Lang Resources & Evaluation (2013) 47:239–268
DOI 10.1007/s10579-012-9196-x

A multidimensional approach for detecting irony in Twitter

Antonio Reyes · Paolo Rosso · Tony Veale

Published online: 24 July 2012

Abstract Irony is a pervasive aspect of many online texts, one made all the more difficult by the absence of face-to-face contact and vocal intonation. As our media increasingly become more *social*, the problem of irony detection will become even more pressing. We describe here a set of textual features for recognizing irony at a linguistic level, especially in short texts created via social media such as Twitter postings or "tweets". Our experiments concern four freely available data sets that were retrieved from Twitter using content words (e.g. "Toyota") and user-generated tags (e.g. "#irony"). We construct a new model of irony detection that is assessed along two dimensions: representativeness and relevance. Initial results are largely positive, and provide valuable insights into the figurative issues facing tasks such as sentiment analysis, assessment of online reputations, or decision making.

Keywords Irony detection · Figurative language processing · Negation · Web text analysis

1 Introduction

Web-based technologies have become a significant source of data in a variety of scientific and humanistic disciplines, and provide a rich vein of information that is easily mined. User-generated Web 2.0 content (such as text, audio and images)

A. Reyes (✉) · P. Rosso
Natural Language Engineering Lab, ELiRF, Universidad Politécnica de Valencia, Valencia, Spain
e-mail: areyes@dsic.upv.es

P. Rosso
e-mail: prosso@dsic.upv.es

T. Veale
School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
e-mail: tony.veale@UCD.ie

provides knowledge that is topical, task-specific, and dynamically updated to broadly reflect changing trends, behavior patterns and social preferences. Consider, for instance, the work described in Pang et al. (2002) which shows the role of implicit knowledge in automatically determining the subjectivity and polarity of movie reviews, or the findings reported in Balog et al. (2006) regarding the role of user-generated tags for analyzing mood patterns among bloggers.

This paper deals with a specific aspect of human communication that relies precisely on this kind of information: irony. This linguistic phenomenon, which is widespread in web content, has important implications for tasks such as sentiment analysis (cf. Reyes et al. 2009 about the importance of determining the presence of irony in order to assign fine-grained polarity levels), opinion mining (cf. Sarmento et al. 2009), where the authors note the role of irony in discriminating negative from positive opinions), and advertising (cf. Kreuz 2001, about the function of irony to increase message effectiveness in advertising), among others.

As described in these research efforts, the problem of irony cuts through every aspect of language, from pronunciation to lexical choice, syntactic structure, semantics and conceptualization. As such, it is unrealistic to seek a computational silver bullet for irony, and a general solution will not be found in any single technique or algorithm. Rather, we must try to identify specific aspects and forms of irony that are susceptible to computational analysis, and from these individual treatments attempt to synthesize a gradually broader solution. The impact of this work thus lies in the way it deals with non-factual information that is linguistically expressed, such as sentiment, attitude, humor and mood. These are inherent to our social activities, and are therefore extremely useful in the automatic mining of new knowledge.

Irony is a topic that has received little serious computational treatment in the past, though this is changing, perhaps because of the prevalence of irony in online texts and social media. On this subject, one of the first computational approaches to formalize irony was described by Utsumi (1996). However, this model is too abstract to represent irony beyond the confines of an idealized hearer-listener interaction. More recently, from the perspective of computational creativity, Veale and Hao (2009) have attempted to throw light on the cognitive processes that underlie verbal irony. By analyzing a large quantity of humorous similes of the form "as X as Y" from the web, they noted how web users often use figurative comparisons as a means to express ironic opinions. Likewise, Carvalho et al. (2009) have presented some clues for automatically identifying ironic sentences by first recognizing emoticons, onomatopoeic expressions, and special punctuation and quotation marks. Furthermore, Veale and Hao (2010) have recently presented a linguistic approach to separating irony from non-irony in figurative comparisons. They note that the presence of ironic markers like "about" can make rule-based categorization of ironic statements a practical reality, at least in the case of similes, and describe a system of linguistically-coded heuristics for performing this categorization. Finally, Reyes and Rosso (2011) have proposed a model that integrates different linguistic layers (from simple n-grams to affective content) to represent irony in customer reviews.

In this current work, we aim to analyze irony in terms of a multidimensional model of textual elements. We thus identify a set of discriminative features to automatically differentiate an ironic text from a non-ironic one. Since irony is common in texts that express subjective and deeply-felt opinions, its presence represents a significant obstacle to the accurate analysis of sentiment in these texts. A successful model of irony can thus play both a direct and an indirect role in tasks as diverse as sentiment analysis, opinion mining, electronic commerce, forum management, online marketing and product tracking.

The rest of the paper is organized as follows: in Sect. 2 we describe the theoretical challenges which underpin any treatment of irony, before then introducing the specific objectives of this current work. In Sect. 3 our new linguistic model is introduced. In Sect. 4 we evaluate the effectiveness of this model, before discussing our results and their implications. In Sect. 5 we then present one further experiment to assess the applicability of the model in the real world. Finally, in Sect. 6 we conclude with some final remarks and present some pointers to future work.

2 Irony in language

Like most creative phenomena, irony is difficult to pin down in formal terms, and no single definition ever seems entirely satisfactory. So to begin with, let us consider three obvious examples of verbal irony in everyday situations:

1. Going to your car in the morning, you notice that one of your tires is completely flat. A friendly neighbor chimes in with: "Looks like you've got a flat". Marveling at his powers of observation, you reply "Ya think?".
2. When having breakfast in a greasy-spoon cafe, you hungrily polish off everything on your plate. Seeing your totally clean plate, your waitress quips: "Well, that must have been terrible". "Yes", you reply, "absolutely awful".
3. A professor explains and re-explains Hegel's theory of the State to his class of undergraduates. "Is it clear now", he asks. "Clear as mud", a student replies.

These examples suggest that pretense plays a key role in irony: speakers craft utterances in spite of what has just happened, not because of it. The pretense in each case alludes to, or echoes, an expectation that has been violated (cf. Clark and Gerrig 1984; Sperber and Wilson 1992), such as the expectation that others behave in a civil fashion, speak meaningfully and with clarity, or not consume every single speck of food on their plate. This pretense may seem roundabout and illogical, but it offers a sharply effective and concise mode of communication. Irony allows a speaker to highlight the expectation that has been violated while simultaneously poking fun at, and often rebuking, the violator.

Sarcasm and irony are two frequently conflated modes of communication (cf. Colston 2007; Gibbs 2007), and several of the above responses are both ironic and sarcastic, e.g. "Ya think?" and "Clear as mud". Broadly speaking, irony tends to be a more sophisticated mode of communication than sarcasm: whereas the former often emphasizes a playful pretense (as in "Well, that must have been

terrible"), the latter is more often concerned with biting delivery and savage put-downs. While irony courts ambiguity and often exhibits great subtlety, sarcasm is delivered with a cutting or withering tone that is rarely ambiguous. Most stock ironies in language are thus instances of sarcasm, from "I could care less" to "great plan, Einstein" to "don't hold back" or "tell us what you really feel" (in response to an emotional outburst). Textual examples of sarcasm lack the sharp tone of an aggressive speaker, so for textual purposes, it is convenient to treat irony and sarcasm as different facets of the same phenomenon.

We might thus expect sarcasm to be easier to detect using superficial linguistic features, and some researchers are successfully focused directly on sarcasm rather than irony. For instance, Tsur et al. (2010) address the problem of finding linguistic elements that mark the use of sarcasm in online product reviews. Based on a semi-supervised approach, they suggest that specific surface features, such as words that convey information about a product, its maker, its name, and so on, as well as very frequent words, and punctuation marks, can be used to identify sarcastic elements in reviews.

Putting sarcasm to one side, textual uses of irony fall into two broad categories: *verbal* irony and verbal reports of *situational* irony. Verbal irony is a playful use of language in which a speaker implies the opposite of what is literally said (Curcó 2007); or expresses a sentiment in direct opposition to what is actually believed (i.e. a kind of indirect negation Giora 1995), as when Raymond Chandler in *Farewell, My Lovely* describes Moose Malloy as "about as inconspicuous as a tarantula on a slice of angel food". According to some pragmatic frameworks, certain authors are focused on fine-grained properties of this concept to correctly determine whether a text is ironic or not.[1] For instance, Grice (1975) requires that an utterance intentionally violate a conversational maxim if it is to be judged ironic. Wilson and Sperber (2007) assume that verbal irony must be understood as echoic, that is, they argue that irony deliberately blurs the distinction between use and mention. Utsumi (1996) suggests that an ironic environment, which establishes a negative emotional attitude, is a prerequisite for considering an utterance as ironic.

Situational irony, in contrast, is an unexpected or incongruous quality in a situation or event (cf. Lucariello 2007), such as a no-smoking sign in the foyer of a tobacco company, or a vegetarian having a heart-attack outside a McDonald's. Moreover, some authors distinguish other types of ironies, such as dramatic irony (Attardo 2007), discourse irony (Kumon-Nakamura et al. 2007), and tragic irony (Colston 2007). In this work we are focused only on verbal irony, but we do not reject the possibility that a linguistic model can be applied to situational irony, not least because much of the irony in online texts (and a great deal of the irony in tweets tagged as #irony) exhibits precisely this type of irony.

[1] Some fine-grained theoretical aspects of these concepts cannot be directly mapped to our framework, due to the idealized communicative scenarios that they presuppose. Nonetheless, we attempt to capture the core of these concepts in our model.

2.1 Irony in social media

Irony is a complex phenomenon that we encounter everyday in a variety of guises and with varying degrees of obviousness. As computational linguists it is verbal irony that chiefly interests us here. However, once one actually views the data itself (in this case, tweets by non-experts who use an intuitive and unspoken definition of irony rather than one sanctioned by a dictionary or a text-book), it becomes clear that casual speakers rarely recognize the pragmatic boundaries outlined above. For instance, the hashtag #irony is used by micro-bloggers in Twitter in order to self-annotate all varieties of irony, whether they are chiefly the results of deliberate word-play or merely observations of the humor inherent in everyday situations. The safest generalization that one can draw is that people perceive irony at the boundaries of conflicting frames of reference, in which an expectation of one frame has been inappropriately violated in a way that is appropriate in the other. Experts can tease apart the fine distinctions between one form of irony and another, in ways that casual speakers and micro-bloggers find it unnecessary to do. Since we wish to avail of the self-annotation #irony in this present work (see below Sect. 2.2), we align ourselves more with the intuitive view of irony (that an expectation has been violated in a way that is both appropriate and inappropriate) than with the strictly scholarly (and perhaps even scholastic) view. Once a broad sense of the ironic has been detected in a text, one can then apply other formal machinery to determine precisely which kind of irony is at work. We relegate this subsequent classification of an irony-laden text into distinct categories of irony to the realm of future work, and focus here on the broad foundations that would support these later efforts.

In this context, our objective then is to propose a model capable of representing the most salient attributes of verbal irony in a text, or at least what speakers believe to be irony, in order to be able to automatically detect it. This objective presupposes three specific tasks: (1) to collect objective data to obtain specific examples of irony and non-irony; (2) to extract a set of features that are suggestive of irony; (3) to evaluate the representativeness of these features and their ability to differentiate ironic texts from non-ironic ones.

2.2 Evaluation corpus

As already noted, the boundaries that differentiate verbal irony from situational irony, or even sarcasm, are very fuzzy indeed. In order to avoid confusion with respect to what constitutes an ironic example, we have opted to collect an evaluation corpus with statements that are a priori labeled as ironic by their users. To this end, we were focused on one of the current trendsetters in social media: the Twitter micro-blogging service. The membership criterion for including a tweet in our corpus is that each should contain a specific *hashtag* (i.e. a tag provided by users when posting their tweets in order to tie their contribution to a particular subject). The hashtags selected are #irony, in which a tweet explicitly declares its ironic nature, and #education, #humor, and #politics, to provide a large sample of potentially non-ironic tweets. These hashtags were selected for three simple reasons:

Table 1 Overall statistics in terms of tokens per set

	#irony	#education	#humor	#politics
Vocabulary	147,671	138,056	151,050	141,680
Nouns	54,738	52,024	53,308	57,550
Adjectives	9,964	7,750	10,206	6,773
Verbs	29,034	18,097	21,964	16,439
Adverbs	9,064	3,719	6,543	4,669

(1) when using the #irony hashtag, bloggers employ (or suggest) a family-resemblance model of what it means (cognitively and socially) for a text to be ironic; a text so-tagged may not actually be ironic by any dictionary definition of verbal irony, but the tag reflects a tacit belief about what constitutes irony; (2) by employing texts with specific hashtags, we avoid the need to manually (and subjectively) collect positive examples; (3) by applying the model to tweets, we broaden our analysis beyond literary uses of irony.

Based on these criteria, we collate an evaluation corpus of 40,000 tweets, which is divided into four parts, comprising one self-described positive set and three other sets that are not so tagged, and thus assumed to be negative. Each set contains 10,000 different tweets (though all tweets may not be textually unique). We assume therefore that our corpus contains 10,000 ironic tweets and 30,000 largely non-ironic tweets.

Duplicate tweets within a particular set were also automatically removed. However, given the presence of syntactic differences, as well as web links and other minor differences, the data sets contain a small number of duplicate tweets. In order to verify the impact of this issue on the experiments, a manual inspection in all the sets was performed. Results indicate that the percentage of duplicate or quasi-duplicate tweets is low (3 %): around 300 of 10,000 tweets. In addition, the case of duplicate tweets which belong to different sets was not considered at all. First, because by selecting different hashtags we aimed to eliminate, or at least, minimizing this issue: if the hashtag focuses the content on specific topics, then the tweets should not appear across the sets.[2]

Some statistics[3] are given in Table 1. It is worth mentioning that only the hashtags were removed. No further preprocessing was applied. The evaluation corpus is available by contacting the authors.

On the other hand, in order to estimate the overlap between the ironic set and each of the three non-ironic ones, the Monge Elkan distance was employed. This measure, according to Monge and Elkan (1996), allows for gaps of unmatched characters, [and thus], it should perform well for many abbreviations, and when fields have missing

[2] A manual comparison shows that a small number of tweets share two or more sets, being the sets #education and #politics the ones that present more of these cases: around 250 tweets (approximately 2 % of the total).

[3] Type-level statistics are not provided because these tweets contain many typos, abbreviations, user mentions, etc. There was no standardization processing to remove such misspelling. Therefore, any statistics regarding types would be biased.

Table 2 Monge Elkan distance among sets

	sim(s,t)
(#irony, #education)	0.596
(#irony, #humor)	0.731
(#irony, #politics)	0.627
(#education, #humor)	0.593
(#education, #politics)	0.605
(#humor, #politics)	0.648

information or minor syntactical differences. Accordingly, the Monge Elkan metric should help us minimize the likelihood of noise arising from the presence of typos, common misspellings, and the abbreviations that are endemic to short texts. Since we are considering tokens instead of types (see Footnote 3), the metric was computed using the approach outlined in Cohen et al. (2003). In such implementation, the authors considered a scheme in which the substrings are precisely tokens. Formula 1 describes the algorithm;[4] whereas results are shown in Table 2.

$$sim(s,t) = \frac{1}{k} = \sum_{i=1}^{K} \max_{j=1}^{L} sim'(A_i, B_j) \tag{1}$$

Similarity is here defined as a recursive matching given by comparing s and t. According to this formula, s and t are substrings $s = a_1, \ldots, a_K$ and $t = b, \ldots, b_L$, whereas sim' is the distance function (see Cohen et al. 2003).

The Monge Elkan distance approaches 1.0 as the data sets share more of their vocabulary. The results in Table 2 thus suggest a difference between the vocabularies of the four tweet sets. As one might expect, this difference is least pronounced between the irony and humor sets. After all, irony is most often used to communicate a humorous attitude or insight, as in the following two tweets from our corpus (both were tagged as #irony):

1. Just think: every time I breathe a man dies.—A friend: Have you tried to do something about bad breath?
2. I find it humorously hypocritical that Jeep advertises on TV about how we shouldn't watch tv in favor of driving their vehicles.

3 The irony model

We propose a model that is organized according to four types of conceptual feature: *signatures, unexpectedness, style*, and *emotional scenarios*. These features capture both low-level and high-level properties of textual irony based on conceptual

[4] Prior to computing the distance between texts, all words were stemmed using the Porter algorithm, and all stopwords were eliminated. Accordingly, the distance measure better reflects the similarity in core vocabularies rather than similarity in superficial forms.

descriptions found in the literature. Each feature is represented in terms of textual elements which appear to represent the core of the phenomenon, and in particular, those aspects that lead a micro-blogger to explicitly tag a tweet as ironic. Each feature, save for unexpectedness, is represented with three dimensions; unexpectedness is represented with just two dimensions.

These dimensions are listed and discussed below:[5]

1. *Signatures*: concerning pointedness, counter-factuality, and temporal compression;
2. *Unexpectedness*: concerning temporal imbalance and contextual imbalance;
3. *Style*: as captured by character-grams (c-grams), skip-grams (s-grams), and polarity skip-grams (ps-grams);
4. *Emotional scenarios*: concerning activation, imagery, and pleasantness.

3.1 Signatures

This feature is focused on exploring irony in terms of specific textual markers or signatures. This feature is largely characterized by typographical elements such as punctuation marks and emoticons, as well as by discursive elements that suggest opposition within a text. Formally, we consider signatures to be textual elements that throw focus onto certain aspects of a text. For instance, from a shallow perspective, quotes or capitals are often used to highlight a concept or an attribute (e.g.“ I HATE to admit it but, I LOVE admitting things”), while from a deeper perspective, adverbs often communicate contradiction in a text (e.g. “Saying we will destroy terrorism is *about* as meaningful as saying we shall annihilate mocking”).

This *signatures* feature is represented in three dimensions: *pointedness, counter-factuality*, and *temporal compression*. **Pointedness** is focused on explicit marks which, according to the most relevant properties of irony (cf. Sect. 2), should reflect a sharp distinction in the information that is transmitted. The set of elements considered here are punctuation marks (such as ., …, **;, ?, !, :,** ,), emoticons,[6] quotes, and capitalized words. **Counter-factuality** is focused instead on implicit marks, i.e. discursive terms that hint at opposition or contradiction in a text, such as *about, nevertheless, nonetheless* and *yet*. We use some adverbs, which hint at negation, as well as their synonyms in WordNet[7] (Miller 1995) to represent this dimension. The last dimension, **temporal compression**, is focused on identifying elements related to opposition in time; i.e. terms that indicate an abrupt change in a narrative. These elements are represented by a set of temporal adverbs such as *suddenly*, *now*, *abruptly*, and so on. The complete list of elements to capture both counter-factuality and temporal compression can be viewed at http://users.dsic.upv.es/grupos/nle.

[5] To aid understanding, “Appendix 1” provides examples from our evaluation corpus.

[6] The complete list with emoticons can be downloaded from http://users.dsic.upv.es/grupos/nle.

[7] Version 3.0 was used.

3.2 Unexpectedness

Irony often exploits incongruity, unexpectedness and the ridiculous to ensure that an insincere text is not taken literally by a listener. Lucariello (2007) suggests the term *unexpectedness* to represent these "imbalances in which opposition is a critical feature". She notes that surprise is a key component of irony, and even goes as far as claiming that unexpectedness underlies all ironic situations. As a consequence, we conceive the *unexpectedness* feature as a means to capture both temporal and contextual imbalances in an ironic text. Lucariello defines these imbalances in terms of oppositions or inconsistencies within contexts or situations, or between roles, or across time-frames (e.g. "The wimp who grows up to be a lion tamer", or "A kiss that signifies betrayal"; cf. Lucariello 2007). This feature is represented in two dimensions. The first, **temporal imbalance**, is used to reflect the degree of opposition in a text with respect to the information profiled in the present and past tenses. Unlike the temporal compression dimension, here we are focused on analyzing divergences related to verbs only (e.g. "I *hate* that when you *get* a girlfriend most of the girls that *didn't* want you all of a sudden *want* you!). **Contextual imbalance**, in contrast, is intended to capture inconsistencies within a context. In order to measure this dimension, we estimate the semantic similarity of a text's concepts to each other. The Resnik measure, implemented in WordNet::Similarity module (Pedersen et al. 2004), is used to calculate the pair-wise semantic similarity of all terms in a text. A normalized semantic relatedness score is then determined by summing the highest scores (across different senses of the words in the text) and dividing the result by the length of the text. The same general processing was performed to calculate the overall semantic relatedness per set (see Table 3). Instead of summing per single text, the semantic relatedness is summed for all the texts within the set, and the result is divided by 10,000. The driving intuition here is: the smaller the semantic inter-relatedness of a text, the greater its contextual imbalance (suggesting an ironic text); the greater the semantic inter-relatedness of a text, the lesser its contextual imbalance (suggesting a non-ironic text). Thus, we calculate the contextual imbalance of a text as the reciprocal of its semantic relatedness (that is, 1 divided by its semantic relatedness score). Overall statistics of semantic relatedness per set are given in Table 3.

3.3 Style

According to one dictionary definition, style is a "distinctive manner of expression". It is this kind of fingerprint, imparted by the stylistic characteristics of text, that allows people (and machines) to discriminate, for instance, Shakespeare's work from that of

Table 3 Semantic relatedness per set

Bold values are used to highlight relevant insights and results of the model

#irony	**0.892**
#education	1.006
#humor	1.270
#politics	1.106

Oscar Wilde. Within the current framework, the concept of style refers to recurring sequences of textual elements that express relatively stable features of how a text is appreciated, and which might thus allow us to recognize stylistic factors that are suggestive of irony. The style feature is captured in the current model using three kinds of textual sequences: **character n-grams (c-grams)**, **skip-grams (s-grams)**, and **polarity s-grams (ps-grams)**. The first, c-grams, captures frequent sequences of morphological information, such as affixes and suffixes (e.g. -ly). In order to obtain the best c-grams sequences, we consider sequences of 3–5 characters.[8]

In the second kind of sequences, s-grams, we widen the scope to consider whole words. But instead of looking for sequences of adjacent words (simple n-grams), we look for word sequences that contain (or skip over) arbitrary gaps; hence the name skip-grams (cf. Guthrie et al. 2006; Chin-Yew and Och 2004). For instance, in the sentence "There are far too many crazy people in my psychology class", a typical 2-gram is represented by the sequences *there are*, whereas a 2-sgram, with a 1 token gap, would be *there far*. Gaps are limited to 2 or 3 word skips, because longer sequences are not very common, especially if we take into account the length of the micro-blogging texts in the evaluation corpus (tweets must contain no more than 140 characters; i.e. $\sim$12 words).

The last sequence type, polarity s-grams, provides sequences of abstract categories based on s-grams; i.e. we can produce an abstract structure for a text from sequences of positive and negative terms instead of specific content words or characters. The intuition here is that one generally employs positive terms to communicate a negative meaning when using irony; for example, there is usually a positive ground in an ironic comparison that conveys a critical meaning (cf. Veale and Hao 2009). As in the case of s-grams, the gaps in ps-grams are limited to 2-word and 3-word skips only.

To provide tags for s-grams, as well as to observe the distribution of positive and negative terms in each text, we use a public resource commonly used in sentiment analysis and opinion mining tasks: the Macquarie Semantic Orientation Lexicon (MSOL) (Saif et al. 2009). As an example of this representation, consider the tweet "I need more than luck. I need Jesus and I'm an atheist…". Using to the MSOL, and considering only 2-word skips, the abstract representation provided by the terms labeled with positive or negative polarity is the following sequence of tags (after removing stop words): $pos_{need}\ pos_{jesus}\ neg_{atheist}$.

It is worth noting that all the *-grams sequences are obtained by generating a reference language map with all the tweets. First, all the sequences of *-grams with

[8] It is obvious that most sequences of c-grams are neutral with respect to irony. Moreover, they are neutral with respect to any topic. For instance, "ack or "acknowledgements are not representative of scientific discourses. However, irony and many figurative devices take advantage of rhetorical devices to accurately convey their meaning. We cite the research works described in Mihalcea and Strapparava (2006a, b) in which the authors focused on automatically recognizing humor by means of linguistic features. One of them is alliteration (which relies on phonological information). Therefore, in "*Infan*ts dont enjoy *infan*cy like *adult*s do *adult*ery" is clear the presence of such linguistic feature to produce the funny effect. Rhetorical devices like the one cited are quite common in figurative language to guarantee the transmission of a message. In this respect, we modified the authors' approach: instead of reproducing their phonological feature, we aimed to find underlying features based on morphological information in such a way we could find sequences of patterns beyond alliteration or rhyme.

Table 4 Example of *-grams sequences per set

	c-grams	s-grams	ps-grams
#irony	ack, ealt	liberti ironi, time enjoi	pos- neg, neg-neg
#education	aca, bit	monei homeschool, child obama	pos-pos, pos-neg
#humor	ber, desi	lol haircut, brilliant sens	neg-neg, neg-pos
#politics	tric, obam	stock monei, congression situat	neg-pos, neg-neg

their respective frequencies are obtained. The sequences with frequency $\leq$50 (in the case of c-grams) and 20 (in the case of s-grams and ps-grams) are removed in order not to bias the representation. Then, the language map is compared to every single tweet. All 40,000 tweets are then numerically represented with the number of sequences that it contains, either 0 or *n*. Such values are computed in order to obtain a global representativeness score per tweet, and later, per set. For instance, if a tweet contains 3 c-grams registered in the reference model ("ack" from acknowledgment, "ealt" from healt, "total" from totally), then such a tweet is assigned the value = 3. This value is normalized by the length of the tweet in terms of its tokens (e.g. 12). In Table 4 are shown some examples of the *-grams representation.

3.4 Emotional scenarios

Language, in all its forms, is one of our most natural and important means of conveying information about emotional states. Textual language provides specific tools on its own, such as the use of emoticons in web-based content to communicate information about moods, feelings, and our sentiments toward others. Online ironic expressions often use such markers to safely realize their communicative effects (e.g. "I feel so miserable without you, it is almost like having you here :P"). Emotional scenarios capture information that goes beyond grammar, and beyond the positive or negative polarity of individual words. Rather, this feature attempts to characterize irony in terms of elements which symbolize abstractions such as overall sentiments, attitudes, feelings and moods, in order to define a schema of favorable and unfavorable contexts for the expression of irony.

Adopting a psychological perspective, we represent emotional contexts in terms of the categories described by Whissell (2009), namely *activation*, *imagery*, and *pleasantness*. These categories (or *dimensions* in current our terminology) attempt to quantify the emotional content of words in terms of scores obtained from human raters. **Activation** refers to the degree of response, either passive or active, that humans exhibit in an emotional state (e.g. *burning* is more active than *basic*). **Imagery** quantifies how easy or difficult is to form a mental picture for a given word (e.g. it is more difficult to mentally depict *never* than *alcoholic*). **Pleasantness** quantifies the degree of pleasure suggested by a word (e.g. *love* is more pleasant than *money*). In order to represent these dimensions, we use Whissell's Dictionary of Affect in Language. This dictionary scores over 8,000 English words along the above three dimensions. The range of scores goes from 1.0 (most passive, or most difficult to form a mental picture, or most unpleasant) to 3 (most active, or easiest to

form a mental picture, or most pleasant). For example, Whissell's Dictionary notes that the word *flower* is passive (activation = 1.0), easily representable (imagery = 3.0), and generally produces a pleasant affect (pleasantness = 2.75); in contrast, *crazy* is more active (1.33), moderately representable (2.16), and quite unpleasant (1.6); whereas *lottery* is very active (3.0), moderately representable (2.14), and mostly pleasant (2.4).

4 Model evaluation

We evaluate the model in two ways: (1) by considering the appropriateness or representativeness of different patterns to irony detection; and (2) by considering the empirical performance of the model on a tweet classification task. Both considerations are evaluated in separate and independent experiments. When evaluating representativeness we look to whether individual features are linguistically correlated to the ways in which users employ words and visual elements when speaking in a mode they consider to be ironic. The classification task, in contrast, evaluates the capabilities of the model as a whole, focusing on the ability of the entire system of features to accurately discriminate ironic from non-ironic tweets.

4.1 Phase 1

In the first phase, each one of the 40,000 tweets is converted into a vector of term frequencies[9] according to a representativeness criterion. This criterion is intended to provide a global insight into the effectiveness of the model for actually identifying patterns in the ways that users employ the four conceptual features when genuinely speaking ironically. We need to know that the model is not simply detecting artifacts of the ways that users employ the #irony hashtag, or worse, artifacts of the way they use the #education, #humor, or #politics hashtags. By characterizing the tweets with this criterion, we obtain global insights about the distribution of features in all sets, allowing us to determine those which are more likely to express ironic meanings.

The representativeness of a given document d_k (e.g. a tweet) is computed separately for every dimension of each feature according to Formula 2:

$$\delta_{i,j}(d_k) = \frac{fdf_{i,j}}{|d|} \tag{2}$$

where i is the i-th feature ($i = 1, \ldots, 4$); j is the j-th dimension of i ($j = 1, \ldots, 2$ for the unexpectedness feature, and $1, \ldots, 3$ otherwise); *fdf* (feature dimension frequency) is the frequency of the dimension j of the feature i; and $|d|$ is the length (in terms of tokens) of the k-th document d_k. To aid understanding our guiding principle, let us take the signatures feature (i). We can compute the representativeness of its three dimensions j (pointedness, counter-factuality, and

[9] All tweets underwent preprocessing, in which terms were stemmed and both hashtags and stop words were removed.

temporal compression) by applying Formula 2 in the following tweet: "HAHA-HAHA!!! now thats the definition of !!! lol…tell him to kick rocks!"

- concerning pointedness, $\delta = 0.85$ (HAHAHAHA, !!!, !!!, lol, …, !) / (hahahaha, now, definit, lol, tell, kick, rock);
- concerning counter-factuality, $\delta = 0$;
- concerning temporal-compression, $\delta = 0.14$ (now) / (hahahaha, now, definit, lol, tell, kick, rock).

This process is applied to all dimensions for all four features.

Once $\delta_{i,j}$ is obtained for every single tweet d_k, a representativeness threshold is established in order to filter the documents that are more likely to have ironic content.[10] In this respect, if $\delta_{i,j}(d_k)$ is ≥ 0.5, then document d_k is assigned a representativeness value of 1 (i.e. dimension j of feature i is representative of d_k); otherwise, a representativeness value of 0 (not representative at all) is assigned. For instance, considering the previous example, only one dimension of the signatures feature exceeds such threshold: counter-factuality ($\delta = 0.85$), thus it is considered as representative; whereas pointedness and temporal-compression do not ($\delta = 0$ and 0.14, respectively), thus they are not considered as representative (at least in this specific tweet). In order not to bias the representativeness threshold, the normalization of d_k is done in terms of tokens rather than in terms of the number of features, due to the former is a constant variable (recall that all tweets must contain no more than 140 characters; i.e. around 12 words); instead, the number of features is an unpredictable and inconstant variable (there are tweets that do not contain any feature).

Lastly, in order to observe the representativeness of the model in terms of sets, an overall representativeness score for each dimension j is calculated by summing $\delta_{i,j}(d_k)$ for all documents in the set they belong to, and normalizing by the size of the set (10,000). Results are shown in Table 5.

As shown by the results in Table 5, all dimensions except pointedness and temporal imbalance appear to be sufficiently indicative to represent ironic tweets from educational, humorous and political tweets. On a set level then, there appear to be patterns of feature use in a text that correlate with the ways in which people use irony. Consider, for instance, the counter-factuality dimension, whose textual elements are terms that suggest contradiction. It is evident from Table 5 that terms which suggest counter-factuality appear most often in our ironic tweets. In contrast, ironic tweets do not score well overall on semantic relatedness, which means they score well on the contextual imbalance dimension. This, in turn, supports our hypothesis about the reduced inter-word semantic relatedness of ironic tweets. With respect to the dimensions of the *style* and *emotional scenarios* features, the scores achieved for each indicate a greater presence of textual elements related to these dimensions in the ironic set, especially as regards the scores for the *s-grams* and *pleasantness* dimensions.

[10] In order to observe the density function of each dimension for all four features, in "Appendix 2" we present the probability density function (PDF) associated with $\delta_{i,j}(d_k)$ prior applying the threshold.

Table 5 Overall feature representativeness per set

	Irony	Education	Humor	Politics
Signatures				
Pointedness	0.314	0.268	**0.506**	0.354
Counter-factuality	**0.553**	0.262	0.259	0.283
Temporal compression	**0.086**	0.054	0.045	0.046
Unexpectedness				
Temporal imbalance	0.769	0.661	**0.777**	0.668
Contextual imbalance	**1.121**	0.994	0.788	0.904
Style				
c-grams	**0.506**	0.290	0.262	0.395
s-grams	**0.554**	0.195	0.144	0.161
ps-grams	**0.754**	0.481	0.494	0.534
Emotional scenarios				
Activation	**1.786**	1.424	1.482	1.324
Imagery	**1.615**	1.315	1.378	1.218
Pleasantness	**1.979**	1.564	1.660	1.394

Bold values are used to highlight relevant insights and results of the model

Finally, the graphs depicted in Fig. 1 show the distribution of positive and negative words in terms of their position in the tweet (*X* axis) and their overall representativeness ratio (*Y* axis). It is interesting to note how the preponderance of negative terms in the ironic set is concentrated in the first 7 words of the texts, whereas the frequency of positive terms is lower but relatively constant across texts. In the education and politics sets, by contrast, the distribution appears to be just the contrary: more positive terms are found in the first 6 words of a text, while negative terms appear with relative constancy and a lower frequency throughout a text. In the humor set, the negative terms tend to appear with higher frequency between word positions 3 and 8, while positive terms tend to occur between word positions 1 and 4. This behavior hints at that part of the utterance in which irony produces its effect, and on which the greatest energy should be placed.

4.2 Phase 2

In the second phase, we use two different classifiers to evaluate the ability of the model to automatically discriminate each text set. We perform a series of binary classifications, between irony versus education; between irony versus humor; and between irony versus politics. In each case, features are added incrementally to the classification processing, to determine their relative value to the classifier. Thus, classification first uses the *signatures* feature; the *unexpectedness* feature is then added; and so on.[11] Two distributional scenarios are evaluated: (1) a balanced distribution, comprising 50 % positive texts and 50 % negative texts; (2) an

[11] It is worth mentioning that we use all 11 dimensions of the four conceptual features by adding them in batches.

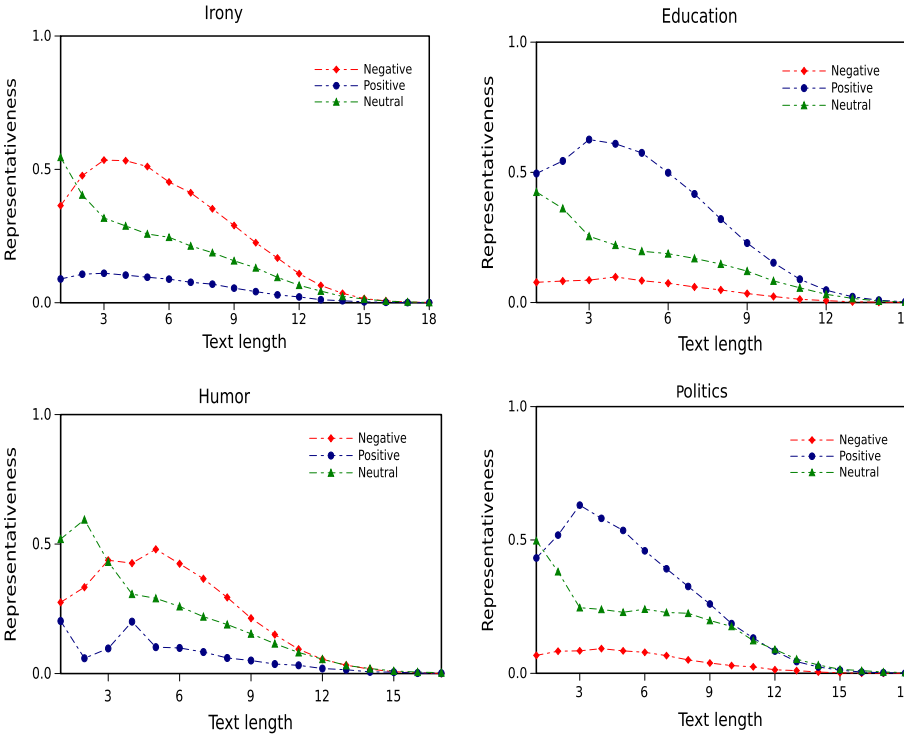

Fig. 1 Distribution of positive, negative and out of vocabulary (neutral) terms per set

imbalanced distribution, with a more realistic mix of 30 % ironic texts and 70 % non-ironic texts. Two popular algorithms from the literature are used to perform classification: the Naïve Bayes approach, and decision trees.[12] We choose these particular algorithms for two reasons: first, we use the Naïve Bayes algorithm since our experiments are focused on the presence or absence of features as represented by boolean attributes (Witten and Frank 2005) that are treated as independent variables, assigned to the class with maximum probability; and second, decision trees are used in order to analyze the sequences of decisions regarding the relevance of such features, and to be able to make further inferences about them.

The classifiers, both for balanced and imbalanced distributions, were tested using tenfold cross validation. The results shown in Fig. 2 indicate an acceptable performance on the automatic classification. The model evidently improves its performance in almost all cases (with the exception of the *emotional scenarios* feature) each time a new feature is added (e.g. the accuracy increases after considering at least two or three features). Focusing on the accuracy, a trivial classifier that labels all texts as non-ironic would achieve an accuracy of 50 %, our entire model, instead, achieves an accuracy higher than the baseline (over 75 %), suggesting that the four conceptual features cohere as a single framework that is

[12] Each algorithm is implemented in Weka toolkit (Witten and Frank 2005). No optimization was performed.

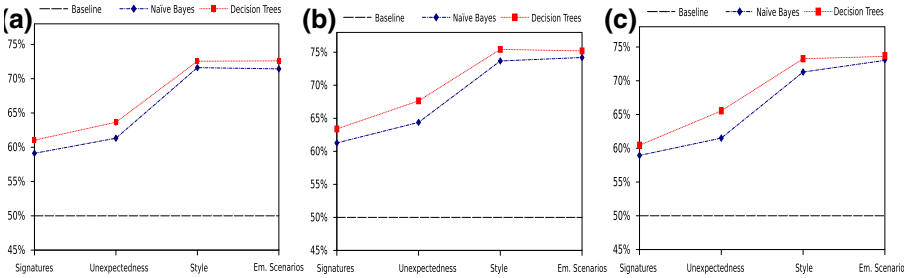

Fig. 2 Classification accuracy regarding irony versus education (**a**), humor (**b**), and politics (**c**), considering a balanced distribution

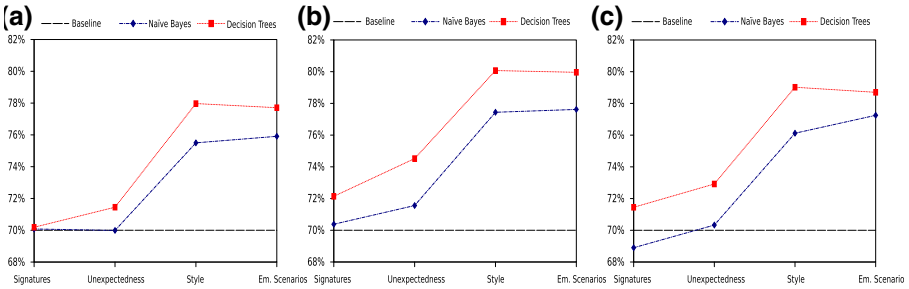

Fig. 3 Classification accuracy regarding irony versus education (**a**), humor (**b**), and politics (**c**), considering an imbalanced distribution

able to clearly discriminate positive (ironic) tweets from negative (non-ironic) tweets. Similar results are reported by Carvalho et al. (2009). By exploring oral and gestural features to detect irony in user comments, authors achieve accuracies ranging from 44.88 to 85.40 %.

With respect to Fig. 3, the results are not as good as in the balanced distribution. A classifier which labels all texts as non-ironic would achieve an accuracy of 70 %, whereas in this figure we see that our model hardly exceeds this baseline when considering just a couple of features (from 68 to 74 %). Even when the entire model is considered the accuracy barely reaches 10 % points beyond than the baseline. This evidences the difficulty of identifying irony in data sets where the positive examples are very scarce; i.e. it is easier to be right with the set that statistically appears quite often than with the set that barely appears. This situation, nonetheless, is the expected when facing tasks in which the absence of positive data, or the lack of labeled examples, is the main practical difficulty. However, this first approach has shown some advances when dealing with distributional issues. Our efforts thus must be addressed to find more discriminating features which allow us to increase current accuracy on both balanced and imbalanced scenarios.

Table 6, on the other hand, presents the results obtained in terms of precision, recall and F-Measure on both the balanced and imbalanced distributions. These results support our intuitions about irony. While the results reported by

Table 6 Precision, Recall and F-Measure regarding i) balanced distribution, and ii) imbalanced distribution

	Precision		Recall		F-Measure	
	i (%)	ii (%)	i (%)	ii (%)	i	ii
Naïve Bayes						
Education	73	60	66	62	0.69	0.61
Humor	**79**	64	68	59	**0.73**	0.62
Politics	75	60	**69**	60	0.72	0.60
Decision trees						
Education	76	70	66	52	0.70	0.60
Humor	**78**	75	**74**	47	**0.76**	0.58
Politics	75	69	71	52	0.73	0.59

Bold values are used to highlight relevant insights and results of the model

Davidod et al. (2010) and Burfoot and Baldwin (2009) relate to analyses of different figurative devices, such as sarcasm and satire respectively, and are thus not entirely comparable to the current results, such a comparison is nonetheless warranted. Our model obtains F-Measures that are comparable to, or better than, either of these previous approaches. For instance, the former study (Davidov et al. 2010) reports a highest F-Measure of 0.545 on a corpus collected from Twitter, while the latter study (Burfoot and Baldwin 2009) reports a highest F-Measure of 0.798 for a corpus of newswire articles. In the current study, the highest F-Measure obtained is a score of 0.768 in the balanced distribution.

To further assess the capabilities of the model, an additional variation of the classification task was undertaken, It is based on considering positive set (irony) against all three negative ones (education, humor, politics). Classification was performed using decision trees, and evaluated using tenfold cross validation. We considered both a balanced distribution (10,000 positive instances and 3,333 of each negative set) and an imbalanced distribution (10,000 positive instances and all 30,000 negative instances). Results show a similar pattern to those previously observed. When using a balanced distribution, the accuracy is lower but precision, recall and F-measure are all clearly higher (72.30 %, 0.736, 0.695, 0.715, respectively). Conversely, when using an imbalanced distribution, the accuracy is higher but precision, recall and F-measure suffer (80.44 %, 0.661, 0.447, 0.533, respectively). These results support our belief that a system of textual features can capture the linguistic patterns used by people when communicating what they believe to be ironic statements.

The model operates with a system of features, yet each feature can be analyzed in terms of information gain to determine its individual contribution to the discrimination power of the system. Figure 4 presents the results of an information gain filter (Y axis) on each of the dimensions of our four features (X axis).[13]

Information gain results show that there are dimensions that appear to be useless in the discrimination of ironic tweets (e.g. the *temporal compression* dimension of

[13] Only the information gain values for the balanced distribution are displayed. The imbalanced case is not considered here since the values follow a similar distribution.

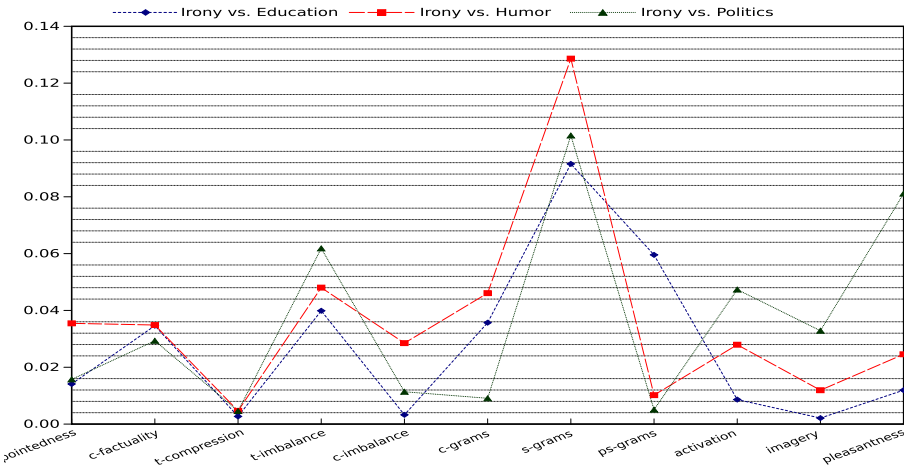

Fig. 4 The relevance of every single dimension according to its information gain value

signatures, the *contextual imbalance* dimension of unexpectedness, the *ps-grams* dimension of style, and the *imagery* dimension of emotional scenarios). However, the apparent uselessness is a function of the kinds of texts that are to be discriminated. Consider, for instance, the *ps-grams* dimension: while it exhibits a very low information gain when discriminating irony versus humor and irony versus politics, this score clearly increases when discriminating irony versus education. A similar situation holds with respect to the *contextual imbalance* dimension: when considering the discrimination of irony versus humor, the score is acceptable, whereas on the remaining two negative sets, the score is unacceptably low. Likewise, there are dimensions that exhibit a strong relevance to the irony task (e.g. the *temporal imbalance* dimension of unexpectedness, the *s-grams* dimension of style, and the *pleasantness* dimension of emotional scenarios). Once again, this relevance is also a function of the kinds of texts that are to be discriminated. This behavior suggests that these features cohere well together, so that while no single feature captures the essence of irony, all four together provide a useful linguistic framework for detecting irony at a textual level.

To conclude, this analysis suggests that the proposed model is capable of representing the most obvious aspects of the way verbal irony is exploited by users of Twitter. Nonetheless, the effectiveness of the model is largely dependent on the kind of text genre that is analyzed. Accordingly, the model can be viewed as a *local optimum* model instead of a *global optimum* model, insider it is provides a good solution for some, but not all, text genres.

5 A case study: Toyota

Irony is a challenge, not only from a computational perspective, but from a communicative one as well. The linguistic and social factors which impact on the perception of irony make the task of automatically identifying ironic texts quite

complex. We have proposed and evaluated a model which efficiently captures a number of the most salient attributes of irony. Yet one can ask whether this model yields actual benefits in real-world applications. Large companies have the most to gain from the appreciation of irony in social media, since these media are increasingly being used to comment on products and services and thereby encourage or discourage new customers. If a company can look beyond the distortional effect of irony, it can more accurately gather valuable marketing knowledge from the opinions of its users.

In this section we explore the utility of such an understanding of irony in the case of a specific enterprise and its marketing problem: Toyota has of late encountered a variety of hardware problems to do with braking and acceleration, real or merely perceived, that have seriously affected its reputation for quality and safety.[14] With respect to this topic, we have collected a corpus of 500 tweets from Twitter via the following attributes:

1. the **#Toyota** tag;
2. the positive emoticon **:)** and the negative emoticon **:(**.

All 500 tweets must contain the #toyota hashtag. To provide further focus, and to help us verify some assumptions regarding the contexts in which irony appears, these 500 tweets should also contain either a positive or negative emoticon. Our test set thus contains 250 tweets with a positive emoticon and 250 labeled with a negative emoticon.

This experiment allows us to test the applicability of the model to tweets that are not explicitly tagged as ironic by their senders. In this respect, this experiment is addressed as an information retrieval task in which the database is integrated with 500 tweets. The query thus is focused on retrieving from such database all the tweets with ironic content. To this end, we first obtain human judgments with respect to the presence or absence of ironic contents in the #toyota set. Such tweets are deemed as the total number of relevant documents to be retrieved (i.e. our benchmark). Annotation is performed by 80 annotators,[15] who manually tagged the 500 tweets. They were asked to assign a value of 1 if they considered a tweet to be ironic, and a value of 0 if they considered it to be non-ironic. No theoretical background was requested or offered, and no dictionary definition of irony was provided. Instead, each annotator was asked to rely on their own intuitions about what constitutes irony in a short text (we expect these intuitions to largely agree with the intuitions that lead a sender to mark a tweet with the hashtag #irony). Every annotator tagged 25 different tweets, and every tweet was tagged by 4 different annotators. In order to estimate the degree of agreement between the four annotators of each tweet, the Krippendorff α coefficient was calculated in each case. According to Artstein and Poesio (2008), this coefficient calculates the expected agreement by looking at the overall distribution of judgments without regard to which annotators produce which judgments. Furthermore, the Krippendorff α is tailored for multiple annotators, allowing for different magnitudes of disagreement. Both of these qualities are particularly suited to this task

[14] This problem affected Toyota during the last months of 2009 and the beginning of 2010.

[15] Only 55 annotators were native speakers of English, while the remaining 25 were post-graduate students with sufficient English skills.

Table 7 Statistics regarding annotators judgments

	Tweets
Total tweets	500
Ironic tweets	147
Non ironic tweets	353
Ironic tweets[a]	
4 annotators agree	28
3 annotators agree	39
2 annotators agree	80

[a] Considering only the 147 tweets annotated as ironic

given the number of annotators involved, and more importantly, because irony relies on subtle nuances which are not always recognized by individual listeners/annotators. Table 7 presents overall statistics for the manual tagging of tweets, for which a Krippendorff α coefficient of 0.264 was noted. This value, according to the criteria exposed in Artstein and Poesio (2008), indicates a fair reliability with respect to the generalization of these annotations. Nonetheless, those authors also indicated that the purpose of reliability studies is not to find out whether annotations can be generalized, but whether they capture some kind of observable reality. According to this point of view, one of the main problems of the task is that irony remains a somewhat subjective concept, so that human annotators tend to disagree substantially. This, of course, is precisely the reason some tweeters feel the need to annotate their messages with an explicit indication of the presence of #irony.

We assume a tweet is ironic when at least two of its four human annotators classify it as such. Following this criterion, 147 of the 500 #toyota tweets are ironic. Of these 147 tweets, 84 belonged to the tweets labeled with the positive emoticon **#:)**, whereas 63 belonged to the ones labeled with the negative emoticon **#:(**. This difference, although not statistically important, supports the general assumption that irony more often relies on a positive ground to produce its critical effect.[16] Moreover, only in 28 tweets was there complete agreement among four annotators with respect to their assigned tags, while in only 39 tweets was agreement observed between three of the four annotators. In 80 tweets there was agreement between just two annotators.[17] Now, taking into consideration both Krippendorff α coefficient and the amount of ironic tweets, we will realize that the difficulty of recognizing irony, which somewhat perversely, is often greater than the difficulty of understanding irony. Quite simply, one does not always need to understand the concept of irony to understand the use of irony. Moreover, because irony requires a knowledge of cultural and social stereotypes and other pragmatic factors, the perception of irony tends to be subjective and personal.

[16] Recall that the #toyota set is artificially balanced, and contains 250 tweets with a positive emoticon and 250 tweets with a negative emoticon, regardless of the overall frequency of these emoticons on Twitter. Each emoticon serves a different purpose in an ironic tweet. Irony is mostly used to criticize, and we expect the negative emoticon will serve to highlight the criticism, while the positive emoticon will serve to highlight the humor of the tweet.

[17] It is important to mention that 141 tweets were tagged as ironic by just single annotators. However, these tweets were not considered in order to not bias the test. It is senseless to take a tweet as ironic when only one annotator tagged it as ironic, if 3 annotators said it was non-ironic.

Table 8 Irony retrieval results

Level	Tweets retrieved	Precision (%)	Recall (%)	F-Measure
A	59	56	40	0.47
B	93	**57**	63	0.60
C	123	54	**84**	**0.66**

Bold values are used to highlight relevant insights and results of the model

Once the ironic tweets (relevant documents) are obtained, our model is applied to all 500 tweets in order to evaluate its performance to retrieve the documents with ironic content (147 tweets according to the human annotation). First, we determine three separate levels of representativeness (A, B, C) in order to cluster the texts into different groups for subsequent analysis. Each level is established by modifying the cutoff threshold in Formula 2 according to the following schema:

- Level A. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.8}$; otherwise = 0.
- Level B. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.6}$; otherwise = 0.
- Level C. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.5}$; otherwise = 0.

Then, for each level, we count how many retrieved documents matched with the relevant documents. Table 8 presents the results in terms of precision, recall and F-Measure. Taking the 147 tweets previously described as the total number of relevant documents to be retrieved, the results concerning precision are really low (they hardly exceed the 50 % for each level); however, the results concerning recall are more satisfactory (from 40 to 84 %). In this respect, such results seem to be very dependent on the level of representativeness. For instance, at the most discriminating level (A), the recall achieved is 40 %, and the number of tweets retrieved is 59, of which 9 are tweets on which all four human annotators are in agreement, 16 are tweets on which three of the annotators agree, and 34 are tweets on which just two of the annotators agree. At the middle discriminating level (B), the number of tweets retrieved increased to 93 (recall = 63 %), of which 14, 26, and 53 agree with the judgments of four, three, and two annotators, respectively. At the lowest discriminating level (C), the number of relevant documents retrieved with ironic content increased to 123 (recall = 84 %), of which 22, 32, and 69 agree with the judgments of respective annotators.

In terms of precision, it is evident the need of improving the model. Due to this experiment was thought as an information retrieval task, it is barely helpful to any user a system that only retrieves 50 % of the relevant documents. However, if considering the results concerning recall, the model shows some applicability to real-world problems. Though the performance of the model is not ideal when the representativeness level is close to 1, it seems clear that some of its features can capture recurrent linguistic patterns that characterize the use of irony in social media. The creation of indexes for obtaining the most ironic topics can be viewed as a trend discovery task, while characterization of information posted by bloggers can be seen as an application of influence modeling. Each perspective in turn requires the ability to extract fine-grained knowledge for decision making. Thus, if our results hold true or improve as new and greater amounts of data are tested, then the implications of processing irony in real applications will be significant.

6 Conclusions and future work

Irony, satire, parody and sarcasm are overlapping figurative phenomena, whose differences are a matter of usage, tone, and obviousness. For instance, sarcasm has an obviously mocking tone that is used against another, while irony is often more sophisticated, more subtle and ambiguous, and even self-deprecating. Our objective in this paper has not been to distinguish between irony from other figurative devices, but to recognize affective statements that have non-literal meanings, where these meanings are the opposite of what a shallow interpretation might normally conclude. This paper has presented an approach to the detection of verbal irony in short online texts, focusing on texts in social media that are produced by what we have dubbed here *the twittering classes*. Though often repetitive and inane, these kinds of social texts are receiving increased attention as a carrier of influential customer opinions and feedback.

Our model goes beyond surface elements to extract four different kinds of features from a text: signatures, degrees of unexpectedness, stylistic features, and emotional scenarios. These features work better when they are used as part of a coherent framework rather than used individually. No single feature captures the essence of irony, but all four kinds together provide a valuable linguistic inventory for detecting irony on a textual level.

We have evaluated the model in an online domain where texts are short and laden with social meaning. A corpus of 40,000 tweets, which was automatically harvested from Twitter, allowed us to evaluate the model on two key fronts: how representative are the features that we use, and how well can they discriminate ironic from non-ironic texts? Our model is promising, and though it clearly leaves room for improvement, it achieves encouraging results in terms of representativeness, classification accuracy, precision, recall and F-Measure.

Our final experiment considered the practical applicability of the model. The comparison of human judgments with automatic classifications yields intriguing insights into how humans think about irony. Certainly, anyone who examines how the #irony hashtag is used in Twitter will know that humans do not have a single, precise notion of irony; rather, we seem to possess a diffuse, fuzzy, family-resemblance model of what it means for a text to be ironic. This suggests that as part of our future work on this approach, we should not just be focused on the quality and value of different linguistic features, though this of course will be a topic of some importance. We shall also have to tackle the problem of how people think about irony, and recognize irony in their own texts and in those of others. This will require that we tease apart the categories of verbal irony and situational irony. Logically these are distinct categories; in real texts however, where people mix ironic remarks with observations about ironic situations, the two are very much intertwined.

Acknowledgments This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+I. The National Council for Science and Technology (CONACyT - Mexico) has funded the research work of Antonio Reyes.

Appendix 1: Examples of the model representation

In this appendix are given some examples regarding how the model is applied over the tweets.

1. *Pointedness*
   - The govt should investigate him thoroughly; do I smell **IRONY**
   - Irony is such a funny thing **:)**
   - Wow the only network working for me today is 3G on my iPhone. **WHAT DID I EVER DO TO YOU INTERNET???????**
2. *Counter-factuality*
   - My latest blog post is **about** how twitter is for listening. And I love the irony of telling you about it via Twitter.
   - Certainly I always feel compelled, obsessively, to write. **Nonetheless** I often manage to put a heap of crap between me and starting...
   - BHO talking in Copenhagen **about** global warming and DC is **about** to get 2ft. of snow dumped on it. You just gotta love it.
3. *Temporal compression*
   - @*ryan*$_c$*onnolly* oh the irony that will occur when they finally end movie piracy and **suddenly** movie and dvd sales begin to decline sharply.
   - I'm seriously really funny when nobody is around. You should see me. But **then** you'd be there, and I wouldn't be funny...
   - RT @*Butler*$_G$*eorge*: **Suddenly**, thousands of people across Ireland recall that they were abused as children by priests.
4. *Temporal imbalance*
   - **Stop** trying to find love, it will find you;...and no, he **didn't** say that to me..
   - Woman on bus **asked** a guy to turn it down please; but his music **is** so loud, he **didn't hear** her. Now she **has** her finger in her ear. The irony
5. *Contextual imbalance*
   - DC's snows coinciding with a conference on global warming proves that God has a sense of humor.
     Relatedness score of **0.3233**
   - I know sooooo many Haitian-Canadians but they all live in Miami.
     Relatedness score of **0**
   - I nearly fall asleep when anyone starts talking about Aderall. Bullshit.
     Relatedness score of **0.2792**
6. *Character n-grams (c-grams)*
   - **WIF**
     More about Tiger—Now I hear his **wif**e saved his life w/ a golf club?
   - **TRAI**
     SeaWorld (Orlando) **trai**ner killed by killer whale. or reality? oh, I'm sorry politically correct Orca whale
   - **NDERS**
     Because common sense isn't so common it's important to engage with your market to really u**nders**tand it.

7. *Skip-grams (s-grams)*
   - 1-skip: *richest . . .mexican*
     Our president is black nd the **richest** man is a **Mexican** hahahaha lol
   - 1-skip: *unemployment . . .state*
     When **unemployment** is high in your **state**, Open a casino tcot tlot lol
   - 2-skips: *love . . .love*
     Why is it the Stockholm syndrome if a hostage falls in **love** with her kidnapper? I'd simply call this **love**. ;)
8. *Polarity s-grams (ps-grams)*
   - 1-skip: *pos-neg*
     Reading **glasses**$_{pos}$ have **RUINED**$_{neg}$ my eyes. B4, I could see some shit but I'd get a headache. Now, I can't see shit but my head feels fine
   - 1-skip: *neg-neg-pos*
     Breaking$_{neg}$ **News**$_{neg}$: New **charity**$_{pos}$ offers people to adopt a banker and get photos of his new bigger house and his wife and beaming mistress.
   - 2kips: *pos-pos-neg*
     Just heard the **brave**$_{pos}$ **hearted**$_{pos}$ English Defence **League**$_{neg}$ thugs will protest for our freedoms in Edinburgh next month. Mad, Mad, Mad
9. *Activation*
   - I enjoy(2.22) the fact(2.00) that I just addressed(1.63) the dogs(1.71) about their illiteracy(0) via(1.80) Twitter(0). Another victory(2.60) for me.
   - My favorite(1.83) part(1.44) of the optometrist(0) is the irony(1.63) of the fact(2.00) that I can't see(2.00) afterwards(1.36). That and the cool(1.72) sunglasses(1.37).
   - My male(1.55) ego(2.00) so eager(2.25) to let(1.70) it be stated(2.00) that I'am THE MAN(1.8750) but won't allow(1.00) my pride(1.90) to admit(1.66) that being egotistical(0) is a weakness(1.75). . .
10. *Imagery*
    - Yesterday(1.6) was the official(1.4) first(1.6) day(2.6) of spring(2.8). . . and there was over a foot(2.8) of snow(3.0) on the ground(2.4).
    - I think(1.4) I have(1.2) to do(1.2) the very(1.0) thing(1.8) that I work(1.8) most on changing(1.2) in order(2.0) to make(1.2) a real(1.4) difference(1.2) paradigms(0) hifts(0) zeitgeist(0)
    - Random(1.4) drug(2.6) test(3.0) today(2.0) in elkhart(0) before 4(0). Would be better(2.4) if I could drive(2.1). I will have(1.2) to drink(2.6) away(2.2) the bullshit(0) this weekend(1.2). Irony(1.2).
11. *Pleasantness*
    - Goodmorning(0), beauties(2.83)! 6(0) hours(1.6667) of sleep(2.7143)? Total(1.7500) score(2.0000)! I love(3.0000) you school(1.77), so so much(2.00).
    - The guy(1.9000) who(1.8889) called(2.0000) me Ricky(0) Martin(0) has(1.7778) a blind(1.0000) lunch(2.1667) date(2.33).
    - I hope(3.0000) whoever(0) organized(1.8750) this monstrosity(0) realizes(2.50) that they're playing(2.55) the opening(1.88) music(2.57) for WWE's(0) Monday(2.00) Night(2.28) Raw(1.00) at the Olympics(0).

Appendix 2: Probability density function

In this appendix are shown 11 graphs in which we depict the probability density function associated with $\delta_{i,j}(d_k)$ for all dimensions according to Formula 2. All these graphs are intended to provide descriptive information concerning the fact that the model is not capturing idiosyncratic features of the negative sets; rather, it is really capturing some aspects of irony. For all the graphs we keep the following representation: #irony (blue line), #education (black line), #humor (green line), #politics (brown line) (Figs. 5, 6, 7, 8).

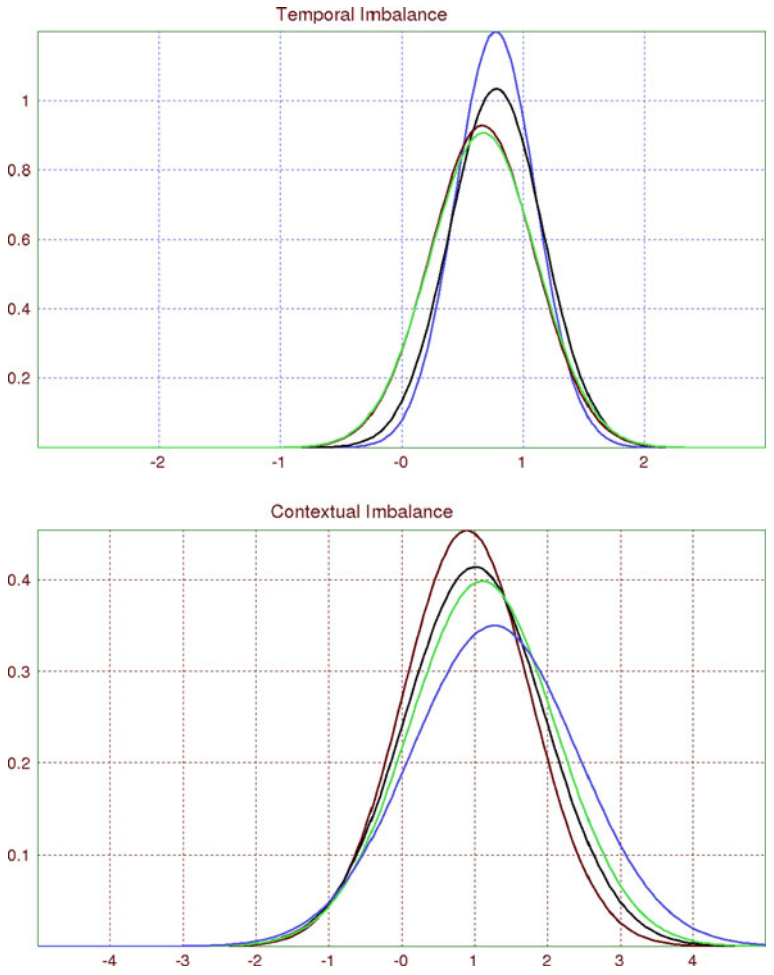

Fig. 5 PDF for dimensions in the unexpectedness feature. (Color figure online)

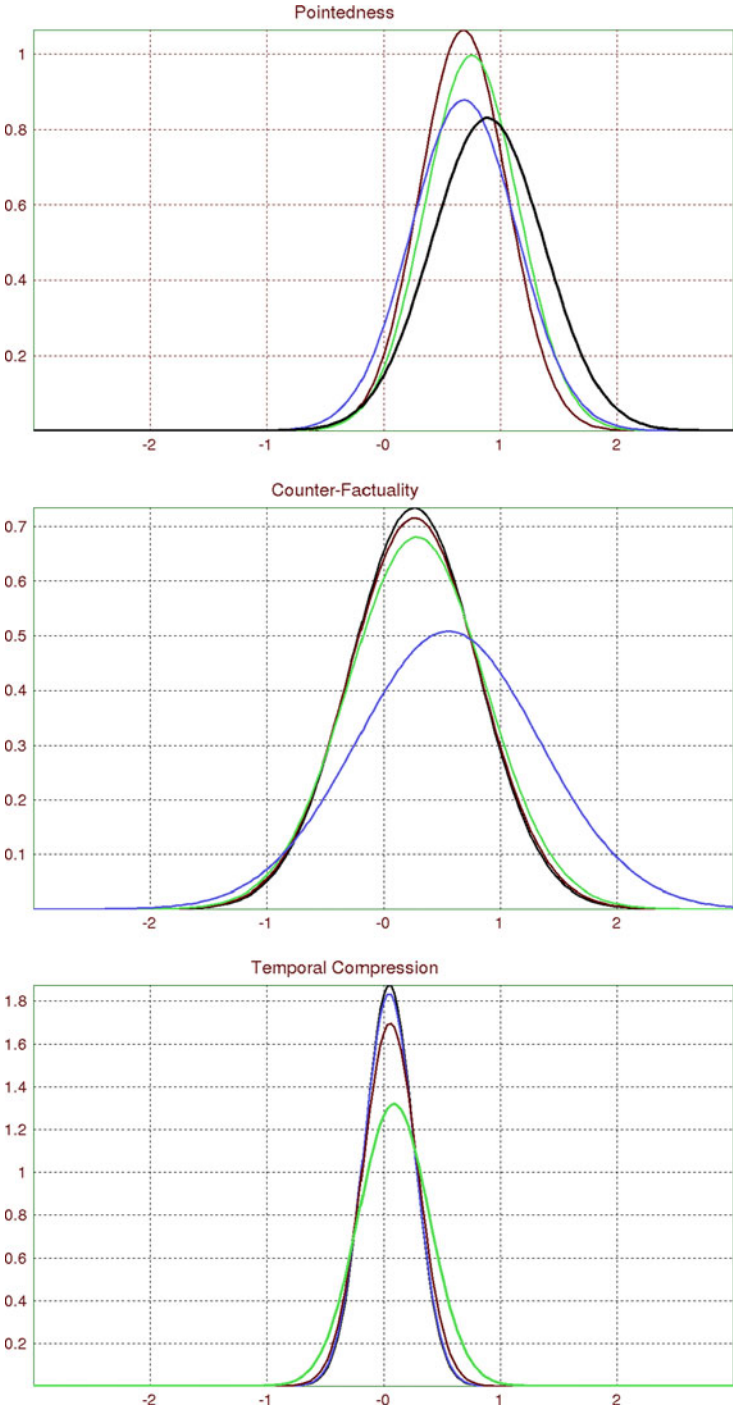

Fig. 6 PDF for dimensions in the signatures feature. (Color figure online)

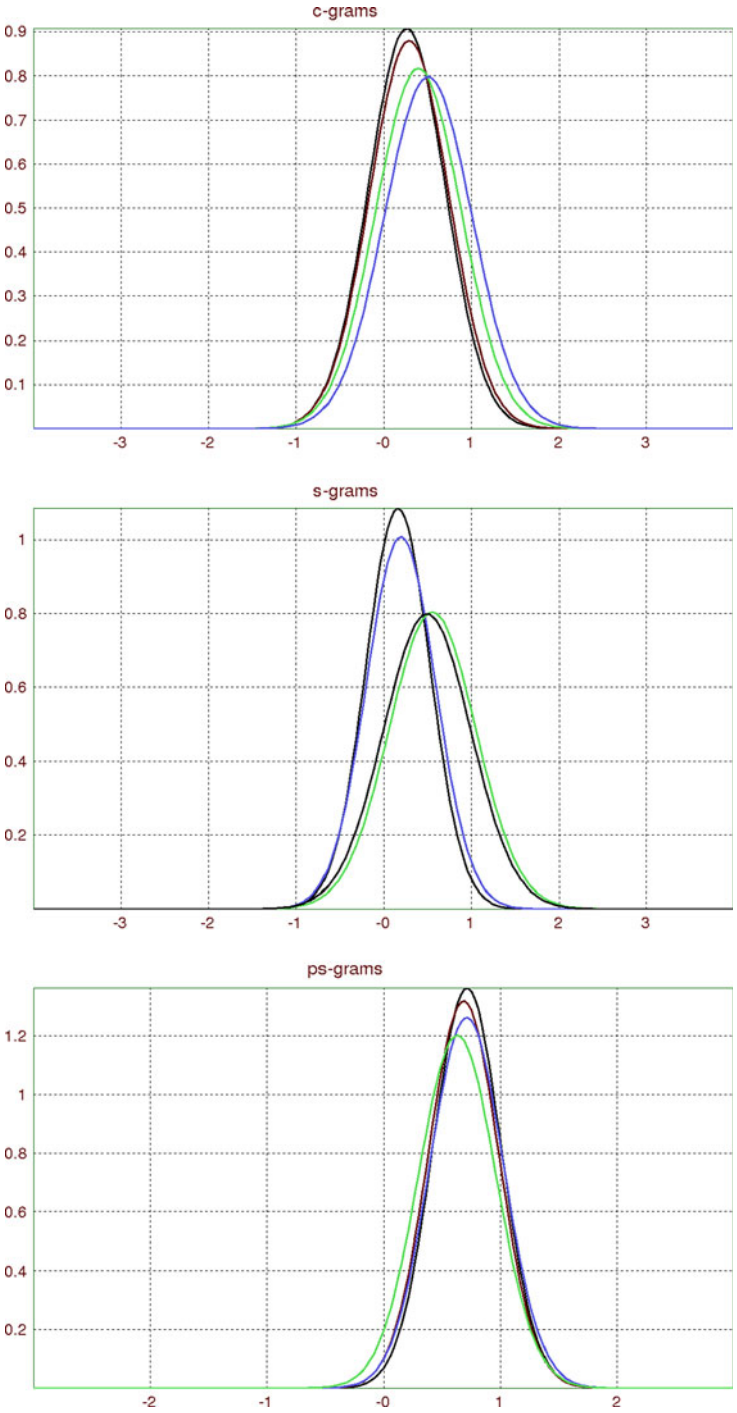

Fig. 7 PDF for dimensions in the style feature. (Color figure online)

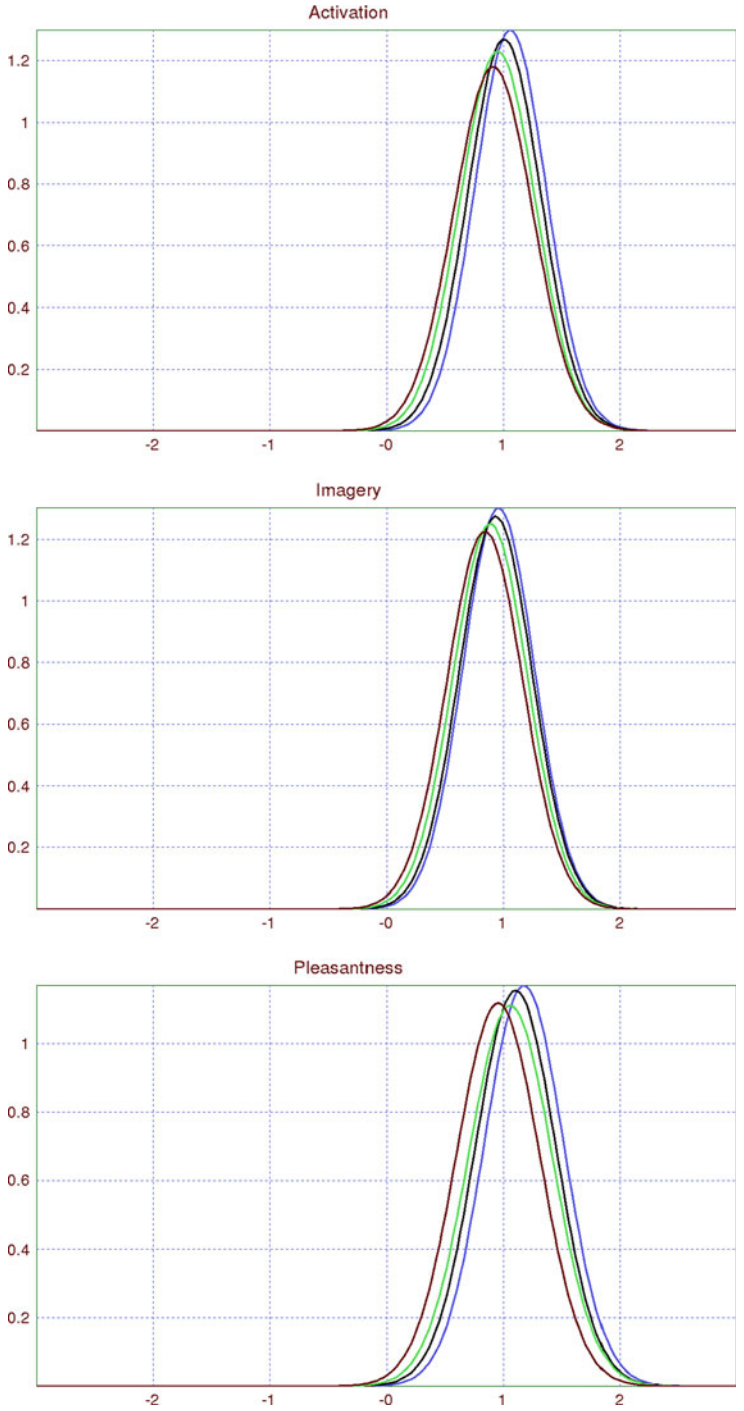

Fig. 8 PDF for dimensions in the emotional scenarios feature. (Color figure online)

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596.

Attardo, S. (2007). Irony as relevant inappropriateness. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 135–174). London: Taylor and Francis Group.

Balog, K., Mishne, G., & Rijke, M. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. In: *European chapter of the association of computational linguistics* (EACL 2006).

Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In: *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 161–164).

Carvalho, P., Sarmento, L., Silva, M., & de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In: *TSA '09: Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion* (pp. 53–56). Hong Kong: ACM.

Chin-Yew, L., & Och, F. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: *ACL '04: Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 605). Morristown, NJ: Association for Computational Linguistics.

Clark, H., & Gerrig, R. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General, 113*(1), 121–126.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string distance metrics for name-matching tasks. In: *Proceedings of IJCAI-03 workshop on information integration* (pp. 73–78).

Colston, H. (2007). On necessary conditions for verbal irony comprehension. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 97–134). London: Taylor and Francis Group.

Curcó, C. (2007). Irony: Negation, echo, and metarepresentation. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 269–296). London: Taylor and Francis Group.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: *Proceedings of the 23rd international conference on computational linguistics (COLING).*

Gibbs, R. (2007). Irony in talk among friends. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 339–360). London: Taylor and Francis Group.

Giora, R. (1995). On irony and negation. *Discourse Processes, 19*(2), 239–264.

Grice, H. (1975) Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). New York: Academic Press.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. In: *Proceedings of the fifth international conference on language resources and evaluation (LREC-2006)* (pp. 1222–1225).

Kreuz, R. (2001) Using figurative language to increase advertising effectiveness. In: *Office of naval research military personnel research science workshop*. Memphis, TN.

Kumon-Nakamura, S., Glucksberg, S., & Brown, M. (2007). How about another piece of pie: The allusional pretense theory of discourse irony. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 57–96). London: Taylor and Francis Group.

Lucariello, J. (2007) Situational irony: A concept of events gone away. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 467–498). London: Taylor and Francis Group.

Mihalcea, R., & Strapparava, C. (2006a). Learning to laugh (automatically): Computational models for humor recognition. *Journal of Computational Intelligence, 22*(2), 126–142.

Mihalcea, R., & Strapparava, C. (2006b). Technologies that make you smile: Adding humour to text-based applications. *IEEE Intelligent Systems, 21*(5), 33–39.

Miller, G. (1995). Wordnet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Monge, A., & Elkan, C. (1996). The field matching problem: Algorithms and applications. In: *In Proceedings of the second international conference on knowledge discovery and data mining* (pp. 267–270).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In: *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP)* (pp. 79–86). Morristown, NJ: Association for Computational Linguistics.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet::Similarity—Measuring the relatedness of concepts. In: *Proceedings of the 9th national conference on artificial intelligence (AAAI-04)* (pp. 1024–1025). Morristown, NJ: Association for Computational Linguistics.

Reyes, A., & Rosso, P. (2011). Mining subjective knowledge from customer reviews: A specific case of irony detection. In: *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (WASSA 2.011)* (pp. 118–124). Association for Computational Linguistics.

Reyes, A., Rosso, P., & Buscaldi, D. (2009). Humor in the blogosphere: First clues for a verbal humor taxonomy. Journal of Intelligent Systems 18(4), 311–331.

Saif, M., Cody, D., & Bonnie, D. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: *Proceedings of the 2009 conference on EMNLP* (pp. 599–608). Morristown, NJ: Association for Computational Linguistics.

Sarmento, L., Carvalho, P., Silva, M., & de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In: *TSA '09: Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion* (pp. 29–36). ACM: Hong Kong, China.

Sperber, D., & Wilson, D. (1992). On verbal irony. *Lingua, 87*, 53–76.

Tsur, O., Davidov, D., & Rappoport, A. (2010). {ICWSM}—A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In W. W. Cohen & S. Gosling (Eds.), *Proceedings of the fourth international AAAI conference on weblogs and social media* (pp. 162–169). Washington, D.C.: The AAAI Press.

Utsumi, A. (1996). A unified theory of irony and its computational formalization. In: *Proceedings of the 16th conference on computational linguistics* (pp. 962–967). Morristown, NJ: Association for Computational Linguistics.

Veale, T., & Hao, Y. (2009). Support structures for linguistic creativity: A computational analysis of creative irony in similes. In: *Proceedings of CogSci 2009, the 31st annual meeting of the cognitive science society* (pp. 1376–1381).

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. In: *Proceedings of 19th European conference on artificial intelligence—ECAI 2010* (pp. 765–770). Amsterdam: IOS Press.

Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports, 105*(2), 509–521.

Wilson, D., & Sperber, D. (2007). On verbal irony. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought* (pp. 35–56). London: Taylor and Francis Group.

Witten, I., & Frank, E. (2005). *Data mining. Practical machine learning tools and techniques*. Los Altos, CA, Amsterdam: Morgan Kaufmann Publishers, Elsevier.