LING 573: JEDi CLan

SARCASM DETECTION

Diana Abagyan, Libby Merchant, Emma McKibbin, Jade Phoreman, and Catherine Ball

TASK DESCRIPTIONS

SARCASM DETECTION TASKS

SARC:
Khodak et al. (2018)

- General subreddits
- Previous comments included for opt. Context
- Single sentence

PARTITION	COUNT
Training	231,374
Development	25,708
Test	64,666

BOTH:

- Binary labels:
  - 0: non-sarcastic
  - 1: sarcastic
- English data from Reddit
- "Self-annotated" using tone indicators (/s)
- Artificially balanced

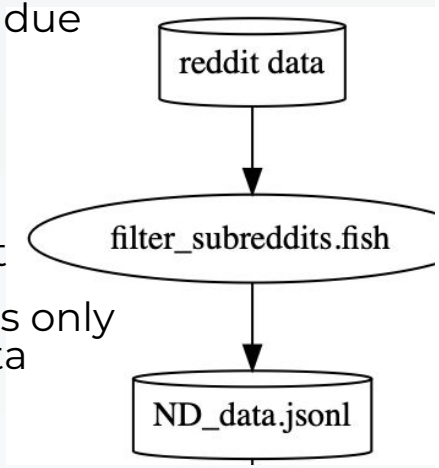
SAND:
Our Dataset

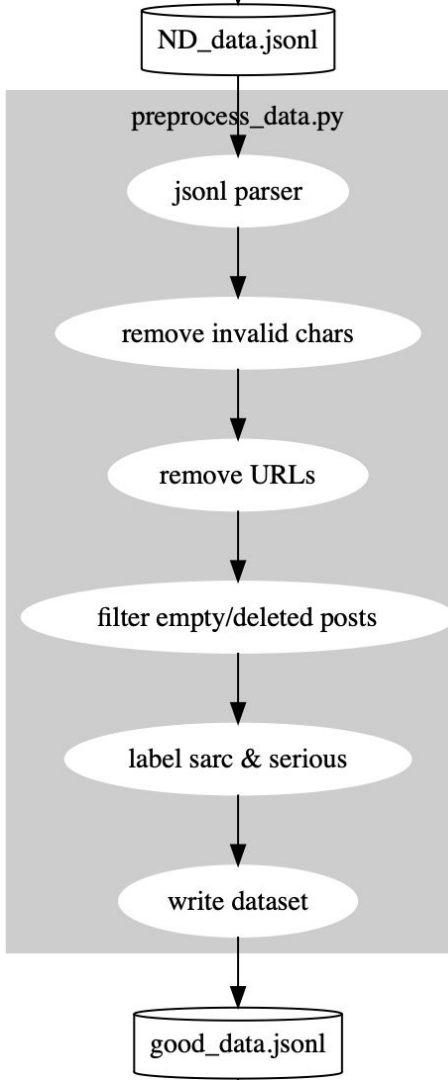
- Neurodivergent-related subreddits
- No context
- Full comment

PARTITION	COUNT
Training	337,782
Development	48,236
Test	48,186

ADAPTATION TASK:
DATA COLLECTION

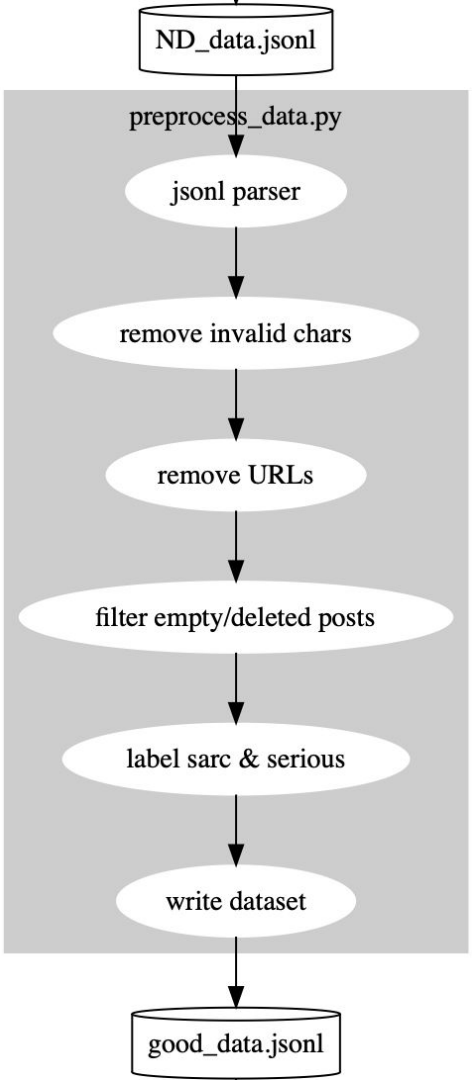
- Reddit data sourced from Pushshift dataset due to API and TOS limitations
- Pipeline approach to filter & preprocess
- Issues
  - Datatrove concurrency mangles output
  - Even with tests, many unforeseen issues only apparent when manually reviewing data
- Successes
  - Rerunning preprocessing from checkpoints with much faster than on whole reddit dataset

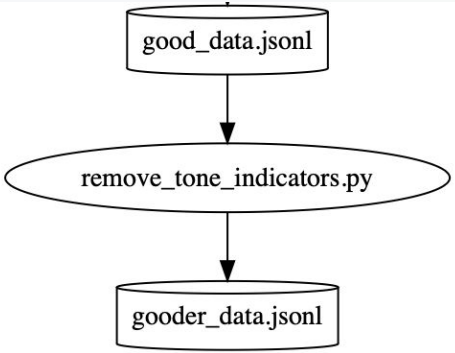




ADAPTATION TASK: PROCESSING DATA

- Included posts without tone markers from authors who had used tone markers in other posts, as examples of non-sarcasm
  - Posts without tone markers from authors who had never used tone markers were excluded
- Filtered out deleted comments, deleted authors, and bot messages
- After filtering, the final dataset included:
  - 878 instances with /serious or /srs tags
  - 246,474 instances with /s or /sarcastic tags
  - 246,474 instances without tone markers
  - We purposely balanced the number of sarcastic and unlabeled instances
- Partitioned gooder_data.jsonl into 80-10-10 (train-dev-test) split





SYSTEM REVISIONS & APPROACH

SYSTEM REVISIONS: OVERVIEW

D2 SYSTEM:

- BERT

D3 REVISIONS:

- BERT with context
- RoBERTa
- RoBERTa with context (revisions combined)

D4 REVISIONS:

- Data preprocessing
- Fine-tuning
- Ensemble

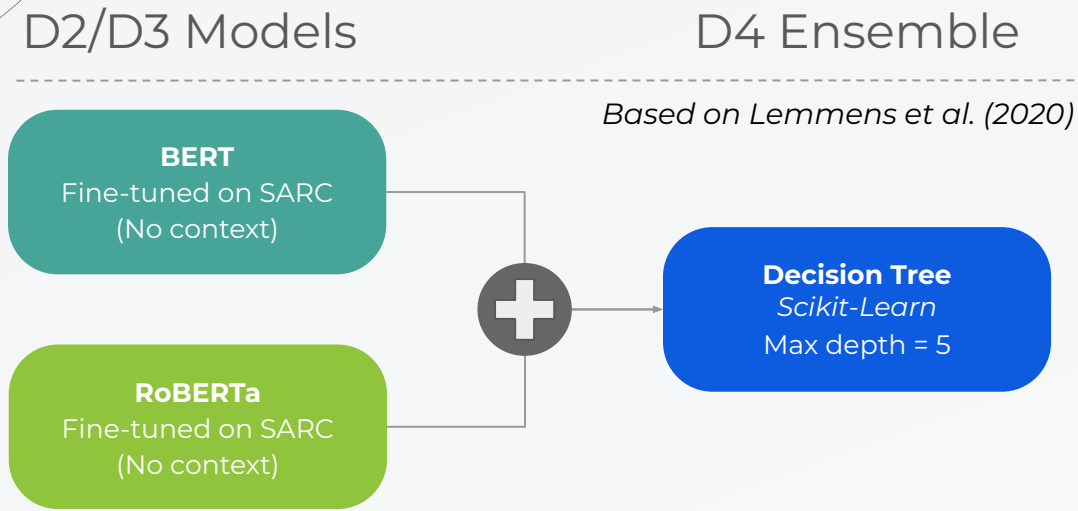
ADAPTATION

PRIMARY

REVISION 1: FINE-TUNING

- Took the best performing non-context model from D3 (RoBERTa finetuned on SARC)
- Finetuned further on SAND, for 2 epochs
- Dramatically improves performance on SAND data
- Truly catastrophic forgetting even after the first epoch, so we chose the first epoch as the final model.

REVISION 2: ENSEMBLE



- Predict on SARC training data using D2/D3 models
- Concatenate predictions
- Use as input to train decision tree classifier

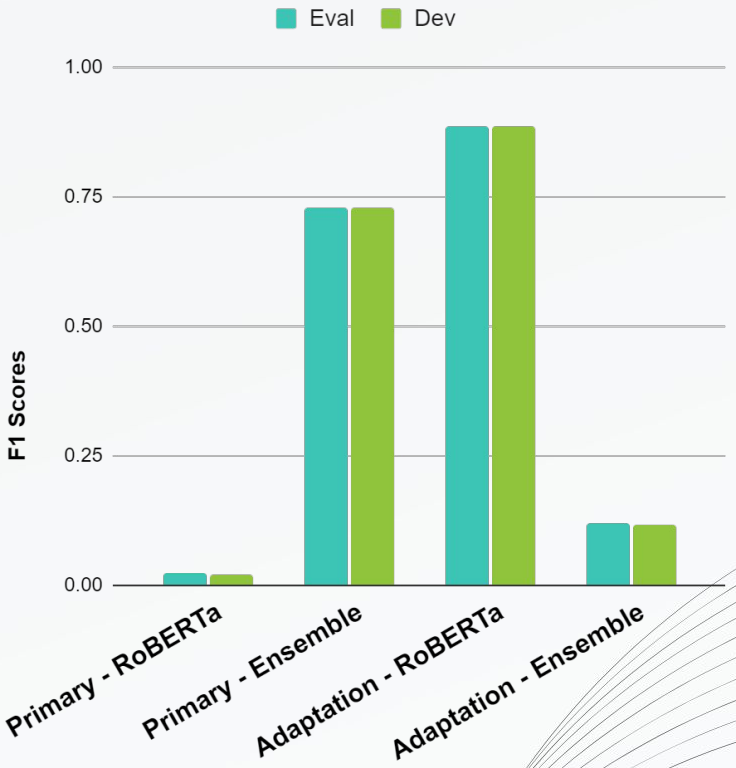
Generalizes well from SARC dev → SARC test

Does *not* generalize well to SAND...

RESULTS & ANALYSIS

RESULTS

	MODEL	F1-SCORE (eval/dev)	
BASELINE	Random	0.500	
PRIMARY	RoBERTa	0.0236	0.0186
	Ensemble	0.729	0.728
ADAPTATION	RoBERTa	0.885	0.884
	Ensemble	0.118	0.116



ISSUES & SUCCESSES

Dataset: SARC vs SAND

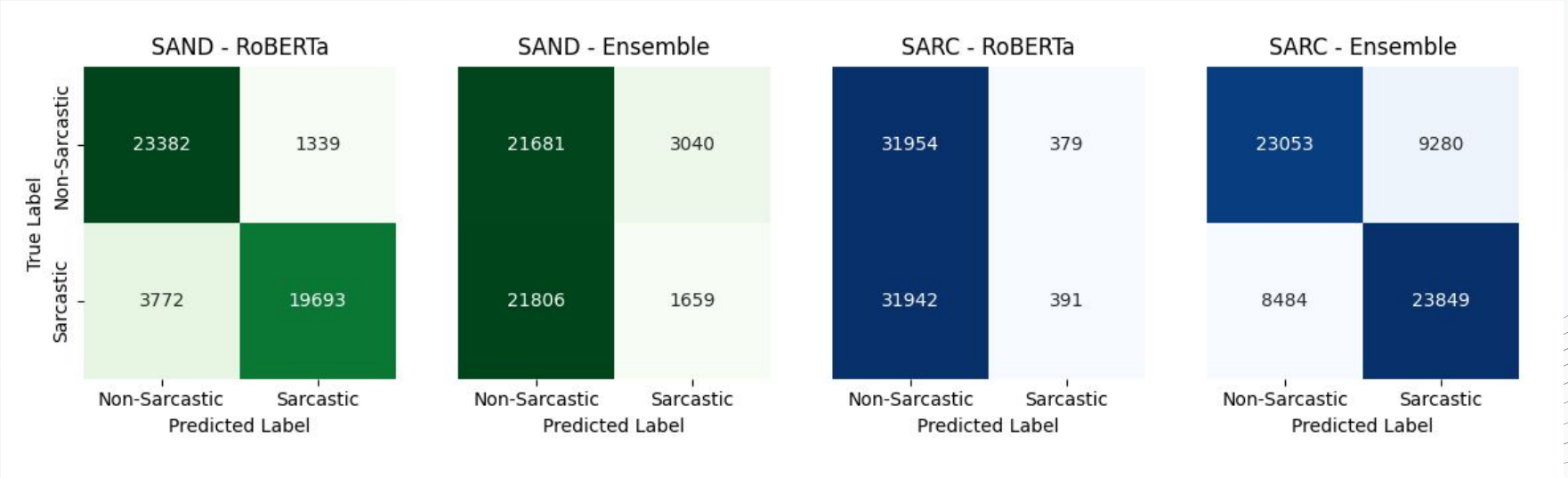
14x longer!

	SARC	**SAND**
Average text length (dev)	55	700
Max text length (dev)	**1114**	**25,183**
Sarcastic (train)	115,687	141,968
Non-sarcastic (train)	115,687	195,814

“Might want to bullet point what you need from us to get better responses, we do struggle with reading lengthy text.**”**

ISSUES & SUCCESSES

Confusion Matrix



FURTHER DIRECTIONS

- Dataset revisions
  - Include context in SAND
  - Try breaking up SAND data by sentence rather than whole post
- Model revisions
  - Finetune RoBERTa on SAND and SARC together
  - Include SAND-fine-tuned model in ensemble
  - Try models with longer max length

QUESTIONS?

BIBLIOGRAPHY

- Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm Detection Using an Ensemble Approach. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 264–269, Online. Association for Computational Linguistics.