledgments. The National Council for Science and Technology (CONACyT Mexico) has funded the research work of the first author (Grant No. 218109/313683, CVU-369616). The research work of third author was carried out in the framework of WIQ-EI IRSES (Grant No. 269180) within the FP 7 Marie Curie, DIANA-APPLI CATIONS (TIN2012-38603-C02-01) projects and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

1. Alba-Juez, L.: Irony and the other off record strategies within politeness theory. J. Engl. Am. Stud. **16**, 13–24 (1995)
2. Attardo, S.: Irony markers and functions: towards a goal-oriented theory of irony and its processing. Rask **12**, 3–20 (2000)
3. Barbieri, F., Saggion, H.: Modelling Irony in Twitter, pp. 56–64. Association for Computational Linguistics (2014)
4. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis: the case of irony and senti-tut. IEEE Intell. Syst. **28**(2), 55–63 (2013)

5. Buschmeier, K., Cimiano, P., Klinger, R.: 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, pp. 42–49. Association for Computational Linguistics (2014)
6. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Reyes, A., Barnden, J.: Sentiment analysis of figurative language in twitter. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM (2015)
7. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 168–177(2004)
8. Maynard, D., Greenwood, M.: 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), European Language Resources Association (ELRA) (2014)
9. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity: measuring the relatedness of concepts. In: Proceedings of the 9th National Conference on Artificial Intelligence, pp. 1024–1025. Association for Computational Linguistics
10. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Lang. Resour. Eval. **47**(1), 239–268 (2013)
11. Wallace, B.C.: Computational irony: a survey and new perspectives. Artif. Intell. Rev. **43**, 467–483 (2013)
12. Wang, A.P.: #irony or #sarcasm – a quantitative and qualitative study based on twitter. In: Proceedings of the PACLIC: the 27th Pacific Asia Conference on Language, Information, and Computation, pp. 349–356. Department of English, National Chengchi University (2013)
13. Whissell, C.: Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages. Psychol. Rep. **2**, 509–521 (2009)
14. Wolf, A.: Emotional expression online: gender differences in emoticon use. CyberPsychology Behavior **3**, 827–833 (2000)