Sarcasm Tones: Reverse Engineering Sarcasm in Twitter

Varin Sikand and Aditya Guin

Computer Science; varin.sikand@utdallas.edu, aditya.guin@utdallas.edu

Faculty Mentor: Prof. Vincent Ng; Program: Computer Science

2023 Jonsson School UG Research Award Recipient

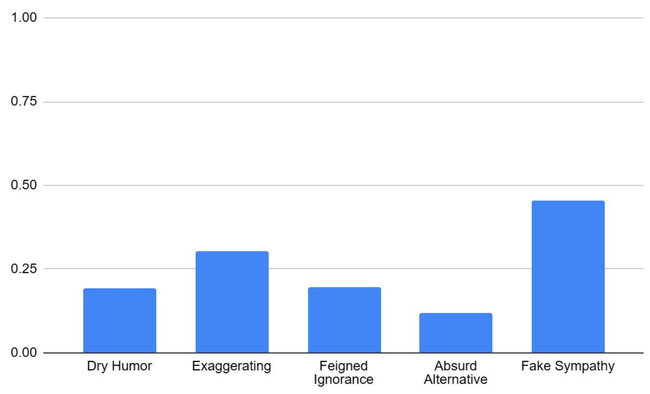
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

Research Project Goals:

- We aim to show the benefits of considering the tone of sarcasm within twitter sarcasm detection datasets
- We aim to provide a clear and concise method of identifying and understanding tones

Sarcasm in Twitter

- The task of Sarcasm Detection is a binary classification task, with a label of 1 given to statements that are sarcastic and 0 otherwise.
- Datasets are constructed by identifying certain “sarcasm hashtags”
  - #sarcasm, #irony, #not, etc.
- This results in a highly biased dataset that relies on these signals to convey sarcasm.
  - This also led to the tweets to have the same tone of sarcasm, since they all depended on the same hashtag to convey sarcasm.



Distributions of tone labels in iSarcasmEval training dataset

- We consider the iSarcasmEval dataset, which was gathered using the authors of the tweets themselves rather than signals
  - We identified 5 distinct tones, and their distributions showed how this method resulted in more diverse tweets.
- We surveyed participants, asking first for a non-sarcastic rephrase of the sarcastic tweets while changing as few of the original words as possible
- Next, the participant considers words in the sarcastic version but not in the rephrase. These “sarcasm words” are then used to identify tone, with tone labels operating as categories for the sarcasm words. It was important that we define an objective method for defining these tones as opposed to the standard.
- This method draws a clear line between the meaning of a tweet and the tone of its sarcasm. The ability of humans to extract a nonsarcastic meaning from a sarcastic statement requires this.

Tone labels	Description	Sarcasm Words
Dry Humor	This tweet's sarcasm is portrayed with a non-serious tone, but still uses a statement that means the opposite of what they believe/mean	Sarcasm words make light of the original point by (for example) likening it to unimportant subjects or by formatting it as a joke
Exaggerating	This tweet depicts sarcasm via an absurdly exaggerated claim or statement, making it obvious that the author meant the opposite of what they said	Sarcasm words take the original (nonsarcastic) statement and exaggerates the user's point such that the final version, while still stating the user's point, is inaccurate in terms of severity
Feigned Ignorance	The user showcases sarcasm by pretending to have no knowledge on the subject matter, but then proceeds to make their informed point afterwards.	Sarcasm words are expressing doubt (for example by posing their point as a question or by making their nonsarcastic point unsure)
Absurd Alternative	This tweet depicts sarcasm by providing an alternative option to their original point. This alternative is so absurd that the implication is that their point is the only valid one	Sarcasm words introduce a new subject to liken the original point to. This new subject is less believable or true than the original point
Fake Sympathy	The user displays sarcasm by using a sympathetic tone and appearing to agree with the subject, while their statements clearly admonish the subject.	Sarcasm words are of positive sentiment, are encouraging, or otherwise enthusiastic about the subject matter they disagree with

Model Setup:

- Baseline Model
  - BERTweet-Large model
    - BERT encoder pretrained on ~860 million Tweets
  - Softmax classification layer outputting labels of 0 or 1
  - Trained on iSarcasmEval and SemEval 2017 Task 3A
- Tone Generator
  - Trained using the tone labels gathered from our surveys
  - Separate XGBoost Tree models for each tone label
    - Binary classification with 1 indicating presence of tone
- ChatGPT w/o tones
  - Prompt asks for sarcasm detection with the following definition of sarcasm: “A form of verbal irony in which the statement's literal meaning is inconsistent with the speaker's true intent”
- ChatGPT w/ tones
  - Prompt gives descriptions of tones along with examples and specifies that a tweet is sarcastic if and only if it belongs to at least one of the tones

Research Project Results:

Model	Precision	Recall	F1-Score
Baseline	0.584	0.68	0.628
Tone Generator	0.183	0.69	0.29
ChatGPT w/o tones	0.213	0.285	0.244
ChatGPT w/ tones	0.320	0.85	0.464

- Using iSarcasmEval test dataset with 1400 tweets; 200 sarcastic
- Training Dataset: 3468 tweets; 25% sarcastic

Project Conclusions/Outcomes:

Although our Baseline still performed the best, this is due to the inclusion of the SemEval 2017 Task 3A dataset in the training data. Without this additional data, the F1 Score was 0.526.
Notably, ChatGPT performed significantly better on the test set when instructed to classify sarcasm in terms of the tones. This is despite the Tone Generator performing poorly even though it was trained using the survey results. This seems to suggest that the tones themselves are useful for identifying sarcasm, as seen by the ChatGPT results, but we would need more training data for a BERT model to fit correctly.

Overall, this methodology for classifying sarcasm tones has merit in more than just the field of information mining. The area of ML explainability is becoming increasingly coveted as the models become increasingly complicated. Therefore, being able to describe why a certain statement is sarcastic, in what way it depicts sarcasm, and even identifying the words that cause it to be sarcastic ought to be of great interest. This work shows that the identification of tones can be beneficial to the task of sarcasm detection.