LING 573: JEDi CLan

SARCASM DETECTION

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

TASK DESCRIPTIONS

PRIMARY TASK

Self-Annotated Reddit Corpus (SARC):
Khodak et al. (2018)

PARTITION	COUNT
Training	231,374
Development	25,708
SARC in Total	~533M

- English data from Reddit
- “Self-annotated” using tone indicators (/s)
  - More reliable than 3rd-party annotator
  - Still noisy
- Artificially balanced
  - Select a sarcastic (1) and non-sarcastic (0) response per thread
- Previous comments included for opt. Context
  - Variable length threads
- Created our own development set (random 10% of train)

PRIMARY TASK - SARC

SAMPLE RESPONSES WITH CONTEXT POST:

Original Post

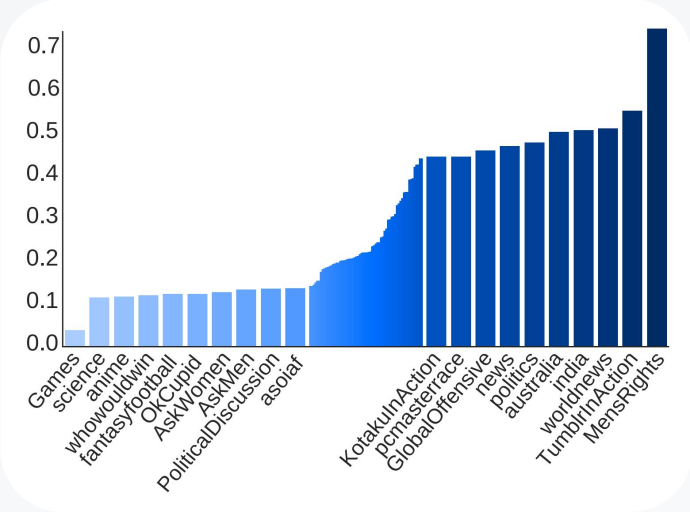
Braided Spaghetti Bread.

Non-Sarcastic Response (0)

Can I come over to your house, OP?

Sarcastic Response (1)

Needs more carbs.



% sarcasm in subreddits with >1M SARC comments *(Khodak et al., 2018)*

- Over 11.5K subreddits
- Also includes: time posted, author, subreddit, up- & down-vote counts

ADAPTATION TASK:
DATA COLLECTION

Goal: sarcasm data from ND population

- Gather data from neurodiverse populations using sarcastic tone indicators
  - Data is essentially pre-labeled for us where tone indicators are present
  - Reddit dataset (sort of) readily available
  - Neurodiverse subreddits allow us to (sort of) identify ND populations
- Assumptions
  - Posters in the selected subreddits self-identify neurodiverse
  - Tone indicators will be accurate and applicable to statements where tone indicators weren't used
- Caveats
  - Posters may not necessarily be neurodiverse
  - Posters have not explicitly consented to their data being used in this project
  - Reddit TOS violated

ADAPTATION TASK: GATHERING DATA

- Reddit API
  - Oops, it's limited to the most recent 1,500 posts for each subreddit
  - Reddit TOS requires special permission from their lawyers for any kind of ML training done on Reddit data or redistributing models trained by reddit data
- Offline Reddit dataset
  - Pushshift dataset of Reddit scrapes to date available from academictorrents.com
  - Doesn't contain all posts or subreddits, but large (~2 TB zstd compressed)
  - Contained posts from ND subreddits we were looking for

ADAPTATION TASK:
PROCESSING DATA

- Filter to subreddits we care about out of dataset
  - ~2.7 GB zstd compressed
- Filter to only posts containing sarcastic / serious tone indicators with regex
  - Tone indicators: /s, /sarcastic, /serious, /srs
  - Plus the above but using \
- Write output in same format as training data for later tuning

- Implementation notes
  - Piped data from compressed archives to filter script directly to avoid needing to write entire uncompressed dataset to disk; avoid extra disk IO operations
  - `rg` for search; memmap files while processing, teddy/aho-corasick for simd string matching

RELATED WORK

RELATED WORK

Neurodiversity & Sarcasm:
(Au-Yeung et al., 2015)
- ND people are less confident in interpreting sarcasm → tone indicators

Previous Sarcasm Detection Systems:
- Many studies use SVMs and/or ensemble models
- GloVe → BERT
- CASCADE – considers *user context:* models behavior of user & their preferences *(Hazarika et al., 2018)*

Context-Aware Sarcasm Detection Using BERT:
(Baruah et al., 2020)
- Last utterance context helps with Twitter data, not Reddit (non-SARC)

BERT vs. RoBERTa:
(Liu et al., 2019)
- RoBERTa claims to be > BERT, as BERT is “undertrained”
- Outperforms BERT on Stanford Sentiment Treebank

SYSTEM OVERVIEW & APPROACH

APPROACH

INITIAL MODEL:
- BERT, using only the response

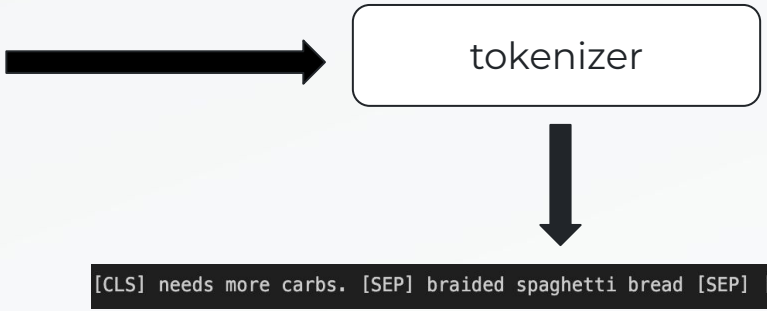
REVISIONS:
- BERT with context
- RoBERTa
- RoBERTa with context (revisions combined)

EVALUATION:
- F1 score

DATA PRE-PROCESSING

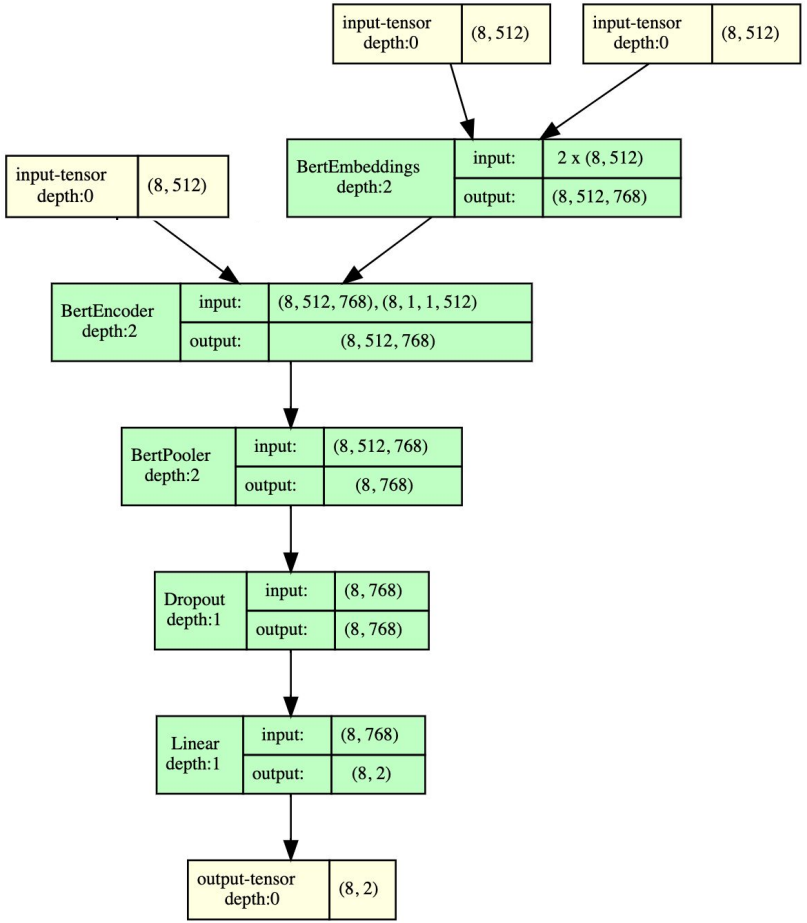
- CSV → JSON format (very large files)
- Context is concatenated after the post, divided by a separator token


```
{
  "posts": [
    "Braided Spaghetti Bread"
  ],
  "post_ids": [
    "yae7e"
  ],
  "context_size": 1,
  "response_id": "c5u2m60",
  "response": "Needs more carbs.",
  "label": "1"
}
```

MODEL ARCHITECTURE

- BERT or RoBERTa model
  - Embeddings
  - Encoder
  - Pooler layer
- Dropout
- Classification (linear) layer



TRAINING

- Trained classification layer and fine-tuned BERT model together
- To avoid catastrophic forgetting:
  - Low learning rate (2e-5) (Sun et al.)
  - Kept the checkpoint from the epoch with best eval F1 score
  - Train for 2-4 epochs only (Devlin et al.)
    - BERT models are from second epoch
    - RoBERTa: third epoch
    - RoBERTa + context: fourth epoch

RESULTS & ANALYSIS

RESULTS

	MODEL	F1-SCORE
BASELINE	**Random**	0.500
OUR MODELS	**BERT**	0.725
	RoBERTa	0.729
	BERT + Context	0.727
	RoBERTa + Context	**0.736**

ERROR ANALYSIS

But this is the *Best Health Care in The World*.

ISSUES

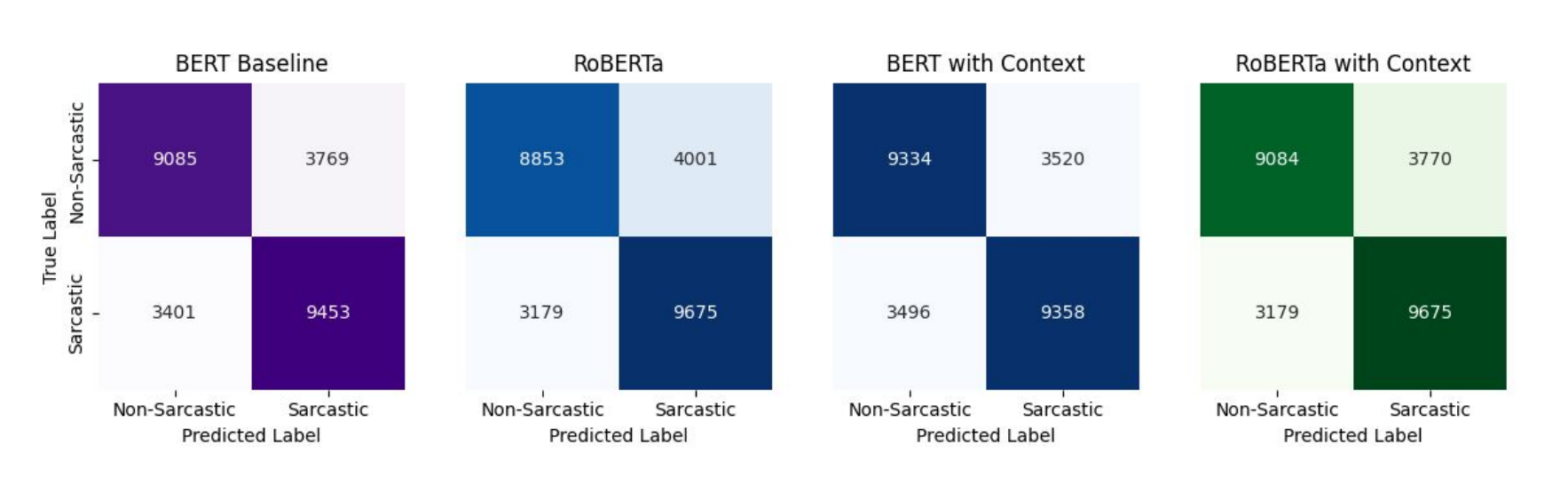
- Different label extraction for + and - class
- Obviously sarcastic instances labeled as not sarcastic
- Short, ambiguous phrases

SUCCESSES

- Fewer false - than false +, as we would expect

ERROR ANALYSIS

/s /srs? /gen? /hj?



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).
- Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.
- Sheena K. Au-Yeung, Johanna K. Kaakinen, Simon P. Liversedge, and Valerie Benson. 2015. Processing of written irony in autism spectrum disorder: An eye-movement study. *Autism Research*, 8(6):749–760.
- Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 83–87, Online. Association for Computational Linguistics.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification? *Preprint*, arXiv:1905.05583.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.