Analyzing Context and User Information in Online Sarcasm Detection

Members: Tusheet Sidharth Goli and Mohan Dodda

College of Computing, Georgia Institute of Technology

Introduction

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

Motivation
- Help distinguish between incorrect information and sarcasm
- Help people prevent sarcastic responses in their comments
  - Sarcasm can give the feeling of being left out (negativity)
- Deeper understanding of human psychology and help linguistic analysis

Data

Self-Annotated Reddit Corpus (SARC) [2] Reddit Dataset

	label	comment	author	subreddit	score	ups	downs	date	created_utc	parent_comment
0	0	NC and NH.	Trumpbart	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...

- Contains parent comment and response (which may or may not be sarcastic)
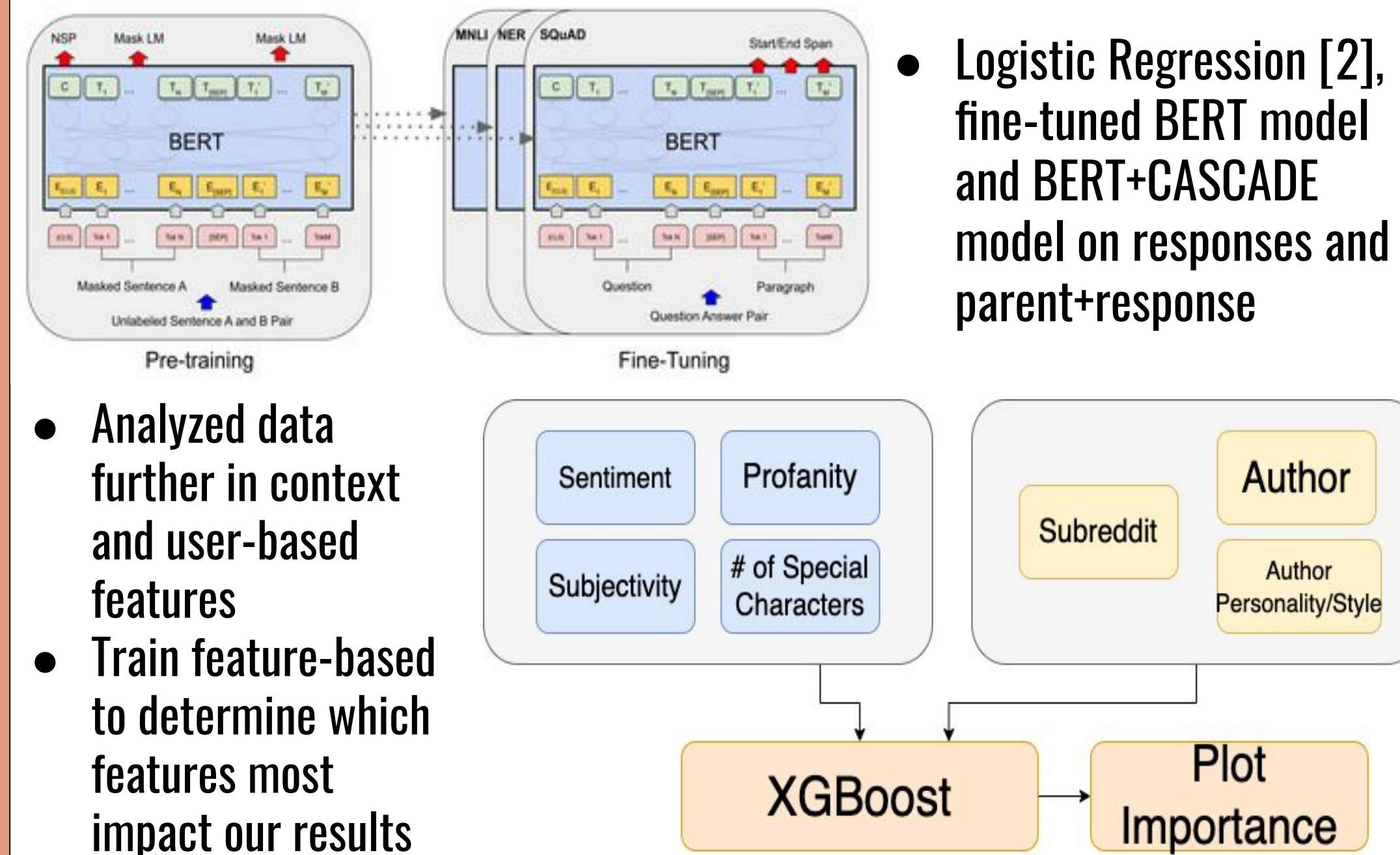
Parent Comment

"Those statistics on Reddit are typically overstated"
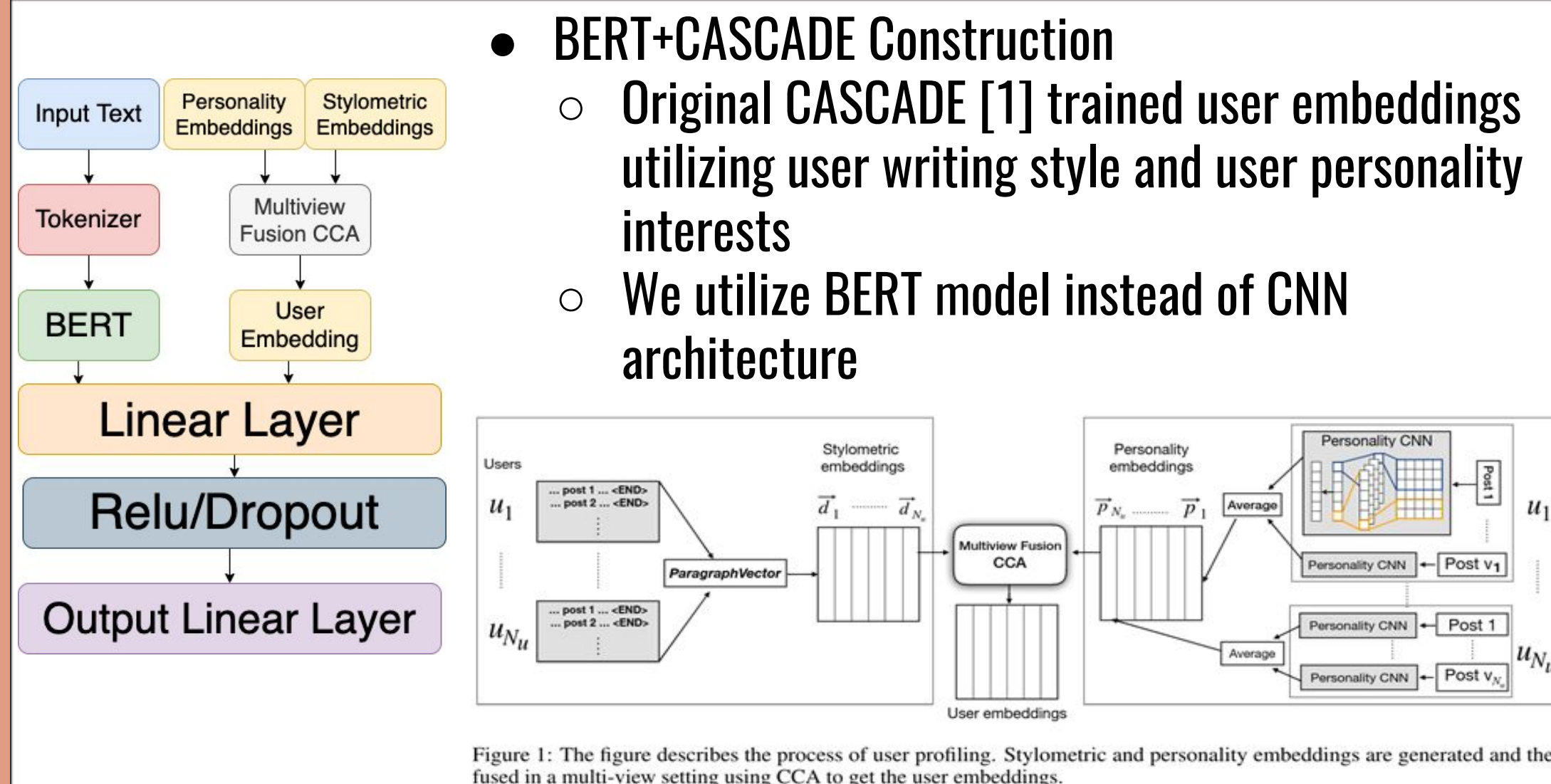
↓

"Got a source on that?"

Response

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

- Logistic Regression [2], fine-tuned BERT model and BERT+CASCADE model on responses and parent+response

- Analyzed data further in context and user-based features
- Train feature-based to determine which features most impact our results

Proposed Architecture - BERT+CASCADE

- BERT+CASCADE Construction
  - Original CASCADE [1] trained user embeddings utilizing user writing style and user personality interests
  - We utilize BERT model instead of CNN architecture

Figure 1: The figure describes the process of user profiling. Stylometric and personality embeddings are generated and then fused in a multi-view setting using CCA to get the user embeddings.
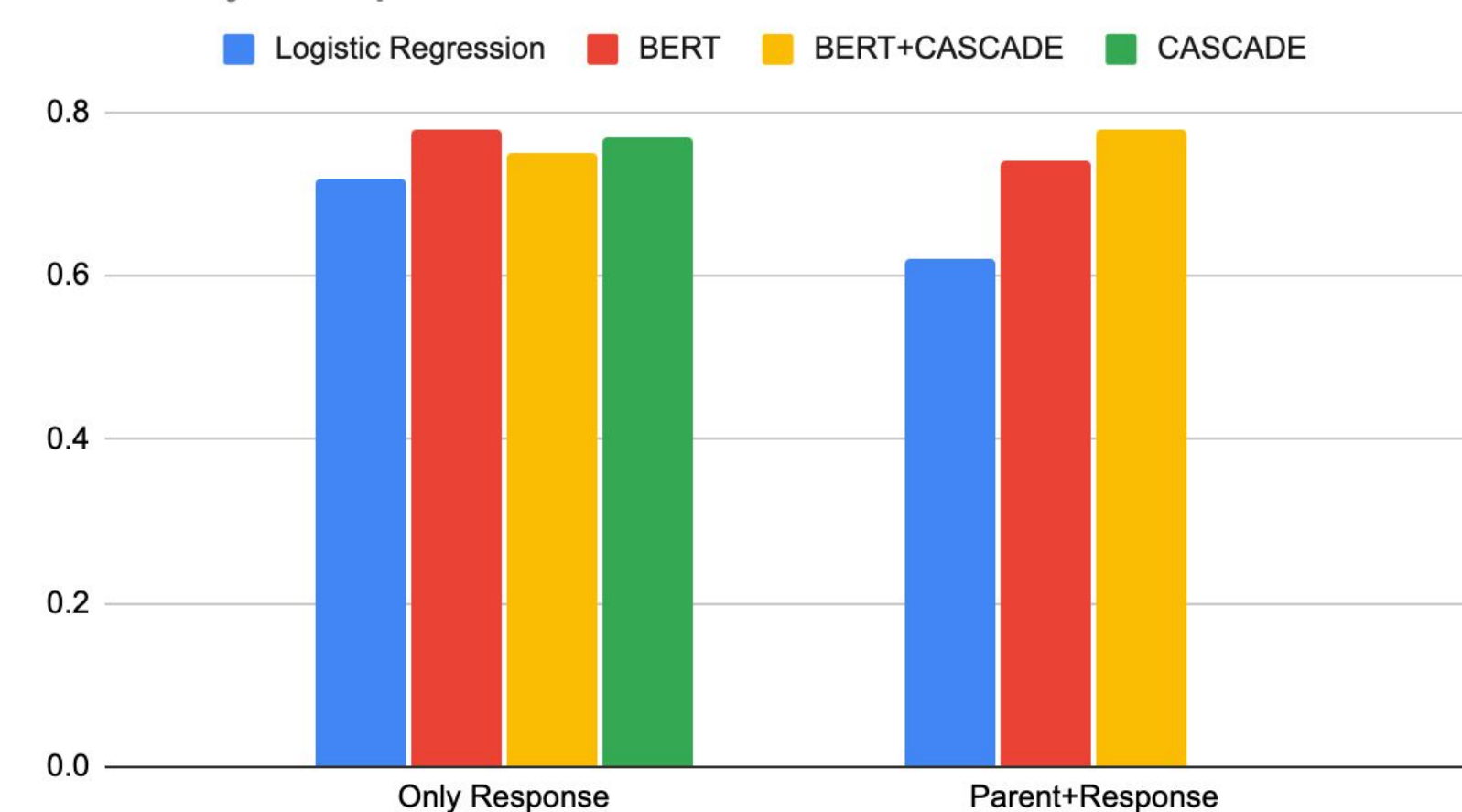
BERT+CASCADE Results

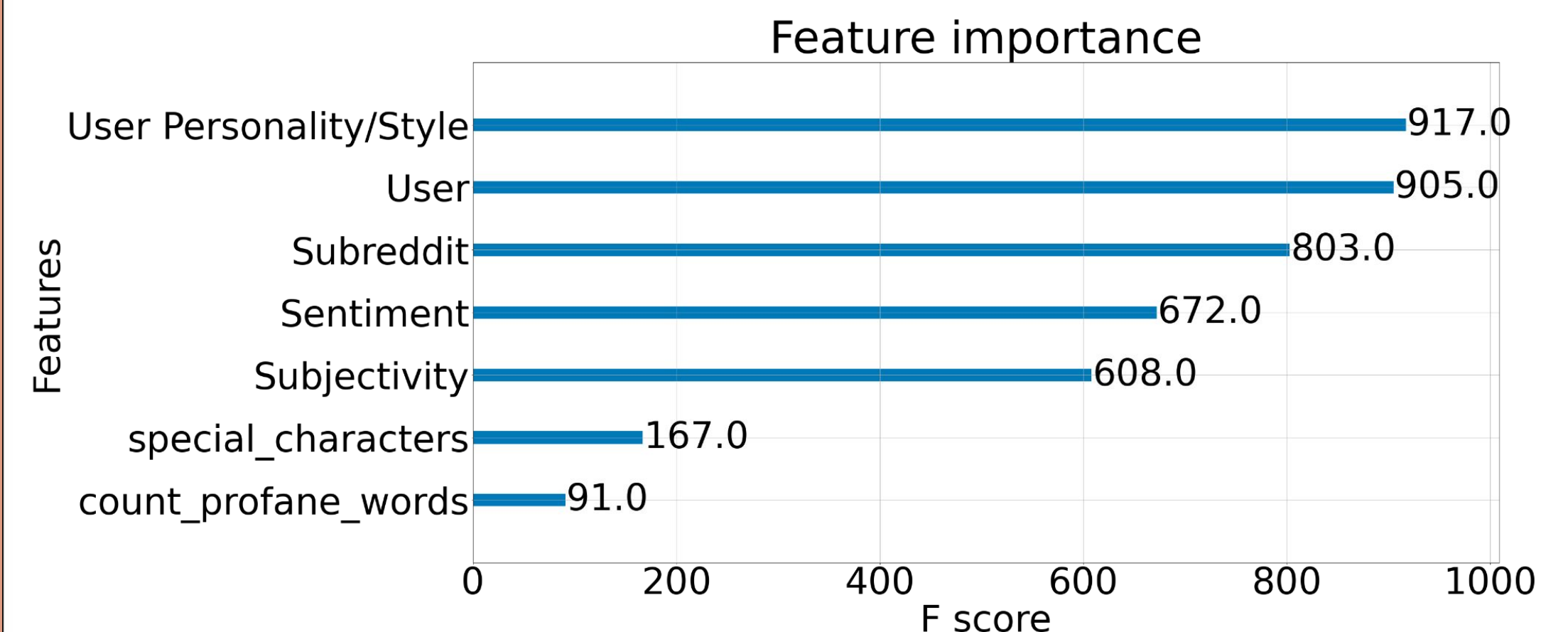Accuracy Comparison Across Models

BERT+CASCADE Results Discussion

- BERT+CASCADE produces better results on Parent+Comment results
- User info properly complements parent comment context

Feature Model Results

- We trained a XGBoost model to predict sarcasm utilizing context, background, and user related features
- User history and writing style are major related features
- 65% prediction accuracy for 7dim vector

Feature importance

Feature	F score
User Personality/Style	917.0
User	905.0
Subreddit	803.0
Sentiment	672.0
Subjectivity	608.0
special_characters	167.0
count_profane_words	91.0

Conclusion and Future Goals

Conclusion
- BERT+CASCADE provides some improvement
- Analyzed user and context related features that influence sarcasm
- User personality and subreddit showed to have the most influence in predicting sarcasm compared to standard context-related features

Future Goals
- Train regression models to generate longer dimension features for subjectivity, word/sentence length, to better capture their features
- Embed context and subreddit features into BERT+CASCADE to have it utilize contextual information

References

[1] 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.

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