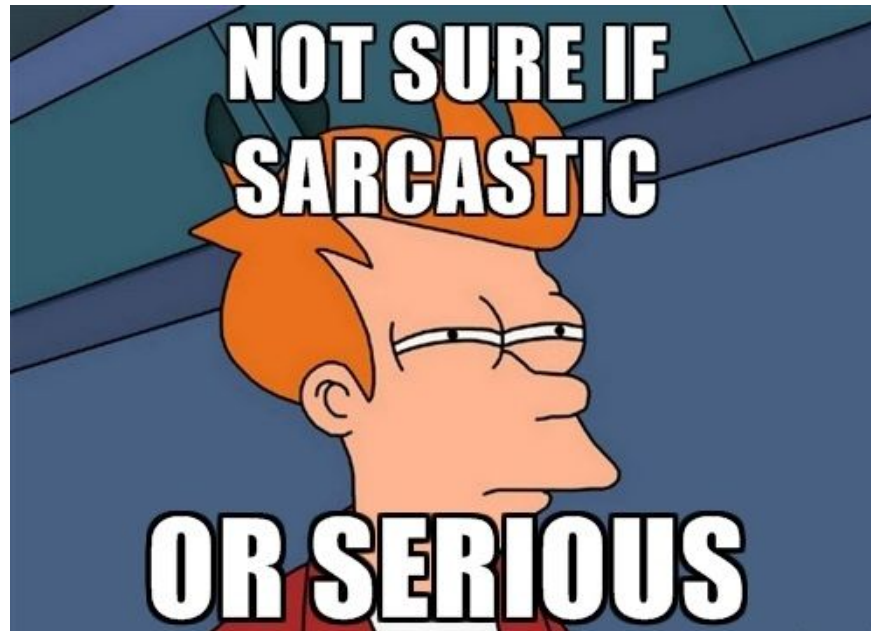


Analyzing Context and User Information in Online Sarcasm Detection

CS 6471 Computational Social Science
Tusheet Goli and Mohan Dodda

Introduction

- Identify sarcasm in conversations
- Examine context and circumstances that provoke sarcastic responses
- Analyze linguistic cues and user attributes that correlate to high sarcastic behaviour



Importance

- Help distinguish between incorrect information and sarcasm

Wow, the Republicans really care about the LBGTQIA+ community.

They really care about my rights!

Importance

- Help people prevent sarcastic responses in their comments
 - Sarcasm can give the feeling being left out
 - Sarcastic responses not normally desired (negativity)
- Borderline cyber-bullying
- Gain a deeper understanding of human psychology and contribute to linguistic analysis

Data

- SARC Reddit Dataset

	label	comment	author	subreddit	score	ups	downs	date	created_utc	parent_comment
0	0	NC and NH.	Trumbart	politics	2	-1	-1	2016-10	2016-10-16 23:55:23	Yeah, I get that argument. At this point, I'd ...
1	0	You do know west teams play against west teams...	Shbshb906	nba	-4	-1	-1	2016-11	2016-11-01 00:24:10	The blazers and Mavericks (The wests 5 and 6 s...
2	0	They were underdogs earlier today, but since G...	Creepeth	nfl	3	3	0	2016-09	2016-09-22 21:45:37	They're favored to win.
3	0	This meme isn't funny none of the "new york ni...	icebrotha	BlackPeopleTwitter	-8	-1	-1	2016-10	2016-10-18 21:03:47	deadass don't kill my buzz
4	0	I could use one of those tools.	cush2push	MaddenUltimateTeam	6	-1	-1	2016-12	2016-12-30 17:00:13	Yep can confirm I saw the tool they use for th...

Data

- SARC Reddit Dataset

- Contains parent comment and response (which can be sarcastic or not be sarcastic)

Parent
Comment:

“Those statistics on
Reddit are typically
overstated”



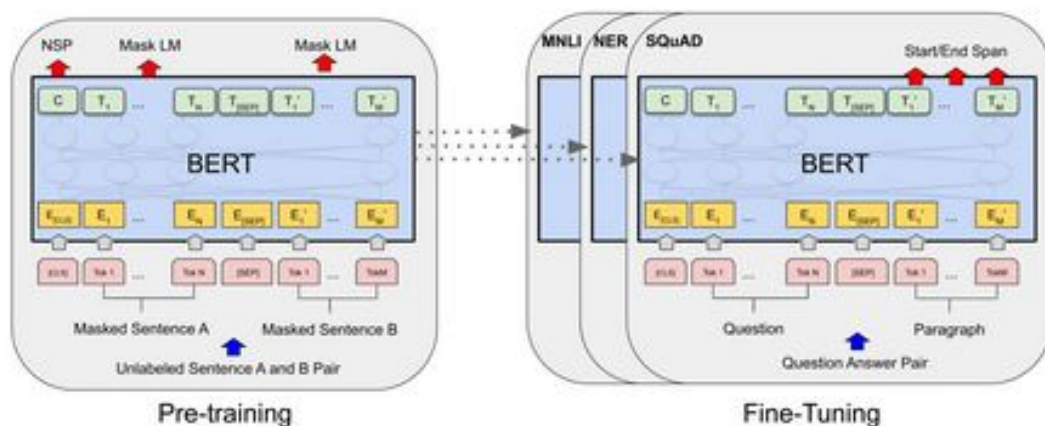
Sarcastic
Response:

“Got a source on
that?”

Statistics vs. comment type	Sarcastic	Non-Sarcastic
average response size (words)	10.33	10.59
average parent comment size	24.21	24.56
mean upvotes	5.22	5.78
mean downvotes	-0.13	-0.17
mean score	6.40	7.37
Number of Documents	505368	505405
label	1	0

Methods

- Fine-tuned BERT model on response, parent comments, and parent+response
- Analyzed data further in subjectivity, number of special characters, word/sentence length, and profanity



Preliminary Results - Model Results

- BERT outperforms Baseline Logistic Regression (Khodak et. al. 2018)
- Utilizing only the parent comment shows promise in predicting sarcasm
- Concatenating parent and response seems to have worse results than just response
 - Too much noise?
 - Concatenation not best method?

Accuracy

Dataset vs model	Logistic Regression	BERT - full dataset
Only Response	0.72	0.78
Only Parent Comment	0.58	0.61
Parent Comment and Response	0.68	0.74

Preliminary Results - Subjectivity

- Sarcasm is more prevalent when the context is antithetical (personal opinion) rather than being factual (Saha et. al. 2017)
- Can be a decent predictor for sarcasm

Type	Sarcastic	Non-sarcastic
Parent comment	18625.377	18114.239
Sub-comment	16510.655	15225.279

Table 6: Subjectivity score of dataset partitions with higher score representing a higher subjectivity

Preliminary Results - Profanity

- Sarcastic comments are generally associated with a larger degree of profanity and vulgarity (Du et. al.)
- Parent comments with a larger degree of profanity tend to provoke sarcastic replies

Sarcastic	Non-sarcastic
5766	5333

Table 6: Number of Profane words by the label of the dataset

Preliminary Results - Special Characters

- Strong correlation for sarcastic text containing the '!' character
- Unfortunately, this was not true for other special characters like '?', etc.

Character	Sarcastic	Non-sarcastic
!	6884	2727
?	5564	6306

Table 5: Amount of times a specific special character appears for sarcastic and non-sarcastic texts

Future Plans

- Specifically analyze user information-related features
 - Common subreddits, popularity, etc.
- BERT+CASCADE model (Hazarika et. al. 2018)
- Train our classifiers by embedding the context and user features
- Verify results using this feature-based model

References

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1837—1848. Association for Computational Linguistics.

Khodak, M., Saunshi, N., Vodrahalli, K. A large self-annotated corpus for sarcasm. In Proceedings of the Linguistic Resource and Evaluation Conference (LREC), 2018.

Questions