Sarcasm detection on Reddit comments DS5110

Praveen Kumar Sridhar, Kapil Sahu, Meghaana Tummapudi

16 November 2021

1 Summary

The aim of our project is to detect if the comments in Reddit threads are sarcastic or not. Sarcasm detection is an arduous task, as it's largely dependent on context, prior knowledge and the tone in which the sentence was spoken or written. It is crucial to know what exactly sarcasm is since its borders are not exactly well defined unlike in sentiment analysis where the sentiment categories are very clearly defined ("love" objectively has a positive sentiment, "hate" a negative sentiment no matter who you ask or what language you speak).

A sarcastic expression or comment can be defined as one that is caustic, bitter, or cutting. Sarcastic words are generally used to communicate the opposite of their surface meaning humorously or to mock someone or something. Although sarcasm is not always ironic, it might include ambivalence. Sarcasm is most perceptible in spoken words and is usually context-dependent. It is mostly identifiable by the inflection with which it is spoken.

2 Dataset

This dataset contains 1.3 million Sarcastic comments from the Internet commentary website Reddit. The dataset was generated by scraping comments from Reddit containing the (sarcasm) tag. This tag is often used by Redditors to indicate that their comment is in jest and not meant to be taken seriously, and is generally a reliable indicator of sarcastic comment content.This is a balanced dataset.

Attribute Information:

1. label: If comment is Sarcastic or not
2. comment: The comment for which we need to determine if its sarcastic or not
3. author: Author of the comment
4. subreddit: The subreddit in which the comment was posted
'5. score: The net of upvote and downvotes
6. ups: The number of upvotes
7. downs: The number of downvotes
8. date: The date comment was posted
9. created_utc: The timestamp when the comment was posted
10. parent_comment: The parent comment to which the comment was posted as a response.

3 Objective

To perform sentiment analysis and detect whether a particular comment on the social news website and forum "Reddit" can be classified as sarcastic or not.

4 Proposed Plan

We will be performing exploratory data analysis (EDA) on the selected dataset to gain interesting insights and answer the following queries:

1. Detecting which sub-reddit has more sarcastic comments?
2. Do sarcastic comments get more upvotes?
3. Does sarcasm come with profanity, as a response to profanity in comments?
4. What are some most frequently used internet slang in a sarcastic comment?
5. Do people on Reddit use emoticons with sarcastic comments?
6. Observing/analysing sarcasm trends over time on popular sub-reddits.

Sarcasm is a sentence that has the opposite meaning from the written words or spoken words. Sarcastic comment often uses different polarity than they express. Hence, one of the biggest **challenge** is to know and understand the context while predicting. A comment that seems to be positive in first sight might pose as sarcasm after knowing the context. We will be exploring if knowing the context from parent comment improves the accuracy of our model in predicting sarcasm.
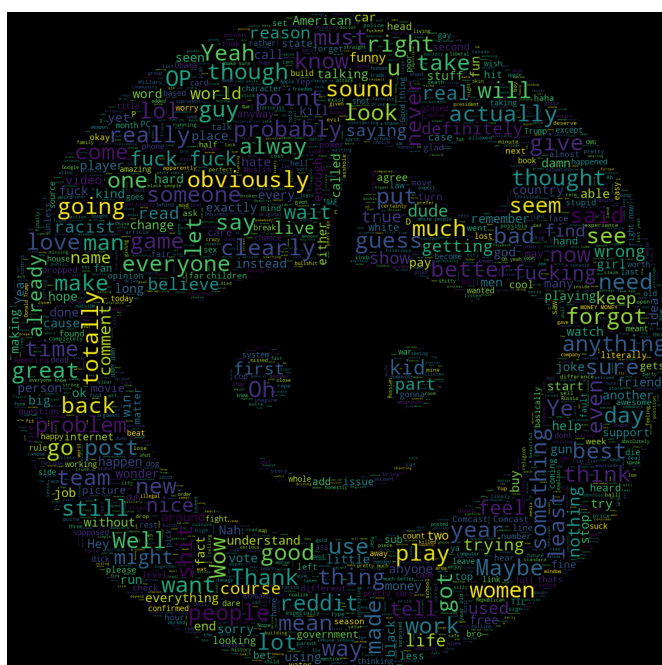
As a part of EDA we will be visualising the missing values(if any), in the dataset. We will be doing univariate/multivariate analysis to find answers to the objectives mentioned above and to find any new relations among the features. We will be using Neural Networks with models like artificial neural networks, CNN, RNNs like LSTMs with the representations/embedding of the words using techniques like bag-of-words with TF-IDF, Word2Vec and/or GloVe to predict whether a comment is sarcastic or not.

5 Preliminary Results

The following plot represents the top 25 subreddits with the most number of comments in the dataset. From this we can see that the r/Askreddit is the most popular subreddit by a lot.

Top 25 popular subreddit
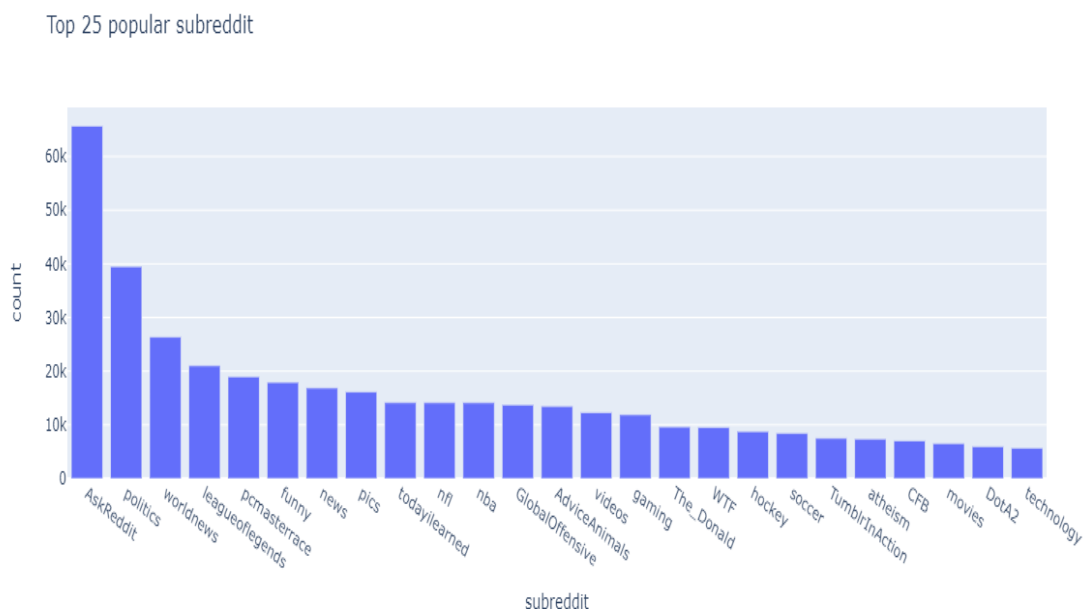
The following is a wordcloud formed from the comments in the comments column in the dataset. A "Wordcloud" is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

6 References

1. https://www.kaggle.com/danofer/sarcasm
2. https://nlp.stanford.edu/projects/glove
3. https://radimrehurek.com/gensim/models/word2vec.html