# Unjust Equivalence: Are Irony and Sarcasm Truly the Same in NLP?

**Bojan Puvača, Florijan Sandalj, Ivan Unković**

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{bojan.puvaca, florijan.sandalj, ivan.unkovic}@fer.hr

**Abstract**

## 1. Introduction

Human communication is a convoluted subject, being a topic of discussion and investigation across multiple fields of research. There are various forms of human communication which are particularly intriguing to linguists and psychologists, with irony and sarcasm being of paramount importance due to their complexity and the depth of insight they provide into human cognition and social interaction.

## 2. Irony and sarcasm in NLP

The relationship between irony and sarcasm is unfortunately a heavily contested subject in NLP. This problem is the easiest to notice when looking at different sarcasm and irony datasets, where we can find cases when they are treated as completely seperate concepts (Nikhil, 2020), when sarcasm is treated as a subset of irony (Van Hee et al., 2018) or even vice versa (Oprea and Magdy, 2020). Searching for a consensus in the realm of linguistics is a futile effort as well, however we have found two distinctions between the two that have merit in the context of NLP.

### 2.1. Sarcasm - irony's meaner cousin

The online Merriam-Webster dictionary defines sarcasm as "a sharp and often satirical or ironic utterance designed to cut or give pain" (Merriam-Webster, 2024). This definition seems to be in line with the general consensus that sarcasm is a form of irony that is more patronizing and mean-spirited. The iSarcasm dataset (Oprea and Magdy, 2020) is a good example of this categorization, as the "sarcasm" label is, in fact, a subset of the unfortunately named "sarcastic" label, which actually indicates any kind of ironic speech.

In this context, irony refers to any type of speech that is based on polarity - whether that be pointing out the polarity between the expected and actual outcome of a situation (situational irony), or expressing with words the opposite of what we mean (verbal irony based on polarity). Although this definition works on paper, there are some pitfalls. Most notably, the line between sarcasm and verbal irony is unclear, as whether or not a statement is mean-spirited is subjective, while a clearly mean spirited statement can also be based on polarity. Also, tweets that contain irony and aren't directed at a specific person can still be considered sarcastic, as they often target a group of people, concepts, ideas or themselves in the form of self-deprecating humor. How the object of the irony affects the classification is a question that remains unanswered.

All things considered, this distinction is a fair starting point for NLP research, and the iSarcasm (Oprea and Magdy, 2020) dataset does a solid job at distinguishing between the two. However, the usefullness of this distinction is limited, as both concepts are based on dishonest speech, meaning that in practice there isn't much use in distinguishing between the two.

In this paper, we will take a closer look at how different models perform on the seperate tasks of irony and sarcasm detection, with the goal of determining the amount of overlap between the two tasks and the potential benefits of treating them as seperate concepts in NLP. We will do so using a combination of the iSarcasm (Oprea and Magdy, 2020) and the SemEval-2018 (Van Hee et al., 2018) datasets, both of which contain tweets that are labeled as either ironic or sarcastic based on this distinction.

### 2.2. Sarcasm - the figure of speech

## 3. Experimental setup

Three seperate datasets were created for the experiment of comparing irony and sarcasm detection, one containing tweets labeled as ironic, one containing tweets labeled as sarcastic and the third one combining the first two.

All three datasets were constructed as a binary classification task, with neutral tweets, not containing any irony or sarasm, being labeled as negative, and the tweets containing either irony or sarcasm being labeled as positive.

We ensured that during training and evaluation, all three datasets were completely balanced, with an equal number of positive and negative examples in the training, validation and test sets. As the SemEval-2018 dataset contained significantly more ironic tweets than sarcastic ones, we merged it with the iSarcasm dataset in order to produce larger and more balanced datasets. We found this approach to be justified, as both datasets discerned between irony and sarcasm in a similar manner, both in their explanations for the labels and upon manual inspection of the tweets.

Various models were trained on all three of these tasks, after which their performance was evaluated on the test sets of all three datasets in order to determine the amount of overlap between irony and sarcasm detection.

In section 3.1., we will describe the process of constructing the datasets and in 3.2. we will describe the models used in this experiment.

### 3.1. Construction of the datasets

The data for the sarcasm detection task

### 3.2. Models

(Oprea and Magdy, 2020)

## 4. Results

## 5. Discussion

## 6. Conclusion

## Acknowledgements

## References


Merriam-Webster. 2024. Merriam-webster dictionary.

John Nikhil. 2020. Tweets with sarcasm and irony.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In Dan Jurafssky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online, July. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana, June. Association for Computational Linguistics.