“Can we reliably detect sarcasm in a comment?”

A PROJECT PROPOSAL

David Kaiser, PhD

Fundamentals of Data Science, Spring 2020

Sarcasm is hard to define, as we see in the four divergent dictionary definitions below:

- Merriam-Webster Dictionary: "a sharp and often satirical or ironic utterance designed to cut or give pain"
- Oxford Dictionary: "the use of irony to mock or convey contempt."
- Cambridge Dictionary: "remarks that mean the opposite of what they say, made to criticize someone or something in a way that is amusing to others but annoying to the person criticized"
- dictionary.com "harsh or bitter derision or irony."

As hard as it can be to define it, it can be even harder to spot it. In spoken conversation, sarcasm relies heavily upon tone, inflection, and facial expression or other gestures. All of these are missing in text, although text can be enhanced with emojis, punctuation or other markers. As a result, it can be very difficult for a reader to pick up on the writer’s intention.

But perhaps it is detectable, based upon features of the text. Perhaps we just haven’t found the correct features or combinations of features.

In this study, we will use machine learning techniques to classify Reddit comments as either Sarcastic or not.

Our data set from Kaggle (see below for URL) contains 1,010,826 comments, evenly split into those which were labeled sarcastic (by users, using the ‘/s’ tag) and those which were not. This which are sarcastic have

the “label” value of “1” in the data set, those which are not have label “0.” We will be trying to predict if a comment should be labeled 0 or 1.

The data set has ten columns, including the label, the comment itself, the author’s username, the total upvotes, downvotes and net (upvotes minus downvotes), the subreddit where it was found, the month and year of creation, the UTC timestamp of creation time, and the parent comment it was posted to. The data file is a .CSV format which is more than 250MG.

The good news is that the data appears to be complete, with no NA values that need to be cleaned or deleted.

The interesting part of the task pertains to what data is used to predict sarcasm.

Ideally, we would like to predict it from the text of the comment alone. Even then, we will face some difficult decisions regarding text preprocessing. While many text-related tasks will put all of the letters into lower case and remove punctuation and non-word characters, and then possibly run the words through a stemmer and / or lemmatizer to standardize them, it may be the case that these “non-sterilized features” are necessary to predict sarcasm, so we may need to keep them.

At another level, we may find that adding other data, such as the sub-reddit where the comment was found, or the number of downvotes, may be predictive of sarcasm.

We expect we will have to run our model-building processes several times in order to determine what type and degree of text-processing is necessary and which variables are informative.

https://www.kaggle.com/sherinclaudia/sarcastic-comments-on-reddit

Data provided by Mikhail Khodak, Nikunj Saunshi and Kiran Vodrahalli: "A Large Self-Annotated Corpus for Sarcasm"