Irony Detection in Twitter: The Role of Affective Content

DELIA IRAZÚ HERNÁNDEZ FARÍAS, PRHLT Research Center, Universitat Politècnica de València; Dipartimento di Informatica, University of Turin
VIVIANA PATTI, Dipartimento di Informatica, University of Turin
PAOLO ROSSO, PRHLT Research Center, Universitat Politècnica de València

Irony has been proven to be pervasive in social media, posing a challenge to sentiment analysis systems. It is a creative linguistic phenomenon where affect-related aspects play a key role. In this work, we address the problem of detecting irony in tweets, casting it as a classification problem. We propose a novel model that explores the use of affective features based on a wide range of lexical resources available for English, reflecting different facets of affect. Classification experiments over different corpora show that affective information helps in distinguishing among ironic and nonironic tweets. Our model outperforms the state of the art in almost all cases.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Social media**

Additional Key Words and Phrases: Irony detection, figurative language processing, affective resources

ACM Reference Format:
Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. ACM Trans. Internet Technol. 16, 3, Article 19 (June 2016), 24 pages.
DOI: http://dx.doi.org/10.1145/2930663

1. INTRODUCTION

The huge amount of information streaming from online social networking and microblogging platforms such as Twitter is increasingly attracting the attention of researchers in the area of sentiment analysis. Twitter communications include a high percentage of ironic devices [Davidov et al. 2010; Veale and Hao 2010; González-Ibáñez et al. 2011; Reyes et al. 2013; Reyes and Rosso 2014], and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification of ironic messages [Bosco et al. 2013; Ghosh et al. 2015]. Indeed, the presence of ironic devices in a text can flip the polarity of an opinion expressed with positive words to the intended negative meaning (one says something "good" to mean something "bad"), or vice versa, working as an unexpected polarity reverser. This can undermine

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Farías (Grant No. 218109/313683 CVU-369616). The work of Viviana Patti was partially carried out at the Universitat Politècnica de València within the framework of a fellowship of the University of Turin cofunded by Fondazione CRT (World Wide Style Program 2). The work of Paolo Rosso has been partially funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).
Authors' addresses: D. I. H. Farías and P. Rosso, Pattern Recognition and Human Language Technology, DSIC, Building 1F - Campus de Vera Universitat Politècnica de València Camino de Vera, s/n - 46022 València, Spain; emails: {dhernandez1, prosso}@dsic.upv.es; V. Patti, Dipartimento di Informatica, University of Turin, Corso Svizzera 185, 10149, Turin, Italy; email: patti@di.unito.it.

DOI: http://dx.doi.org/10.1145/2930663

systems' accuracy. The automatic detection of irony is, therefore, crucial for the development of irony-aware sentiment analysis systems, but at the same time it is also an interesting conceptual challenge from a cognitive point of view and can help to shed some light on how human beings use irony as a communicative tool.

Irony has been a topic studied by various disciplines, such as linguistics, philosophy, and psychology, but it is difficult to define it in formal terms. There is no consensus on a single definition, and different accounts shed light on relevant aspects of a creative and complex linguistic phenomenon. However, most theorists would agree that emotions play a role in the use of irony in different respects, and the important role of affective information for irony communication-comprehension is also emphasized by recent psychological findings [Leggitt and Gibbs 2000; Shamay-Tsoory et al. 2005].

Linguistic devices such as irony and sarcasm allow users to express themselves by using words in a creative and nonliteral sense. They are intimately connected with the expression of affective contents such as feelings, emotions, attitudes, or evaluations [Grice 1975; Wilson and Sperber 1992; Alba-Juez and Attardo 2014] toward a particular target (e.g., a person, an event, but also a product or a movie when we consider social media texts). In irony, people express affective contents in an indirect way, since the critical or praising attitudinal load they communicate is on top of what they explicitly say. According to the Gricean tradition [Grice 1975], the function of irony is to effectively communicate the opposite of the literal interpretation of the utterance. Furthermore, an ironic statement can elicit affective reactions. For instance, ironic criticism (or sarcasm) has been recognized in Bowes and Katz [2011] with a specific target to attack, as offensive [Lee and Katz 1998], and as "intimately associated with particular negative affective states" [McDonald 2007]. It may enhance the negative emotions felt by the recipient, such as anger, irritation, or disgust [Leggitt and Gibbs 2000], and it can be hypothesized that the use of such a figurative device also conveys information on the speaker's attitude toward the target. On the other hand, there are cases where irony may reduce the strength of a statement; that is, criticism becomes gentler or less negative, and praise less positive or more ambivalent, if phrased ironically [Dews et al. 1995]. Overall, the affective information involved in ironic communications is multifaced, involving aspects related to the emotional state of the ironist and of the recipient, and issues related to the evaluative meaning of the ironic utterance, that is, to the expression of a positive or negative opinion toward a target.

There is now a consistent body of work on computational models for irony and sarcasm detection in social media [González-Ibáñez et al. 2011; Reyes et al. 2013; Wang 2013; Riloff et al. 2013; Barbieri et al. 2014; Ptáček et al. 2014; Hernández Farías et al. 2015], and in particular in Twitter, which can be considered the most widely used source of information to experiment with irony detection. In this article, we also address the task of detecting irony in tweets, by identifying a set of discriminative features to automatically differentiate an ironic text from a nonironic one. In line with most of the current approaches and with some theoretical accounts [Gibbs 2000; Whalen et al. 2013], irony is here considered an umbrella term that also covers sarcasm, with the issue of discriminating between the two devices being a further challenge for figurative language processing. Our irony detection model, called emotIDM, extends the model proposed in Hernández Farías et al. [2015] with new features, in particular experimenting with the use of a wide range of psycholinguistic and affective features concerning affective information, with the main aim to answer our main research questions: (1) Does information about different facets of affect help in distinguishing among ironic and nonironic tweets? (2) Which facets of affect seem to be more important in order to address our classification task? Affective information expressed in our texts is multifaceted. Both sentiment and emotion lexicons, and psycholinguistic resources available for English, refer to various affective models. In our view, all such resources

represent a rich and varied lexical knowledge about affect, under different perspectives. Therefore, we propose here a comprehensive study of their use in the context of our analysis, in order to test if they convey relevant knowledge to distinguish between ironic and nonironic messages. To our knowledge, this is the first work that addresses the issue by considering different facets of the affective content, taking advantage of the wide availability of lexical resources for English covering the various perspectives. Such facets include sentiment polarity aspects related to the polarity of words, but also finer-grained ones, related to the writer's emotional state or to emotions evoked in the reader, which can be captured according to different categorical or dimensional models of emotions.

Another novelty of our proposal is that we evaluated our model over six different Twitter corpora developed in previous work on irony and sarcasm detection, without creating our own dataset. This is important not only in order to carry out a fair evaluation of our model against the state-of-the-art approaches but also to test the robustness under different datasets, where samples of ironic utterances were collected by using different criteria (i.e., different hashtags).

The evaluation of our model for irony detection over a set of Twitter corpora already used in the same task confirms the significance of affective features for irony detection. Experimental results show that emotIDM outperforms the irony detection models presented in Riloff et al. [2013], Reyes et al. [2013], Barbieri et al. [2014], and Hernández Farías et al. [2015] over the same datasets.

Contributions. Summarizing, the main contributions of this article are the following: (1) We propose a new approach to irony detection emotIDM based on Hernández Farías et al. [2015] that exploits affective information as features to represent ironic tweets; (2) we evaluate emotIDM carrying out a battery of binary classification experiments over a set of Twitter corpora, developed in different ways for both what concerns the selection criteria for samples of irony/sarcasm and the annotation methodology—this is important in order to validate the robustness of the model and to better compare results with the state of the art; (3) we demonstrate that affective information helps in distinguishing among ironic and nonironic tweets, presenting a comparative evaluation of the performances over the various corpora, and a feature analysis in order to identify the most useful features in emotIDM.

Organization. The article is structured as follows. Section 2 describes related work in irony detection. Section 3 presents a set of Twitter corpora developed in the literature for evaluating previously proposed models in irony detection. Section 4 introduces our starting point, the IDM model in Hernández Farías et al. [2015], and the new proposal, emotIDM, which enriches IDM with affective features. In Section 5, we describe a set of experiments carried out over the set of corpora by using both models for irony detection, as well as an information gain analysis to identify the most relevant features in emotIDM. Finally, in Section 6, we conclude with final remarks and future work.

2. RELATED WORK

Different approaches to the task of recognizing verbal irony in texts have been developed. The majority of them take advantage only of the textual content itself, since in textual messages other paralinguistic cues, like for instance the tone or corporal movements, are not available. Twitter is the most widely used source of information to experiment with irony detection. This is mainly due to availability of a large set of samples of ironic texts, which are easy to be collected relying on the behavior of Twitter users, who often explicitly mark their ironic messages by using hashtags such as #irony or #sarcasm. The pretty good reliability of the user-generated hashtags as golden labels for irony has been experimentally confirmed by Kunneman et al. [2015]. Moreover, it

seems that, due to the interaction model underlying the microblogging platform, irony expressed here could be somehow easier to analyze. Indeed, Twitter users have to be sharp and short, having only 140 characters for expressing their comments, and most of the time the ironic posts do not require knowledge about the conversational context to be understood. Several works have been carried out using tweets for experimental purposes [Davidov et al. 2010; González-Ibáñez et al. 2011; Reyes et al. 2013; Wang 2013; Riloff et al. 2013; Barbieri et al. 2014; Ptáček et al. 2014; Hernández Farías et al. 2015; Rajadesingan et al. 2015; Bamman and Smith 2015; Joshi et al. 2015; Karoui et al. 2015]. Furthermore, there are some efforts in other social media such as customer reviews from Amazon[1] [Filatova 2012; Buschmeier et al. 2014], comments from the online debate sites such as 4forums.com[2] [Abbott et al. 2011; Lukin and Walker 2013], and, recently, Reddit[3] [Wallace et al. 2015].

The majority of the research in irony detection has been addressed in English, although there is some research in other languages, such as Dutch [Kunneman et al. 2015], Italian [Bosco et al. 2013], Czech [Ptáček et al. 2014], French [Karoui et al. 2015], Portuguese [Carvalho et al. 2009], and Chinese [Tang and Chen 2014]. A shared task for English on sentiment analysis of figurative language in Twitter has been organized at SemEval-2015 for the first time [Ghosh et al. 2015], and a pilot shared task for Italian on irony detection has been proposed in Sentipolc-2014 within the periodic evaluation campaign EVALITA [Basile et al. 2014; Attardi et al. 2015]. This confirms the growing interest for this task in the research community, especially for understanding the impact of the ironic devices on sentiment analysis.

Irony detection has been modeled as a binary classification problem, where mostly tweets labeled with certain hashtags (i.e., #irony, #sarcasm, #sarcastic, #not) have been considered as ironic utterances. Following this framework, different approaches have been proposed [Davidov et al. 2010; González-Ibáñez et al. 2011; Reyes et al. 2013; Riloff et al. 2013; Barbieri et al. 2014; Ptáček et al. 2014; Hernández Farías et al. 2015; Fersini et al. 2015]. The authors proposed models that exploit mainly textual content such as punctuation marks, emoticons, part-of-speech labels, discursive terms, and specific patterns (e.g., according to Riloff et al. [2013], a common form of sarcasm in Twitter consists of a positive sentiment contrasting with a negative situation), among others.

Another key characteristic for irony is *unexpectedness* [Attardo 2000]. According to many theoretical accounts, people infer irony when they recognize an incongruity between an utterance and what is known (or expected) about the speaker and/or the environment. This is something that can be referred to as the pragmatic context. Recent approaches started to address such an issue, taking into account information about context [Rajadesingan et al. 2015; Bamman and Smith 2015; Wallace et al. 2015].

For what concerns the affective information, some approaches already used some kind of sentiment and emotional information in their models. Reyes et al. [2013] included in their model some features to characterize irony in terms of elements related to sentiments, attitudes, feelings, and moods exploiting the Dictionary of Affect in Language proposed by Whissell [2009]. Barbieri et al. [2014] considered the amount of positive and negative words by using SentiWordNet [Baccianella et al. 2010]. Hernández Farías et al. [2015] exploited two widely applied sentiment lexicons, Hu&Liu and AFINN,[4] as features in their model. However, no previous work focused specifically on studying the role of affective information in a comprehensive manner, by exploring

[1]http://www.amazon.com/.

[2]http://www.4forums.com/political/.

[3]http://www.reddit.com.

[4]Hu&Liu: http://www.cs.uic.edu/~liub/FBS; AFINN: http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt.

the use of a wide range of lexical resources available for English, reflecting different aspects of a multifaceted phenomenon.

3. EVALUATION DATASETS

Annotated data are a crucial source of information to capture the real use of irony in social media. Large corpora providing annotations marking whether an expression is ironic or not are scarce [Buschmeier et al. 2014; Tang and Chen 2014]. Therefore, in general, the authors have built their own corpora for evaluating the proposed models. This constitutes a problem for establishing a fair comparison, and this is the reason we decided to follow here a different approach, by evaluating our model against a set of already available Twitter corpora that have been developed in related work on irony detection. We observed that there are two main approaches that have been used for creating corpora for irony detection: self-tagging and crowdsourcing.

Self-Tagging. Twitter allows users to communicate ideas in short messages and to assign labels (i.e., hashtags) to their own messages. The "self-tagging" approach considers as positive instances those tweets in which the author points out his or her intention using an explicit label. For instance, the hashtags #irony and #sarcasm can be considered as markers of irony, which rely on the author's definition about what irony is. The underlying assumption is that the best judge of whether a tweet is intended to be ironic is the author of the tweet [González-Ibáñez et al. 2011; Reyes et al. 2013]. Furthermore, some experiment shows that self-labeled tweets allow one to produce good-quality gold standards [Kunneman et al. 2015]. However, it is worth noticing that Twitter users do not use hashtags to mark explicitly the intention to be ironic in all languages. For instance, both Czech and Italian users generally do not use the sarcasm (i.e., #sarkasmus in Czech, #sarcasmo in Italian) or irony (#ironie in Czech or #ironia in Italian) hashtag variants; thus, in such cases, relying on simple self-tagging is not an option [Ptáček et al. 2014; Bosco et al. 2013].

Crowdsourcing. The "crowdsourcing" approach involves human interaction by labeling the content as ironic or nonironic. Mainly, the labeling process is carried out without any strict definition or guideline. Therefore, it represents a subjective task, where the agreement between annotators is often very low.

Next, we describe six corpora that have been created by using the methodologies depicted earlier. In Reyes et al. [2013], Barbieri et al. [2014], and Ptáček et al. [2014], the authors took advantage of the presence of hashtags to create the corpus and evaluate their models. Likewise, in Mohammad et al. [2015], data were manually annotated by using crowdsourcing with information related to irony, and annotators were asked to decide whether a tweet was ironic or not, whereas in Riloff et al. [2013] and Mohammad et al. [2015], a mixed approach was taken.

TwReyes2013. Reyes et al. [2013] retrieved a set of 40,000 tweets by using the "self-tagging" criterion. They selected four hashtags: #irony to get ironic instances (or at least tweets written by Twitter users with an intuitive definition of what irony is) and #education, #humor, and #politics to retrieve a large sample of nonironic tweets on different topics. This corpus (henceforth TwReyes2013) contains 10,000 ironic tweets and 30,000 nonironic tweets.

TwBarbieri2014. Barbieri et al. [2014] introduced a Twitter dataset constructed following a methodology similar to Reyes et al. [2013]. Overall, it includes 60,000 tweets equally divided into six different classes: education, humor, politics, newspaper, irony, and sarcasm. For what concerns the first three categories (education, humor, and politics), the authors reused samples from the TwReyes2013. The irony and sarcasm tweets were collected by using the #irony and #sarcasm hashtags, respectively.

In the following, we will use TwIronyBarbieri2014 to refer to a corpus where irony-laden tweets are sampled by the irony class of TwBarbieri2014, whereas we will use TwSarcasmBarbieri2014 to denote a different corpus where they are sampled by the sarcasm class. In both corpora, the nonironic samples are tweets from the education, humor, politics, and newspaper classes.

TwRiloff2013. Riloff et al. [2013] created a manually annotated corpus from Twitter including 3,200 tweets (henceforth TwRiloff2013). They followed a mixed approach for developing a corpus of samples including ironic and nonironic tweets. First, a set of tweets tagged with the #sarcasm and #sarcastic hashtags as well as tweets without these hashtags were retrieved (self-tagging methodology). Then, three annotators were asked to manually annotate the collected tweets by omitting the hashtags. Annotation guidelines asked users to label a tweet as sarcastic if it contains comments judged to be sarcastic based solely on the content of that tweet.

TwPtáček2014. In the work by Ptáček et al. [2014], two datasets were collected: in Czech and English. The first one involved manual annotation of tweets.[5] Instead, for the English dataset, the hashtag #sarcasm was used as an indicator of sarcastic tweets (henceforth TwPtáček2014); for the nonsarcastic samples, the authors collected tweets from the general Twitter stream using as a parameter only the language (English). Two different distribution scenarios were created for the English dataset: balanced (composed by 50,000 sarcastic and 50,000 nonsarcastic tweets) and imbalanced (composed by 25,000 sarcastic and 75,000 nonsarcastic tweets).

TwMohammad2015. The TwMohammad2015 corpus [Mohammad et al. 2015] contains a set of tweets with a multilayer annotation concerning different aspects: sentiment (positive or negative), emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), purpose (to point out a mistake, to support, to ridicule, etc.), and style (simple statement, sarcasm, hyperbole, understatement). Note that only 23.01% of the tweets were labeled with a style tag pertinent to the expression of irony, whereas most of them were annotated with the label *simple statement*, which can be interpreted as a tag for marking nonironic expressions. The authors collected tweets labeled with a set of hashtags pertaining to the 2012 US presidential elections.[6] The tweets were annotated by relying on crowdsourcing platforms.

The next sections describe the experimental setting and results obtained over these corpora. A summary of their features is reported in Table I.[7] Most of the previously described corpora were created for evaluating irony detection models presented in related work. TwMohammad2015 is the only one designed for purposes that go beyond irony detection, that is, for predicting emotion and purpose labels in tweets. Most of the corpora rely on self-annotation of tweets, but we have also samples of corpora manually annotated by using crowdsourcing platforms. The datasets were developed based on criteria that are different for what concerns the choice of the hashtags or the guidelines for manual annotation. Such variety of aspects makes it particularly interesting to use all the datasets in order to evaluate our proposal, which is described in the next section. Indeed, our model will be evaluated by using tweets coming from different scenarios (e.g., tweets in TwMohammad2015 pertain to the political domain), collected with different methodologies. This allows us to test the robustness of the approach across a wide set of irony samples, which represent a rich variety of use of ironic devices.

[5]For more details about the Czech dataset, see Ptáček et al. [2014].

[6]Some of the hashtags used are #election2012, #election, #campaing2012, and #president2012.

[7]Note that for some corpora only the IDs of the tweet coupled with the annotation were available. Thus, we had to retrieve again the text of the tweet by Twitter API at experiment time, but some data were not available anymore (deleted tweets or canceled accounts).

Table I. Evaluation Datasets

Ironic	Nonironic	Labeling Criterion	Hashtag
TwReyes2013			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics)	Self-tagging	#irony
TwIronyBarbieri2014			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics) 10,000 (#newspaper)	Self-tagging	#irony
TwSarcasmBarbieri2014			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics) 10,000 (#newspaper)	Self-tagging	#sarcasm
TwPtáček2014			
19,026	51,860	Self-tagging	#sarcasm
TwMohammad2015			
532	1,397	Crowdsourcing	-
TwRiloff2013			
474	1,689	Self-tagging/Crowdsourcing	#sarcasm #sarcastic

4. OUR APPROACH: THE EMOTIDM MODEL

We addressed the task of irony detection as a classification problem, applying supervised machine learning to the set of corpora described in the previous section. To represent each tweet, we use different group of features: some of them (structural features, henceforth) are designed to detect common patterns in the structure of the ironic tweets (e.g., type of punctuation, length, emoticons); others are designed to detect affective information (affective features, henceforth).

In this section, we will recall the main characteristics of the irony detection model to identify ironic tweets [Hernández Farías et al. 2015], which is our starting point (IDM henceforth). Then, we will present emotIDM, which enriches IDM with additional features, with a special focus on features that exploit information about affect.

It is important to highlight that in this work, irony and sarcasm are considered as synonyms, a common assumption in computational linguistic approaches to irony detection [Davidov et al. 2010; Filatova 2012; Reyes et al. 2013; Maynard and Greenwood 2014; Ptáček et al. 2014]. Moreover, the approach proposed here does not rely on bag-of-words (BOW). We consider that irony detection should be addressed by models based mainly on features that allow one to capture irony disregarding domain or topic, since our aim is to develop a model able to identify irony in social media texts capturing inherent characteristics of this kind of device. Some authors share a similar perspective on this issue [Barbieri et al. 2014; Buschmeier et al. 2014; Wallace 2015].

4.1. Irony Detection Model (IDM)

Let us describe the set of features used in IDM [Hernández Farías et al. 2015].

4.1.1. Structural Features. Structural features are the following:

Punctuation Marks. Punctuation marks have been widely applied in irony detection [Carvalho et al. 2009; Davidov et al. 2010; Reyes et al. 2013]. Some lexical marks help the writer to point out the sense and meaning in a text. According to Kreuz and Caucci [2007], the use of some textual factors (e.g., punctuation marks) may provide reliable

clues for identifying ironic intent in social media content. In short texts like tweets, this kind of visual cue can help to achieve the real intention behind the literal content in the utterance. In IDM, the punctuation marks and uppercase words are considered as lexical markers to distinguish ironic from nonironic utterances.

Length of Words. Twitter users must communicate their messages in 140 characters and express their ideas in a concise and direct manner. We consider a feature to catch the length in words (lengthWords) of each tweet, under the assumption that, thanks to a creative use of language, ironic tweets may achieve a communicative goal probably with fewer words than nonironic tweets.

Emoticons. In social media, emoticons ("emotional icons") are used to display a feeling in as few characters as possible. They can be used as visual cues to show the real intention of the speaker in order to achieve a particular effect: humor, sadness, despair, confusion, to apologize, or to express solidarity/support. Sometimes the emoticons are required under certain circumstances in text-based communication, where the absence of some sort of cue can hide what was originally intended (to be humorous, sarcastic, ironic, and often negative) [Wolf 2000]. In IDM, the frequency of emoticons is considered as a feature.

Discourse Markers. People use different discourse markers for writing. They have certain functions and help to express ideas. In IDM, there are two different kinds of discursive terms:[8] Counterfactuality and Temporal Compression. A list of terms that hint an opposition or contradiction in a text (such as "nevertheless") was considered to calculate a Counterfactuality score. Furthermore, the frequency of terms that identify elements related to opposition in time (i.e., terms that indicate an abrupt change in a narrative, like "suddenly") refers to the Temporal Compression score.

Part of Speech. To capture the structure used in a tweet, we consider the frequency of different part-of-speech (POS) labels. According to Kreuz and Caucci [2007], adjectives and adverbs can also be considered as lexical markers in ironic expressions. In IDM, four POS tags were taken into account: verbs, nouns, adjectives, and adverbs. These sets of labels allow us to identify the presence of certain kinds of words in ironic utterances.

Semantic Similarity. Ironic texts are often expressed by using words with a different meaning. According to Giora and Fein [1999], at the initial stage irony comprehension involves getting the literal sense of the words and then involves incompatible meanings. In order to obtain the degree of inconsistency in a tweet, [Wu and Palmer 1994] the semantic similarity measure was calculated using the WordNet::Similarity module.[9]

4.1.2. Affective Features. The use of some features related to affect was already investigated by Hernández Farías et al. [2015] in IDM:

Dictionary of Affect in Language (DAL). Such resource (see Table VI) was exploited in a first attempt to capture some kind of affective information related to a tweet. Three different values were calculated: Activation (degree of response that humans have under an emotional state), Imagery (how difficult it is to form a mental picture of a given word), and Pleasantness (degree of pleasure produced by words).

[8]These discursive terms have been used previously by Reyes et al. [2013]. Both lists are available at http://users.dsic.upv.es/grupos/nle.

[9]This module allows one to calculate a set of seven different similarity measures. According to the experiments carried out in Hernández Farías et al. [2015], this semantic similarity performed better than the others.

Sentiment Lexicons: Hu&Liu and *AFINN.* Giving negative (or positive) evaluations toward some targets is inherent to ironic utterances [Alba-Juez and Attardo 2014]. In this sense, the sentiment score of a tweet may help to distinguish between different types of tweets [Wang 2013], that is, ironic and nonironic. In order to catch the writer's attitude, two features were considered: (1) the score, which refers to the overall sentiment (positive, negative, or neutral) expressed in a tweet, taking into account a well-known sentiment analysis resource developed by Hu&Liu, and (2) the valence, which is used to compute the rate of evaluation expressed, that is, a criticism (negative) or a praise (positive), by using the AFINN lexicon.[10] Both features related to the sentiment score and to the polarity value were strongly relevant to irony classification, according to an information gain analysis reported in Hernández Farías et al. [2015]. This encouraged us to better investigate the use of features related to affect.

In Hernández Farías et al. [2015], some experiments were carried out with the corpus developed by Reyes et al. [2013], obtaining encouraging results. As experimental setting five different classifiers were applied (Naïve Bayes, Decision Tree, Support Vector Machine, Multilayer Perceptron, and Maximum Entropy) under a 10-fold cross-validation. The results outperformed those from Reyes et al. [2013]. In Section 5, we will extend the evaluation for this model, by presenting the results obtained applying the IDM model over all the other corpora mentioned in Section 3, for comparison purposes with the results obtained by using the extended model emotIDM.

4.2. emotIDM: Irony Detection Model + Emotional Information

In this section, we introduce emotIDM, which extends IDM considering a much wider set of features exploiting information related to emotions for irony detection. In particular, as a novelty with respect to other approaches, we sought what could be useful to incorporate in emotIDM information about the psychological and emotional content of tweets by means of (1) a variety of sentiment and emotion lexicons that can offer information about sentiment and emotions expressed in text according to different levels of granularity (e.g., referring simply to positive or negative sentiment, or to emotional categories such as joy, sadness, fear, and so on) and (2) a variety of psycholinguistic resources that could give some additional measure about the emotional disclosure in our sample, according to different theoretical perspectives on emotions. We organize the description of affect-related features to catch such different aspects in three groups: the first group is related to information about *sentiment polarity*, the second group is related to information about *emotions* by referring to a finer-grained categorization model (beyond the polarity valence), and the third one to different perspectives related to emotions according to dimensional approaches to emotion modeling. Affect-related features rely on the use of various lexical resources. This is needed with the purpose to increase the coverage of different affective aspects in textual content. Moreover, new structural features were also considered. Next we describe in detail each group of features as well as the resources involved.

4.2.1. Structural Features. This group includes structural features in the IDM model and, in addition, eight new features: the length in characters (lengthChars), colon, exclamation, question, and the amount of uppercase characters (upperCaseChars), as well as a set of specific markers of Twitter content: hashtagsFreq, mentionsFreq, and rt (retweets). The complete group of features is described and summarized in Table II.[11] As we are proposing a model specifically for Twitter, we consider that in ironic tweets these markers could provide important clues.

[10]See Table III for a description of the sentiment lexicons mentioned.
[11]PM is defined as the sum of colon, exclamation, and question marks.

Table II. Structural Features In emotIDM

Features	Description
colon exclamation question PM	The frequency of each punctuation mark in a tweet
lengthWords lengthChars	The total amount of words and characters in a tweet, respectively
verbs nouns adjectives adverbs	The frequency of each pos-tag in the tweet
upperCaseChars	The total amount of uppercase characters in a tweet
totalEmoticons	The total number of emoticons in a tweet
val_counter val_temporal	Frequency of Counterfactuality and Temporal compression terms defined in Section 4.1
semantic_similarity	The degree of inconsistency in a tweet (Wu&Palmer semantic similarity measure)
hashtagsFreq mentionsFreq rt	The frequency of each specific Twitter marker in a tweet

4.2.2 Affective Features.

Sentiment-Related Features. As we already mentioned, irony can be used to express an evaluative judgment and sentiment resources can be useful in order to capture the positive or negative polarity of words in a sentence. Three different scores were used to catch the sentiment expressed in tweets: positive, negative, and a total value (that considers both positive and negative values). The sentiment resources we exploited can be split into two categories: those composed by simple lists of positive and negative words, and those where each word is labeled with a sentiment strength in a range of polarity values (from positive to negative). In the first case, in order to obtain the positive and negative score for each tweet, we sum the number of words belonging to each category (positive or negative expressions). For resources assigning a numerical score varying in a range of intensity for the polarity valence, the positive/negative score is the sum of all the positive/negative values in a tweet. In both cases, the total value is defined as the difference between the positive and negative score. In total, 24 sentiment features were obtained from nine different resources. Table III summarizes the features and the resources exploited to calculate their values.[12]

Emotional Categories. Theories in the nature of emotion suggested the existence of basic or fundamental emotions such as anger, fear, joy, sadness, and disgust. Different approaches propose different sets of basic or fundamental emotions, each having its own specific eliciting conditions and its own specific physiological, expressive, and behavioral reaction patterns. The emotional categories included in emotIDM are based on four resources: EmoLex, EmoSenticNet, SentiSense, and LIWC (see Table IV). Different resources related to various theories were considered with the purpose to increase the coverage of emotions in textual content. Indeed, the resources we used refer to different emotion models well grounded in psychology, such as the ones proposed by Plutchik [2001], Ekman [1992], Arnold 1960], and Parrot [2001]. In particular, emotional labels of EmoLex refer to the eight basic emotions of the Plutchik circumplex model, the ones of EmoSenticNet to the six emotions from the Ekman model, whereas

[12]Normalization was carried out in order to adjust the values of all resources in a range between 0 and 1.

Table III. Sentiment Features in emotIDM

Features	Description
AFINN_total AFINN_pos AFINN_neg	AFINN[1] is a resource collected by Finn Arup Nielsen [Nielsen 2011]. The most recent available version of the dictionary contains 2,477 English words. Each one has been manually labeled with a sentiment strength in a range of polarity from −5 up to +5. The list includes a number of words frequently used on the Internet, like obscene words and Internet slang acronyms such as LOL (laughing out loud).
HL_total HL_pos HL_neg	The Hu&Liu lexicon (HL) is a well-known resource originally developed for opinion mining [Hu and Liu 2004]. The final version of the dictionary includes 6,789 words divided into two groups: 4,783 negative (HL_neg) and 2,006 positive (HL_pos).[2]
GI_total GI_pos GI_neg	The Harvard General Inquirer (GI)[3] developed by Stone and Hunt [1963] is a resource for content analysis that attaches syntactic, semantic, and pragmatic information to 11,788 part-of-speech tagged words. A total of 182 categories are included in the GI. Two of them, that is, positive words (1,915) and negative words (2,291), are exploited in our model (GI_pos and GI_neg, respectively).
SWN_total SWN_pos SWN_neg	SentiWordNet[4] (SWN) is a lexical resource based on WordNet developed by Baccianella et al. [2010]. It assigns to each of the about 117,000 synsets of WordNet three sentiment numerical scores (in a range between 0 up to 1): positivity, negativity, and objectivity.
EWN_total EWN_pos EWN_neg	EffectWordNet,[5] developed by Choi and Wiebe [2014], is a lexicon created on the basis of WordNet. The main idea is that the expressions of sentiment are often related to states and events that have positive or negative (or null) effects on entities. It contains more than 11,000 events distributed in three groups: positive (3,288), negative (2,427), and null (5,296).
SO	Taboada and Grieve [2004][6] annotated a list of adjectives with Semantic Orientation (SO) values. The resource is made of 1,720 adjectives and their "near bad" and "near good" values according to the Pointwise Mutual Information-Information Retrieval measure (PMI-IR).
SUBJ_str_pos SUBJ_weak_pos SUBJ_str_neg SUBJ_weak_neg	The Subjectivity lexicon (SUBJ) includes 8,222 terms (labeled as subjective expressions) collected by Wilson et al. [2005]. It contains a list of words, along with their POS tagging, labeled with polarity (positive, negative, neutral) and intensity (strongly or weakly subjective). This resource is part of the Multi-Perspective Question-Answering (MPQA) lexicon.[7]
EmoLex_positive EmoLex_negative	EmoLex[8] is a word-emotion association lexicon developed by Mohammad and Turney [2013], which includes also manual annotations about the polarity value of words, negative or positive. The dictionary contains 14,182 words.
SN_Pol SN_Formula	SenticNet[9] (SN) is a recent semantic resource for concept-level sentiment analysis [Cambria et al. 2014]. The current version (SenticNet 3) contains 30,000 words. A value of polarity is provided directly by the resource for each word (SN_Pol). Each concept is associated with the four dimensions of the Cambria's Hourglass of Emotions model [Cambria et al. 2012], and a polarity measure can be defined in terms of the four affective dimensions, according to the formula in Cambria et al. [2012]. We will also consider such measure in our study (SN_Formula).

[1]https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt.
[2]http://www.cs.uic.edu/~liub/FBS.
[3]http://www.wjh.harvard.edu/~inquirer/homecat.htm.
[4]http://sentiwordnet.isti.cnr.it/download.php.
[5]http://mpqa.cs.pitt.edu/.
[6]We considered the "near good" as positive and "near bad" as negative to calculate the SO value. http://www.sfu.ca/~mtaboada/research/nserc-project.html.
[7]http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.
[8]http://www.saifmohammad.com/WebPages/lexicons.html.
[9]http://sentic.net/.

Table IV. Emotional Categories Features in emotIDM

Features	Description
EMOLEX_emotion[a]	EmoLex[1] is a word-emotion association lexicon [Mohammad and Turney 2013] containing 14,182 words labeled according to the eight Plutchik primary emotions [Plutchik 2001]: joy, sadness, anger, fear, trust, surprise, disgust, and anticipation.
EmoSN_emotion[a]	EmoSenticNet[2] (EmoSN) is a lexical resource [Poria et al. 2013] that assigns WordNet-Affect emotion labels related to the Six Ekman's basic emotions to SenticNet concepts. The whole list includes 13,189 entries annotated with the six Ekman emotions: disgust, sadness, anger, joy, fear, and surprise.
SentiSense_emotion[a]	SentiSense,[3] developed by Carrillo de Albornoz et al. [2012], attaches emotional meanings to concepts from the WordNet lexical database. It is composed by a list of 5,496 words tagged with emotional labels from a set of 14 emotional categories, which refer to a merge of models by Arnold, Plutchik, and Parrot. In emotIDM we considered a subset[b] composed by joy, fear, surprise, anger, disgust, love, anticipation, sadness, and like.
LIWC_total LIWC_pos LIWC_neg	The Linguistic Inquiry and Word Counts[4] dictionary [Pennebaker et al. 2001] (LIWC) contains 4,500 words distributed in categories for analyzing psycholinguistic features in texts. One of the categories is related to positive and negative emotions.

[a]"Emotion" is parametric to the various instances of emotion, i.e., anger, joy, etc.
[b]Due to the very limited size of word lists related to emotions, some of them were removed from SentiSense features.
[1]http://www.saifmohammad.com/WebPages/lexicons.html.
[2]http://www.gelbukh.com/emosenticnet/.
[3]http://nlp.uned.es/~jcalbornoz/SentiSense.html.
[4]http://www.liwc.net.

Table V. Emotions in emotIDM

Emotion	Resource
Anger	EmoLex, EmoSenticNet, and SentiSense
Anticipation	EmoLex and SentiSense
Disgust	EmoLex, EmoSenticNet, and SentiSense
Fear	EmoLex, EmoSenticNet, and SentiSense
Joy	EmoLex, EmoSenticNet, and SentiSense
Sadness	EmoLex, EmoSenticNet, and SentiSense
Suprise	EmoLex, EmoSenticNet, and SentiSense
Trust	EmoLex
Like	SentiSense
Love	SentiSense

SentiSense proposes a wider set of emotional labels inspired by different models, including Arnold and Parrot. We compute the frequency of words in a tweet belonging to an emotional category according to information encoded in the various resources. In total, 10 different emotions were considered as features (see Table V). Moreover, we also consider in this group of features the coarser-grained classification of emotional words w.r.t. positive and negative emotions provided by LIWC. Table IV summarizes the resources included in emotIDM.

Dimensional Models of Emotions. There are some theories proposing that the nature of an emotional state is determined by its position in a space of independent dimensions. According to a dimensional approach, emotions can be defined as a coincidence of values on a number of different strategic dimensions [Bradley and Lang 1999]. Dimensional views of emotions have been advocated by a large number of theorists. emotIDM considers the Pleasantness-Activation-Imagery dimensions of the Dictionary of Affect in Language (DAL), already exploited in IDM. Moreover, it considers

Table VI. Emotional Dimensions Features in emotIDM

Features	Description
ANEW_val ANEW_aro ANEW_dom	Affective Norms for English Words[a] (ANEW) is a set of words associated with emotional ratings [Bradley and Lang 1999]. In ANEW, each concept in the dictionary is rated in terms of the Valence-Arousal-Dominance (VAD) model.
DAL_ple DAL_act DAL_ima	The Dictionary of Affect in Language[a] (DAL) developed by Whissell [2009] contains 8,742 English words rated on a three-point scale along three dimensions: Pleasantness, Activation, and Imagery.
SN_Pleas SN_Atten SN_Sensit SN_Apti	SenticNet[a,1] (SN) is a semantic resource where each concept is associated with the four dimensions of the Cambria Hourglass of Emotions model [Cambria et al. 2012]: Pleasantness, Attention, Sensitivity, and Aptitude.

[a]Normalization was carried out in order to adjust the values of all resources in a range between 0 and 1.
[1]http://sentic.net/.

dimensions from the ANEW resource, which refers to the the VAD model (Valence-Arousal-Dominance), and from SenticNet, which relies on the Hourglass of Emotions model [Cambria et al. 2012] and reinterprets the Plutchik model by organizing primary emotions around four independent but concomitant dimensions (Pleasantness-Attention-Sensitivity-Aptitude). In Table VI, the resources related to dimensional models used in emotIDM are summarized. In emotIDM, 10 features related to dimensional models of emotions were considered. It is important to mention that ANEW and DAL were constructed by human-manual rating of words, while SenticNet was by an automatic process that merges different resources. To calculate the degree of each dimension, the sum of the values for each word in a tweet was considered.

5. EXPERIMENTS

We carried out a set of experiments in order to evaluate and compare the effectiveness of both models, IDM and emotIDM, in automatically distinguishing between ironic and nonironic tweets over the set of corpora described in Section 3. Using the IDM model, a tweet is represented as a vector composed by 16 features, while in emotIDM the vector has 78 features. As we mentioned before, in this work irony is considered as an umbrella term that covers sarcasm. Both IDM and emotIDM were designed to identify ironic content in this general sense. However, some authors developing the datasets used in our experiments used the term "sarcasm" to refer to their irony-laden textual samples [Barbieri et al. 2014; Ptáček et al. 2014; Riloff et al. 2013], depending on the hashtags used for collecting the samples (see Table I, fourth column). Therefore, in order to be consistent with the original terminology, in the following we describe the experiments using the labels "ironic" or "sarcastic" depending on the term used by the authors during the corpora development. But let us remark that we will use the same model to identify both the phenomena in tweets.

Different experimental setting were evaluated:

(1) *TwReyes2013*. Three binary classifications: irony-vs-education, irony-vs-humor, and irony-vs-politics. Each combination is balanced with 10,000 ironic and 10,000 nonironic samples (balanced distribution).
(2) *TwIronyBarbieri2014*. Four binary classifications: irony-vs-education, irony-vs-humor, irony-vs-newspaper, and irony-vs-politics Each combination is balanced with 10,000 ironic and 10,000 nonironic samples (balanced distribution). Let us remark again that here the nonironic samples are the same that are used in the previous item, whereas the ironic samples are the new ones introduced in Barbieri et al. [2014].

Table VII. Results in F-Measure Obtained by Applying Both IDM and emotIDM

Corpus		F-Measure					
		IDM			emotIDM		
		NB	DT	SVM	NB	DT	SVM
TwReyes2013	*Irony-vs-Education*	0.70	0.83	0.85	0.74	0.90	0.89
	Irony-vs-Humor	0.71	0.81	0.83	0.76	0.90	0.90
	Irony-vs-Politics	0.71	0.84	0.86	0.74	0.92	0.91
TwIronyBarbieri2014	*Irony-vs-Education*	0.67	0.84	0.85	0.75	0.90	0.89
	Irony-vs-Humor	0.74	0.84	0.85	0.77	0.91	0.90
	Irony-vs-Politics	0.74	0.85	0.86	0.80	0.92	0.91
	Irony-vs-Newspaper	0.76	0.85	0.87	0.82	0.91	0.93
TwMohammad2015		0.65	0.64	0.62	0.66	0.64	0.60
TwSarcasmBarbieri2014	*Sarcasm-vs-Education*	0.75	0.84	0.85	0.81	0.90	0.90
	Sarcasm-vs-Humor	0.74	0.83	0.85	0.80	0.92	0.90
	Sarcasm-vs-Politics	0.78	0.86	0.88	0.86	0.94	0.93
	Sarcasm-vs-Newspaper	0.8	0.88	0.90	0.88	0.96	0.96
TwRiloff2013		0.73	0.75	0.71	0.74	0.75	0.73
TwPtáček2014		0.68	0.74	0.75	0.70	0.78	0.82

The underlined values are not statistically significant (t-test with 95% of confidence value).

(3) *TwMohammad2015*. Binary classification: ironic-vs-nonironic (imbalanced distribution).
(4) *TwSarcasmBarbieri2014*. Four binary classifications: sarcasm-vs-education, sarcasm-vs-humor, sarcasm-vs-newspaper, and sarcasm-vs-politics. Each combination is balanced with 10,000 sarcastic and 10,000 nonsarcastic samples (balanced distribution).
(5) *TwRiloff2013*. Binary classification: sarcastic-vs-nonsarcastic (imbalanced distribution).
(6) *TwPtáček2014*. Binary classification: sarcastic-vs-nonsarcastic (imbalanced distribution).

Three of six sets of experiments used corpora with an imbalanced distribution, as can be seen by observing Table I. Because of the perishability of Twitter data, in some cases we could rely only on a subset of the tweets originally collected.

For what concerns classifiers, irony detection mainly relies on traditional supervised methods. The two most widely applied have been the Support Vector Machine (SVM) and Decision Tree (DT) [González-Ibáñez et al. 2011; Reyes et al. 2013; Riloff et al. 2013; Barbieri et al. 2014; Ptáček et al. 2014; Buschmeier et al. 2014; Hernández Farías et al. 2015]. We evaluated our models by applying Weka[13] implementations of three standard classifiers: Naïve Bayes (NB), Decision Tree, and Support Vector Machine.[14] We believe that at this stage the most important issue to address for irony detection as a classification problem is the feature engineering one, not the one related to the optimization of the performance of the classifier [Ptáček et al. 2014; Wallace et al. 2015; Barbieri et al. 2014], which can be an issue to address in a second stage. All experiments were conducted in a 10-fold cross-validation setting. Results obtained are shown in Table VII.

[13]http://www.cs.waikato.ac.nz/ml/index.html.
[14]We used default values of Weka as parameters for each classifier.

Table VIII. Comparison of Results with the State-of-the-Art

Corpus	State of the Art			Our Results	
	Reference	Classifier	F-Measure	IDM	emotIDM
TwReyes2013					
Irony-vs-Education	Reyes et al. [2013]	DT	0.70	0.83	**0.90**
	Barbieri et al. [2014]		0.73		
	Hernández Farías et al. [2015]		0.78		
Irony-vs-Humor	Reyes et al. [2013]	DT	0.76	0.81	**0.90**
	Barbieri et al. [2014]		0.75		
	Hernández Farías et al. [2015]		0.79		
Irony-vs-Politics	Reyes et al. [2013]	DT	0.73	0.84	**0.92**
	Barbieri et al. [2014]		0.75		
	Hernández Farías et al. [2015]		0.79		
TwSarcasmBarbieri2014					
Sarcasm-vs-Education	Barbieri et al. [2014]	DT	0.88	0.84	**0.90**
Sarcasm-vs-Humor			0.88	0.83	**0.92**
Sarcasm-vs-Politics			0.90	0.86	**0.94**
Sarcasm-vs-Newspaper			0.97	0.88	0.96
TwRiloff2013					
	Riloff et al. [2013]	SVM	0.51	0.71	**0.73**
	Joshi et al. [2015]		0.61		
TwPtáček2014[a]					
	Ptáček et al. [2014]	SVM	0.90	0.75	0.82

[a]We have selected the imbalanced distribution for evaluation.

5.1. Discussion

As a preliminary remark, let us notice that in case of the TwIronyBarbieri2014 and TwMohammad2015 corpora, it is not possible to compare our results with results achieved in related work. In fact, this is the first time TwMohammad2015 is used in the context of the irony detection task, whereas the set of ironic samples in TwIronyBarbieri2014 (collected relying on the #irony hashtag) was not used by Barbieri et al. [2014] for evaluating their irony detection model, but it has been created and exploited only in a pilot attempt to distinguish sarcasm from irony, which is a different task. IDM improves the state of the art over the TwReyes2013 corpus, as already highlighted in Hernández Farías et al. [2015].[15] For what concerns the other corpora, which were already used for the evaluation of irony detection models, by observing Table VIII, we can see that IDM outperforms the state of the art in TwRiloff2013, whereas results regarding TwSarcasmBarbieri2014 are not higher than those reported in Barbieri et al. [2014]. It is interesting to note that in general, results obtained over the "Self-tagged" corpora (TwReyes2013, TwIronyBarbieri2014, TwSarcasmBarbieri2014, and TwPtáček2014) are higher than those from "Crowdsourced" ones (TwMohammad2015 and TwRiloff2013). This can be an aspect to be further investigated, reflecting on the differences that exist in corpora construction. In terms of performance over "Crowdsourced" corpora, there is much less difference between IDM and emotIDM than in "Self-tagged" corpora.

Overall, emotIDM outperforms IDM. The results show that emotional information helps to achieve higher F-measure rates in order to distinguish irony-laden tweets. emotIDM seems to be able to capture relevant features from these kinds of tweets. This may confirm our hypothesis about the important role of emotional information for

[15]As a main difference with the partial results reported in Hernández Farías et al. [2015], we use a normalized version of two resources: AFINN and DAL.

irony detection. Both IDM and emotIDM show a consistent performance even working with different-size corpora. The higher results are achieved in balanced distribution (TwReyes2013, TwIronyBarbieri2014, TwSarcasmBarbieri2014). The NB classifier presents the worst performance as in other approaches to irony detection [Reyes et al. 2013; Buschmeier et al. 2014; Fersini et al. 2015]. The SVM classifier obtains slightly better results than DT using IDM, while for emotIDM the DT achieves the best performance.

We compare the performance in terms of F-measure of IDM and emotIDM against the reported results for each corpus (see Table VIII). For what concerns the state of the art, together with the F-measure we mention the classifier used, and we report our results, both for IDM and emotIDM, by using the same classifier. Overall, emotIDM outperforms the state of the art (values in bold). All experiments except two were improved. Let us comment on such cases. For what concerns the results achieved on Sarcasm-vs-Newspaper, it is the only outcome where our approach does not improve the state of the art on the TwSarcasmBarbieri2014. However, notice that our set of features does not consider the presence of a URL, unlike the proposal in Barbieri et al. [2014], where the authors themselves report that nine of 10 tweets in the Newspaper category contain a URL.

The comparison with the results of Ptáček et al. [2014] over the TwPtáček2014 corpus deserves further investigation. Ptáček et al. propose a model to identify sarcastic tweets that include as features information referring to a bag-of-words (BOW) representation of text, whereas our system does not. Their result by using only BOW (0.90 in F-measure) is almost the same as using the whole set of features (including bag-of-words). It is difficult to compare the performance of our system with the one proposed by Ptáček et al. [2014] for two main reasons: (1) TwPtáček2014 contains sufficient data to train a successful bag-of-words classifier, but the same approach could be not adequate for irony detection across different datasets, and (2) no results without bag-of-words were reported, whereas our system shows consistent results even without the presence of bag-of-words features. Furthermore, more importantly, as explained also in Wallace [2015] and Barbieri et al. [2014], the risk for BOW approaches is to be topic dependent, since they work as a topic-based classifier and not as an irony detection procedure. Instead, the advantage of approaches that are not relying on bag-of-words, like the one we propose, is that they are able to capture ironic style disregarding domain, as is proved by our evaluation across different datasets that cover different topics.

5.2. Feature Analysis: Information Gain

We used many features to detect ironic utterances. An Information Gain analysis of features was carried out in order to identify which features are useful in emotIDM. The 10 best-ranked features for each binary classification can be seen in the appendix (Table IX). In order to have an overall view, we computed the frequency of each best-ranked feature for all the binary classifications, with the aim to evaluate which features were ranked as the best. A total of 34 features emerged as the most frequent. Figure 1 shows the results obtained. For sake of readability, structural features are grouped on the left. The following three groups are related to affective features and refer to sentiment features, emotional dimensions features, and emotional categories features, respectively.

We observe that features derived from the structural group rank high. This validates once again the importance of lexical markers in Twitter ironic contents [Kreuz and Caucci 2007; Carvalho et al. 2009; Davidov et al. 2010; Reyes et al. 2013; Barbieri et al. 2014]. Both sentiment features and the ones related to emotional dimensions captured by ANEW, DAL, and SenticNet appear to be useful to identify ironic tweets. In particular, AFINN emerges as an efficient sentiment resource for irony detection,

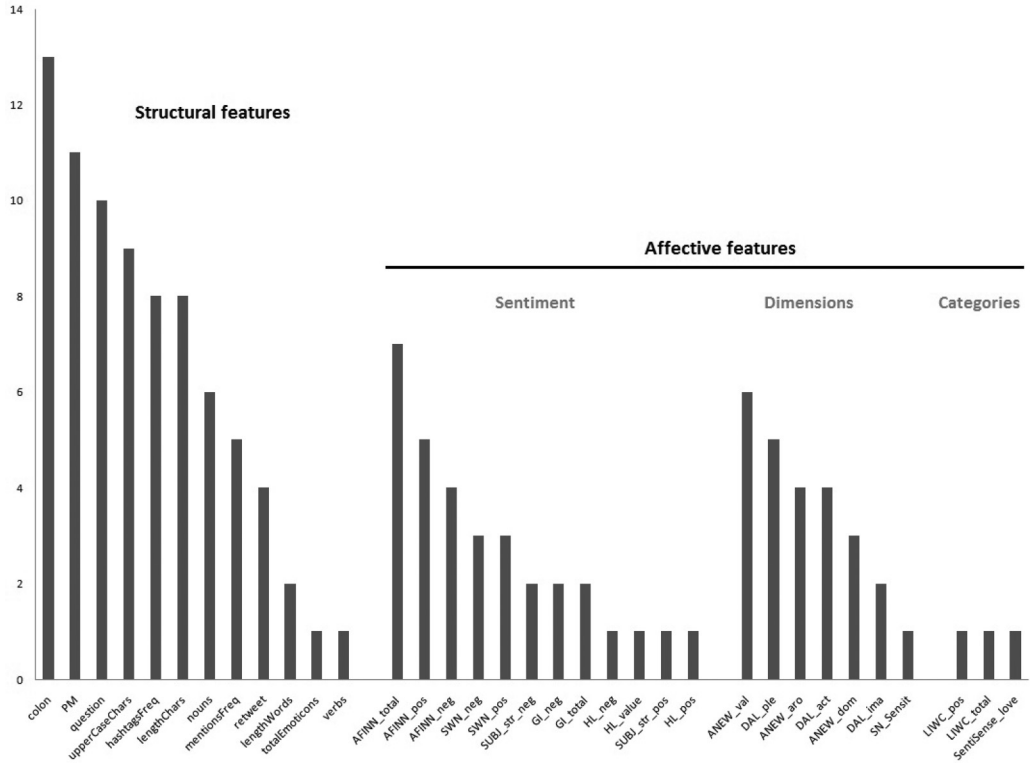

Fig. 1. Best-ranked features according to Information Gain.

but SentiWordNet, General Inquirer, Hu&Liu, and SUBJ also play a role. All the dimensions in ANEW and DAL have a relevant discriminative power, whereas for what concerns SenticNet, the "Sensitivity" dimension seems to be the most useful. Nevertheless, features related to emotional categories also help in the classification performance, even if they are not among the best-ranked features. In this group, we can see coarser-grained features related to LIWC, but also the features related to words expressing the emotion "Love" (SentiSense).

Additionally, in order to investigate if some differences could emerge by keeping separate corpora where users were marking the intention to be "ironic" and the ones where they marked the intention to be "sarcastic" (see Table I), we calculated the same frequency on the best-ranked features according to Information Gain considering on the one hand ironic-vs-nonironic tweets and on the other hand sarcastic-vs-nonsarcastic tweets. The outcome, shown in Figure 2, is interesting and introduces new data-driven arguments for a possible separation between irony and sarcasm.[16] Information from dimensional models of emotions (in particular from DAL and ANEW) is very important to distinguish tweets belonging to the ironic class. In both tasks, features related to sentiment are in the top 10. Some authors consider that one of the main differences between irony and sarcasm is based on the evaluation they express [Alba-Juez and Attardo 2014]. Irony may be positive (i.e., noncritical), while sarcasm is not [Giora and Attardo 2014]. Sarcasm is considered more aggressive and offensive than irony. According to Wang [2013], the tweet with more aggressive intention should be sugar coated

[16]Features are grouped as in the previous figure.

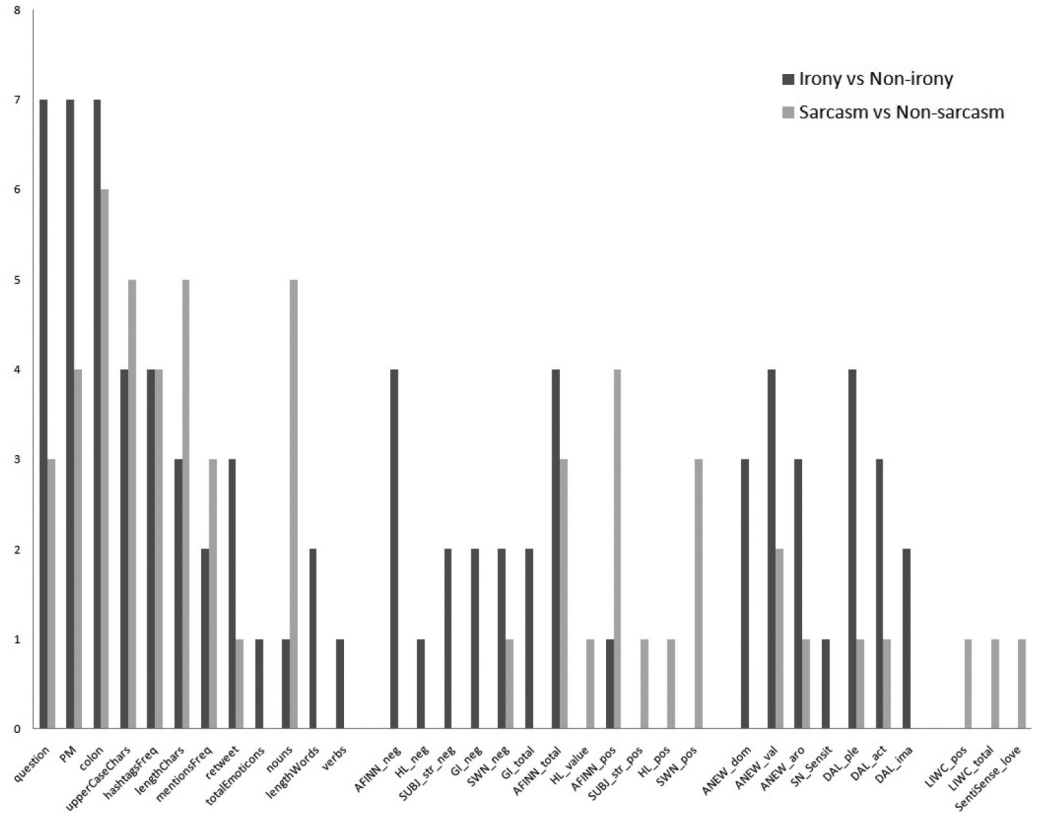

Fig. 2. Best-ranked features according to Information Gain, differentiating between tweets tagged as ironic and sarcastic.

with more positive words. Such hypothesis seems to be well supported here. Indeed, it can be clearly noticed as the discriminative power of the sentiment features related to the positive and negative polarity values of words varies in the two cases (positive words are more relevant for identifying sarcasm and vice versa). These could be indicators of the fact that such features could help in differentiating sarcasm from irony. Moreover, it is worth noticing that features related to emotional categories seem to be more discriminative in corpora self-tagged with #sarcasm and #sarcastic. In particular, a preliminary analysis for what concerns the feature related to words expressing "Love" suggests that it could be related to the higher frequencies of constructions such as "I just love . . . ," "I love when . . . ," "I love being . . . " in tweets tagged with #sarcasm. This will be a further data-driven element to investigate to address the finer-grained task of distinguishing different types of irony.

For what concerns the structural features, interestingly, the feature related to frequency of nouns seems to be particularly relevant in tweets containing the #sarcasm hashtag. Besides, the mentionsFreq is also relevant for sarcastic tweets; one possible explanation is that this kind of feature can be considered as a way to point out the target by a specific Twitter marker, that is, the mention. This is in line with Lee and Katz [1998]: "Sarcasm conveys ridicule of a specific victim whereas irony does not." In this sense, sarcastic utterances may contain a noun or a mention to refer to the target. Finally, the lenghtChars feature also seems to be especially relevant in sarcastic tweets. A possible hypothesis is that sarcastic tweets are sharper, and then shorter.

6. CONCLUSION AND FUTURE WORK

In this article, we presented emotIDM, a novel model for irony detection in Twitter that includes information on affect encompassing different aspects of this multifaceted phenomenon. We have performed several experiments over a set of corpora already used in the same task, outperforming previous results both for what concerns IDM, the previous model we used as a starting point, and results obtained on the same datasets by previous authors, in almost all cases.

To the best of our knowledge, this is the first work in irony detection where the robustness of the model is evaluated on a set of representative Twitter corpora including samples of ironic and nonironic messages, which were different along various dimensions: size, balanced versus imbalance distribution, collection methodology, and criteria (i.e., self-tagging vs. crowdsourcing, hashtags used for collecting samples, etc.). Dealing also with imbalanced distributions is, indeed, important, since, as highlighted also in Reyes et al. [2013] and Ptáček et al. [2014], real world does not resemble the balanced distribution. Results show that our model achieves good performances in classification terms across all these dimensions. It performs better in cases of datasets with balanced distribution, where a self-tagging methodology has been applied, but it has to be noticed that it achieves good results, improving the state of the art, also with the TwRiloff2013 dataset, with fewer data and imbalanced distribution. A more detailed reflection on the better performances related to corpora developed by using self-tagging is matter of future work.

Overall, results confirm that affective information helps in distinguishing among ironic and nonironic tweets. In particular, a first analysis of the affective features via information gain highlights the discriminating power, on the one hand, of sentiment-related features based on resources such as AFINN, SentiWordNet, General Inquirer, and Subjectivity Lexicons, and, on the other hand, of features related to resources such as ANEW, DAL, and SenticNet, which refer to dimensional models of emotions. For what concerns features related to emotion words such as joy, anger, and so on, they seem to have a minor role, with the exception of the one related to the emotion "love."

Comparative results on corpora collected by using different self-tagging criteria (i.e., on the one hand hashtags such as #irony, and on the other hand hashtags such as #sarcasm and #sarcastic) introduce new data-driven arguments for a possible separation between irony and sarcasm. The issue of distinguishing between such devices is very challenging, still poorly understood, and only rarely addressed from computational linguistics [Wang 2013; Barbieri et al. 2014], deserving further investigation [Sulis et al. 2016].

A cross-language study of our model could be an interesting line of future research, even if some of the features could be language dependent. Moreover, it could be interesting to apply this model to other languages apart from English also to see if it would assist the state of the art in going beyond irony detection, leading to an improvement of emotion forecast. Finally, it will be interesting to investigate also the effect of using word embeddings as features (extracted from a selected large corpus, e.g., a large corpus of tweets) in the classification system, in order to evaluate their effectiveness and to test if the features extracted from the lexical resources still play a positive role.

APPENDIX

In Table IX, the rank for each binary classification mentioned in Section 5 is shown.

Table IX. Ten Best-Ranked Features According to Information Gain

	TwReyes2013		**TwIronyBarbieri2014**			
Iro-vs-Edu	*Iro-vs-Hum*	*Iro-vs-Pol*	*Iro-vs-Edu*	*Iro-vs-Hum*	*Iro-vs-Pol*	*Iro-vs-News*
question	PM	PM	PM	PM	PM	colon
PM	question	question	question	colon	colon	PM
colon	colon	colon	colon	question	question	upperCaseChars
AFINN_neg	AFINN_neg	ANEW_dom	hashtagsFreq	hashtagsFreq	hashtagFreq	mentionsFreq
HL_neg	GI_total	upperCaseChars	lenghtChar	rt	upperCaseChars	hashtagFreq
SUBJ_str_neg	GI_neg	DAL_ple	upperCaseChars	AFINN_pos	noun	lenghtChar
GI_neg	ANEW_val	ANEW_val	mentionsFreq	AFINN_total	DAL_ple	lenghtWords
AFINN_total	ANEW_aro	ANEW_aro	rt	ANEW_val	rt	DAL_act
ANEW_dom	AFINN_total	AFINN_neg	SWN_neg	emoticons	ANEW_val	DAL_ple
GI_total	SN_Sensit	SUBJ_str_neg	DAL_act	ANEW_arousal	ANEW_dom	DAL_ima

	TwSarcasmBarbieri2014			**TwRiloff2013**	**TwPtáček2014**	**TwMohammad2015**
Sar-vs-Edu	*Sar-vs-Hum*	*Sar-vs-Pol*	*Sar-vs-News*	HL_pos	colon	DAL_ple
colon	colon	colon	colon	AFINN_pos	lenghtChar	DAL_act
PM	PM	PM	upperCaseChars	mentions_Freq	DAL_ple	verbs
question	question	question	PM	LIWC_pos	DAL_act	lenghtWords
hashtagsFreq	hashtagsFreq	lenghtChar	lenghtChar	colon	PM	question
upperCaseChars	upperCaseChars	upperCaseChars	mentionsFreq	LIWC_total	DAL_ima	DAL_ima
lenghtChar	rt	hashtagsFreq	hashtagsFreq	HL_value	SWN_pos	AFINN_neg
nouns	lenghtChar	nouns	AFINN_pos	SUBJ_str_pos	AFINN_total	SWN_neg
AFINN_total	ANEW_val	AFINN_pos	SWN_pos	upperCaseChars	nouns	SWN_neg
AFINN_pos	nouns	ANEW_val	AFINN_total	SentiSense_love	SWN_neg	DAL_ima
mentionsFreq	ANEW_aro	SWN_pos	nouns			

REFERENCES

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media (LSM'11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2–11.

Laura Alba-Juez and Salvatore Attardo. 2014. The evaluative palette of verbal irony. In *Evaluation in Context*, Geoff Thompson and Laura Alba-Juez (Eds.). John Benjamins Publishing Company, Amsterdam/ Philadelphia, 93–116.

Magda B. Arnold. 1960. *Emotion and Personality*. Vol. 1. Columbia University Press, New York, NY.

Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the art language technologies for italian: The EVALITA 2014 perspective. *Journal of Intelligenza Artificiale* 9, 1 (2015), 43–61.

Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics* 32, 6 (2000), 793–826.

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 7th International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta, 2200,2204.

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *Proceedings of the 9th International Conference on Web and Social Media, (ICWSM'15)*. AAAI, Oxford, UK, 574–577.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Baltimore, Maryland, 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)*. Pisa University Press, Pisa, Italy, 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 (March 2013), 55–63.

Andrea Bowes and Albert Katz. 2011. When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal* 48, 4 (2011), 215–236.

Margaret M. Bradley and Peter J. Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical Report. Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

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*. Association for Computational Linguistics, Baltimore, Maryland, 42–49.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In *Cognitive Behavioural Systems*. Lecture Notes in Computer Science, Vol. 7403. Springer, Berlin, 144–157.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of AAAI Conference on Artificial Intelligence*. AAAI, Québec, Canada, 1515–1521.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 3562–3567.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! It's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion (TSA'09)*. ACM, New York, NY, 53–56.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Doha, Qatar, 1181–1191.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL'10)*. Association for Computational Linguistics, Uppsala, Sweden, 107–116.

Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? The social functions of irony. *Discourse Processes* 19, 3 (1995), 347–367.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3–4 (1992), 169–200.

Elisabetta Fersini, Federico Alberto Pozzi, and Enza Messina. 2015. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA'15)*. IEEE Xplore Digital Library, Paris, France, 1–8.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, 392–398.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. Association for Computational Linguistics, Denver, Colorado, 470–478.

Raymond W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol* 15, 1–2 (2000), 5–27.

Rachel Giora and Salvatore Attardo. 2014. Irony. In *Encyclopedia of Humor Studies*. SAGE, Thousand Oaks, CA.

Rachel Giora and Ofer Fein. 1999. Irony: Context and salience. *Metaphor and Symbol* 14, 4 (1999), 241–257.

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 (HLT'11)*. Association for Computational Linguistics, Portland, OR, 581–586.

H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, P. Cole and J. L. Morgan (Eds.). Academic Press, San Diego, CA, 41–58.

Irazú Hernández Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Pattern Recognition and Image Analysis*. Lecture Notes in Computer Science, Vol. 9117. Springer International Publishing, Santiago de Compostela, Spain, 337–344.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. ACM, Seattle, WA, 168–177.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings 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)*. Association for Computational Linguistics, Beijing, China, 757–762.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich-Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings 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)*. Association for Computational Linguistics, Beijing, China, 644–650.

Roger J. Kreuz and Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language (FigLanguages'07)*. Association for Computational Linguistics, Rochester, NY, 1–4.

Florian Kunneman, Christine Liebrecht, Margot van Mulken, and Antal van den Bosch. 2015. Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management* 51, 4 (2015), 500–509.

Christopher Lee and Albert Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol* 13, 1 (1998), 1–15.

John S. Leggitt and Raymond W. Gibbs. 2000. Emotional reactions to verbal irony. *Discourse Processes* 29, 1 (2000), 1–24.

Stephanie Lukin and Marilyn Walker. 2013. Really? Well. Apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*. Association for Computational Linguistics, Atlanta, GA, 30–40.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (26-31). European Language Resources Association (ELRA), Reykjavik, Iceland, 4238–4243.

Skye McDonald. 2007. Neuropsychological studies of sarcasm. In *Irony in Language and Thought: A Cognitive Science Reader*, H. Colston and R. Gibbs (Eds.). Lawrence Erlbaum, 217–230.

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

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management* 51, 4 (2015), 480–499.

Finn Årup Nielsen. 2011. 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 (CEUR Workshop Proceedings)*, Vol. 718. CEUR-WS.org, Heraklion, Crete, Greece, 93–98.

W. Gerrod Parrot. 2001. *Emotions in Social Psychology: Essential Readings*. Psychology Press, Philadelphia, PA.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates, 71.

Robert Plutchik. 2001. The nature of emotions. *American Scientist* 89, 4 (2001), 344–350.

Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems* 28, 2 (2013), 31–38.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English twitter. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 213–223.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM'15)*. ACM, 97–106.

Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: Beyond a simple case of negation. *Knowledge Information Systems* 40, 3 (2014), 595–614.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47, 1 (2013), 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 2013 Conference on Empirical Methods in Natural Language Processing, (EMNLP'13)*. Association for Computational Linguistics, Seattle, Washington, 704–714.

Simone Shamay-Tsoory, Rachel Tomer, B. D. Berger, Dorith Goldsher, and Judith Aharon-Peretz. 2005. Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cognitive Behavioral Neurology* 18, 1 (2005), 55–67.

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, Spring Joint Computer Conference (AFIPS'63 (Spring))*. ACM, New York, NY, 241–256.

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*. In Press. Available online.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. AAAI, Stanford, CA, 158–161.

Yi-jie Tang and Hsin-Hsi Chen. 2014. Chinese irony corpus construction and ironic structure analysis. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. Association for Computational Linguistics, Dublin, Ireland, 1269–1278.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *Proceedings of the 19th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, The Netherlands, 765–770.

Byron C. Wallace. 2015. Computational irony: A survey and new perspectives. *Artificial Intelligence Review* 43, 4 (2015), 467–483.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings 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)*. Association for Computational Linguistics, Beijing, China, 1035–1044.

Angela P. Wang. 2013. #Irony or #sarcasm—a quantitative and qualitative study based on twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC'13)*. Department of English, National Chengchi University, Taipei, Taiwan, 349–356.

Juanita M. Whalen, Penny M. Pexman, J. Alastair Gill, and Scott Nowson. 2013. Verbal irony use in personal blogs. *Behaviour & Information Technology* 32, 6 (2013), 560–569.

Cynthia Whissell. 2009. 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.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua* 87, 1–2 (1992), 53–76.

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

Alecia Wolf. 2000. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior* 3, 5 (2000), 827–833.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL'94)*. Association for Computational Linguistics, Stroudsburg, PA, 133–138.

Received December 2015; revised March 2016; accepted April 2016