Sulis, E.; Hernandez-Farias, DI.; Rosso, P.; Patti, V.; Ruffo, G. (2016). Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not. Knowledge-Based Systems. 108:132-143. doi:10.1016/j.knosys.2016.05.035.

Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not

Emilio Sulis[a,1], Delia Irazú Hernández Farías[b,a], Paolo Rosso[b], Viviana Patti[a], Giancarlo Ruffo[a],

[a] *University of Turin, Italy*
[b] *Universitat Politècnica de València, Spain*

Abstract

The use of irony and sarcasm has been proven to be a pervasive phenomenon in social media posing a challenge to sentiment analysis systems. Such devices, in fact, can influence and twist the polarity of an utterance in different ways. A new dataset of over 10,000 tweets including a high variety of figurative language types, manually annotated with sentiment scores, has been released in the context of the task 11 of SemEval-2015. In this paper, we propose an analysis of the tweets in the dataset to investigate the open research issue of how separated figurative linguistic phenomena irony and sarcasm are, with a special focus on the role of features related to the multi-faceted affective information expressed in such texts. We considered for our analysis tweets tagged with #irony and #sarcasm, and also the tag #not, which has not been studied in depth before. A distribution and correlation analysis over a set of features, including a wide variety of psycholinguistic and emotional features, suggests arguments for the separation between irony and sarcasm. The outcome is a novel set of sentiment, structural and psycholinguistic features evaluated in binary classification experiments. We report about classification experiments carried out on a previously used corpus for #irony vs #sarcasm, outperforming the state-of-the-art in terms of F-measure. Overall results confirm the difficulty of the task, but introduces new data-driven arguments for the separation between #irony and #sarcasm. Interestingly, #not emerges as a distinct phenomenon.

[1] Corresponding author: sulis@di.unito.it

Keywords:
Figurative Language, Affective Knowledge, Irony, Sarcasm, Twitter

1. Introduction

The use of figurative devices such as irony and sarcasm has been proven to be a pervasive phenomenon on social media platforms such as Twitter and poses a significant challenge to sentiment analysis systems, since irony-laden expressions can play the role of polarity reversers [1]. Irony and sarcasm can influence and twist the affect of an utterance in complex and *different* ways. They can elicit different affective reactions, and can behave differently with respect to the polarity reversal phenomenon, as shown in [12]. However, the issue of distinguishing between such devices is still poorly understood. In particular, the question of whether irony and sarcasm are separated or similar linguistic phenomena is a controversial issue in literature and no clear consensus has already been reached. Although some researchers consider them strongly related figurative devices, other authors proposed a separation: sarcasm is offensive, more aggressive than irony [2, 3] and delivered with a cutting tone (rarely ambiguous), whereas irony often exhibits great subtlety and has been considered more similar to mocking in a sharp and non-offensive manner [4].

Furthermore, there is a consistent body of work on computational models for sarcasm detection [5] and irony detection [6] in social media, but only preliminary studies addressed the task to distinguish sarcasm and irony [7, 8].

In this paper we contribute to the debate of whether irony and sarcasm are similar or distinct phenomena by investigating how hashtags marking a figurative intent are used in Twitter. Our experiments concern a rich corpus of figurative tweets. We considered tweets marked with the user-generated tags #irony and #sarcasm, as such tags reflect a tacit belief about what constitutes irony and sarcasm, respectively [6]. We extend our analysis also to tweets tagged with hashtag #not, previously used to retrieve sarcastic tweets [5, 9], in order to investigate further their figurative meaning. Samples of tweets marked with different hashtags follows:

(tw1) *Fun fact of the day: No one knows who invented the*

fire hydrant because its patent was destroyed in a fire.
#irony
(tw2) *I just love it when I speak to folk and they totally*
ignore me!!! #Sarcasm!
(tw3) *So I just colored with Ava for an hour. Yeah my*
summer so far has been so fun [smiling face emoji] #not

Our methodology comprehends two steps. First, we performed a distribution and correlation analysis relying on the dataset of SemEval2015-Task11 [1], which includes samples of the kinds of figurative messages under consideration here (step 1). We explored the use of the three hashtags including structural as well as psycholinguistic and affective features concerning emotional information.

The affective information expressed in our texts is multi-faceted. Both sentiment and emotion lexicons, and psycholinguistic resources available for English, refer to various affective models and capture different facets of affect, such as sentiment polarity, emotional categories and emotional dimensions. Some of such resources, i.e., SenticNet [26] and EmoSenticNet [36], are not flat vocabularies of affective words, but include and model semantic, conceptual and affective information associated with multi-word natural language expressions, by enabling concept-level analysis of sentiment and emotions conveyed in texts. 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 distinguishing different kinds of figurative messages.

The analysis provided valuable insights on three kinds of figurative messages, including different ways to influence and twist the affective content. The outcome is a novel set of features evaluated in binary classification experiments (step 2). To better understand the impact of each feature, we evaluated our model performing experiments with different subset combinations, proceeding also by feature ablation, i.e. removing one feature at time in order to evaluate its contribution on the results.

To sum up, our experiments address the following research questions:

1. Is it possible to distinguish irony from sarcasm?
2. What is the role of the #not hashtag as a figurative language device? Is it a synonym of irony, of sarcasm, or something in between?
3. Does information about sentiment and psycholinguistics features help in distinguishing among #irony, #sarcasm and #not tweets?
4. What is the role of the polarity reversal in the three kinds of figurative messages?

Overall results confirm the difficulty of the task, but introduce new data-driven arguments for the separation between #irony and #sarcasm. As shown in the next sections, we outperform the state-of-the-art from 0.62 [8] to 0.70 in F-measure in #irony vs #sarcasm classification.

As for the separation of #irony vs #not and #sarcasm vs #not, interestingly, #not emerges as a distinct phenomenon. Analysis of the relevance of each feature in the model confirms the significance of sentiment and psycholinguistics features. Finally, an interesting finding about polarity reversal is given by correlation study presented in Section 4.2.1: the polarity reversal phenomenon seems to be relevant in messages marked with #sarcasm and #not, while it is less relevant for messages tagged with #irony.

The paper is structured as follows. Section 2 surveys main issues in literature about irony and the like. In Section 3 we describe the corpus and the resources exploited. Section 4 presents the feature analysis and Section 5 describes the experiments. Section 6 concludes the paper.

2. Irony, sarcasm *et similia*

Many authors embrace an overall view on irony, as expressing an opposite or different meaning from what is literally said [10]. Under this perspective, the presence of irony-related figurative devices is becoming one of the most interesting aspects to check in social media corpora since it can play the role of polarity reverser with respect to the words used in the text unit [11]. However, a variety of typologies of figurative messages can be recognized in tweets: from irony to sarcastic posts, and to facetious

tweets that can be playful, aimed at amusing or at strengthening ties with other users. Ironic and sarcastic devices can express different interpersonal meaning, elicit different affective reactions, and can behave differently with respect to the polarity reversal phenomenon [12]. Therefore to distinguish between them can be important for improving the performances of systems in sentiment analysis. According to the literature, boundaries in meaning between irony, sarcasm et similia are fuzzy. While some authors consider irony as an umbrella term covering also sarcasm [13, 14, 10], others provides insights for a separation. Sarcasm has been recognized in [15] with a specific target to attack [3], more offensive [2], and "intimately associated with particular negative affective states" [16], while irony has been considered more similar to mocking in a sharp and non-offensive manner [4].

The use of figurative language has been studied also in social media, but most researchers focus on irony or sarcasm separately. Computational models for sarcasm detection [17, 18, 5, 8, 19] and irony detection [6, 20] in social media has been proposed, mostly focussed on Twitter. Only a few preliminary studies addressed the task to investigate the differences between irony and sarcasm [7, 8]. The current work aims to further contribute to this subject. Furthermore a little studied form of irony that can be interesting to investigate in social media is *self-mockery*: "Self-mockery usually involves a speaker making an utterance and then immediately denying or invalidating its consequence, often by saying something like *'No, I was just kidding'*" [21]. Self-mockery seems to be different from other forms of irony, also from sarcasm, because it does not involve contempt for others, but the speaker wishes to dissociate from the content of the utterance. Investigations on the role of the #not hashtag as figurative language device could maybe provide insights into this phenomenon.

3. Dataset and lexical resources

In this section we describe the resources used in our work. First, the corpus of tweet messages in English developed for Task 11 of SemEval-2015[2] has been studied

[2]We consider the training, the trial and the test set: `http://alt.qcri.org/semeval2015/task11`

Description	N	MP	SD	ML
With #irony	1,737	-1.77	1.41	83
With #sarcasm	2,260	-2.33	0.77	66
With #not	3,247	-2.16	1.04	71

Table 1: Corpus description: Number of tweets (N), Mean (MP) and Standard Deviation (SD) of the Polarity, Median of the Length (ML)

extensively [1]. It consists in a set of tweets containing creative language that are rich in metaphor and irony. This is the only available corpus where a high variety of figurative language tweets has been annotated in a fine-grained sentiment polarity from -5 to +5. We finally rely on a dataset of 12,532 tweets[3]. Among the 5,114 different hashtags in the corpus, the most used ones are #not (3,247 tweets), #sarcasm (2,260) and #irony (1,737). Table 1 shows some introductory statistics over the dataset. The whole distribution of the polarity has a mean value of -1.73, a standard deviation of 1.59 and a median of -2.02. We consider the median as it is less affected by extreme values, instead of mean values. These results confirm that messages using figurative language mostly express a negative sentiment [11].

To cope with emotions and psycholinguistic information expressed in tweets, we explore different lexical resources developed for English. Finally, these can be grouped into three main categories related to "Sentiment polarity", to "Emotional categories" or to "Dimensional models of emotions".

Sentiment polarity. In order to gather information about sentiment polarity expressed in the corpus, we exploited lexicons including positive and negative values associated to terms.

(i) AFINN: This affective dictionary has been collected by Finn Årup Nielsen starting from most frequent words used in a corpus of tweets [22]. Each one has been manually labelled 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

[3]Due to the perishability of the tweets we were not able to collect all the 13,000 messages of the corpus.

Internet slang acronyms such as LOL (laughing out loud). The most recent available version of the dictionary contains 2,477 English words[4]. A bias towards negative words (1,598, corresponding to 65%) compared to positive ones (878) has been observed.

(ii) HL: The Hu-Liu's lexicon is a well-known resource originally developed for opinion mining [23]. The final version of the dictionary includes an amount of 6,789 words divided in 4,783 negative (HL_neg) and 2,006 positive (HL_pos)[5].

(iii) GI: The Harward General Inquirer is a resource for content analysis of textual data originally developed in the 1960s by Philip Stone [24]. The lexicon attaches syntactic, semantic, and pragmatic information to 11,788 part-of-speech tagged words. It is based on the Harvard IV-4 dictionary and Lasswell dictionary content analysis categories. Words are labelled with a total of 182 dictionary categories and subcategoriess[6]. The positive words (GI_pos) are 1,915, while the negative ones are 2,291 (GI_neg).

(iv) SWN: SentiWordNet [25] is a lexical resource based on WordNet 3.0. Each entry is described by the corresponding part-of-speech tag and associated to three numerical scores which indicate how positive, negative, and "objective" (i.e., neutral) the terms contained in the synset are. Each of the three scores ranges in the interval [0,1] and their sum is 1. Synsets may have different scores for all the three categories: it means the terms have each of the three opinion-related properties to a certain degree. In SentiWordNet 3.0[7] all the entries are classified as belonging to these three sentiment scores including a random-walk step for refining the scores in addition to a semi-supervised learning step. The first two categories (SWN_pos and SWN_neg) will be considered in our analysis.

(v) SN: SenticNet is a recent semantic resource for concept-level sentiment analysis [26]. The current version (SenticNet 3) contains 30,000 words, mainly unambiguous adjectives as stand-alone entries, plus multi-word expressions. The dictionary exploits an energy-based knowledge representation (EBKR) formalism to provide the affective

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

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

[6] http://www.wjh.harvard.edu/~inquirer/homecat.htm

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

semantics of expressions. Each concept is associated with the four dimensions of the Cambria's hourglass of emotions model [27]: Pleasantness, Attention, Sensitivity and Aptitude. We refer to these four values as SN_dim in our experiments in Section 5. A value of polarity is provided directly by the resource (SN_polarity henceforth). Moreover, since polarity is strongly connected to attitude and feelings, a further polarity measure is proposed, which can can be defined in terms of the four affective dimensions, according to the formula:

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

where c_i is an input concept, N is the total number of concepts of the tweet, 3 is a normalization factor. We will also consider such polarity measure in our study. In the following we will use 'SN_formula' to refer to the value p obtained by using the equation above.

(vi) EWN: The EffectWordNet lexicon has been recently developed by Choi [28] as a sense-level lexicon created on the basis of WordNet. The main idea is that the expressions of sentiment are often related to states and events which have positive or negative (or null) effects on entities. This lexicon includes more than 11k events in three groups: positive, negative and null. By exploiting the corresponding synset in WordNet, it is possible to collect a larger list of 3,298 positive, 2,427 negative and 5,298 null events[8].

(vii) SO: Semantic Orientation is a list of adjectives annotated with semantic-orientation values by Taboada and Grieve [29]. 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) as proposed by Turney [30]. In this analysis, the values of Semantic Orientation for each term is obtained by the difference between the corresponding "near good" and "near bad" values.

(viii) SUBJ: The subjectivity lexicon includes 8,222 clues collected by Wilson and colleagues [31] from a number of sources. Some were culled from manually developed resources and others were identified automatically. Each clue can be strongly or

[8] http://mpqa.cs.pitt.edu/

weakly subjective, or positive and negative. A clue that is subjective in most contexts is considered strongly subjective, while those that may only have certain subjective usages are considered weakly subjective. This resource is part of the Multi-Perspective Question-Answering (MPQA) lexicons[9].

Emotional categories. In order to gather information about the emotions expressed by referring to a finer-grained categorization (beyond the polarity valence), we considered the following resources which rely on categorical approaches to emotion modeling:

(ix) LIWC: Linguistic Inquiry and Word Counts dictionary[10] contains 127,149 words distributed in categories that can further be used to analyse psycho-linguistic features in texts. We selected two categories for positive and negative emotions: LIWC_PosEmo, with 12,878 entries and LIWC_NegEmo, with 15,115 entries [32].

(x) EmoLex: The resource EmoLex is a word-emotion association lexicon[11] developed at the National Research Council of Canada by Saif Mohammad [33]. The dictionary contains 16,862 words labelled according to the eight Plutchik's primary emotions [34]: sadness, joy, disgust, anger, fear, surprise, trust, anticipation.

(xi) EmoSN: EmoSenticNet is a lexical resource developed by Poria and colleagues [35] [36] that assigns WordNet Affect emotion labels to SenticNet concepts. The whole list includes 13,189 entries for the six Ekman's emotions: joy, sadness, anger, fear, surprise and disgust.

(xii) SS: SentiSense[12] is a concept-based affective lexicon that has been developed by Carrillo de Albornoz [37]. It attaches emotional meanings to concepts from the WordNet lexical database and consists of 5,496 words and 2,190 synsets labelled with an emotion from a set of 14 emotional categories, which are related by an antonym relationship.

[9] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

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

[11] http://www.saifmohammad.com/WebPages/lexicons.html

[12] nlp.uned.es/~jcalbornoz/SentiSense.html

Dimensional models of emotions. To provide some additional measures of the emotional disclosure in the corpus, according to different theoretical perspectives on emotions, we exploited the following resources developed which refer to dimensional approaches to emotion modelling:

(xiii) ANEW: Affective Norms for English Words is a set of normative emotional rating [38]. Each word in the dictionary is rated from 1 to 9 in terms of the Valence-Arousal-Dominance (VAD) model. This work considers the three dimensions separately.

(xiv) DAL: Dictionary of Affective Language developed by Whissell [39] contains 8,742 English words rated in a three-point scale[13]. We employed the following three dimensions: Activation (degree of response that humans have under an emotional state); Imagery (how difficult is to form a mental picture of a given word); Pleasantness (degree of pleasure produced by words).

Finally, we include among the *dimensional models of emotions* also the measures related to the Pleasantness, Attention, Sensitivity and Aptitude dimensions from SenticNet.

4. Features: a quantitative analysis

In this section, we identify the main characteristics of the tweets tagged with #irony, #sarcasm and #not from the SemEval 2015-Task 11 corpus. Our main interest is to find differentiating traits among these three kinds of figurative messages.

First, we focus our attention on polarity value which clearly shows a first regularity: the distribution of sarcastic tweets is more positively skewed, as the long "tail" shows, than the ironic ones (Figure 1). Moreover, the mean value of tweets marked with #irony is -1.73 instead of -2.33 for the #sarcasm ones. These differences show that sarcasm is perceived as more negative than irony by the hashtag adopters in our corpus. A first suggestion is that Twitter users consider irony as a more nuanced and varied phenomenon in terms of the associated sentiment. These distributions signal initially

[13] `ftp://perceptmx.com/wdalman.pdf`

that messages tagged with #not can be considered somehow different from #sarcasm and #irony ones.

Structural and affective features are considered. We perform a distribution analysis in each subgroup for every feature, as well as a correlation study taking into account the fine-grained polarity of the messages.

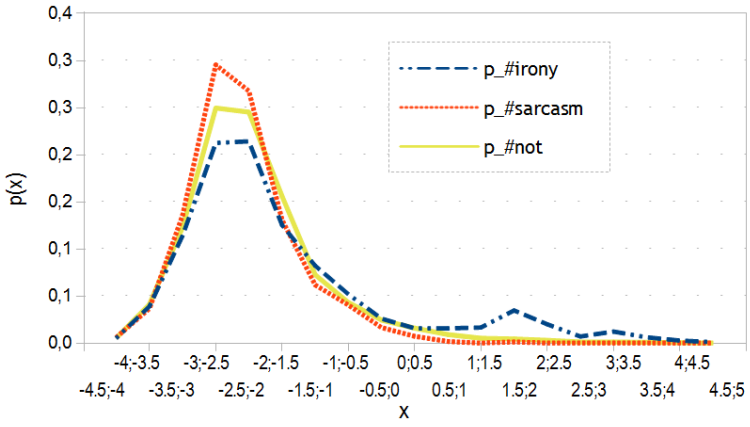

Figure 1: Distribution of tweets by polarity, p(x) is the probability that a tweet has polarity x

4.1. Structural and tweet features

Investigating the distributions of most traditional features is our first step. In addition to the analysis of the frequency of the part-of-speech (POS), emoticons, capital letters, URLs, hashtags, re-tweet and mentions, we report here two features showing interesting differences in the three subgroups: tweet length and punctuation marks.

Tweet length. The relation between the length of the tweets and the value of their polarity shows a Pearson's correlation of 0.13, with a statistically significant p-value $p<0.001$. We observe also that shorter messages (5% of tweets with less than 50 characters) are mostly negative with an average value of -2.1 and a standard deviation of 1.2. On the contrary, longer messages (5% of tweets with at least 138 characters) have a mean of -1.6 and a larger standard deviation of 1.7. This suggests that the length could play a role on the polarity of tweets when figurative language is employed. Tweets tagged with #sarcasm are shorter (mean of 66 characters), less than #not (71

char.) and #irony (83 char.). To sum up, it seems that sarcasm expresses in just a few words its negative content (see tweet *tw2* in the Introduction).

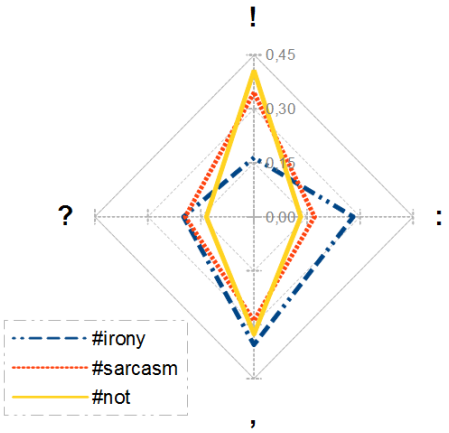

Figure 2: Distribution of punctuation marks in the corpus: colons are most used in #irony tweets, exclamation marks in #sarcasm and #not ones, question marks is less used in #not tweets

Punctuation marks. Figure 2 summarizes the frequency of commas, colons, exclamation and question marks in the three groups of tweets. Given the observed difference in the length of messages, counts are normalized by the length of tweets. While the use of colons is most frequent in #irony tweets and exclamation marks in #sarcasm and #not ones, the frequency of question marks is lower in #not tweets (e.g. tweets *tw1* and *tw2*).

4.2. Affective features

Some important regularities can be detected by analyzing the use of affective words. First, in order to investigate differences in the use of emotions among the three figurative language groups, EmoLex has been used to compute the frequency of words related to emotions, normalized by the number of words. As the distribution in Figure 3 shows, tweets marked with #irony contain fewer words related to *joy* and *anticipation*, than tweets marked with #sarcasm or #not. The same is for *surprise*, although to a lesser extent. On the other hand, in #irony words related to *anger*, *sadness* and *fear* (and to less extent *disgust*) are more frequent. Interestingly, tweets tagged with #not

and #sarcasm overlap quite perfectly with respect to the use of emotion words, while #irony shows a different behaviour.

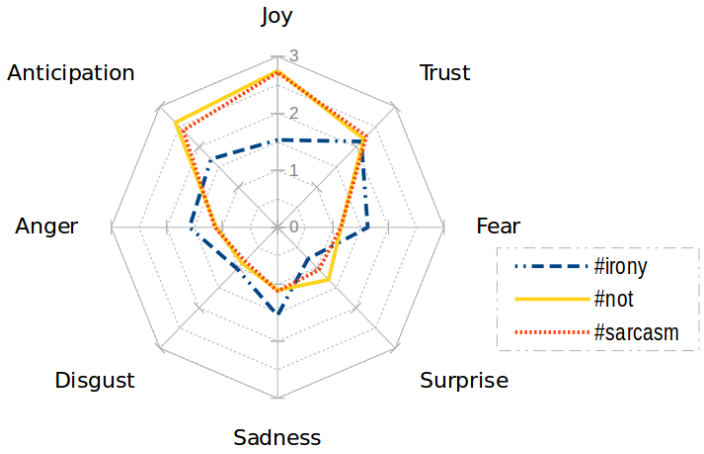

Figure 3: Distribution of emotion words (EmoLex [33]) in the SemEval Task 11 corpus: #not and #sarcasm tweets overlap, while #irony shows a different behaviour.

To further investigate the affective content, we extended the quantitative analysis to all the affective resources mentioned in Section 3: ANEW, DAL and the SenticNet's four singular dimensions (Dimensional Models of Emotions); EmoSenticNet, EmoLex, SentiSense and LIWC (Emotional Categories); AFINN, the Hu-Liu's lexicon, General Inquirer, SentiWordNet (SWN), EffectWordNet (EWN), Semantic Orientation (SO), SUBJ, and both the SenticNet (SN) polarity values mentioned above (Sentiment Polarity).

These resources have been previously normalized in the range from 0 to 1. For each group of tagged messages we compute two kinds of measures, depending on the kind of resource. When the lexicon is a list of terms (i.e., HL, GI, LIWC, EmoLex), we computed the mean value of the occurrences in each group. Instead, for lexicons containing a list of annotated entries (i.e., SN, AFINN, SWN, SO, DAL and ANEW), we calculated the sum of the corresponding values over all the terms, averaged by the total number of words in tweets. Formally, given a group T of n tagged messages where each single tweet t $\in T$ is composed by up to m words, and a lexical resource L assigns to each word w for every tweet in T a corresponding value $L(w)$, we calculated

the value $a(T, L)$ according the following equation:

$$a(T, L) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} L(w_{i,j})}{n} \quad (1)$$

Results of this analysis are shown from Table 2 to Table 4, where final values have been multiplied by 100 to improve the readability.

Sentiment Polarity features (Table 2) seem to be promising. While #sarcasm and #not messages contain more positive words, ironic messages are generally characterized by the use of more words with negative polarity. In fact, we can observe that *all* the lexical resources concerning the polarity of terms we considered (HL, AFINN, General Inquirer, SentiWordNet, SUBJ, SenticNet and SO) confirm that sarcastic and #not messages contain more positive terms than ironic ones; on the other hand, ironic messages contain more negative terms. Furthermore, also if we consider the polarity of terms related to *events*, detected by EffectWordNet, we obtain similar findings for what concerns irony and sarcasm. In fact, as shown in the last rows of Table 2, #not messages always contain more terms related to events (both positive, negative and null ones), but positive events are more frequent in sarcastic messages than in ironic ones, whereas negative events are more frequent in ironic than in sarcastic messages. Finally, let us observe that the objectivity measure from SWN highlights that messages tagged with #irony and #not contain more objective terms than sarcastic messages.

Lexicons related to *Dimensional Models of Emotions* (Table 3) also introduce interesting patterns: messages marked with #irony almost always contain a smaller amount words belonging to these resources. In contrast, #not messages always have a large number of words belonging to these dimensions, i.e. Arousal, Dominance from ANEW or Imagery from DAL. We can also notice a larger frequency of terms related to Imagery and Sensitivity (SN) in #irony than in #sarcasm, whereas we observe a higher use of words related to Dominance (DAL), Attention and Aptitude (SN) in #sarcasm than in #irony. These findings support the idea that irony is more subtle than sarcasm, while a higher degree of aggressiveness can be detected in sarcasm. Results related to the degree of pleasantness produced by words (DAL and SN) and valence (ANEW) of words are higher in sarcastic and #not messages than in ironic ones. This is in tune

	Resource	#irony	#sarcasm	#not
Sentiment Polarity	AFINN*	33.63	**47.89**	47.14
	SN_polarity*	51.28	55.54	**56.59**
	SN_formula*	26.11	37.31	**41.05**
	SO*	39.53	45.32	**45.54**
	GI_pos	1.68	**2.65**	2.53
	HL_pos	2.33	**4.97**	4.62
	SWN_pos*	11.52	**15.43**	14.12
	SUBJ_weak_pos	2.18	**2.69**	2.62
	SUBJ_strong_pos	2.46	**4.83**	4.44
	GI_neg	**1.26**	1.00	0.91
	HL_neg	**3.15**	2.53	2.31
	SWN_neg*	**11.98**	10.49	10.20
	SUBJ_weak_neg	**1.78**	1.51	1.49
	SUBJ_strong_neg	**1.77**	1.7	1.34
	SWN_obj*	**87.97**	84.64	87.05
	EWN_pos	7.61	8.54	**9.61**
	EWN_neg	4.34	4.20	**4.89**
	EWN_null	8.40	9.21	**10.26**

Table 2: Normalized counts for *sentiment polarity* features: values for resources with * are based on scores according to Equation 1. Higher scores are in bold.

	Resource	#irony	#sarcasm	#not
Dimentional Models of Emotions	ANEW_val*	51.24	54.81	**60.03**
	ANEW_arousal*	44.84	45.44	**48.63**
	ANEW_dominance*	46.14	47.59	**52.07**
	DAL_pleasantness*	61.72	63.46	**64.09**
	DAL_activity*	56.25	56.55	**57.22**
	DAL_imagery*	51.81	50.21	**52.12**
	SN_pleasantness*	50.61	55.54	**56.70**
	SN_attention*	50.83	52.10	**52.24**
	SN_sensitivity*	51.11	49.56	**51.19**
	SN_aptitude*	52.44	56.82	**57.80**

Table 3: Normalized counts for *Dimensional Models of Emotions*: values for resources with * are based on scores according to Equation 1. Higher scores are in bold.

with the *sentiment polarity* values, confirming what we already noticed before.

Lexicons related to *Emotional Categories* (Table 4) allow to detect further regularities. While terms related to positive emotions (joy, love, like) are nearly always more frequent in #sarcasm and #not messages, whereas negative emotions terms (anger, fear, disgust, sadness) in EmoLex and LIWC are more frequent in #irony ones. This confirms, at a finer level of granularity, that ironic tweets contain more positive words than sarcastic ones.

To sum up, the quantitative analysis carried out above suggests the following considerations concerning the distinction between irony and sarcasm, the role of the #not hashtag and the polarity reversal phenomenon.

Irony is more subtle than sarcasm. Analysis over affective content shows that irony is more creative and less evident than sarcasm. We observed traces of it in the values of ANEW and DAL affective lexicons. In particular higher values of Imagery, Activation, Arousal and Dominance show that irony is more subtle than sarcasm, and a more in-depth cognitive process is activated. On the other hand, lower values for sarcasm of Valence, Imagery and Pleasantness suggest that it is more direct and less creative than irony. Words related to fear, anger, and sadness are more frequent in #irony than in

	Resource	#irony	#sarcasm	#not
Emotional Categories	EmoLex_anger	**1.59**	1.13	1.10
	EmoLex_anticipation	1.70	2.41	**2.60**
	EmoLex_disgust	**1.03**	0.83	0.90
	EmoLex_fear	**1.62**	1.14	1.14
	EmoLex_surprise	0.78	1.05	**1.30**
	EmoLex_joy	1.54	2.72	**2.75**
	EmoLex_sadness	**1.55**	1.12	1.10
	LIWC_PosEmo	1.71	**3.71**	3.59
	LIWC_NegEmo	**1.25**	1.13	1.08
	EmoSN_joy	21.63	20.5	**21.99**
	EmoSN_sadness	**2.30**	2.21	2.21
	EmoSN_surprise	**1.61**	1.38	1.45
	SS_anticipation	0.84	0.91	**1.06**
	SS_joy	0.40	**0.89**	0.72
	SS_disgust	1.56	1.67	**1.81**
	SS_like	1.73	**2.91**	2.65
	SS_love	0.33	0.89	**0.94**

Table 4: Normalized counts for *emotional categories*: values for resources with * are based on scores according to Equation 1. Higher scores are in bold.

#sarcasm.

#not is a category on its own. Values of both affective and polarity largely suggest that tweets tagged with #not are a category on their own. Although #not is used quite often with a figurative meaning closer to sarcasm from a perspective of polarity and emotional contents, from a cognitive viewpoint it shows a certain similarity with irony. In fact the values obtained in terms of Pleasantness, Activation, Imagery, Valence, Arousal and Dominance are higher than in the case of #sarcasm. On the contrary, sentiment polarity values are very similar to sarcasm ones. By using the tag #not the speaker manifests the intention of dissociating himself from the literal content of the post, as in *self-mockery*. The impression is that such explicit dissociation introduces an attenuation with respect to the aggressiveness typical of sarcasm (e.g. tweet *tw3* in the Introduction).

4.2.1. Polarity reversal

Sentiment polarity values and the use of emotion words related to positive emotions discussed above show that sarcastic and #not messages contain more positive words than the ironic ones. This finding is in line with what was empirically shown also in [7], where the following hypothesis has been tested: "Given the fact that sarcasm is being identified as more aggressive than irony, the sentiment score in it should be more positive".

In this section, we further investigated the role of the polarity reversal in the three kinds of figurative messages, also in order to understand when the expressed sentiment is only superficially positive. A correlational study is presented in Table 5. The results offer further interesting suggestions related to the polarity reversal phenomenon. No relation exists between the polarity values calculated by lexical resources (RES) and the annotation, considering the whole Corpus (C). Our experiment consists in forcing the reversal of RES polarity values for one kind of tweets at a time. Then, we calculate the correlations between these groups and the annotated values. Thus, in $revI$ group we only forced the reversal of the RES values for messages tagged with #irony. The same is for #sarcasm ($revS$), #not ($revN$), and both #sarcasm and #not ($revSN$). This clearly states how the correlation improves with the reversal of #sarcasm and

RES	C	revI	revS	revN	revSN
AFINN	0.032	0.018	0.096	0.096	0.160
GI	0.116	0.109	0.168	0.175	0.228
HL	0.128	0.118	0.188	0.172	0.236
SN_pol	0.006	0.001	0.158	0.145	0.268
SN	0.058	0.049	0.179	0.180	0.297
SWN	0.062	0.065	0.115	0.115	0.168

Table 5: Correlation (p-value < 0.001) between scores from lexical resources (RES) and polarity of the annotation in the Corpus (C), forcing the reversal for Irony (revI), Sarcasm (revS), Not (revN), and both Sarcasm and Not (revSN). Darker\lighter shades indicate higher\lower values.

F-1	Iro - Sar	Iro - Not	Sar - Not
Naïve Bayes	65.4	67.5	57.7
Decision Tree J48	63.4	69.0	62.0
Random Forest	**69.8**	**75.2**	**68.4**
SVM	68.6	74.5	66.9
LogReg	68.7	72.4	64.6

Table 6: F-measure values (multiplied by 100) for each binary classification with all features. The underlined values are not statistically significant (t-test with 95% of confidence value)

#not, while the polarity reversal phenomena is less relevant for ironic messages.

We also carried out a qualitative analysis, showing that sarcasm is very often used in conjunction with a seemingly positive statement, to produce a negative one. Very rarely sarcasm involves a negative statement, to produce a positive one.

This is in accordance with theoretical accounts stating that expressing positive attitudes in a negative mode are rare and harder to process for humans [3].

5. Classification experiments

On the basis of the results obtained in identifying differences among the three kinds of figurative messages, we formulate an experimental setting in terms of a classification task. A novel set of structural and affective features is proposed to perform binary

	#irony-vs-#sarcasm					#irony-vs-#not					#sarcasm-vs-#not				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Each set individually															
Str	59.6	60.3	60.9	61.2	61.3	**66.0**	**68.0**	68.6	**69.6**	67.2	**58.9**	**66.2**	**64.5**	**66.1**	**62.6**
SA	**64.1**	**64.4**	**66.2**	**65.1**	**68.0**	63.8	64.4	**70.2**	68.7	**68.0**	54.0	55.5	58.2	57.9	57.4
EC	61.6	62.1	61.7	52.9	63.4	65.0	65.8	64.4	66.2	66.1	54.1	55.3	54.7	56.9	56.4
DM	54.0	57.7	59.9	60.0	59.5	56.9	60.8	63.3	62.6	62.2	53.5	55.1	54.2	56.1	55.5
Combination between sets															
SA+EC	64.4	62.2	67.9	66.1	66.0	67.0	65.3	70.1	68.8	68.5	54.5	54.7	59.7	58.8	58.0
SA+DM	63.5	60.4	66.6	65.7	65.3	64.1	66.6	69.9	67.7	67.6	54.4	54.7	58.8	58.3	58.6
SA+Str	**64.7**	63.2	**69.3**	**67.3**	**67.6**	**67.9**	**69.8**	**75.2**	**73.4**	**71.7**	**58.9**	62.7	**68.3**	66.5	**64.3**
Str+EC	**64.7**	**63.6**	67.5	65.9	66.8	**67.9**	69.7	74.0	72.6	70.3	**58.9**	63.7	67.8	65.5	63.1
DM+EC	62.6	60.7	64.8	64.9	64.5	63.0	63.7	68.1	67.7	66.8	54.5	54.1	56.6	57.5	56.8
DM+Str	59.4	59.6	64.9	64.0	64.6	64.9	67.1	72.7	71.9	69.7	58.2	**64.0**	67.7	**66.9**	63.7

Table 7: Comparison of classification methods using different feature sets. The underlined values of F-measure (multiplied by 100) are not statistically significant (t-test with 95% of confidence value)

classification experiments: #irony-vs-#sarcasm (Iro - Sar), #irony-vs-#not (Iro - Not) and #sarcasm-vs-#not (Sar - Not). The best distinguishing features have been grouped in four sets, including common patterns in the structure of the messages (Str), sentiment analysis (SA), emotional ($Emot$) features. Structural features include: length, count of colons, question and exclamation marks (PM), part-of-speech tags (POS). Tweet features ($TwFeat$) refer to the frequency of hashtags, mentions and a binary indicator of retweet. Emotional features belong to two kinds of groups: "Emotional Categories" (EC) and "Dimensional Models" (DM) of emotions. The first group includes LIWC (positive and negative emotions), EmoSenticNet (surprise, joy, sadness), EmoLex (joy, fear, anger, trust) and SentiSense (anticipation, disgust, joy, like, love). The second group includes ANEW (Valence, Arousal, Dominance), DAL (Pleasantness, Activation and Imagery) and SenticNet four dimensions (Pleasantness, Attention, Sensitivity and Aptitude). In addition, the Sentiment Analysis set is composed by

	#irony-vs-#sarcasm					#irony-vs-#not					#sarcasm-vs-#not				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Structural + each resource from SA and Emotional															
Str+AFINN	**63.7**	64.8	66.4	65.6	65.7	**67.3**	**70.8**	72.7	**71.8**	**70.1**	**58.8**	**65.7**	66.4	**66.5**	62.8
Str+HL	63.3	**64.9**	66.3	**66.0**	**66.1**	66.7	70.4	71.6	71.7	68.9	58.6	65.0	65.3	66.1	62.5
Str+GI	59.5	60.5	60.8	61.4	62.2	65.0	67.0	68.2	68.7	66.4	58.6	64.9	64.4	66.0	62.5
Str+SWN	60.0	61.4	**65.1**	62.2	64.5	66.3	69.1	**73.0**	70.8	69.8	58.7	64.7	66.9	66.1	63.1
Str+SN_dim	59.1	58.6	62.9	61.4	62.1	65.0	65.9	70.1	69.8	67.3	58.5	64.6	66.1	65.9	62.9
Str+EWN	57.8	58.1	61.1	60.5	61.4	64.5	65.9	68.8	68.2	65.7	**58.8**	64.3	66.0	65.0	62.6
Str+SO	58.0	60.2	61.6	61.4	60.6	63.7	67.3	69.1	69.0	65.6	56.7	65.4	65.3	66.1	62.5
Str+LIWC	62.7	63.7	64.2	64.8	64.9	66.6	69.6	70.8	70.9	68.6	58.4	64.7	65.1	66.2	62.5
Str+EmoLex	58.6	59.5	61.8	61.2	61.9	65.0	67.5	69.5	69.5	66.5	58.5	64.6	65.3	66.1	62.5
Str+EmoSN	58.3	58.2	60.7	60.2	60.9	66.0	67.1	70.2	68.9	67.2	**58.8**	63.7	65.7	64.9	62.5
Str+SS	61.6	62.4	63.8	63.1	64.1	65.7	68.3	70.1	69.9	67.6	**58.8**	64.4	65.8	66.3	62.6
Str+ANEW	58.1	59.1	62.2	60.9	61.1	64.7	66.6	69.3	68.8	66.2	58.3	65.4	66.2	66.1	62.5
Str+DAL	57.6	58.7	63.1	62.5	63.3	64.7	66.7	70.6	70.0	68.1	58.6	65.0	**67.0**	66.4	**63.2**
Str+SUBJ	60.5	61.7	64.6	63.6	64.0	65.7	68.7	71.3	70.3	67.8	58.6	63.6	66.4	65.8	62.5

Table 8: Comparison of classification methods using different feature sets. The underlined F-measure values (multiplied by 100) are not statistically significant (t-test with 95% of confidence)

features extracted from SN (SN_polarity and SN_formula), referred as SN_pol in the following tables, positive, negative and polarity values[14] from AFINN, HL, General Inquirer, SentiWordNet, SUBJ, SO and EffectWordNet. Finally, our tweet representation is composed of 59 features ($AllFeatures$ henceforth) that have been evaluated over a corpus of 30,000 tweets equally distributed in three categories: 10,000 tweets labeled with #irony and 10,000 with #sarcasm retrieved by [8]. In addition, a novel dataset of 10,000 tweets with the #not hashtag has been retrieved. The criteria adopted to automatically select only samples of figurative use of #not were: having the #not in the last position (without considering urls and mentions) or having the hashtag fol-

[14]We consider polarity values as the difference between the positive and the negative scores.

	#irony-vs-#sarcasm					#irony-vs-#not					#sarcasm-vs-#not				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
SA + each resource from Emotional															
SA+LIWC	64.2	**61.3**	66.7	65.5	65.2	**65.0**	64.5	**70.7**	69.3	68.1	53.8	**55.2**	58.3	58.3	57.5
SA+EmoLex	64.2	60.6	66.7	65.2	65.2	63.3	64.3	70.3	68.9	67.9	52.3	54.2	57.8	56.4	56.9
SA+EmoSN	64.0	60.0	**66.8**	65.2	65.0	64.2	**64.8**	70.6	**69.0**	68.2	54.9	54.4	58.8	58.2	58.2
SA+SS	64.2	61.2	66.7	65.2	65.4	64.6	64.6	70.4	69.0	**68.2**	**55.0**	**55.2**	**59.3**	**58.5**	58.2
SA+ANEW	64.2	60.6	66.5	65.3	65.0	63.6	64.5	70.6	68.8	68.0	53.9	**55.2**	58.7	58.3	57.4
SA+DAL	63.8	60.2	66.6	**65.7**	**65.5**	63.9	64.4	70.2	69.0	68.0	54.6	**55.2**	58.6	58.1	**58.5**
SA+SN_dim	**64.3**	60.6	66.5	65.1	65.0	63.4	64.4	70.6	68.8	68.0	53.8	54.9	58.5	58.0	57.7
Emotional (EC+DM) + each one of the resources from SA															
Emot+AFINN	63.8	**61.8**	65.8	65.3	64.9	**64.4**	64.1	68.9	67.8	67.3	54.4	54.4	57.0	57.7	57.3
Emot+HL	**64.1**	**61.8**	66.2	**65.6**	**65.7**	**64.4**	**65.1**	69.1	**68.6**	67.6	54.5	**54.6**	56.7	57.7	57.0
Emot+GI	62.6	60.9	65.2	64.7	64.8	63.1	63.4	68.0	67.7	67.0	54.5	54.3	56.6	57.8	57.1
Emot+SWN	63.2	60.7	66.0	**65.6**	65.4	63.3	63.7	68.9	68.3	67.6	54.9	53.8	57.1	57.7	56.9
Emot+SN_pol	62.4	61.3	64.7	64.5	64.6	64.1	63.5	69.1	67.8	**67.7**	**55.1**	54.4	**57.8**	57.8	**58.6**
Emot+EWN	62.1	60.5	65.4	64.6	64.6	63.0	63.5	67.7	67.4	66.4	55.0	53.9	57.5	**58.6**	57.4
Emot+SO	62.4	61.1	65.8	64.8	64.5	61.8	64.9	68.3	67.6	66.5	53.1	54.1	56.4	57.6	56.8
Emot+SUBJ	63.4	61.1	**66.5**	**65.6**	65.6	63.5	63.7	**69.5**	68.1	67.3	54.5	54.0	56.9	57.9	56.9

Table 9: Comparison of classification methods using different feature sets. Best performances for each classifier are in bold. The underlined F-measure values (multiplied by 100) are not statistically significant (t-test with 95% of confidence)

lowed by a dot or an exclamation mark. Only a small percentage of tweets selected according to such criteria resulted to be unrelated to a figurative use of #not[15].

The classification algorithms used are: Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM)[16]. We performed a 10-fold cross-validation for each binary classification task. F-measure values are reported in Table 6. Generally, our model is able to distinguish among

[15]The dataset with the IDs of the #not tweets is available upon request.

[16]We used the Weka toolkit: `http://www.cs.waikato.ac.nz/ml/weka/`

Structural - one of the resources each time

	#irony-vs-#sarcasm					#irony-vs-#not					#sarcasm-vs-#not				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Str	59.6	60.3	60.9	61.2	61.3	66.0	68.0	68.6	69.6	67.2	58.9	66.2	64.5	66.1	62.6
Str-lenght	59.2	59.9	58.0	61.1	60.6	**62.8**	66.9	**64.8**	68.0	66.9	**55.7**	63.6	62.0	64.0	61.7
Str-PM	**57.9**	**58.1**	**57.8**	**59.3**	**59.9**	64.8	**66.1**	66.0	**67.7**	**65.2**	58.2	**62.3**	**59.6**	**62.1**	**58.9**
Str-POS	59.2	60.5	58.2	60.7	60.5	65.1	70.0	67.4	69.9	67.1	56.7	66.9	64.8	66.8	62.4
Str-TwFeat	59.8	60.5	58.8	59.9	60.8	66.2	69.0	67.3	69.4	67.0	58.6	65.7	62.7	64.7	60.7

Table 10: Comparison of classification methods with feature ablation. Worst performances for each classifier are in bold, to underline the more relevant role of the feature removed. The underlined values are not statistically significant (t-test with 95% of confidence value)

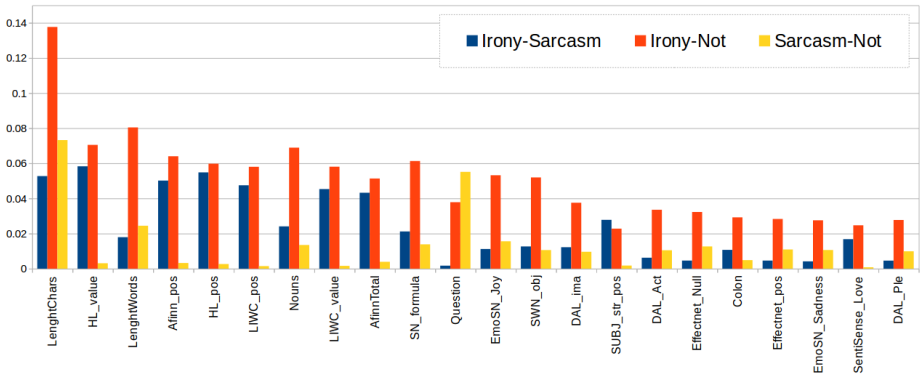

Figure 4: Information Gain values for the 22 best ranked features in binary experiments.

the three kinds of figurative messages. The best result is achieved in #irony vs #not classification using Random Forest (0.75). In the #irony vs #sarcasm task, we improve the state-of-the-art F-measure (same dataset of [8]) from 0.62 to 0.70.

5.1. Analysis of features

To investigate the contribution of the different features further experiments were performed. We divided features into the four main sets already mentioned. Table 7 shows the results for ten different configurations. The first experiment involves the use of each set individually (1st row in Table 7). From the results, we clearly observe that

using only one category of features is not enough. At the same time, we state which group of features are more interesting. Let us comment each subtask. In the *Irony vs Sarcasm* subtask, while the most relevant subsets are *Sentiment Analysis* (0.68 with Logistic Regression) and *Emotional Categories* (0.634), the worst are the $Structural$ and *Dimentional Model of Emotions* ones. These results clearly confirm the usefulness of adopting affective resources in the distinction of irony and sarcasm. This is not so evident in the *Irony vs Not* subtask, where the *Structural* set is the most relevant in the *Sarcasm vs Not* subtask.

A second experiment presents all possible pair combinations constructed from the four sets (i.e., six different pairs). One of the best results, very similar to those reached by *AllFeatures*, is achieved using the "*Sentiment Analysis + Structural*" pair for the *Irony vs Sarcasm* task. In this task, we notice that, while *Structural* features alone are not important as detailed in the previous experiment, the result increases just adding features from *Emotional Categories* or *Sentiment Analysis*. Furthermore, the *Emotional Categories* set, combined both with *Sentiment Analysis* and with *Structural* features, obtains relevant results in all the three subtasks. A strong indicator to distinguish #not tweets, in particular, seems to be the *Structural* feature set. In preliminary analysis, we coherently identified "structural" differences in messages looking at length or punctuation marks. The *Sarcasm vs Not* task is the only one where the *Emotional Categories* set is better than the *Sentiment Analysis* one (i.e., $Str + EC$ is better than $Str + SA$).

To further investigate the obtained results from the perspective of the importance of the affective resources, we took into consideration the individual contribution of individual features. A third experiment includes all pair combinations between the *Structural* features (which seems to be a strong indicator in all the binary classification tasks at issue) and each one of the *Sentiment Analysis* and *Emotional* resources (Table 8).

First, it is important to note that in most cases, an improvement with respect to the state-of-the-art is achieved for *#irony vs #sarcasm*. The higher contribution is given by resources AFINN, HL, LIWC, SentiSense and Subjectivity. In *#irony vs #not*, the F-measure is higher when the *Structural* set is applied together with AFINN, HuLiu,

SentiWordNet, and LIWC, including also Subjecivity, SenticNet, SentiSense, DAL, and EmoSenticNet. In the *#sarcasm vs #not* task, where only DAL slightly improves the results for each classifier, measures are not as clear.

Further experiments are specifically related to *Sentiment Analysis* and *Emotional* sets. Each resource in the *Emotional* set is combined with the *Sentiment Analysis* one and vice versa (Table 9). Generally, adding an *Emotional* resource to the *Sentiment Analysis* set in *#irony vs #not* and *#sarcasm vs #not* tasks, most of the times allows to obtain better results than adding a *Sentiment Analysis* feature to the *Emotional* one. This does not happen in *#irony vs #sarcasm* task.

In a last experiment, we performed feature ablation by removing one feature or one group of features (i.e. all the features belonging to a particular resource) at a time in order to evaluate the impact on the results. First, we investigated the effects of each structural features, in Table 10, where bold values highlight the most important results. A drop in performance for each subtask can be observed when Punctuation Marks (PM) are removed. Furthermore, removing the length features also significantly affects the overall performance for *#irony vs #not* and *#sarcasm and #not* tasks. These results confirm the role of punctuation marks and length, as described by Figures 1 and 2 in Section 4.

Moreover, to measure the contribution of each resource in the *Sentiment Analysis* and *Emotional* sets, we proceeded by feature ablation in Table 11. The most relevant resources are HuLiu in *#irony vs #sarcasm* and *#irony vs #not* tasks, and EffectWordNet in *#sarcasm and #not* task. The most relevant emotional resources are LIWC in *#irony vs #sarcasm* and EmoSenticNet in *#sarcasm and #not* task. Both of them are relevant in the *#irony vs #not* task. As we have already noted, the Dictionary of Affective Language is the most relevant among the *Dimentional Model of Emotions* ones, in the three tasks.

5.2. Information Gain

In order to measure the relevance that a single feature provides in our classification model, we calculated the Information Gain for each binary experiment. According to Figure 4, most features among the best ranked ones (17 over 22) are related to senti-

ment and emotion resources (e.g. HL, AFINN, SN, LIWC, DAL, SWN). This clearly confirms the importance of this kind of features in figurative language processing.

Sentiment and affective features are more relevant in the *Irony vs Sarcasm* task, including terms with positive valence from different lexicons. In particular, 6 over the first 7 features are related to the HL, AFINN and LIWC lexicons.

Structural features are more relevant in the *Irony vs Not* task, together with the Sentiment Analysis ones. In particular, the length of messages both in characters and in words plays an important role. Interestingly, besides the structural features, the three emotional dimensions of DAL are useful to discriminate between figurative messages. Imagery is the most relevant dimension in this task. A special mention is reserved for Objectivity terms from SWN and neutral events from EWN: we think that their relevance could be related to the larger presence of events in #not, detected thanks to the quantity analysis related to EffectWN reported in Table 2.

In the *Sarcasm vs Not* subtask, the structural features play a relevant role, outperforming the other subsets. This is true also for *Irony and Not*, coherently with previous analysis (i.e., punctuation marks play an important role, as observed also in Figure 2). The relevance of question marks is notable. This is coherent with our preliminary analysis and with the idea that a sort of self-mockery is expressed by this kind of messages

The three subtasks clearly indicate the usefulness of adopting lexical resources that linked to semantic information, such as the emotional categories and the dimensional models of emotion groups.

6. Conclusions

In this paper, we investigated the use of figurative language in Twitter. Messages explicitly tagged by users as #irony, #sarcasm and #not were analysed in order to test the hypothesis to deal with different linguistic phenomena. In our experiments we took into account emotional and affective lexical resources, in addition to structural features, with the aim of exploring the relationship between figurativity, sentiment and emotions at a finer level of granularity. Classification results obtained confirm the important role of affective content. In the impact analysis, when sentiment analysis and emotional

resources are used as features, an improvement in the state-of-the-art is achieved in terms of F-measure for #irony vs #sarcasm.

As for the separation of #irony vs #not and #sarcasm vs #not, our results contribute to shed light on the figurative meaning of the #not hashtagging, which emerges as a distinct phenomenon. They can be considered as a baseline for future research on this topic. We also created a dataset to study #not as a category on its own.

In this work we intended to focus on the new task of differentiating between tweets tagged with #irony, #sarcasm and #not. However, since our analysis shows that different kinds of figurative messages behave differently with respect to the polarity reversal phenomenon (see Table 5, Section 4.2), in future work we will experiment the impact of our findings on the sentiment analysis task, investigating if our classification outcome can be a useful precursor to the analysis. In particular, findings reported here about the *polarity reversal* phenomenon in tweets tagged as #sarcasm and #not have been already exploited in a sentiment analysis task by the ValenTo system, obtaining good results [40].

[1] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, A. Reyes, Semeval-2015 task 11: Sentiment analysis of figurative language in twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 470–478.
URL http://www.aclweb.org/anthology/S15-2080

[2] C. J. Lee, A. N. Katz, The differential role of ridicule in sarcasm and irony, Metaphor and Symbol 13 (1) (1998) 1–15.

[3] S. Attardo, Irony as relevant inappropriateness, in: H. Colston, R. Gibbs (Eds.), Irony in language and thought: A cognitive science reader, Lawrence Erlbaum, 2007.

[4] L. Alba-Juez, S. Attardo, The evaluative palette of verbal irony, in: G. Thompson, L. Alba-Juez (Eds.), Evaluation in Context, John Benjamins Publishing Company, 2014, pp. 93–116.

[5] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as contrast between a positive sentiment and negative situation., in: Proceedings of the EMNLP: Conference on Empirical Methods in Natural Language Processing, 2013, pp. 704–714.

[6] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in Twitter, Language Resources and Evaluation 47 (1) (2013) 239–268.

[7] A. P.-Y. Wang, #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 27), Department of English, National Chengchi University, 2013, pp. 349–356.
URL http://aclweb.org/anthology/Y13-1035

[8] F. Barbieri, H. Saggion, F. Ronzano, 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, 2014, pp. 50–58.
URL http://www.aclweb.org/anthology/W/W14/W14-2609

[9] C. Liebrecht, F. Kunneman, 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, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 29–37.

[10] R. W. Gibbs Jr, J. O'Brien, Psychological aspects of irony understanding, Journal of pragmatics 16 (6) (1991) 523–530.

[11] A. Reyes, P. Rosso, On the difficulty of automatically detecting irony: Beyond a simple case of negation, Knowl. Inf. Syst. 40 (3) (2014) 595–614.

[12] C. Bosco, V. Patti, A. Bolioli, Developing corpora for sentiment analysis: The case of irony and Senti-TUT, IEEE Intelligent Systems 28 (2) (2013) 55–63.

[13] R. L. Brown, The pragmatics of verbal irony, Language use and the uses of language (1980) 111–127.

[14] R. J. Kreuz, R. M. Roberts, The empirical study of figurative language in literature, Poetics 22 (1) (1993) 151–169.

[15] A. Bowes, A. Katz, When sarcasm stings, Discourse Processes: A Multidisciplinary Journal 48 (4) (2011) 215–236.

[16] S. McDonald, Neuropsychological studies of sarcasm, in: H. Colston, R. Gibbs (Eds.), Irony in language and thought: A cognitive science reader, Lawrence Erlbaum, 2007, pp. 217–230.

[17] D. Davidov, O. Tsur, A. Rappoport, Semi-supervised recognition of sarcastic sentences in Twitter and Amazon, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10, Association for Computational Linguistics, 2010, pp. 107–116.

[18] R. González-Ibáñez, S. Muresan, N. Wacholder, Identifying sarcasm in Twitter: a closer look, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics, 2011, pp. 581–586.

[19] D. Maynard, M. A. Greenwood, 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, European Language Resources Association, 2014, pp. 4238–4243.

[20] E. Filatova, Irony and sarcasm: Corpus generation and analysis using crowdsourcing, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 392–398, aCL Anthology Identifier: L12-1386.
URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/661_Paper.pdf

[21] R. W. Gibbs, H. L. Colston (Eds.), Irony in Language and Thought, Routledge (Taylor and Francis), New York, 2007.

[22] F. Å. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, Proceedings of the Workshop on Making Sense of Microposts.

[23] M. Hu, 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, pp. 168–177.

[24] P. J. Stone, E. B. Hunt, 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, USA, 1963, pp. 241–256. doi:10.1145/1461551.1461583.
URL http://doi.acm.org/10.1145/1461551.1461583

[25] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[26] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis, 2014.
URL https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8479

[27] E. Cambria, A. Hussain, Sentic Computing: Techniques, Tools, and Applications, Springer Briefs in Cognitive Computation Series, Springer-Verlag GmbH, 2012.

[28] Y. Choi, J. Wiebe, +/-effectwordnet: Sense-level lexicon acquisition for opinion inference, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1181–1191.
URL http://www.aclweb.org/anthology/D14-1125

[29] M. Taboada, J. Grieve, Analyzing appraisal automatically, 2004.

[30] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[31] T. Wilson, J. Wiebe, P. Hoffmann, 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, USA, 2005, pp. 347–354. doi:10.3115/1220575.1220619.
URL http://dx.doi.org/10.3115/1220575.1220619

[32] J. W. Pennebaker, M. E. Francis, R. J. Booth, Linguistic inquiry and word count: Liwc 2001, Mahway: Lawrence Erlbaum Associates 71 (2001) 2001.

[33] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, Computational Intelligence 29 (3) (2013) 436–465.

[34] R. Plutchik, The Nature of Emotions, American Scientist 89 (4).

[35] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, Emosenticspace: A novel framework for affective common-sense reasoning, Knowledge-Based Systems 69 (2014) 108–123.

[36] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, S. Bandyopadhyay, Enhanced senticnet with affective labels for concept-based opinion mining, IEEE Intelligent Systems 28 (2) (2013) 31–38. doi:http://doi.ieeecomputersociety.org/10.1109/MIS.2013.4.

[37] J. C. de Albornoz, L. Plaza, P. Gervás, Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk,

S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[38] M. M. Bradley, P. J. Lang, Affective norms for english words (anew): Instruction manual and affective ratings, Tech. rep., Center for Research in Psychophysiology, University of Florida, Gainesville, Florida (1999).

[39] 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.

[40] D. I. Hernández Farías, E. Sulis, V. Patti, G. Ruffo, C. Bosco, Valento: Sentiment analysis of figurative language tweets with irony and sarcasm, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 694–698.
URL http://www.aclweb.org/anthology/S15-2117

	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	NB	DT	RF	SVM	LR	NB	DT	RF	SVM	LR	NB	DT	RF	SVM	LR
SA - one of the resources each time															
SA	64.1	64.4	66.2	65.1	68.0	63.8	64.4	70.2	68.7	68.0	54.0	55.5	58.2	57.9	57.4
SA-AFINN	63.0	60.9	65.8	64.8	64.8	62.9	64.3	**69.4**	68.4	67.8	53.9	54.6	58.6	57.7	57.2
SA-HL	**62.7**	60.9	**65.2**	**63.8**	**63.8**	62.7	**63.5**	69.8	**67.5**	**66.9**	54.4	54.1	58.2	57.6	57.3
SA-GI	64.2	61.1	66.2	65.2	65.0	64.0	65.3	69.9	68.9	68.0	54.2	55.4	58.5	57.9	57.4
SA-SWN	63.8	61.2	65.6	64.8	64.6	63.4	64.4	69.8	68.3	67.6	53.4	55.0	57.3	57.4	57.2
SA-SN	64.1	**60.7**	66.2	65.3	65.1	**62.6**	64.5	69.5	68.5	67.5	53.1	54.7	57.6	57.9	**55.8**
SA-EWN	63.8	62.1	66.5	64.8	65.0	63.7	65.4	**69.4**	68.5	67.8	**52.5**	**53.3**	**57.1**	**56.2**	57.0
SA-SO	64.1	61.0	66.1	64.4	65.0	64.2	66.0	69.6	68.0	67.5	55.5	55.3	58.2	58.0	57.4
SA-SUBJ	64.0	61.8	65.5	65.1	64.5	64.2	64.8	70.0	68.7	67.9	53.9	55.3	58.0	57.7	57.4
EC - one of the resources each time															
EC	61.6	62.1	61.7	52.9	63.4	65.0	65.8	64.4	66.2	66.1	54.1	55.3	54.7	56.9	56.4
EC-LIWC	**60.0**	**60.0**	**59.3**	**61.4**	**60.9**	**62.1**	**64.6**	**62.9**	64.6	64.6	54.5	55.4	54.9	57.7	56.5
EC-EmoLex	61.6	62.0	60.2	65.1	63.1	65.2	66.2	64.1	65.8	65.8	54.9	56.3	53.7	57.0	56.6
EC-EmoSN	61.5	62.1	61.5	62.2	62.2	63.1	63.9	63.4	**64.0**	**63.8**	**50.1**	**52.3**	**52.2**	**53.4**	**52.7**
EC-SS	61.7	61.9	59.7	62.5	62.8	64.0	66.1	63.6	66.1	65.7	54.1	56.5	54.3	56.8	56.4
DM - one of the resources each time															
DM	54.0	57.7	59.9	60.0	59.5	56.9	60.8	63.3	62.6	62.2	53.5	55.1	54.2	56.1	55.5
DM-ANEW	54.4	57.6	59.0	59.4	59.3	57.7	60.5	62.7	62.2	61.6	53.9	55.3	54.2	55.6	55.3
DM-DAL	**51.9**	**54.3**	**58.2**	**54.9**	**54.9**	**53.3**	**57.2**	**60.8**	**57.2**	**57.1**	**51.6**	**53.6**	**52.8**	**53.7**	**53.3**
DM-SN_dim	53.7	57.4	58.9	59.4	59.0	57.5	60.7	61.8	62.0	61.8	53.7	55.1	55.0	56.2	55.4

Table 11: Comparison of classification methods with feature ablation. Lowest performances for each classifier are in bold, indicating the greater contribution of the feature removed. The underlined values are not statistically significant (t-test with 95% of confidence value).

Complete list of features

Click here to download Software/code (.ZIP): List_of_features.txt