How Challenging is Sarcasm versus Irony Classification?: An Analysis From Human and Computational Perspectives

Aditya Joshi[1,2,3], **Vaibhav Tripathi**[1], **Pushpak Bhattacharyya**[1], **Mark James Carman**[2]
Meghna Singh[1], **Jaya Saraswati**[1], **Rajita Shukla**[1]
[1]Indian Institute of Technology Bombay, India, [2]Monash University, Australia
[3]IITB-Monash Research Academy, India
{adityaj, pb}@cse.iitb.ac.in, mark.carman@monash.edu
mohan.meghnasingh@gmail.com, jaya.saraswati@gmail.com
rajita.shukla38@gmail.com

Abstract

Sarcasm and irony, although similar, differ in that sarcasm has an impact on sentiment (because it is used to ridicule a target) while irony does not. Past work treats the two interchangeably. In this paper, we wish to validate if sarcasm versus irony classification is indeed a challenging task. To this end, we use a dataset of quotes from English literature, and conduct experiments from two perspectives: the human perspective and the computational perspective. For the former, we show that three human annotators have lower agreement for sarcasm versus irony as compared to sarcasm versus philosophy. Similarly, sarcasm versus irony classification performs with a lower F-score as compared to another classification task where labels are not related: sarcasm versus philosophy classification. Another key point that our paper makes is that features designed for sarcasm versus non-sarcasm classification do not work well for sarcasm versus irony classification.

1 Introduction

Irony is a situation in which something which was intended to have a particular result has the opposite or a very different result[1]. On the other hand, sarcasm is a form of verbal irony that is intended to express contempt or ridicule. In other words, sarcasm has an element of ridicule and a target of ridicule, which is absent in irony (Lee and Katz, 1998). For example, '*He invented a new cure for a heart disease but later, died of the same disease himself*' is ironic but not sarcastic. However, '*I didn't make it big in Hollywood because I don't write bad enough*' is sarcastic where the target of ridicule is Hollywood.

Past work in sarcasm detection considers it as a sarcastic versus non-sarcastic classification (Kreuz and Caucci, 2007; Davidov et al., 2010; González-Ibáñez et al., 2011). Alternately, Reyes et al. (2012) consider classification of irony/sarcasm versus humor. In many past approaches, sarcasm and irony are treated interchangeably (Buschmeier et al., 2014; Joshi et al., 2015; Maynard and Greenwood, 2014). However, since sarcasm has a target that is being ridiculed, it is crucial that sarcasm be distinguished from mere irony. This is because when the target is identified, the sentiment of the target can be appropriately assigned. Owing to the two above reasons, sarcasm versus irony detection is a useful task.

In this paper, we investigate sarcasm versus irony classification. To do so, we compare sarcasm versus irony classification with sarcasm versus philosophy classification. In case of former, the two classes are similar (where sarcasm is hurtful/contemptuous). In case of sarcasm versus philosophy, the two classes are likely to be diverse. Thus, the goal of this paper is to *to establish the challenging nature of sarcasm versus irony detection*. The novelty of this paper is that we present our findings from two perspectives: human and computational perspectives. We first describe agreement statistics and challenges faced by human annotators to classify between sarcasm and irony. Then, to validate computational challenges of the task, we compare sarcasm versus irony classification with sarcasm versus philosophy classification. Our dataset consists of book snippets, annotated with one among three labels: sarcasm, irony and philosophy. The dataset is available on request. Our results show that for both humans and computers, detecting sarcasm versus irony in literature is more challenging than detecting sarcasm versus philosophy. Our experiments also suggest that the set of features generated in past works for sarcasm versus non-sarcasm tasks work well for sarcasm versus philosophy but not as much for sarcasm versus irony.

2 Related Work

Several approaches have been proposed for sarcasm detection, with context incongruity as the basis of sarcasm detection. Joshi et al. (2015) present features based on explicit and implicit incongruity for sarcasm detection. Maynard and Greenwood (2014) use contrasting sentiment between hashtag and text of a tweet as an indicator of sarcasm. Davidov et al. (2010) rely on Wallace and Do Kook Choe (2014) use properties of reddit comments to add contextual information. Recent work uses deep learning-based techniques for sarcasm detection (Poria et al., 2016; Joshi et al., 2016).

The work closest to ours is by Ling and Klinger

[1]Source: The Cambridge Dictionary

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati and Rajita Shukla. 2016. How Challenging is Sarcasm versus Irony Classification?: A Study With a Dataset from English Literature. In *Proceedings of Australasian Language Technology Association Workshop*, pages 123−127.

		Original Labels		
		Sarcasm	Philosophy	Irony
A1	Sarcasm	27	18	6
	Philosophy	5	222	17
	Irony	2	7	12
	Cannot say	13	24	14

Table 1: Confusion matrix for Annotator 1

		Original Labels		
		Sarcasm	Philosophy	Irony
A2	Sarcasm	30	20	9
	Philosophy	9	227	18
	Irony	4	16	16
	Cannot say	3	7	6

Table 2: Confusion matrix for Annotator 2

(2016). They present an empirical analysis of difficulty of understanding sarcasm and irony. They use a wide set of features for the classification task, and show that word-based features are good candidate features for the task. However, our analysis from both human and computational perspectives is novel, along with our observations.

Another novelty of this work is the domain of our dataset. Majority of the past works in sarcasm detection use tweets. Riloff et al. (2013a) and Maynard and Greenwood (2014) label these tweets manually whereas Bamman and Smith (2015) and Davidov et al. (2010) rely on hashtags to produce annotations. Some works in the past also explore long text for the task of sarcasm detection. Wallace and Do Kook Choe (2014) download posts from Reddit [2] for irony detection, whereas Lukin and Walker (2013) work with reviews. One past work by Tepperman et al. (2006) performs sarcasm detection on spoken dialogues as well. There are past works using literary quotes corpora for a variety of other NLP problems. Elson and McKeown (2010) extract quotes from popular literary work, for the task of quote attribution[3]. Søgaard (2012) perform the task of detecting famous quotes in literary works, gathered from the *Gutenberg Corpus*. Skabar and Abdalgader (2010) cluster famous quotations by improving sentence similarity measurements. This dataset consists of quotes from a quotes website. In terms of a dataset of literary snippets for sarcasm detection, we use the approach by Joshi et al. (2016) by using user-defined tags as labels.

[2]Reddit (`www.reddit.com`) is an entertainment, social news networking service, and news website.

[3]Quote Attribution is the computational task of attributing a quote to the most likely speaker.

		Original Labels		
		Sarcasm	Philosophy	Irony
A3	Sarcasm	16	13	6
	Philosophy	6	180	17
	Irony	4	33	11
	Cannot say	22	44	15

Table 3: Confusion matrix for Annotator 3

	All Three	Sarcasm-Irony	Sarcasm-Philosophy
A1	0.532	0.624	0.654
A2	0.479	0.537	0.615
A3	0.323	0.451	0.578

Table 4: Cohen's Kappa Values for the three annotators

3 Our Dataset of Quotes from English Literature

Goodreads[4] is a book recommendation website that allows users to track their friends' reads and obtain recommendations. We use a specific section of the website. The website has a section containing quotes from books added by the users of the website. These quotes are accompanied with user-assigned tags such as philosophy, experience, crying, etc. We download a set of 4306 quotes with three tags: philosophy, irony and sarcasm. These tags are assigned as the labels. The label-wise distribution is: (a) Sarcasm: 753, (b) Irony: 677, (c) Philosophy: 2876. We ensure that quotes marked with one of the three labels are not marked with another label. The dataset is available on request. Some examples in our dataset are:

1. **Sarcasm**: *A woman needs a man like a fish needs a bicycle*.
2. **Irony**: *You can put anything into words, except your own life*.
3. **Philosophy**: *The best way to transform a society is to empower the women of that society*.

The first quote is sarcastic towards a man, and implies that women do not need men. The victim of sarcasm in this case is 'a man'. On the other hand, the second quote is ironic because the speaker thinks that a person's own life cannot be put in words. It, however, does not express contempt or ridicule towards life or another entity. Finally, the last quote is philosophical and talks about transforming a society.

It is interesting to note that a sarcastic quote can be converted to a philosophical quote by word replacement. For example, converting the first quote to '*A woman needs a man like a fish needs water*' makes it

[4]`www.goodreads.com`

non-sarcastic (and arguably philosophical). The converse is also true. A philosophical quote can be converted to a sarcastic quote by word replacement. For example, converting the third (*i.e.*, philosophical) quote to '*The best way to transform a society is to empower the criminals of that society*' makes it sarcastic.

4 The Human Perspective

This section describes the human perspective of sarcasm versus irony classification. In the forthcoming subsections, we describe our annotation experiments followed by the quantitative and qualitative observations from these experiments.

4.1 Annotation Experiment

Three annotators, with annotation experience of 8k+ hours each, participate in our annotation experiment. We refer to them as A1, A2, and A3.

For a subset of 501 quotes as described in the previous subsection[5], we obtain exactly one label out of four labels: sarcasm, irony, philosophy and 'cannot say'. The last label 'cannot say' is a fall-back label that indicates that the annotator could not determine the label as one among sarcasm, philosophy and irony. The three annotators annotate the dataset separately. The annotators are provided definitions of the three classes as from the Free Dictionary. They are aware that sarcasm has an element of ridicule which irony lacks. In addition to these definitions, the annotators are instructed that a statement 'about' sarcasm/irony/philosophy (e.g. 'People use sarcasm when they are tired') must not be marked as sarcastic/ironic/philosophical.

4.2 Evaluation

The confusion matrices for the three annotators are shown in Tables 1, 2 and 3. The rows indicate labels assigned by an annotator, whereas columns indicate the 'gold' label *i.e.*, the label as extracted from the source website.

Table 4 compares Cohen's Kappa values for the three annotators with the gold label. In case of annotator A1, the agreement of the annotator with the gold labels for the three-label task is 0.532. The agreement of A1 with the gold labels for the Sarcasm-Irony task is 0.624. The corresponding value for sarcasm-philosophy task is higher: 0.654. This trend holds for the two other annotators as well. In general, an annotator agrees with the gold label in case of sarcasm versus philosophy classification, as compared to sarcasm versus irony classification.

4.3 Error Analysis

The following situations are where our annotators did not agree with the gold label, for each of the two pairs. These categories highlight the difficulties they faced during annotation.

[5]This subset was selected randomly.

	Precision (%)	**Recall (%)**	**F-Score (%)**
Sarcasm versus irony			
Average	65.4	65.4	65.4
Weighted Average	65.2	65.3	65.2
(b) Sarcasm versus philosophy			
Average	85	84.8	84.6
Weighted Average	76.5	77.7	77
(c) Sarcasm versus philosophy (class-balanced)			
Average	80.2	80	80
Weighted Average	80.2	80.1	80.1

Table 5: Average and weighted average values for three configurations: sarcasm versus philosophy, sarcasm versus philosophy (class-balanced) and sarcasm versus irony; Weighted average indicates that the values were weighted according to class label skews.

1. **Confusion between sarcasm and irony**: Consider the example '*... And I wondered if we had disappointed God so much, that he wrote us off as pets, just alive to entertain.*' Annotator A1 labeled this quote as sarcastic whereas the gold label was ironic. The annotators felt that the quote was a self-deprecating post where the speaker was being sarcastic towards themselves.

2. **Confusion between sarcasm and philosophy**: Consider another example '*Business people - Your business - is your greatest prejudice: it ties you to your locality, to the company you keep, to the inclinations you feel. Diligent in business - but indolent in spirit, content with your inadequacy, and with the cloak of duty hung over this contentment: that is how you live, that is how you want your children to live!*'. This example was labeled as philosophical according to the gold labels. However, Annotator A2 labeled it as sarcastic towards business people. Although the quote expresses contempt towards business people, it does not use positive words to express this contempt.

Config.	**Precision (%)**	**Recall (%)**	**F-Score (%)**
(a)	67.2	66.6	67.2
(b)	62.2	65.8	63.8
(c)	80.2	80.2	80.2

Table 6: Average Precision, Recall, and F-score values for the label 'sarcasm' for the three configurations

5 The Computational Perspective

In this section, we describe our results from training automatic classifiers to perform the two classification tasks: sarcasm versus irony and sarcasm versus philosophy.

5.1 Classifier & Features

We use LibSVM[6] to train our classifier. We use default parameters, and report five-fold cross-validation values. For features, we use features given in Joshi et al. (2015). These features were used to distinguish between sarcastic and non-sarcastic text. It is interesting to note that in our case, sarcasm versus philosophy is likely to indicate the 'sarcastic versus non-sarcastic' divide, but sarcasm versus irony is not as distant.

The features proposed by Joshi et al. (2015) are:

1. Unigrams
2. Pragmatic features: Capitalization, emoticons, punctuation marks
3. Implicit sentiment phrases: These are phrases that are indicative of sarcasm. They are extracted from a separate dataset of sarcastic tweets based on algorithm given in Riloff et al. (2013b).
4. Explicit sentiment features: # positive and negative words, largest positive/negative subsequences, lexical polarity

We consider three classification tasks: (a) Sarcasm versus irony, (b) Sarcasm versus philosophy, and (c) Sarcasm versus philosophy (data-balanced). The configuration in (c) neutralizes the effects of data skew on performance of classification since it is known that the performance on skewed datasets may not be reliable (Akbani et al., 2004). This configuration is motivated by the fact that (a) does not contain substantial skew. In case of (c), we undersample from philosophy class by randomly eliminating some training instances, as given in (Tang et al., 2009). This ensures that there are equal number of training and test instances from both classes for all folds.

5.2 Evaluation

Table 5 shows average and weighted average Precision, Recall, and F-score values for three sets of experiments. Weighted average indicates that the average is computed by weighting according to the class imbalance. On the other hand, average indicates that class imbalance is not taken into consideration.

The average F-score for sarcasm versus philosophy is 84.6%. In the class-balanced configuration as well, the F-score reduces to 80%. This F-score is 15% higher than that for sarcasm versus irony, where it is 65.4%. Also, the weighted average is 77% in case of sarcasm versus philosophy and 80.1% in case of sarcasm versus philosophy (class-balanced). The value is 12% higher than that for sarcasm versus irony, where it is 65.2%. The trend is common for both precision and recall. It must be noted that the features used in the classifier were generated initially for sarcasm versus non-sarcasm task. The results show that these features can be used for sarcasm versus philosophy as well, but not for sarcasm versus irony task. This points to the fact that for sarcasm versus irony classification, a new set of features will be required in the future.

To understand how the three configurations compare, it is also important to compare their performance for the sarcasm label. We, therefore, show the average values for the three configurations for the sarcasm class in Table 6. For the class-balanced sarcasm versus philosophy configuration, the F-score for sarcasm class is 80.2%. The corresponding value for sarcasm versus irony is 67.2%. This highlights that sarcasm versus irony proves to be challenging in general and specifically for sarcastic quotes.

6 Conclusion & Future Work

The focus of this paper is to highlight challenges of the sarcasm versus irony classification task, because sarcasm and irony are closely related to one another. We compare this classification formulation with sarcasm versus philosophy. To describe the challenging nature of the sarcasm versus irony classification task, we present our findings from two perspectives: human and computational perspective.

In terms of the human perspective, our three annotators have a lower Kappa score for sarcasm-irony as compared to sarcasm-philosophy and sarcasm-philosophy-irony classification. In the computational perspective, we observe that for the features reported for sarcasm versus non-sarcasm classification, sarcasm versus irony classification performs 12-15% lower than sarcasm versus philosophy. Even in case of the sarcasm class, the difference is 13%. Our findings show that although these features work well for sarcasm versus philosophy classification, they do not work well for sarcasm versus irony classification. This means that novel features are imperative for the task of sarcasm versus irony classification.

Our findings show the non-triviality and challenges underlying sarcasm versus irony classification. Since a key distinction between sarcasm and irony is a target of ridicule, having techniques for the detection of sarcasm targets, like in the case of sentiment target identification, may be helpful. Our results will also act as a baseline for future work in sarcasm versus irony classification. Additionally, features that distinguish between the two will be useful.

References

Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz, 2004. *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa,*

[6]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Italy, September 20-24, 2004. Proceedings, chapter Applying Support Vector Machines to Imbalanced Datasets, pages 39–50. Springer Berlin Heidelberg, Berlin, Heidelberg.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

David K Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *AAAI*. Citeseer.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 757–762.

Aditya Joshi, Kevin Patel, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *EMNLP*.

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

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

Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *Semantic Web: ESWC 2016 Satellite Events*.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue.

Diana Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.*, 74:1–12, April.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013a. Sarcasm as contrast between a positive sentiment and negative situation.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013b. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, volume 13, pages 704–714.

Andrew Skabar and Khaled Abdalgader. 2010. Improving sentence similarity measurement by incorporating sentential word importance. In *AI 2010: Advances in Artificial Intelligence*, pages 466–475. Springer.

Anders Søgaard. 2012. Mining wisdom. *NAACL-HLT 2012*, page 54.

Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. 2009. Svms modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):281–288.

Joseph Tepperman, David R Traum, and Shrikanth Narayanan. 2006. " yeah right": sarcasm recognition for spoken dialogue systems.

Byron C Wallace and Laura Kertz Do Kook Choe. 2014. Humans require context to infer ironic intent (so computers probably do, too).