Irony detection in Twitter with imbalanced class distributions

Delia Irazú Hernández Farías[a,*], Ronaldo Prati[b], Francisco Herrera[c] and Paolo Rosso[d]
[a] *División de Ciencias e Ingenierías Campus León, Universidad de Guanajuato, Mexico*
[b] *Universidade Federal do ABC, Brazil*
[c] *Department of Computer Science and Artificial Intelligence, University of Granada, Spain*
[d] *Universitat Politècnica de València, Spain*

Abstract. Irony detection is a not trivial problem and can help to improve natural language processing tasks as sentiment analysis. When dealing with social media data in real scenarios, an important issue to address is data skew, i.e. the imbalance between available ironic and non-ironic samples available. In this work, the main objective is to address irony detection in Twitter considering various degrees of imbalanced distribution between classes. We rely on the emotIDM irony detection model. We evaluated it against both benchmark corpora and skewed Twitter datasets collected to simulate a realistic distribution of ironic tweets. We carry out a set of classification experiments aimed to determine the impact of class imbalance on detecting irony, and we evaluate the performance of irony detection when different scenarios are considered. We experiment with a set of classifiers applying class imbalance techniques to compensate class distribution. Our results indicate that by using such techniques, it is possible to improve the performance of irony detection in imbalanced class scenarios.

Keywords: Irony detection, class imbalance, imbalanced learning

1. Introduction

Users of social media platforms tend to formulate points of view, opinions, and judgments concerning almost everything surrounding them: from a given event up to a personal experience. Social media allow the users to employ language in its literal sense but sometimes figurative language devices are also exploited. Among them, there is one that serves to express opinions in a witty (and often funny) way: *irony*.

Irony serves to express an evaluative judgment or attitude towards a particular target [2]. It allows us to convey subjective ideas by using the non-literal meaning of the words. Several theories have been proposed attempting to describe what irony is. Perhaps the most common one is that from the Gricean tradition [14], where irony is defined as a trope where the speaker intends to communicate the opposite meaning of what is literally said. Other authors consider that an ironic utterance serves to reveal the speaker's position (approval or disapproval) on the result of something [25, 43].

Besides the different theories defining irony, such a language device is also related to another concept: sarcasm. Both terms are perceived as synonyms due to the subtle distinction between them. When irony involves stressed negative evaluation towards a particular target with the intention of a given offense, it is considered as sarcasm [4, 30]. Under a computational linguistic perspective, irony and sarcasm are considered either as synonyms or being irony an umbrella term covering sarcasm.

In social media, people use irony, having (most of the time) only an intuitive definition of this concept. Consequently, the ironic content in these platforms reflects what people consider as this kind of figurative language device. As can be noticed in the following tweets[1], irony can be used with different purposes: to express an evaluation indirectly (example *(i)*) or to reveal a failed expectation (example *(ii)*).

(i) I seriously loveeeee how much you care
(ii) My train got cancelled.. Good way to start the day! -.- #Västtrafik

Interest in detecting the presence of irony in social media has grown significantly in the past years. Understanding the real meaning of a given message is an ongoing task for computational linguistics; therefore, such an intriguing figurative language device represents a big challenge.

In particular, Twitter data have become popular for irony detection [21]. Twitter represents an interesting source of information regarding how people perceive events, products, and so on. It provides a huge amount of user-generated data (easily accessible via Twitter API[2]) that allows capturing a wide variety of real uses of irony in this kind of short texts. Several approaches have been proposed to deal with irony detection relying on different perspectives. The authors in [6, 23, 40] addressed irony detection relying mainly on textual-based features. Others [5, 24, 44] took advantage of information regarding the context surrounding a given comment to determine whether or not an ironic meaning is intended. Exploiting the affective property of irony, in [17] the authors proposed an approach considering mainly such kind of information. Moreover, novel techniques such as word-embeddings and deep learning models have also been exploited [22, 34,36].

Despite data skew has been recognized as a critical issue for irony detection [21], related work addressing this task as a class imbalance problem is scarce. The Multi-Strategy Ensemble Learning Approach (MSELA) was proposed by Liu et al. [28] to deal with irony detection in imbalanced class datasets. The authors experimented with ironic comments written in English and Chinese. The MSELA combines sample-ensemble, classifier-ensemble, and weighted voting strategies together with a set of different features for each language. Punctuation marks, n-grams, and POS tags were used as features for English, whereas extreme positive and negative nouns, adjectives, adverbs of degree and proverbs were exploited for Chinese. Results on different settings exploiting MSELA were reported achieving in overall 0.8 in AUC (Area Under the ROC curve) terms.

A corpus of manually annotated Twitter conversation was used by Abercrombie and Hovy [1] for experimenting with irony detection on balanced and imbalanced class scenarios. Furthermore, the authors compare the performance of recognizing the irony of both human and machine learning algorithms. A logistic regression classifier with features such as n-grams and POS tags was employed. The performance of the proposed approach suffers from significant drops on the imbalanced class data in both F-measure and AUC terms when compared with those from the balanced one. Cervone et al. [10] experimented with ironic tweets written in Italian by applying balancing techniques to address the data skew. They exploited different sets of features combined with random oversampling, undersampling, and cost-sensitive learning. The best performance was obtained by the last one when it was used together with bag-of-words representation.

The progress so far achieved in irony detection has been focused on the development of models able to automatically capture potential cues for identifying such kind of figurative language device. However, even when the data skew has been recognized as an inherent factor related to the presence of ironic content on Twitter, the majority of the related work fails to have regard to the role of imbalanced class degree.

In this paper, we address irony detection from a perspective of imbalanced distributions aim at evaluating the impact of applying different preprocessing techniques for detecting irony in Twitter. As far as we know, this is the first work where irony detection is addressed by considering many factors related to the imbalanced learning problem. It is important to highlight that we are not proposing a novel technique for dealing with imbalanced class data, instead, we are experimenting with the use of existing techniques for evaluating irony detection in an imbalanced class scenario. We are considering several factors related to the class imbalance problem described by [20]. Another important point arises from the fact that we are not introducing a new approach to detect irony rather, we are using the model described in [17]. Furthermore, we carried out an extensive experimental study with a large benchmark of irony detection corpora that covers several aspects ranging from developing criteria to imbalanced class degree.

Summarizing, our main contributions are the following:

1. We exploited the irony detection model described in [17] in order to carry out a set of classification experiments aimed to determine the impact of class imbalance for detecting irony.
2. We experimented with several treatment techniques in order to assess its impact on the performance of irony detection.
3. We developed a new set of corpora (denoted as *TwImbData*) retrieved considering criteria for simulating a realistic scenario. This dataset could serve as a starting point for expanding the research on irony detection considering an imbalanced class scenario.

Organization. The paper is organized as follows. Section 2 introduces irony detection from a class imbalanced problem perspective; besides, the irony detection model and the corpora we used for experimental purposes are also presented. In Section 3 we describe the experimental setting and the obtained results of addressing irony detection as a class imbalance problem. Section 4 summarizes the main findings of our study. Finally, in Section 5 we draw some conclusions and some directions for future work.

2. Irony Detection as an imbalanced class classification problem

Every day millions of tweets are posted on Twitter[3]. Even when the use of irony in this social media platform is quite common, the difference between the amount of non-ironic, i.e., tweets expressing any other kind of intention, and ironic tweets is enormous. Therefore, an important issue to address in irony detection is data skew, i.e., the imbalance between ironic and non-ironic samples available, as it reflects the realistic distribution of the use of irony in Twitter [1, 27, 39]. This problem also happens in many other real-world problems such as biology, medicine, economy, etc. The role of data skew for detecting the presence of irony in social media has been recognized as an important challenge [1, 21] that needs to be considered when designing irony detection models.

Furthermore, according to [3], a class imbalance distribution problem could occur in two situations: (i) when class imbalance occurs naturally in the problem in hand, and (ii) when the data is not imbalanced by definition, instead is very expensive to acquire such data for minority class due to factors such as cost, effort, etc. The irony detection problem fits with both situations. First, there is a big difference between the ironic and non-ironic tweets published in a given time frame. Secondly, retrieving potentially ironic data from a given data source is not a trivial task. Two different methodologies for acquiring data for irony detection have been recognized in [18]: self-labeling and crowdsourcing. The former one involves the use of certain labels such as hashtags that are added by the authors of the texts, while in the second one a manual annotation procedure needs to be performed; then, the task becomes more complex involving the inherent subjectivity of irony not only from the author of the text in hand but also from the human annotator.

Generally speaking, irony detection has been addressed as a binary classification task. The main aim is to distinguish ironic from non-ironic texts. In a nutshell, when an irony detection approach is proposed, the principal goals are: *(i)* to propose a set of relevant features helping to capture the ironic intention in a given text and, *(ii)* to assess the performance of the model usually in an in-house dataset collected by the authors. Most of these approaches do not consider other related aspects such as the impact of the inherent imbalanced nature of the presence of irony in social media platforms.

In the next section, we introduce in detail the irony detection model we exploited for detecting ironic tweets in imbalanced class scenarios.

2.1. emotIDM: an irony detection model based on affective information

According to several theorists, affect plays an important role in the use of irony [2, 47]. Therefore, considering the presence of affective content involved in ironic texts represents an interesting starting point. Attempting to take advantage of such kind of information, we rely on *emoIDM* proposed in [17].

emotIDM addresses irony detection as a classification task by considering different facets of affective content as well as structural markers. To represent a tweet, *emotIDM* uses three different groups of features:

1. *Structural*. It includes punctuation marks, length of words, part-of-speech labels, Twitter Marks (i.e., hashtags, mentions, etc.), among others.

2. *Sentiment*. In order to capture sentiment information, *emotIDM* takes advantage of a wide range of English lexical resources such as: AFINN [33], Hu&Liu [19], among others.
3. *Emotions*. Attempting to cover as much information related to emotions as possible, *emotIDM* considers features regarding the main theories in the nature of emotions: Categorical and Dimensional models of emotions. The Categorical model suggests the existence of basic emotions such as anger, fear, joy, disgust, etc., that in *emotIDM* are considered from: EmoLex [31], EmoSenticNet [37], and LIWC [35]. In the Dimensional model, an emotion can be defined according to its position in a space of independent dimensions. *emotIDM* includes the dimensions defined in: ANEW [8], Dictionary of Affect in Language [46], and SenticNet [9].

Exploiting affective information for detecting irony also allows to capture this kind of information, disregarding domain. Besides, in line with most of current approaches in computational linguistics, irony here is considered as an umbrella term that also covers sarcasm. Tackling differences between these devices in social media is a further challenge for figurative language processing [42, 45], which is very interesting but beyond the scope of this work.

Most of the time, when an irony detection model is proposed, it is evaluated over an in-house dataset retrieved by its authors. Instead, the performance of *emotIDM* was assessed by using a set of corpora in the state of the art. The obtained results outperformed those in the related work and validated the importance of affect-related information for detecting ironic content in tweets.

2.2. Irony detection corpora

In a similar fashion than in other natural language processing tasks, collecting user-generated data containing ironic instances is not a simple task. As mentioned before, two main approaches have been adopted for collecting Twitter data:

i By taking advantage of hashtags (such as “#irony" and “#sarcasm") that allow users to explicitly marking their tweets as ironic. The readability of using hashtags as golden labels has been experimentally confirmed [26].

ii By exploiting crowdsourcing techniques to determine whether a tweet is ironic or not.

Interest in investigating the use of irony in Twitter has led into having a wide set of available corpora for addressing irony detection. Nevertheless, there are not specific corpora developed considering imbalanced class scenarios, i.e., a dataset which keeping the inherent imbalanced class ratio of this problem in a real scenario is considered. We experimented with two different groups of corpora: (a) *Benchmark corpora*, and (b) *Imbalanced Class Twitter data for Irony Detection*. Next, we describe both groups of corpora as well as its main characteristics.

Benchmark corpora

As mentioned before, there are several Twitter corpora developed for evaluating different irony detection approaches. We took advantage of the five corpora described below:

- *TwReyes2013*. Reyes et al. [40] collected a set of tweets by taking advantage of specific hashtags. They selected three hashtags for collecting non ironic tweets: #education, #humor, and #politics. Concerning to the ironic instances, they relied on the use of the hashtag #irony by Twitter users.
- *TwRiloff2013*. Riloff et al. [41] created a Twitter corpus of 3,200 tweets following a hybrid approach involving the presence of specific markers as well as crowdsourcing. They retrieved tweets containing sarcastic hashtags (such as #sarcasm and #sarcastic) and also some regular tweets. Then, they asked three annotators to manually annotate the presence of sarcastic content in the tweets after removing the aforementioned hashtags (if any).
- *TwBarbieri2014*. Barbieri et al. [6] adopted a similar methodology to the one of [40]. The non ironic tweets are composed by those equivalents in the *TwReyes2013* together with 10,000 tweets collected by exploiting the #newspaper hashtag. Regarding the ironic tweets, the authors took advantage of two hashtags: #irony and #sarcasm[4].
- *TwPtáček2014*. Ptáček et al. [39] introduced two sarcastic datasets: in Czech[5] and English. For collecting the sarcastic tweets the authors used the hashtag #sarcasm, while the non sarcastic tweets were collected using only the

[4]In the rest of the paper, we will use *TwIronyBarbieri2014* and *TwSarcasmBarbieri2014* to refer which set of tweets are used as ironic tweets those with #irony or #sarcasm, respectively.

[5]More details about this dataset can be found in [39].

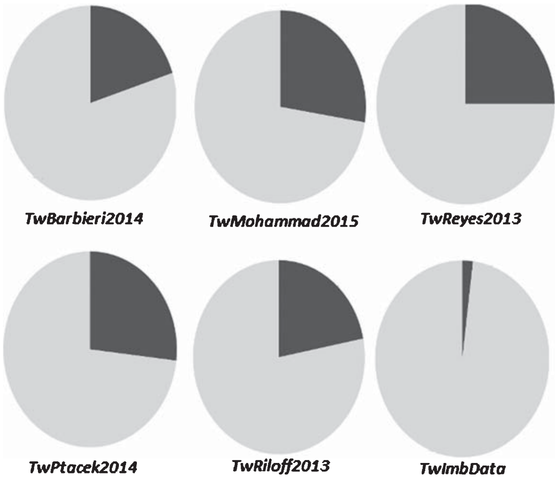

Fig. 1. "Ironic" and "non-ironic" tweets distribution in the corpora.

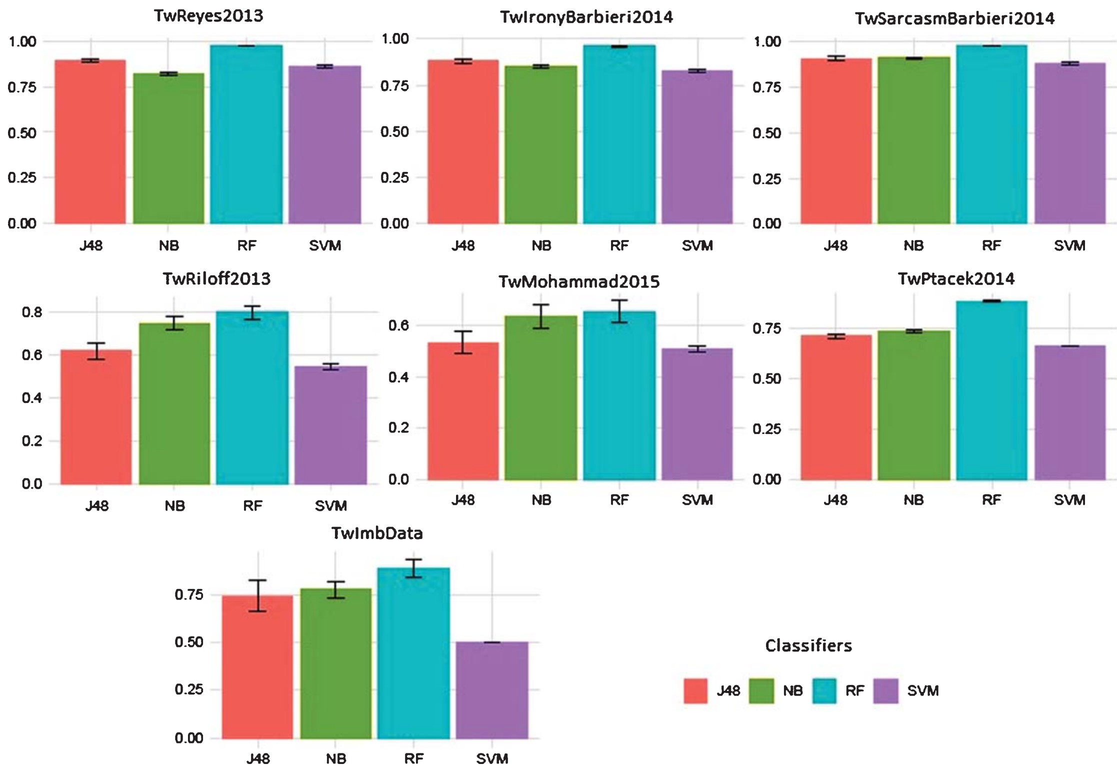

Fig. 2. Obtained results in AUC terms using the original distribution of the corpora.

language (English) as a filter parameter. The *TwPtáček2014* comprises two different distribution scenarios: balanced, and imbalanced[6].

– *TwMohammad2015*. Mohammad et al. [32] collected a set of tweets related to the 2012 US presidential elections[7]. They defined a multi-

[6]In this paper, we used a subset of the tweets in the imbalanced class distribution because of the perishability of Twitter data.

[7]Some hashtags such as #election2012, #election, #president2012, among others were used for retrieving data from Twitter.

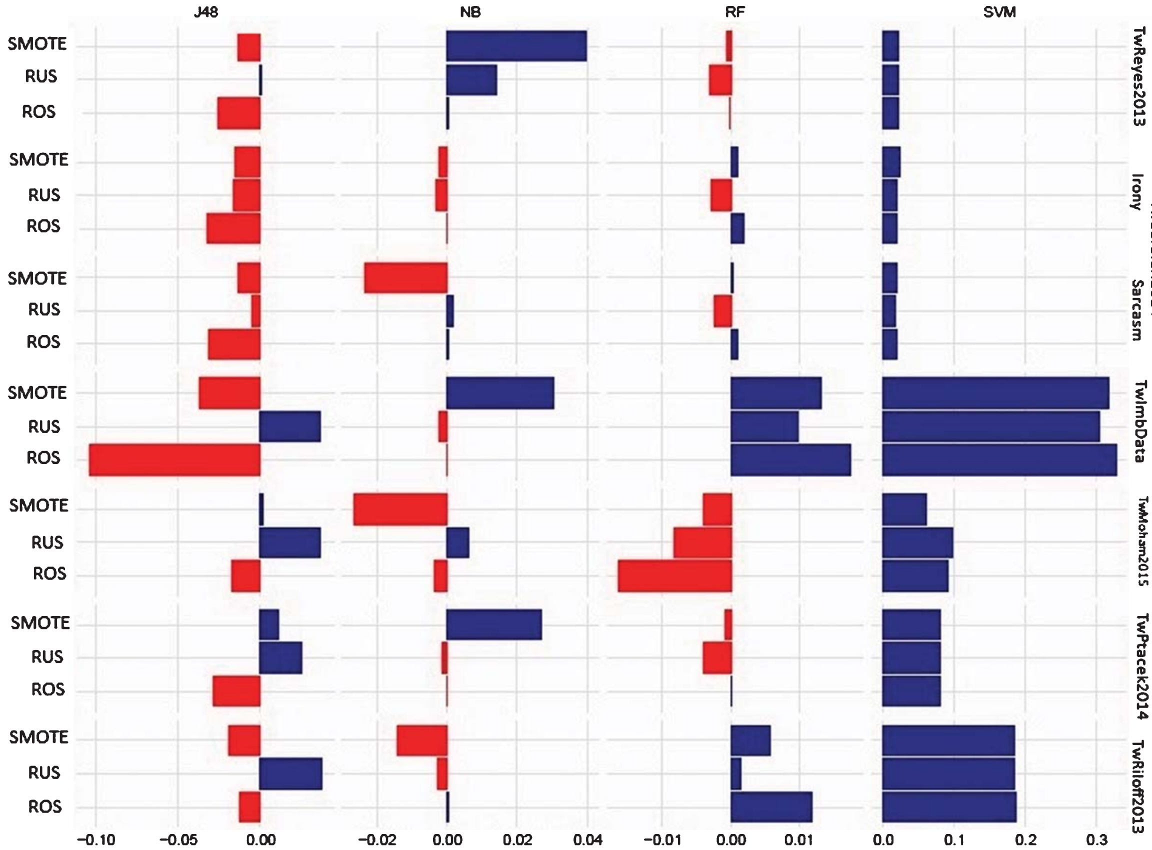

Fig. 3. Differences in terms of AUC with respect to the results on the ORIGINAL distribution after applying treatment techniques.

layer annotation schema concerning different aspects such as sentiment, emotions, purpose, and style. The last one includes sarcasm, hyperbole, understatement, and simple statement as labels.

To sum up, the *TwReyes2013*, *TwBarbieri2014*, and *TwPtáček2014* were retrieved by relying on the presence of specific labels used by the users to point out an ironic (or sarcastic) intention. Instead, *TwRiloff2013* and *TwMohammad2015* involve manual annotation of ironic tweets by exploiting crowdsourcing techniques. Regarding to *TwReyes2013* and *TwBarbieri2014*, we have merged all "non-ironic" samples into a unique class.

New Imbalanced Twitter Corpora for Irony Detection

With the aim to simulate a "realistic scenario", i.e. a dataset that resembles a hypothetical proportion of "ironic" tweets with respect to the "non-ironic" ones, we retrieved data from Twitter by exploiting the Streaming API. Many factors are influencing the number of tweets posted in a day. Therefore, providing a fixed or approximate quantity of ironic tweets posted in a day is not possible. We collected a sample of the tweets posted in a day (from 8th up to 18th November 2016) applying a two-step filtering criteria:

1. The tweets must contain at least one of the following hashtags: "#irony" and "#sarcasm".
2. The tweets must be written in English.

The ironic instances were retrieved by following both criteria. Instead, the "non-ironic" instances are those tweets collected only with the second criterion. A total of eleven datasets were created, the "ironic" instances are those collected with the first criterion while in the case of the "non-ironic" we randomly selected a subset of tweets according to a fixed imbalance ratio between-class of 1 : 50 (i.e., for each "ironic" tweet, there are 50 "non-ironic" tweets). Such datasets were grouped into a single one denoted as *TwImbData*, were each subset has the same imbalance ratio.

To sum up, Fig. 1 shows the distribution of ironic (in black color) and non-ironic (in gray color) tweets in the set of corpora described in this section. As can be noticed, the distribution among classes in *TwImbData* is very different from the other datasets.

3. Addressing irony detection with imbalanced data

In supervised classification, the prediction of rare events is known as the class imbalance problem [12, 38]. Class imbalance may imply great challenges for machine learning algorithms. Most of them tend to misclassify the minority instances more often than the majority instances on imbalanced class datasets. Aimed to determine the impact of class imbalance for detecting irony, we performed an experimental setting considering several aspects.

We carried out a set of experiments to evaluate the performance of *emotIDM* under different degrees of class imbalance by applying different methods for compensating class distribution. To deal with the class imbalance, many solutions have been proposed in the past few years [15]. These solutions can be broadly categorized into two groups: (i) *data level approaches* and (ii) *algorithm level approaches*.

Data level approaches work in the preprocessing phase. They are independent of the learning algorithm, and in general, aim to re-balance the data distribution by discarding (undersampling) majority or replicating (oversampling) minority instances. Simple approaches to do this include random under sampling (hereafter RUS) and random oversampling (denoted as ROS) [7]. There are some disadvantages related to the use of these techniques, for example, with RUS, there is a possibility of discarding useful data for the learning process. On the

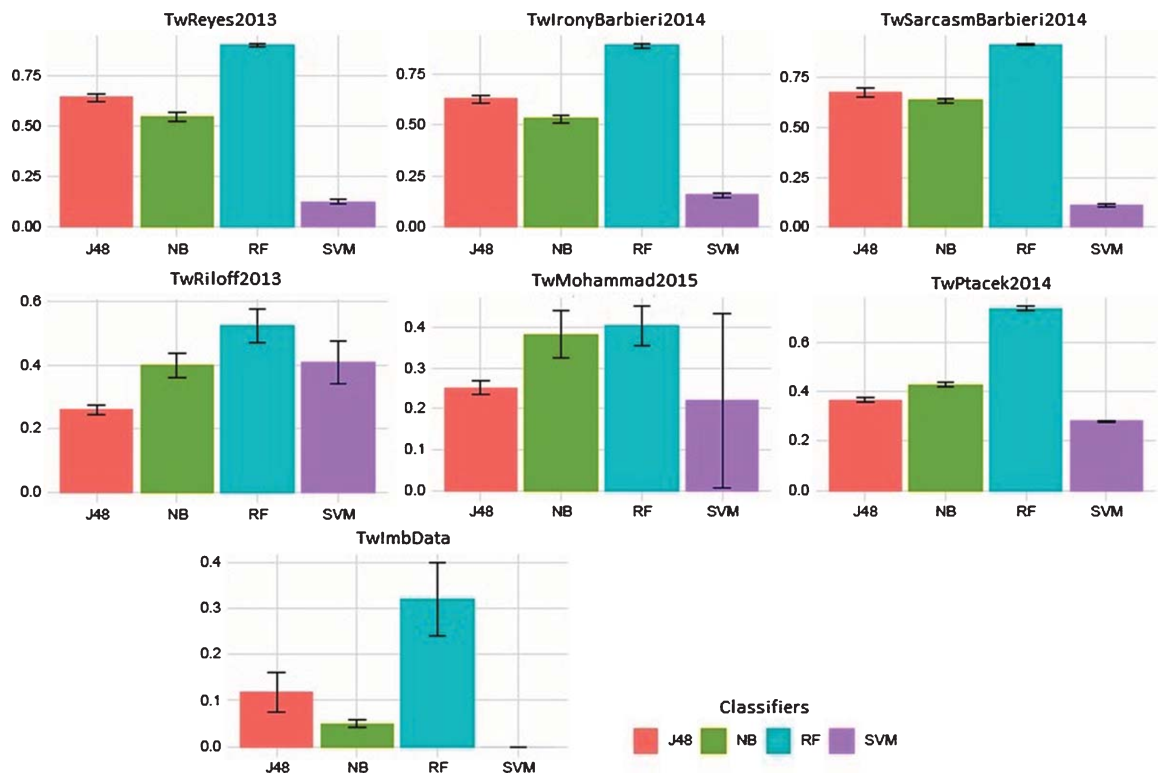

Fig. 4. Obtained results in AUPR terms using the original distribution of the corpora.

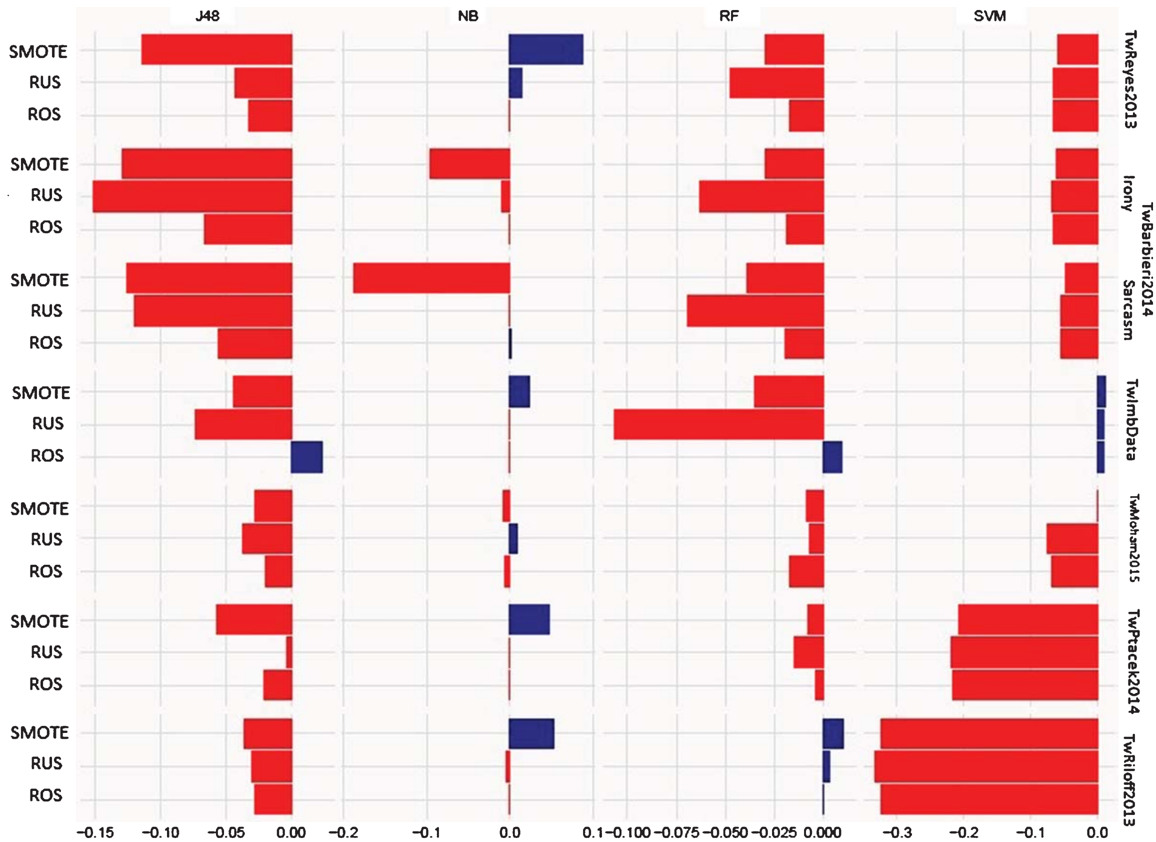

Fig. 5. Differences in terms of AUPR with respect to the results on the ORIGINAL distribution after applying treatment techniques.

other hand, with ROS the probability of provoking overfitting increases. An approach that synthetically generates instances from the minority class is the Synthetic Minority Oversampling Technique, denoted as SMOTE [11]. The main idea of SMOTE is to create new instances of the minority class by interpolating them in order to oversample the training set. Apart from that, *algorithm level approaches* involve the adaptation of learning algorithms to deal with class imbalance. These modifications generally involve the adjustment of some optimization criteria to trade-off frequent and infrequent classes differently.

We addressed the classification between "ironic" and "non-ironic" tweets by exploiting the Weka[8] implementation of the following machine learning classifiers (the default parameters were used for experimental purposes): Naive Bayes (NB), Decision Tree (J48), Support Vector Machine (SVM), and Random Forest (RF). We ran the experiments using five-fold cross validation within each dataset from the corpora. The experiments were paired, that is, the same training and test partitions were used for all learning algorithms.

In order to compensate for different class imbalance distributions in irony detection, we applied three class imbalance treatment techniques, namely ROS, RUS, and SMOTE. The aforementioned data level techniques were applied in order to achieve a balanced (50% of instances in each class) proportion in the training set. The class imbalance treatment methods were applied to the training set, and the test set was left untouched. For the sake of comparison, we also used the original distribution (denoted as ORIGINAL) as presented in each of the corpora described in Section 2.2.

We are interested in assessing the performance of irony detection when imbalance treatment techniques are used in order to compensate for the differences in terms of instances per class. As evaluation metrics we considered five different namely: Area Under the Curve (AUC), Area Under the Precision-Recall

[8]http://www.cs.waikato.ac. nz/ml/index.html

Curve (AUPR), Balanced Accuracy (BAC), Predictive Positive (PPOS), and F1 score. Being the last one the most common used for assessing the performance of irony detection in Twitter.

In the following paragraphs, the obtained results of applying the aforementioned experimental setting are described. For each evaluation metric, we present two figures with the obtained results. The first one reflects the outcomes over the ORIGINAL distribution. After applying the treatment techniques, we calculate the difference between the obtained result over the ORIGINAL distribution and the corresponding performance when using a given preprocessing method. Therefore, when this difference is positive (i.e., there is an improvement of the results), it is represented as a bar towards the right side. On the contrary, when the difference is negative (i.e., the obtained result over the original distribution decreased), it is represented as a bar towards the left side.

Each of the *Benchmark* corpus is presented individually, while in the case of the *TwImbData* we present the average result of considering each dataset individually. All the experiments were performed in each of the datasets composing *TwImbData*, however for the sake of the readability, we decided to group the obtained results since those corpora share similar proprieties.

Area Under the Curve

Figure 2 shows the obtained results over the ORIGINAL distribution considering AUC as evaluation metric. In all corpora, the highest results were obtained using RF as the classifier. SVM emerged as the classifier with the lowest performance in the ORIGINAL distribution.

As it is shown in Fig. 3, when the treatment techniques were applied together with SVM in all corpora, there is a positive impact on the results with respect to the performance of the ORIGINAL distribution. On the other hand, there is a negative impact of using J48 with treatment techniques, except with RUS when it is used for experimental purposes on most of the corpora. Regarding the use of NB, there are some cases where using SMOTE allows improving its performance against the ORIGINAL distribution. The overall performances in terms of AUC of the imbalance treatment techniques are lower in those datasets

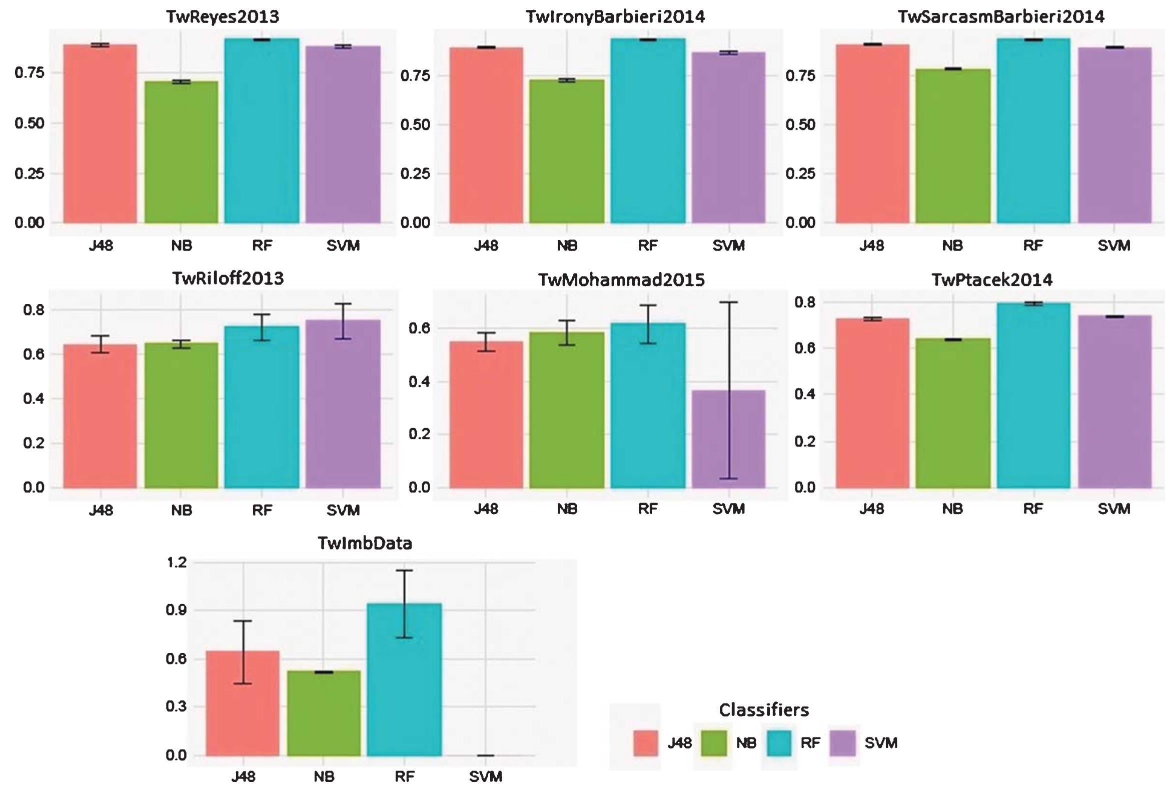

Fig. 6. Obtained results in Balanced Accuracy terms using the original distribution of the corpora.

where crowdsourcing was involved for developing corpora, in line with the findings of [17]. Generally speaking, the performance of the model in terms of AUC across the corpora reveals an improvement in the performance in most cases when treatment techniques are applied. The lowest rates were achieved in *TwMohammad2015* while the best ones were in *TwBarbieri2014* and *TwReyes2013*.

As can be observed, the most noticeable differences are in corpora with a higher imbalanced class degree rate. In *TwImbData* the increase is around 0.3 for all the treatment techniques. In the case of J48, in most of the experiments, there is a negative impact in terms of AUC. Applying treatment techniques together with RF helps to enhance the performance of the classifiers in *TwImbData* and *TwRiloff2013*.

Area Under the Precision-Recall Curve

In Fig. 4 we present the outcomes of the experimental setting when AUPR was considered as evaluation metric. AUPR is considered as a useful measure of success of prediction when the classes are very imbalanced, as this case. The best performance in terms of AUPR was achieved by the RF; while the SVM has the lowest rates.

In most of the cases, there is a drawback in the performance in terms of AUPR of the classifiers when treatment techniques were applied (as shown in Fig. 5). Considering those experiments where there is an improvement, it can be observed that it was achieved by either SMOTE or ROS. In terms of AUPR, when SVM was used the obtained results over the benchmark corpora were not improved by applying treatment techniques. This is not the case of *TwImbData*, where using all the preprocessing techniques there is a slight improvement with respect to the ORIGINAL distribution. The most significant improvement considering AUPR was obtained when NB is used together with SMOTE in the *TwReyes2013*.

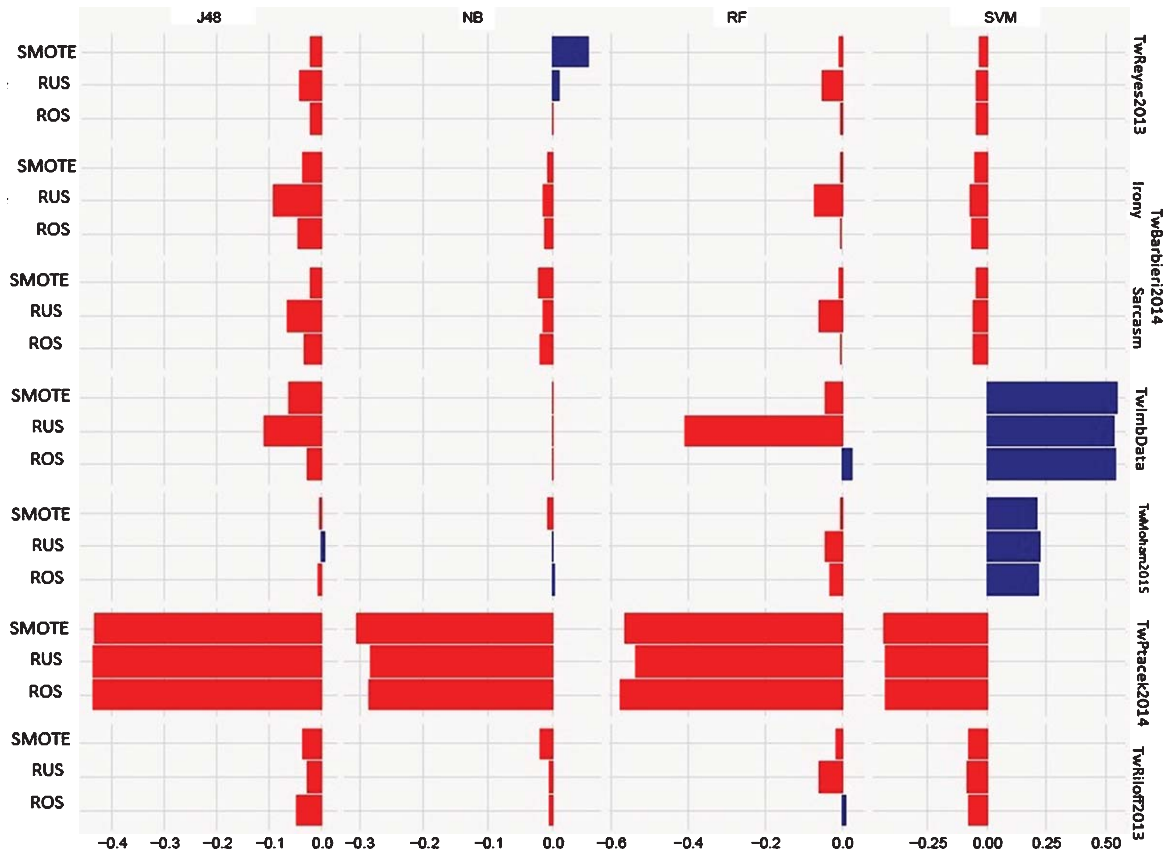

Fig. 7. Differences in terms of Balanced Accuracy with respect to the results of the ORIGINAL distribution after applying treatment techniques.

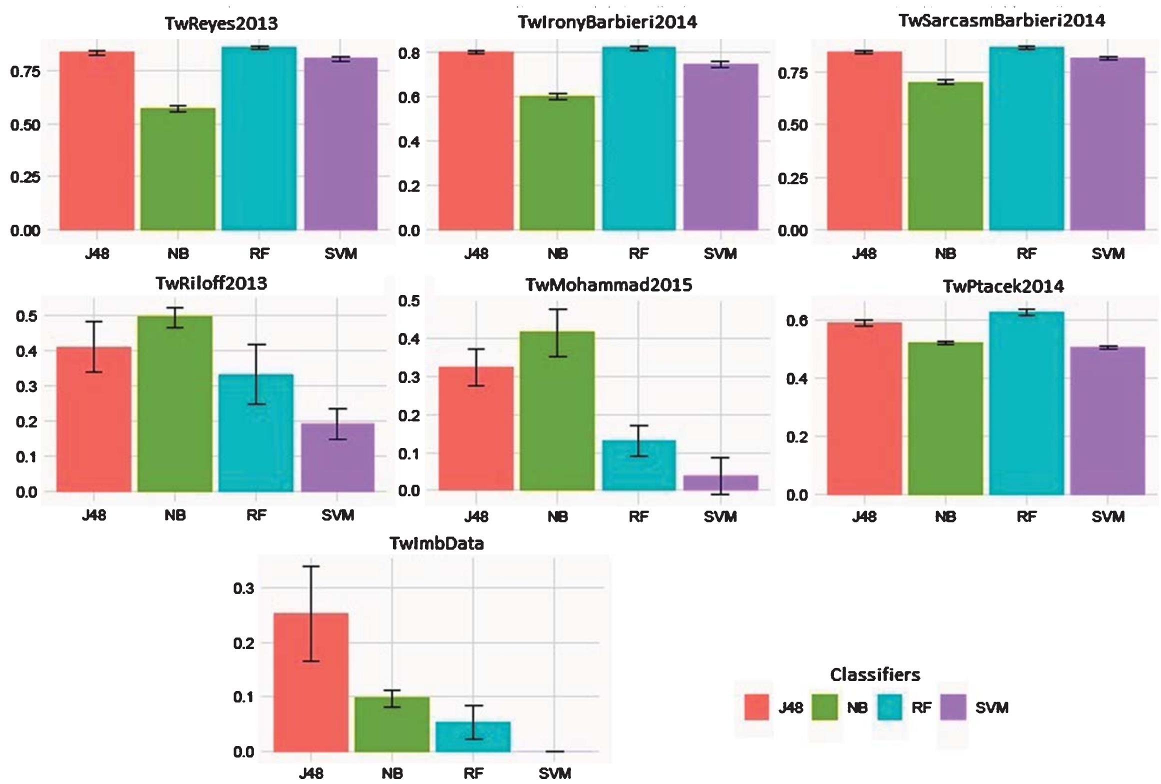

Fig. 8. Obtained results in F-score terms using the original distribution of the corpora.

Balanced Accuracy

Figure 6 shows the results in terms of BAC over the ORIGINAL distribution of the corpora. Overall, the best results were obtained with RF, except in *TwRiloff2013*, where the highest rate of BAC was achieved by applying SVM.

The obtained results after applying imbalance treatment techniques in terms of BAC in most of the cases bring a drawback in the performance of the classifiers. However, in the case of *TwMohammad2015* and *TwImbData* (as it can be observed in Fig. 7) there is a positive impact when the three preprocessing techniques were applied. The performance of the classifiers after applying treatment methods on the *TwPtáček2014* shows the most significative drawback in terms of BAC compared with the ORIGINAL distribution. In terms of BAC, the most noticeable improvement was found over *TwImbData*.

F-score

In Fig. 8 we present the obtained results in terms of F-score (it is the most widely applied evaluation metric in irony detection) when the experimental setting was applied using the ORIGINAL distribution. As it can be noticed, the best performing algorithm in the ORIGINAL distribution was RF in most of the benchmark corpora, particularly in those that have been developed using the self-labeled approach. For what concerns the corpora involving a manual annotation process, the best performing classifier is NB. Concerning *TwImbData*, the J48 classifier obtains the highest results. As can be noticed, the F-score rates on *TwImbData* are lower than in the rest of the corpora reaching only 0.25 in F-score terms, while the highest score was near to 0.80 in *TwReyes2013* and *TwBarbieri2014*.

Figure 9 shows the obtained differences in terms of F-score. When applying the treatment techniques in *TwMohammad2015*, it is possible to improve the results of all classifiers, particularly of SVM. Regarding *TwRiloff2013*, the treatment techniques seem to have a positive impact on most of the experiments except when SMOTE was applied with NB and ROS with J48. Applying treatment techniques together with RF and SVM has a positive impact on the results involving *TwImbData*, while there is a drop in the results in both NB and J48. It is important to highlight that when RUS is used with J48 (the

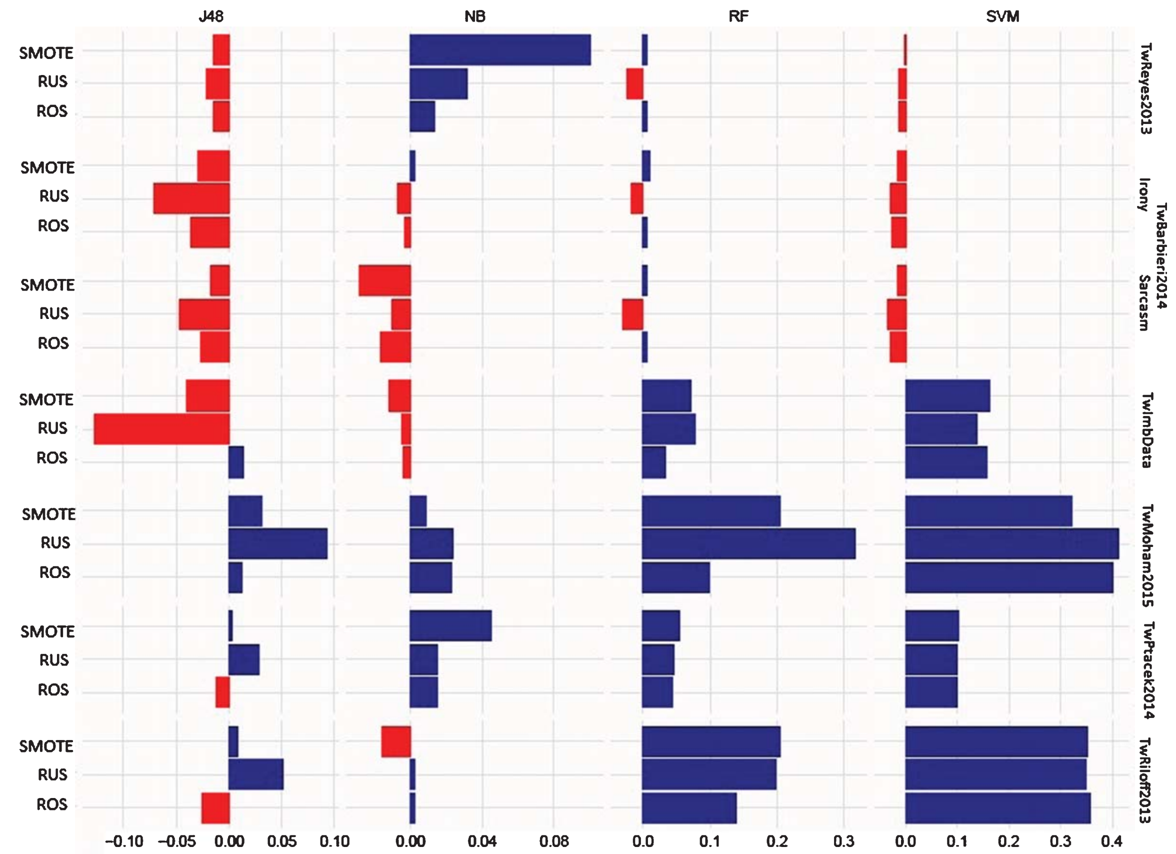

Fig. 9. Differences in terms of F-score with respect to the results of the ORIGINAL distribution after applying treatment techniques.

best performing classifier in the ORIGINAL distribution), its performance decreases. This could serve to validate the fact of the probability of losing useful information due to the nature of this treatment technique.

As already mentioned, F-score has been the most widely applied evaluation metric in the literature on irony detection. Therefore, by using this metric, it is possible to compare the performance of *emotIDM* when applying imbalance treatment techniques. Furthermore, unlike the rest of evaluation metrics used in this paper, it is possible to compare the obtained results against the related work.

Regarding the *TwReyes2013*, it is important to highlight that the experimental setting carried out in [17] for this dataset was different than in this paper. When *emotIDM* was evaluated over the aforementioned corpus, the authors considered a set of binary classifications between the ironic class and as negative instances each of the different subsets of tweets (labeled with #education, #politics, etc.). For comparison purposes on the *TwReyes2013* the results reported in [13, 34] were considered; in these papers, the authors applied a similar setting than ours (i.e., the tweets belonging to the non-ironic classes were merged into a single class, and then a binary classification was carried out). The best performance on the ORIGINAL distribution outperforms the state of the art. Besides, when applying treatment techniques there are other classifiers obtaining better results than in the related work with a rate higher than 0.90 in F-score terms.

For what concerns to *TwBarbieri2014*, it is important to mention that there are not available results considering the same setting than in this paper, therefore it is not possible to compare the obtained results against the literature. In both subsets of *TwBarbieri2014*, the F-score rates are in some way similar to the ones obtained in *TwReyes2013*. Considering the ORIGINAL distribution, the best results were obtained with RF in both cases (irony and sarcasm).

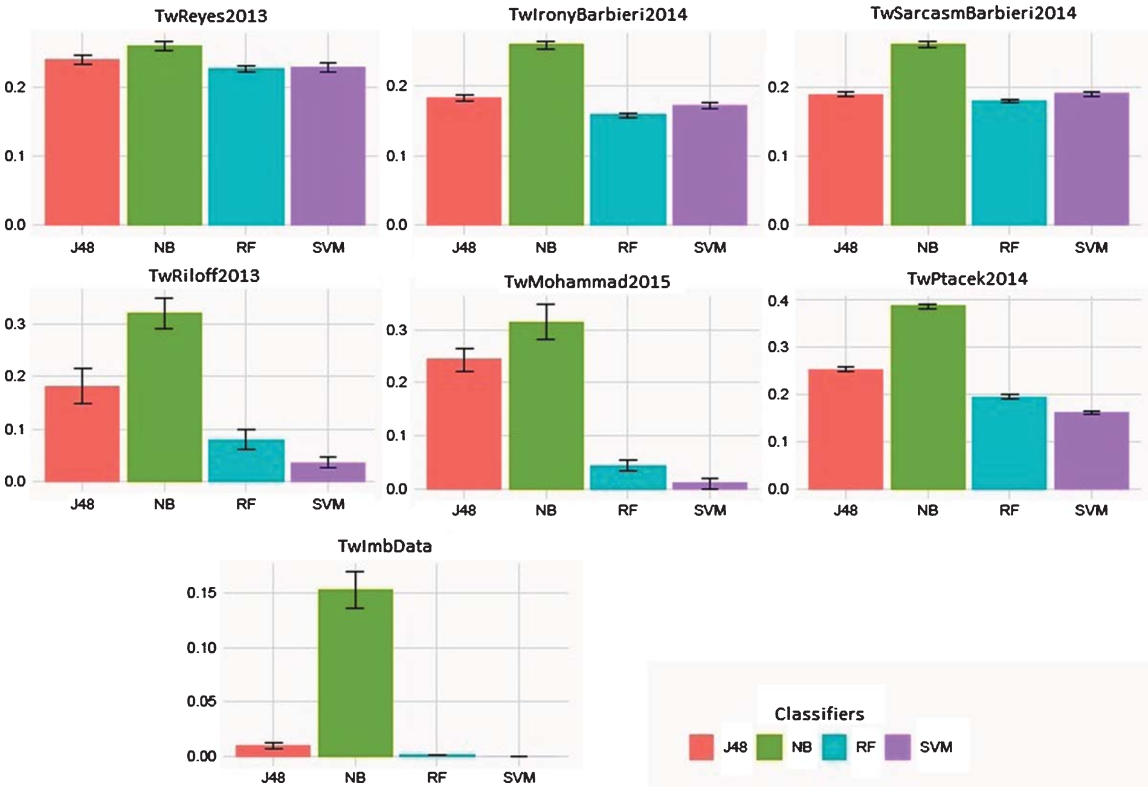

Fig. 10. Obtained results in Predictive Positives Percentage rate terms using the original distribution of the corpora.

Being the results in the *sarcasm-vs-non-sarcasm* experiments slightly better than in the case of *irony-vs-non-irony*.

Finally, in *TwImbData* the worst performing classifier in the ORIGINAL distribution is SVM. However, when the treatment techniques were applied, SVM emerges as the classifier having the best results.

Predictive Positive Percentage

Finally, PPOS was used to show the percentage of instances classified as irony in each experiment. Figure 10 shows the performance in terms of Predictive Positive rate over the ORIGINAL distribution. The best performing classifier in terms of PPOS is NB. While in *TwMohammad2015*, *TwRiloff2013*, and *TwImbData*, SVM shows the worst results.

Figure 11 shows the obtained results after applying the imbalance treatment techniques. As can be observed, in all experiments there is an improvement in terms of PPOS. Overall, the highest results were achieved when applying RUS. While, the worst performance in terms of PPOS was obtained by using RF over *TwImbData* even after applying SMOTE or ROS.

4. Discussion of the results

In this section, we summarize the main findings of the experimental setting carried out applying different imbalance treatment techniques for addressing irony detection.

The RF classifier achieved the best results in terms of AUC, AUPR, and F-score in the case of self-labeled benchmark corpora. On the other hand, SVM showed the worst performance across the experiments, especially in those corpora with a high imbalanced class rate. In a similar fashion than in other domains, applying imbalance treatment techniques to the irony detection corpora before classifying with SVM, leads to an improvement in the performance, particularly in terms of Balanced Accuracy. However, there are some cases where applying imbalance treatment techniques provokes a drop in the performance of some classifiers. In terms of PPOS, it is possible to observe a positive impact on the performance of the classifiers, especially for *TwMohammad2015*, *TwRiloff2013*, *TwPtáček2014*, and *TwImbData*.

According to the results presented before, for each of the evaluation metrics, different imbalance treat-

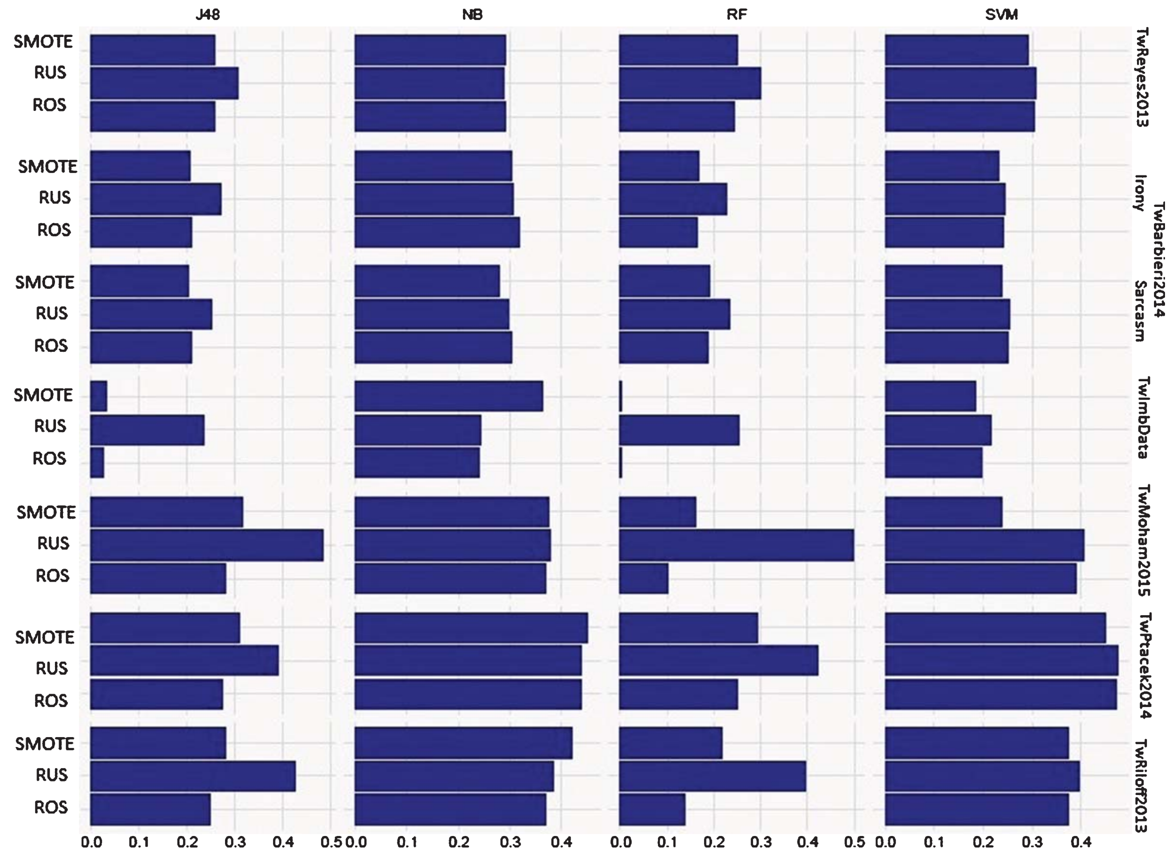

Fig. 11. Obtained results in terms of Predictive Positives Percentage rate considering the ORIGINAL distribution as well as applying treatment techniques.

ment techniques allow to improve the results of the ORIGINAL distribution. SMOTE obtains the best performance in terms of both AUPR and Balanced Accuracy. Considering F-score, RUS is the method allowing the best results. In terms of AUC, ROS obtained the highest outcomes.

The corpora we used for experimental purposes could be divided according to different aspects, for example, considering the criteria used for retrieving the data. The results in terms of PPOS in the ORIGINAL distribution seem to be higher when #sarcasm is considered for retrieving data than in the case or #irony.

Another aspect that can be considered within the corpora we used concerns exploiting author's self-labeled intention of being ironic (*TwReyes2013*, *TwBarbieri2014*, *TwPtáček2014*, and *TwImbData*), and the use of a manual annotation process (*TwRiloff2013* and *TwMohammad2015*). In this case, the results in terms of F-score in self-labeled corpora are higher than in manually annotated data. This could serve to validate the similar findings observed in [16, 17] with reference to the impact of the corpora construction methodology. However, on the other hand, the most noticeable improvements on applying imbalance treatment techniques to compensate the imbalance degree were achieved in those corpora involving manual annotation.

Regarding the obtained results over *TwImbData*, it is important to highlight that in all the evaluation metrics considered in this paper, there is a positive impact on the performance of at least one of the classifiers. *TwImbData* was developed having in mind to resemble a realistic scenario where the difference between ironic and non-ironic instances is very big. Therefore, by improving the results over the ORIGINAL distribution when the treatment methods were applied we confirm the usefulness of using such techniques for irony detection in imbalanced class scenarios.

Being irony a complex phenomenon, it is important to assess the performance of different preprocessing methods for compensating imbalance degree. As it can be noticed, there is not a single imbalance treatment technique allowing to have the best performance across the evaluation metrics and corpora. This could be related to the nature of each method and also to the aim of the metrics.

As already mentioned, the most widely used evaluation metric in irony detection is F-score. Considering such a measure, the best performing technique was RUS, which serves to remove, in this case non-ironic samples. In our experimental setting, after applying imbalance treatment techniques both classes became balanced. Therefore, by applying RUS we are neither losing representative ironic instances nor generating synthetic instances.

Finally, it is important to highlight that the experiments were carried out only considering an irony detection model (*emotIDM*) relying mainly on affective features. It could be interesting to evaluate the performance of imbalance treatment techniques when ironic instances are represented by other kinds of features.

5. Conclusions

In this paper, we have evaluated the impact of class imbalance on detecting irony. We have performed several experiments over a set of Twitter corpora for irony detection covering different aspects such as corpora construction methodology and differences in data skew. Besides, we developed a set of irony corpora[9] aimed to resemble a more realistic scenario where the difference between the ironic and non-ironic class is very big. We employed *emotIDM*, an irony detection model based mainly on the presence of affective content. To the best of our knowledge, this is the first work in irony detection where a model for detecting such figurative language device is evaluated by considering many aspects related to the class imbalance problem.

In our research, we evaluated the performance of *emotIDM* together with a variety of classifiers when different imbalance treatment techniques were applied. Several metrics were used to compare the effectiveness of different classifiers and imbalance treatment techniques. Our results also allow us to compare the obtained results against those of the state of the art.

The main objective of this paper was to show that some treatment techniques can improve the performance of classifiers dedicated to detect irony in Twitter particularly under an imbalanced class scenario. The results of this study indicate that the best performing imbalance treatment technique for addressing irony detection in imbalanced class scenarios depends on the evaluation metric used. However, considering the most widely used metric, i.e. F-score, the best performance was achieved by applying RUS.

We identified some directions for future work. It could be interesting to carry out some experiments using not only *data level approaches* (such as ROS, RUS, and SMOTE) but also *algorithm level approaches* (such as for example cost sensitive learning). Furthermore, experiments with other imbalance degree rates over the set of corpora used is part of the following steps of our research in irony detection in imbalanced class scenarios. On the other hand, it could be interesting to analyze the role of some of the data intrinsic characteristics described in [29] such as small disjuncts, lack of density and information as well as the overlapping between the classes on the irony detection corpora.

[9]The data will be released for research purposes.

Acknowledgments

The first author was funded by CONACYT project FC-2016/2410. Ronaldo Prati was supported by the São Paulo State (Brazil) research council FAPESP under project 2015/20606-6. Francisco Herrera was partially supported by the Spanish National Research Project TIN2017-89517-P. The work of Paolo Rosso was partially supported by the Spanish MICINN under the research project MISMIS (PGC2018-096212-B-C31) and by the Generalitat Valenciana under the grant PROMETEO/2019/121.

References

[1] G. Abercrombie and D. Hovy, Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations. In *Proceedings of the ACL 2016 Student ResearchWorkshop*, pp. 107–113, Germany, August (2016). ACL.

[2] L. Alba-Juez and S. Attardo, The Evaluative Palette of Verbal Irony, In Geoff Thompson and Laura Alba-Juez, editors, Evaluation in Context, pages 93–116. John Benjamins Publishing Company, Amsterdam/Philadelphia, (2014).

[3] A. Ali, S.M. Shamsuddin and A.L. Ralescu, Classification with Class Imbalance Problem: A Review, *Intertional Jour-*

nal Of Advances In Soft Computing And Its Applications **7**(3) (2015), 176–204.

[4] S. Attardo, Irony as Relevant Inappropriateness, In H. Colston and R. Gibbs, editors, Irony in language and thought: A cognitive science reader, pages 135–172. Lawrence Erlbaum, (2007).

[5] D. Bamman and N.A. Smith, Contextualized Sarcasm Detection on Twitter, In *Proceedings of the Ninth International Conference on Web and Social Media*, (2015), pp. 574–577.

[6] F. Barbieri, H. Saggion and F. Ronzano, Modelling Sarcasm in Twitter, A Novel Approach. In *Proc. of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 50–58, USA, June (2014). ACL.

[7] G.E. de Almeida Prado Alves Batista, R.C. Prati and M.C. Monard, A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, *ACM Sigkdd Explorations Newsletter* **6**(1) (2004), 20–29.

[8] M.M. Bradley and P.J. Lang, Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings, Technical Report, Center for Research in Psychophysiology, University of Florida, Florida, (1999).

[9] E. Cambria, D. Olsher and D. Rajagopal, SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis, In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 1515–1521, Canada, (2014). AAAI.

[10] A. Cervone, E.A. Stepanov, F. Celli and G. Riccardi, Irony Detection: From the Twittersphere to the News Space, In *Proceedings of Fourth Italian Conference on Computational Linguistics* (CLiC-it 2017), (2017).

[11] N.V. Chawla, K.W. Bowyer, L.O. Hall and W. Philip Kegelmeyer, SMOTE: Synthetic Minority Oversampling TEechnique, *Journal of Artificial Intelligence Research* **16** (2002), 321–357.

[12] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk and F. Herrera, Learning from imbalanced data sets, *Springer*, (2018).

[13] E. Fersini, F.A. Pozzi and E. Messina, Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers, In *International Conference on Data Science and Advanced Analytics*, (DSAA 2015), pages 1–8, France, (2015). IEEE Xplore Digital Library.

[14] H. Paul Grice, Logic and Conversation. In P. Cole and J.L. Morgan, editors, Syntax and Semantics: Vol. **3** Speech Acts, pages 41–58. Academic Press, (1975).

[15] H. He and E.A. Garcia. Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering,* **21**(9) (2009), 1263–1284.

[16] D.I.H. Farías, M. M.-y. Gómez, H.J. Escalante, P. Rosso and V. Patti, A Knowledge-based Weighted KNN for Detecting Irony in Twitter, In *17th Mexican International Conference on Artificial Intelligence (MICAI 2018)*, Mexico, (2018).

[17] D.I.H. Farías, V. Patti and P. Rosso, Irony detection in Twitter: The Role of Affective Content, *ACM Trans Internet Technol* **16**(3) (2016), 19:1–19:24.

[18] D.I.H. Farías and P. Rosso, Irony, Sarcasm, and Sentiment Analysis. Chapter 7. In Federico A. Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 113–127. Morgan Kaufmann, (2016).

[19] M. Hu and B. Liu, Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (2004), pages 168–177, USA, ACM.

[20] N. Japkowicz and S. Stephen, The Class Imbalance Problem: A Systematic Study, *Intelligent Data Analysis* **6**(5) (2002), 429–449.

[21] A. Joshi, P. Bhattacharyya and M.J. Carman, Automatic Sarcasm Detection: A Survey. CoRR, abs/1602.03426, (2016).

[22] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya and M.J. Carman, Are Word Embeddingbased Features Useful for Sarcasm Detection? In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, (2016), pages 1006–1011.

[23] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles and L. Hadrich-Belguith, Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 644–650, China, (2015). ACL.

[24] A. Khattri, A. Joshi, P. Bhattacharyya and M. Carman, Your Sentiment Precedes You: Using an author's Historical Tweets to Predict Sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–30, Portugal, September (2015). ACL.

[25] S. Kumon-Nakamura and S. Glucksberg, How About Another Piece of Pie: The Allusional Pretense Theory of Discourse Irony, *Journal of Experimental Psychology: General* **124**(1) (1995), 3.

[26] F. Kunneman, C. Liebrecht, M. van Mulken and A. van den Bosch, Signaling Sarcasm: From Hyperbole to Hashtag, *Information Processing & Management* **51**(4) (2015), 500–509.

[27] C. Liebrecht, F. Kunneman and A. Van den Bosch, The Perfect Solution for Detecting Sarcasm in Tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, USA, June (2013). ACL.

[28] P. Liu, W. Chen, G. Ou, T. Wang, D. Yang and K. Lei, Sarcasm Detection in Social Media Based on Imbalanced Classification. In Feifei Li, Guoliang Li, Seungwon Hwang, Bin Yao, and Zhenjie Zhang, editors, *Proceedings of the Web-Age Information Management: 15th International Conference*, pages 459–471, China, (2014). Springer International Publishing.

[29] V. López, A. Fernández, S. García, V. Palade and F. Herrera, An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics, *Information Sciences* **250** (2013), 113–141.

[30] S. McDonald, Neuropsychological Studies of Sarcasm. In H. Colston and R. Gibbs, editors, *Irony in language and Thought: A Cognitive Science Reader*, pages 217–230. Lawrence Erlbaum, (2007).

[31] S.M. Mohammad and P.D. Turney, Crowdsourcing a Word–emotion Association Lexicon, *Computational Intelligence* **29**(3) (2013), 436–465.

[32] S.M. Mohammad, X. Zhu, S. Kiritchenko and J. Martin, Sentiment, Emotion, Purpose, and Style in Electoral Tweets, *Information Processing & Management* **51**(4) (2015), 480–499.

[33] F.Å. Nielsen, A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs, In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume **718**

of CEUR Workshop Proceedings, pages 93–98, Greece, (2011). CEUR-WS.org.

[34] D. Nozza, E. Fersini and E. Messina, Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings, In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages (2016), 68–76.

[35] J.W. Pennebaker, M.E. Francis and R.J. Booth, Linguistic Inquiry and Word Count: LIWC 2001, *Mahway: Lawrence Erlbaum Associates* **71** (2001), 2–23.

[36] S. Poria, E. Cambria, D. Hazarika and P. Vij, A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks, In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, page 1601âĂŞ1612, Japan, December (2016).

[37] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das and S. Bandyopadhyay, Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining, *IEEE Intelligent Systems* **28**(2) (2013), 31–38.

[38] R.C. Prati, Gustavo Enrique de Almeida Prado Alves Batista, and Diego F. Silva. Class Imbalance Revisited: A New Experimental setup to Assess the Performance of Treatment Methods, *Knowledge and Information Systems* **45**(1) (2015), 247–270.

[39] T. Ptáček, I. Habernal and J. Hong, Sarcasm Detection on Czech and English Twitter, In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 213–223, Ireland, August (2014). Dublin City University and ACL.

[40] A. Reyes, P. Rosso and T. Veale, A Multidimensional Approach for Detecting Irony in Twitter, *Language Resources and Evaluation* **47**(1) (2013), 239–268.

[41] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert and R. Huang, Sarcasm as Contrast between a Positive Sentiment and Negative Situation, In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (EMNLP 2013), pages 704–714. USA, October (2013). ACL.

[42] E. Sulis, D.I.H. Farías, P. Rosso, V. Patti and G. Ruffo, Figurative Messages and Affect in Twitter: Differences between #irony, #sarcasm and #not, *Knowledge-Based Systems* **108** (2016), 132–143.

[43] A. Utsumi, Verbal Irony as Implicit Display of Ironic Environment: Distinguishing Ironic Utterances from Nonirony, *Journal of Pragmatics* **32**(12) (2000), 1777–1806.

[44] B.C. Wallace, D.K. Choe and E. Charniak, Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, *Entities and Sentiment. In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, China, July 2015. ACL.

[45] A.P. Wang. #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*, pages 349–356, Taiwan, (2013).

[46] C. Whissell, Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Languages, *Psychological Reports* **2**(105) (2009), 509–521.

[47] D. Wilson and D. Sperber, On Verbal Irony, *Lingua* **87**(1-2) (1992), 53–76.