Knowledge-Based Systems 108 (2016) 132–143

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 18 May 2016

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-