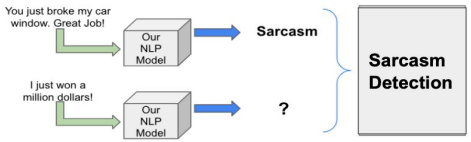
Evaluating the Performance of Statistical and Transformer-Based Language Models for Sarcasm Detection

Authors: Lauren Sampson, William Potts, John Goulart, Stephen Schreder

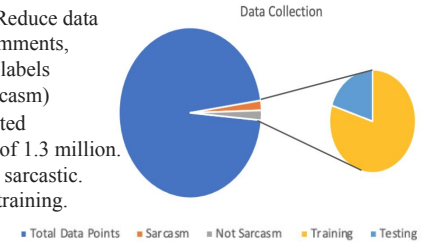
Intro/Motivation

- **Objective**: Build a sarcasm detection model that predicts whether a Reddit message is either sarcastic or not sarcastic
- **What makes this model unique?** Analyzes both semantic and lexical elements of the comment in sarcasm determination as well as looks at the role of the parent comment in classification
-Sarcasm detection is a **difficult task** for both humans and machines



Data

Data/Target Sample: Reduce data set to 50,000 Reddit comments, parents comments, and labels (0: Not Sarcasm, 1: Sarcasm)
- 50k is randomly selected from the total corpus of 1.3 million.
- 25k sarcastic, 25k not sarcastic.
- 80% of data used for training.
- 20% used for testing



Parent Comment: "This is the obvious endgame"
Sarcasm Detected: "Yeah because becoming worse than the place they're fleeing is the right direction to go in."

Parent Comment: "What is the best way to waste 5 hours right now?"
No Sarcasm Detected: "Taking a nap"

Method

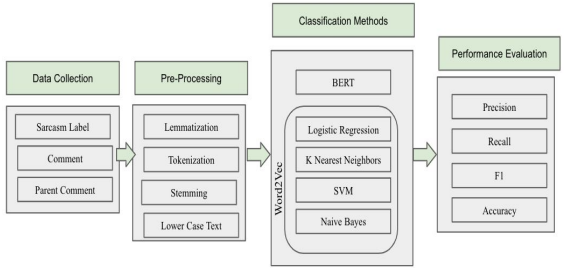
-Approach: Try each method with and without the inclusion of parent comments.

Method 1

-Tokenize the data with AutoTokenizer to generate unique integer IDs and the attention mask.
-Train and test the data with DistilBERT.

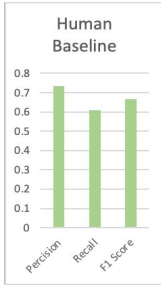
Method 2

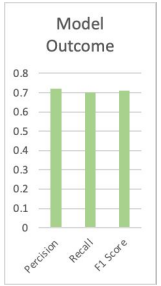
-Tokenize the data with NLTK and generate word embeddings with Word2Vec.
-Train multiple classifiers with the summed embeddings of each comment as features.



Results

Human Classification Baseline:
-4 judges classified 500 comments as either sarcastic or not sarcastic, based on their own understanding within the context of the parent comments
-The human baseline was better than many of the classifiers but relatively comparable to BERT





Without Parent Comments

Methods	Precision	Recall	F1
KNN	.54	.54	.54
NB	.53	.51	.45
LinSVC	.62	.62	.62
LogReg	.62	.62	.62
BERT	.73	.70	.71

With Parent Comments

Methods	Precision	Recall	F1
KNN	.53	.53	.53
NB	.52	.51	.41
LinSVC	.60	.60	.60
LogReg	.60	.60	.60
BERT	.73	.71	.72

Conclusion

- Sarcasm is **difficult to classify** whether it is a machine or human
- Transformer models perform better than statistical methods

Future Work:

- Use model to develop system that integrates with messaging applications to flag messages as potential sarcasm, especially helpful for users with ASD (Autism Spectrum Disorder)
- Emoticons in the comments might be a good way to expand this study