Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not

Knowledge-Based Systems 000 (2016) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not

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

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

ARTICLE INFO

Article history:
Received 16 November 2015
Revised 16 May 2016
Accepted 17 May 2016
Available online xxx

Keywords:
Figurative language
Affective knowledge
Irony
Sarcasm
Twitter

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. We outperform in terms of F-measure the state-of-the-art results on this dataset. Overall, our results confirm the difficulty of the task, but introduce new data-driven arguments for the separation between #irony and #sarcasm. Interestingly, #not emerges as a distinct phenomenon.

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 various affective reactions, and can behave differently with respect to the polarity reversal phenomenon, as shown in [2]. 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 [3,4] 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 [5]. Furthermore, there is a consistent body of work on computational models for sarcasm detection [6] and irony detection [7] in social media, but only preliminary studies addressed the task to distinguish sarcasm and irony [8,9].

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 messages. 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 [7]. We extend our analysis also to tweets tagged with hashtag #not, previously used

* Corresponding author.
E-mail address: sulis@di.unito.it (E. Sulis).
[1] The first two authors equally contributed to this work.

http://dx.doi.org/10.1016/j.knosys.2016.05.035
0950-7051/

to retrieve sarcastic tweets [6,10], in order to investigate further their figurative meaning. Samples of tweets marked with different hashtags follow:

(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 the dataset is multifaceted. Both sentiment and emotion lexicons, as well as 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 [11] and EmoSenticNet [12], 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 results in #irony vs #sarcasm classification from 0.62 [9] to 0.70, in terms of F-measure.

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 4.2.3: 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 in our approach. Section 4 presents the feature analysis and Section 5 describes our experiments. Section 6 concludes the paper.

2. Irony, sarcasm *et similia*

Many authors embrace an overall view on irony. Broadly speaking, under the umbrella term of irony one can find distinct phenomena such as *situational irony* or *verbal irony* [13–15]. Situational irony (or "irony of fate") refers to the state of affairs or events which is the reverse of what has been expected, while the term verbal irony is applied to refer to a figure of speech, characterized by the possibility of distinguishing between a literal and an intended/implied meaning. In particular, according to many theoretical accounts in ironic utterances the speaker intends to communicate the opposite of what is literally said [16,17], but since such definition does not allow to account for many samples of utterances which are considered ironic, we prefer to refer to a more general position, on which different authors in literature would tacitly agree: "Regardless of the type, or absence, of meaning negation/reversal, the literal import of an ironic utterance differs from the implicit meaning the speaker intends to communicate" [15]. Moreover, we can have an ironic statement, meant as utterance of a speaker which refers to certain aspects of an ironic situation [13].

In linguistics, verbal irony is sometimes used as a synonym of sarcasm [18–20]. 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 [16,21,22], others provide insights for a separation. Sarcasm has been recognized in [23] with a specific target to attack [4,15], more offensive [3] and "intimately associated with particular negative affective states" [24]. According to [3] hearers perceive aggressiveness as the feature that distinguishes sarcasm. Instead, irony has been considered more similar to mocking in a sharp and non-offensive manner [5].

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 [25]. 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 [2]. Therefore, to distinguish between them can be important for improving the performances of systems in sentiment analysis.

For computational linguistics purposes irony and sarcasm are often viewed as the same figurative language device. Computational models for sarcasm detection [6,9,26–28] and irony detection [7,29,30] 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 [8,9]. The current work aims to further contribute to this subject.

Furthermore, a rarely investigated form of irony that can be interesting to study in social media is self-mockery. 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. According to some theoretical accounts: "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'*" [31]. Moreover, the analysis of complex forms of self-mockery in spontaneous conversations in [32] highlighted in-

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

Table 1
Corpus description: number of tweets (N), Mean (MP) and standard deviation (SD) of the polarity, median of the length (ML).

Description	N	MP	SD	ML
With #irony	1737	−1.77	1.41	83
With #sarcasm	2260	−2.33	0.77	66
With #not	3247	−2.16	1.04	71

teresting practices related to *narrative self-mockery*, where people, in particular women, jokingly tell a story about a personal experience, only apparently offering themselves as object of laughing. The same study shows that, in the conversational contexts analyzed, makings jokes about their own (sometime negative) experience provided the narrator with a way to share the experience and jointly create a distance through the mocking: "The narrators are not laughed at and do not invite others to do. [...] They seem to be saying, 'I had such an awful experience', or 'I was so dumb', but it is all done with a narrative strategy which prevents regret, pity or even laughter at their expense." [...] "in the episodes there is no invitation to laugh about the teller, but rather with her" [32]. Investigations on the role of the #not hashtag as a figurative language device could maybe provide insights into this phenomenon by relying on social data, where such data, when connected with information about genre and age, could be also be an interesting new research line for studying the relationship between gender and different forms of irony.

People often use specific markers for communication purposes. Research on the use of different hashtags (particularly #irony, #sarcasm and #not) could be useful in order to investigate if they can be low-salience cues [33], i.e. if Twitter users may use these kinds of markers in order to highlight their non-literal intention. This could be the case especially in short texts (such as tweets), where the lack of context could provoke misunderstanding.

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 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 5114 different hashtags in the corpus, the most used ones are #not (3247 tweets), #sarcasm (2260) and #irony (1737). 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 [25].

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 [34]. 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 Internet slang acronyms such as LOL (laughing out loud). The most recent available version of the dictionary contains 2477 English words.[4] A bias towards negative words (1598, 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 [35]. The final version of the dictionary includes an amount of 6789 words divided in 4783 negative (HL_neg) and 2006 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 [36]. 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 subcategories.[6] The positive words (GI_pos) are 1915, while the negative ones are 2291 (GI_neg).

(iv) SWN: SentiWordNet [37] 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 [11,38,39]. 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 formalism to provide the affective semantics of expressions. Each concept is associated with the four dimensions of the hourglass of emotions [40]: 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.

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

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

[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.

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

(vi) EWN: The EffectWordNet lexicon has been recently developed by Choi [41] 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 3298 positive, 2427 negative and 5296 null events.[8]

(vii) SO: Semantic Orientation is a list of adjectives annotated with semantic-orientation values by Taboada and Grieve [42]. 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 [43]. 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 8222 clues collected by Wilson and colleagues [44] from a number of sources. Some were culled from manually developed resources and others were identified automatically. Each clue can be strongly or 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 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 about 4500 entries distributed in categories that can further be used to analyse psycholinguistic features in texts. We selected two categories for positive and negative emotions: LIWC_PosEmo, with 405 entries, and LIWC_NegEmo, with 500 entries [45].

(x) EmoLex: The resource EmoLex is a word-emotion association lexicon[11] developed at the National Research Council of Canada by Saif Mohammad [46]. The dictionary contains 14,182 words labelled according to the eight Plutchik's primary emotions [47]: sadness, joy, disgust, anger, fear, surprise, trust, anticipation.

(xi) EmoSN: EmoSenticNet is a lexical resource developed by Poria and colleagues [48] [12] that assigns WordNet Affect emotion labels to SN concepts. The whole list includes 13,189 entries for the six Ekman's emotions: joy, sadness, anger, fear, surprise and disgust.[12]

(xii) SS: SentiSense[13] is a concept-based affective lexicon that has been developed by Carrillo de Albornoz [49]. It attaches emotional meanings to concepts from the WordNet lexical database and consists of 5496 words and 2190 synsets labelled with an emotion from a set of 14 emotional categories, which are related by an antonym relationship.

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 which refer to dimensional approaches to emotion modeling:

(xiii) ANEW: Affective Norms for English Words is a set of normative emotional rating [50]. Each word in the dictionary is rated from 1 to 9 in terms of the Valence-Arousal-Dominance (VAD) model [51]. The first dimension concerns the valence (or pleasantness) of the emotions invoked by the word, going from unhappy to happy. The second one addresses the degree of arousal evoked by the word, whereas the third one refers to the dominance/power of the word, the extent to which the word denotes something that is weak/submissive or strong/dominant. This work considers the three dimensions separately.

(xiv) DAL: Dictionary of Affective Language developed by Whissell [52] contains 8742 English words rated in a three-point scale.[14] 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 (Fig. 1). Moreover, the mean value of tweets marked with #irony is −1.73 instead of −2.33 for the #sarcasm ones. The differences between the means are statistically significant according to one-way ANOVA (p-value of $3.24e^{-97}$).

These differences show that sarcasm is perceived as more negative than irony by the hashtag adopters in our corpus. On the contrary, ironic messages are more positive as suggested by the above mentioned mean values as well as the little "hill" in the slope. This is a signal that #irony is also used positively (as in positive evaluative irony, i.e. ironic praise), whereas #not and #sarcasm are usually not.

A first hypothesis coming from these results is that Twitter users consider irony as a more nuanced and varied phenomenon in terms of the associated sentiment (see Section 4.2.1 for further remarks on this issue).

These distributions also signal initially that messages tagged with #not can be considered somehow different from #sarcasm and #irony ones.

In the following, we will 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. Structural and affective features are considered.

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, retweets 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

[8] http://mpqa.cs.pitt.edu/.
[9] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.
[10] http://www.liwc.net.
[11] http://www.saifmohammad.com/WebPages/lexicons.html.
[12] http://www.gelbukh.com/emosenticnet/.
[13] http://nlp.uned.es/~jcalbornoz/SentiSense.html.
[14] ftp://perceptmx.com/wdalman.pdf.

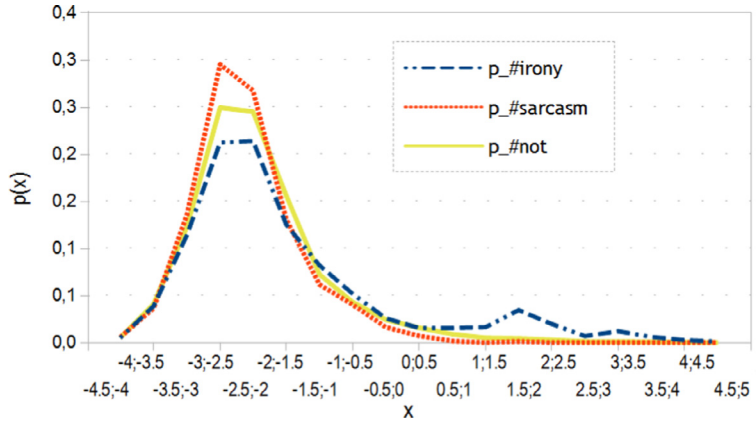

Fig. 1. Distribution of tweets by polarity, p(x) is the probability that a tweet has polarity x.

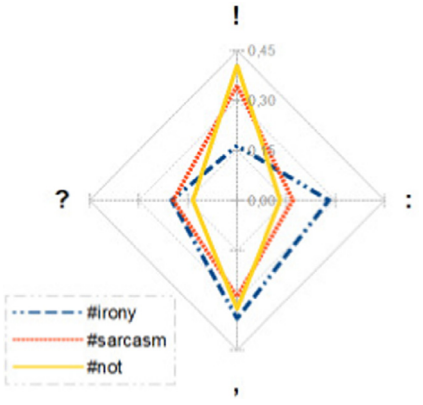

Fig. 2. Distribution of punctuation marks in the corpus: colons are most used in #irony tweets, exclamation marks in #sarcasm and #not ones, question marks are less used in #not tweets.

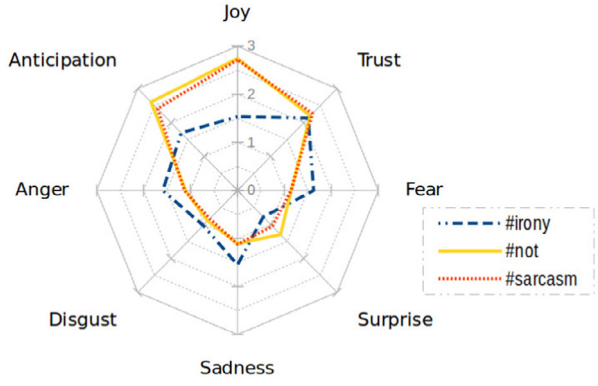

Fig. 3. Distribution of emotion words (EmoLex [46]) in the SemEval Task 11 corpus: #not and #sarcasm tweets overlap, while #irony shows a different behavior.

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).

Punctuation marks. Fig. 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*). This can be linked to the typical grammatical construction of this kind of messages: first a statement, and then the reversal of this statement by the marker #not. Obviously, questions are not easily reversed.

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 Fig. 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 behavior.

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); EmoSN, EmoLex, SS and LIWC (emotional categories); AFINN, HL, GI, SWN, EWN, SO, SUBJ and both the SenticNet sentiment polarity values mentioned above.

The values of 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 4, where final values are multiplied by 100 to improve the readability. To investigate the statistical significance on the difference between the

Table 2
Normalized counts for *sentiment polarity* features: values for resources with ∗ are based on scores according to Eq. 1. For each resource, higher scores are in bold if they are statistically significant.

	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**	**1.70**	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 3
Normalized counts for *dimensional models of emotions*: values for resources with ∗ are based on scores according to Eq. 1. For each resource, higher scores are in bold if they are statistically significant.

	Resource	#irony	#sarcasm	#not
Dimensional 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_activation∗	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 4
Normalized counts for *emotional categories*. For each resource, higher scores are in bold if they are statistically significant.

	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**

mean scores, we performed an ANOVA on our three distributions for each individual resource. Moreover, we computed a Z-test on each pair of distributions [53]. Tables contain in bold, for each lexical resource the highest values which are also statistically significant. In some cases the uncertainty is due to the high variance.

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, GI, SWN, SUBJ, SN 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 EWN, 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, 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 of 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 in #irony than in #sarcasm, whereas we observe a higher use of words related to Dominance (DAL) in #sarcasm than in #irony. These findings support the idea that irony is more creative than sarcasm (see Section 4.2.1 for a deeper discussion on this issue). Results related to the degree of pleasantness produced by words (DAL and SN) and valence of words (ANEW) are higher in sarcastic and #not messages than in ironic ones. This is in tune with the *sentiment polarity* values, confirming what we already noticed before.

Lexicons related to *emotional categories* (Table 4) allow to detect further regularities. 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 granularity level (i.e. the one of *emotional categories*), our findings at the sentiment polarity level, e.g., ironic tweets contain more negative words than the 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.

4.2.1. Irony is more creative and implicit than sarcasm

Analysis over affective content suggests that irony is more creative than sarcasm, and it is used to convey implicit emotions, whereas sarcasm messages are far more explicit. For what concerns the first aspect, we observed traces of it in the values of *dimensional models of emotion* lexica. In particular, we observe higher values for the dimension Imagery of DAL. Such dimension gives a measure of how difficult is to form a mental picture of a given term. In other words, it provides an estimate for a lexical items efficacy in activating mental images associated with the concept. We think that these results can be interpreted as indicating that irony is more creative than sarcasm. Focusing on sarcasm, we observe not only lower values of Imagery but also higher values of Dominance. Let us recall that the latter dimension from ANEW gives a measure about the fact that the word denotes something that is weak/submissive or strong/dominant. Higher values of Dominance are signals of the fact that words making people feel in control are more frequent in #sarcasm messages than in #irony messages.

For what concerns the second aspect, i.e. the use of the different hashtags #irony and #sarcasm for conveying explicit or implicit figurative messages by Twitter users, when we look at those resources which include information about emotions (see

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

for instance the distribution in Fig. 3) we can observe that words related to negative emotions (fear, anger, and sadness) are more frequent in #irony than in #sarcasm, but, more importantly, #sarcasm is usually accompanied by emotions with higher intensity than irony. For instance, the intensity of some emotions such as joy and anticipation in #sarcasm messages is clearly higher. This could be also meant as a signal of the fact that ironic messages are used to convey implicit messages, whereas sarcasm is more explicit.

Finally, focusing on sentiment lexica, we observe that sarcasm tends to involve more positive words than irony. However, as shown by Fig. 1, #irony messages are also used positively, when we look at the figurative, intended meaning, whereas #sarcasm messages are usually not. A first hypothesis is that Twitter users consider irony as a more nuanced and varied phenomenon in terms of the associated sentiment. Another interesting hypothesis could be that Twitter users exploit the hashtag #irony for marking situational irony. In fact, in such cases normally speakers humorously lament a situation, without intending to negate the literal meaning of the utterance, in other words without disengaging from what is said. This would be in tune with the lower frequency of negative polarity terms and lower values for intensity of emotions observed in messages marked with #irony. In fact, ironic utterances referring to certain aspects of an ironic situation can also come without evaluative remarks, but only with the observation that something in a situation is ironical.

4.2.2. Is #not a category on its own? Comparison with # irony and #sarcasm

Values related to affect and polarity suggest that tweets tagged with #not could be considered as a category on their own. On the one hand, #not is used quite often with a figurative meaning closer to sarcasm from a perspective of sentiment polarity and finer-grained emotional contents. Tweets marked by #sarcasm and #not are usually accompanied by explicit emotions with higher intensity. Moreover, sentiment polarity values are very similar to sarcasm ones and tend to involve words with positive sentiment and emotions, intending the opposite of what they mean. These results are consistent with findings showing that sarcasm is easier to derive with positive than with negative concepts, and with the idea that people tend to use positive terms to express indirectly that something is negative [54,55], think for instance to the verbal politeness issue: asserting directly that a particular person has an unfavourable quality is not polite.

However, the #not messages show some peculiarities. By using the tag #not the speaker explicitly manifests the intention of dissociating herself from the literal content of the post, as in certain forms of self-mockery. The impression is that such explicit dissociation introduces an attenuation with respect to the aggressiveness which apparently characterize messages marked with #sarcasm (e.g. tweet *tw3* in the Introduction). Moreover, #not messages differ from #sarcasm messages in that they use negation to invite a sarcastic interpretation of the message. Overall, this seems to be in line with the findings in [18–20], where the role of negation, as low-salience marker that can affect sarcastic non-literal interpretations is studied, and the role of negation as a "mitigator, retaining in memory the concept within its scope while slightly attenuating it" [18] is highlighted. Referring to this theoretical framework, we can hypothesize to consider the #not hashtag as a negation marker used to achieve a non-literal interpretation of the messages, which characterize, in Twitter negative constructions, expressions of more implicit form of sarcasm or self-mockery. Let us also observe that, although #not is used quite often with a figurative meaning closer to sarcasm, when we look at the information related to resources such as DAL, which include dimensions referring to cognitive processes, such as Imagery, it shows a certain similarity with irony. For instance, the values obtained in terms of Imagery, Valence, and Dominance are higher than in the case of #sarcasm.[15]

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.

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

4.2.3. 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 [8], 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 investigate 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 #not, while the polarity reversal phenomena is less relevant for ironic messages.

A preliminary manual analysis of the corpus has been performed by two human evaluators with the aim to explore the direction of the polarity reversal phenomenon in sarcastic tweets (i.e., from the positive literal polarity to the negative intended one, or *vice versa*). Such analysis shown that sarcasm is very often used in conjunction with a seemingly positive statement, to reflect a negative one, but very rarely the other way around. In fact, tweets marked with the hashtag #sarcasm and tagged with a positive polarity score were very few in the Semeval2015-Task11 corpus (only 18). Among them, human evaluators could detect only three tweets expressing a literally negative statement, that finally reverted to an intended positive one, as for instance:"RT GregCooper: These annoying home buyers want to purchase my listings before the sign actually goes up. How inconvenient. #sarcasm #grate". This is in accordance with theoretical accounts stating that expressing positive attitudes in a negative mode are rare and harder to process for humans [4]. On the contrary, our evaluators have found many tweets expressing a literally positive statement, that was finally

[15] For what concerns higher values of Imagery in words occurring in #not messages than in #sarcasm posts, since such factor is commonly known to affect brain activity, and it is generally accepted, as regards linguistic competence, that visual load facilitates cognitive performance, we can hypothesize that from a cognitive point of view the lexical processing of #not and #sarcasm messages will be different.

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).

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

reverted to an intended negative one, as for instance: "There is nothing better than Pitbull singing 'playoffs' as Timber plays in the background. #sarcasm" or "YAY A TEST AND A BUNCH OF HOMEWORK DUE TOMORROW! I LOVE SCHOOL! #sarcasm".

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 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 three 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), EmoSN (surprise, joy, sadness), EmoLex (joy, fear, anger, trust) and SS (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 features extracted from SN (SN_polarity and SN_formula), referred as SN_pol in the following tables, as well as positive, negative and polarity values[16] from AFINN, HL, GI, SWN, SUBJ, SO and EWN. 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 [9]. 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 followed 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.[17]

The classification algorithms used are: Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM)[18]. 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 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 in terms of F-measure the state-of-the-art results (same dataset of [9]) from 0.62 to 0.70 approximately.

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 *dimensional 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. Notice also that the *structural* set is the most relevant in the *#sarcasm vs #not* subtask. This is coherent with the findings of our preliminary analysis, where "structural" differences in messages have been identified looking at length or punctuation marks.

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* (see Table 6), is achieved using the "*Sentiment Analysis + Structural*" pair for the *#irony vs #sarcasm* task. In this task, it can be noticed 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.

To further investigate the obtained results from the perspective of the importance of the affective resources, we took into consideration the 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 many cases, an improvement with respect to the results in [9] is achieved for *#irony vs #sarcasm*. The higher contribution is given by resources AFINN, HL, LIWC, SS and SUBJ. In *#irony vs #not*, the F-measure is higher when the *structural* set is applied together with AFINN, HL, SWN, and LIWC, including also SUBJ, SN, SS, DAL, and EmoSN. 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. Fur-

[16] We consider polarity values as the difference between the positive and the negative scores.
[17] The dataset with the IDs of the #not tweets is available upon request.
[18] We used the Weka toolkit: http://www.cs.waikato.ac.nz/ml/weka/.

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

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

	#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 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 value).

	#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 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 value).

	#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

thermore, removing the length features also significantly affects the overall performance for *#irony vs #not* and *#sarcasm vs #not* tasks. These results confirm the role of punctuation marks and length, as described by Fig. 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 HL in *#irony vs #sarcasm* and *#irony vs #not* tasks, and EWN in *#sarcasm vs #not* task. The most relevant emotional resources are LIWC in *#irony vs #sarcasm* and EmoSN in *#sarcasm vs #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 *dimensional model of emotions* ones, in the three tasks.

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

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).

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 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).

	#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

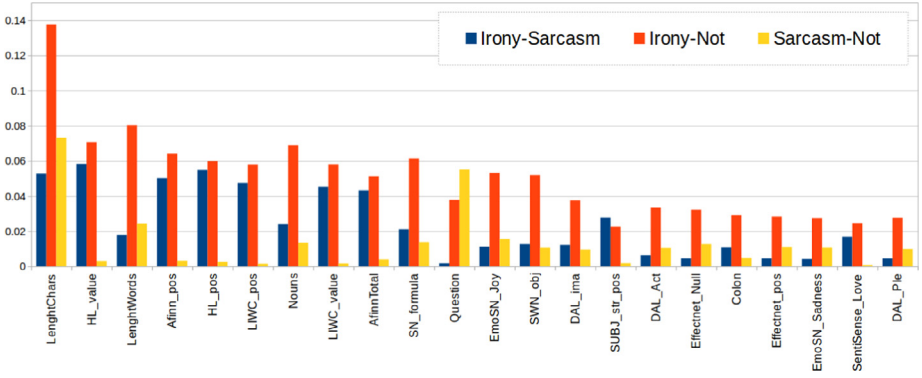

Fig. 4. Information Gain values for the 22 best ranked features in binary experiments.

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 Fig. 4, most features among the best ranked ones (17 over 22) are related to sentiment 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 EWN reported in Table 2.

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035

In the *#sarcasm vs #not* subtask, the structural features play a relevant role, outperforming the other subsets. This is true also for *#irony vs #not*, coherently with previous analysis (i.e., punctuation marks play an important role, as observed also in Fig. 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 one encoded in emotional categories and dimensional models of emotion.

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 particular, when sentiment analysis and emotional resources are used as features, for #irony vs #sarcasm an improvement w.r.t. state-of-the-art results is achieved in terms of F-measure.

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.[19]

An assumption underlying our proposal concerns the reliability of the user-generated hashtags #irony and #sarcasm as labels exploited by Twitter users in English speaking countries to really mark *distinguished* phenomena. Let us notice that the use of hashtags marking irony and sarcasm can be language-specific. It can vary in different languages and cultures, and similar markers in different languages could have different distributions. For what concerns English tweets, in order to get an idea about the distribution of the three hashtags investigated in our study, we collected a sample of English tweets posted on a single day.[20] After some pre-processing steps inspired by [56], mainly devoted to discard re-tweets and to filter out tweets where the hashtags were not used to invite an ironic or sarcastic interpretation of the post, we counted 1461 tweets: 411 marked with #irony, 698 with #sarcasm and 352 with #not. We can observe that the distribution in case of English tweets seems to be not very imbalanced. This is in favor of the hypothesis that users, in this linguistic context, really exploit the three hashtags in order to mark different phenomena. Different findings have been reported about Dutch tweets in [56], where a similar experiment shown that irony-tweets (i.e., tweets marked with #ironie, the Dutch equivalent of #irony) were very rare; in such a scenario it would be hard to state that irony tweets are really exploited by Dutch users in order to mark a phenomenon which is different from sarcasm. A cross-language study of markers for irony and sarcasm could be an interesting strand of future research.

Another interesting direction to further investigate is the educational and socio-demographic background of irony-users and sarcasm-users. Unfortunately, in Twitter explicit meta data about age and gender of users are not provided, thus extracting such information is a further issue that needs to be addressed. Nevertheless, for some authors it is possible to manually inspect the information that they may have published in other social media, e.g. LinkedIn,[21] on their user's profile. For what concerns age, in case the information is not published in the user's profile, it could be approximated taking into account, if present, the information included in the education section, for instance, the degree starting date. For what concerns the information about gender, it could be inferred form the user's photography and name, by following a methodology similar to the one exploited in [57].

In this work we focused on the new task of differentiating between tweets tagged with #irony, #sarcasm and #not, in order to provide some useful insights on the use of these hashtags to label what users consider as ironic or sarcastic content in a social media platform such as Twitter. Investigating the application of our approach in distinguishing between ironic and sarcastic tweets in absence of the explicit hashtags could be also an interesting matter of future work. Moreover, 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 further experiment the impact of our findings on the sentiment analysis task, investigating if our classification outcome can be a useful precursor to the analysis. Some of the results 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 promising results [58].

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Farias (Grant No. 218109/313683 CVU-369616). Paolo Rosso has been partially funded by *SomEMBED* MINECO research project (TIN2015-71147-C2-1-P) and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030). 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 co-funded by Fondazione CRT (WWS Program 2).

References

[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, 2015, pp. 470–478.
[2] C. Bosco, V. Patti, A. Bolioli, Developing corpora for sentiment analysis: the case of irony and Senti-TUT, IEEE Intell. Syst. 28 (2) (2013) 55–63.
[3] C.J. Lee, A.N. Katz, The differential role of ridicule in sarcasm and irony, Metaphor Symbol 13 (1) (1998) 1–15.
[4] S. Attardo, Irony as relevant inappropriateness, in: H. Colston, R. Gibbs (Eds.), Irony in Language and Thought: A Cognitive Science Reader, Lawrence Erlbaum, 2007, pp. 135–172.
[5] 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.
[6] 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 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 704–714.
[7] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in Twitter, Lang. Resour. Eval. 47 (1) (2013) 239–268.
[8] A.P. 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, Department of English, National Chengchi University, 2013, pp. 349–356.
[9] 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, 2014, pp. 50–58.

[19] Available under request.

[20] We retrieved from Twitter Streaming API all tweets in English language (lang: 'en') from 2016-02-01 12:00:00 to 2016-02-02 12:00:00.

[21] http://www.linkedin.com.

[10] 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, 2013, pp. 29–37.
[11] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: C.E. Brodley, P. Stone (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014, pp. 1515–1521.
[12] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, S. Bandyopadhyay, Enhanced SenticNet with affective labels for concept-based opinion mining, IEEE Intell. Syst. 28 (2) (2013) 31–38.
[13] D.C. Littman, J.L. Mey, The nature of irony: Toward a computational model of irony, J. Pragmatics 15 (2) (1991) 131–151.
[14] S. Attardo, Irony, Encyclo. Lang. Linguistics 6 (2006) 26–28.
[15] M. Dynel, Linguistic approaches to (non) humorous irony, Humor - Int. J. Humor Res. 27 (6) (2014) 537–550.
[16] R.W. Gibbs Jr, J. O'Brien, Psychological aspects of irony understanding, J. Pragmatics 16 (6) (1991) 523–530.
[17] D. Sperber, D. Wilson, Relevance: Communication and Cognition, Harvard University Press, Cambridge, MA, USA, 1986.
[18] R. Giora, E. Livnat, O. Fein, A. Barnea, R. Zeiman, I. Berger, Negation generates nonliteral interpretations by default, Metaphor Symbol 28 (2) (2013) 89–115.
[19] R. Giora, S. Givonin, O. Feian, Defaultness reigns: the case of sarcasm, Metaphor Symbol 30 (4) (2015a) 290–313.
[20] R. Giora, A. Drucker, O. Fein, I. Mendelson, Default sarcastic interpretations: on the priority of nonsalient interpretations, Discourse Process. 52 (3) (2015b) 173–200.
[21] R.L. Brown, The pragmatics of verbal irony, Lang. Use. (1980) 111–127.
[22] R.J. Kreuz, R.M. Roberts, The empirical study of figurative language in literature, Poetics 22 (1) (1993) 151–169.
[23] A. Bowes, A. Katz, When sarcasm stings, Discourse Process. 48 (4) (2011) 215–236.
[24] 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.
[25] A. Reyes, P. Rosso, On the difficulty of automatically detecting irony: beyond a simple case of negation, Knowl. Inform. Syst. 40 (3) (2014) 595–614.
[26] 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, in: CoNLL '10, Association for Computational Linguistics, 2010, pp. 107–116.
[27] 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.
[28] D. Maynard, M.A. Greenwood, Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, European Language Resources Association, 2014, pp. 4238–4243.
[29] 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, European Language Resources Association, 2012, pp. 392–398.
[30] I. Hernández-Farías, J.M. Benedí, P. Rosso, Applying basic features from sentiment analysis for automatic irony detection, in: R. Paredes, S.J. Cardoso, M.X. Pardo (Eds.), Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis, Springer International Publishing, 2015, pp. 337–344.
[31] R.W. Gibbs, H.L. Colston (Eds.), Irony in Language and Thought, Routledge (Taylor and Francis), New York, 2007.
[32] H. Kotthoff, Gender and joking: On the complexities of women's image politics in humorous narratives, J. Pragmatics 32 (1) (2000) 55–80.
[33] S. Givoni, R. Giora, D. Berberbest, How speakers alert addresses to multiple meanings, J. Pragmatics 48 (2013) 29–40.
[34] F.Å. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, in: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, 718, CEUR-WS.org, 2011, pp. 93–98.
[35] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '04, 2004, pp. 168–177.
[36] P.J. Stone, E.B. Hunt, A computer approach to content analysis: Studies using the general inquirer system, in: Proceedings of the Spring Joint Computer Conference, in: AFIPS '63 (Spring), ACM, 1963, pp. 241–256.
[37] 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, European Language Resources Association, 2010, pp. 2200–2204.
[38] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, T.S. Durrani, Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis, in: Signal Processing (ICSP), 2012 IEEE 11th International Conference on, 2, IEEE, 2012a, pp. 1251–1255.
[39] S. Poria, A.F. Gelbukh, E. Cambria, D. Das, S. Bandyopadhyay, Enriching SenticNet polarity scores through semi-supervised fuzzy clustering., in: J. Vreeken, C. Ling, M.J. Zaki, A. Siebes, J.X. Yu, B. Goethals, G.I. Webb, X. Wu (Eds.), ICDM Workshops, IEEE Computer Society, 2012b, pp. 709–716.
[40] E. Cambria, A. Hussain, Sentic computing: a common-sense-based framework for concept-level sentiment analysis, 1, Springer, 2015.
[41] 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, Association for Computational Linguistics, 2014, pp. 1181–1191.
[42] M. Taboada, J. Grieve, Analyzing appraisal automatically, in: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004, pp. 158–161.
[43] 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.
[44] 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, Association for Computational Linguistics, 2005, pp. 347–354.
[45] J.W. Pennebaker, M.E. Francis, R.J. Booth, Mahway: Lawrence Erlbaum Associates 71 (2001) 2–23.
[46] S.M. Mohammad, P.D. Turney, Crowdsourcing a word–emotion association lexicon, Comput. Intell. 29 (3) (2013) 436–465.
[47] R. Plutchik, The nature of emotions, Am. Scientist 89 (4) (2001) 344–350.
[48] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, EmoSenticSpace: a novel framework for affective common-sense reasoning, Knowl. Based Syst. 69 (2014) 108–123.
[49] J. Carrillo de Albornoz, J. Plaza, P. Gervás, Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, European Language Resources Association, 2012, pp. 3562–3567.
[50] M.M. Bradley, P.J. Lang, Affective norms for english words (ANEW): instruction manual and affective ratings, Technical Report, Center for Research in Psychophysiology, University of Florida, 1999.
[51] J.A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of Research in Personality 11 (3) (1977) 273–294.
[52] C. Whissell, Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages, Psychol. Reports 2 (105) (2009) 509–521.
[53] S.M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, Academic Press, 2014.
[54] H.L. Colston, "not good" is "bad", but "not bad" is not "good": an analysis of three accounts of negation asymmetry, Discourse Process. 28 (3) (1999) 237–256.
[55] T. Fraenkel, Y. Schul, The meaning of negated adjectives, Intercult. Pragmat. 5 (4) (2008) 517–540.
[56] F. Kunneman, C. Liebrecht, M. van Mulken, A. van den Bosch, Signaling sarcasm: from hyperbole to hashtag, Inform. Process. Manage. 51 (4) (2015) 500–509.
[57] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daelemans, Overview of the 2nd author profiling task at PAN 2014, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (Eds.), CLEF 2014 Labs and Workshops, Notebook Papers, 1180, CEUR-WS.org, 2014, pp. 898–927.
[58] 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, 2015, pp. 694–698.

Please cite this article as: E. Sulis et al., Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.035