IS THIS ENOUGH SARCASM?

GROUP 20: JOE FARRINGTON, JEREMY DANG, BEN KLASMER

DEPARTMENT OF COMPUTER SCIENCE

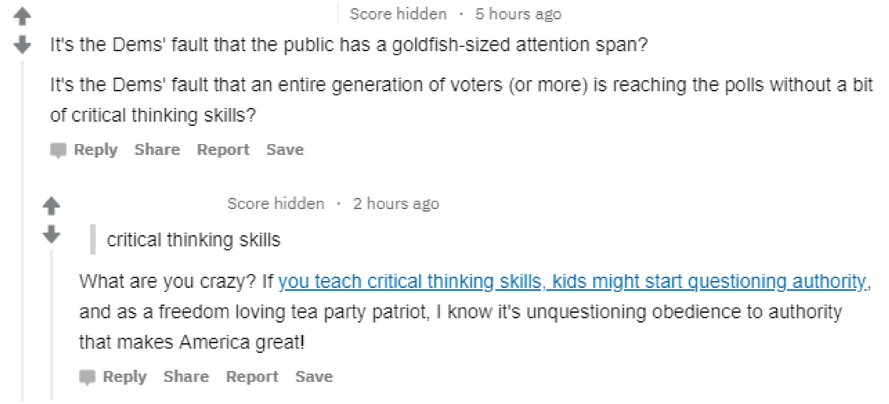
UCL

Why does it matter?

- Obtaining labelled data can be difficult and expensive. Labelling sarcastic data where there are no self-annotations, for example, is difficult because the people labelling the data may not identify the sarcastic intent.
- Establishing an adequate amount of training data which can be used to train a model may therefore prevent wasted effort labelling additional data that makes little difference to model performance, and allow effort to be used ensuring that the labels are of a high quality.

The SARC Dataset

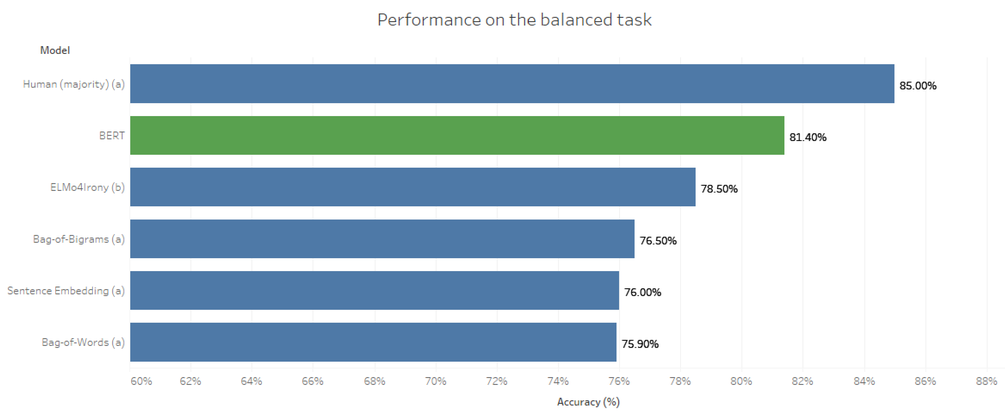
- We used the balanced version of the /pol subset of the SARC 2.0 dataset which consists of 'ancestor' comments and a pair of responses for each: one labelled as sarcastic and one labelled as not-sarcastic.
- **Ancestor: "Jindal being super smug about volcano monitoring"**
  - *Sarcastic response: "Who ever heard of a volcano causing problems for people?"*
  - *Non-sarcastic response: "you'd think someone from Louisiana, of all places, would understand the importance of early warnings for natural disasters"*
- To assess the impact of reducing the number of training examples, we randomly sampled (without replacement) subsets consisting of 50%, 25%, 12.5% and 6.25% of the ancestors in the project training set and their corresponding responses.

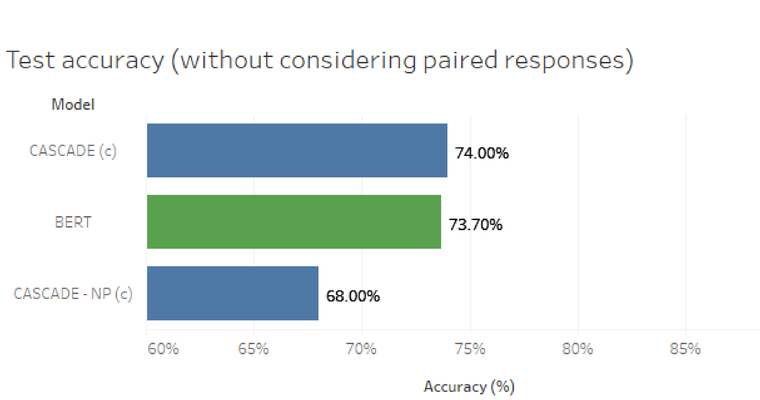


What we've shown

- Using pre-trained BERT model and fine tuning on the SARC /pol dataset, we have outperformed current state of the art results on models that don't include personality embeddings. Furthermore we've achieved competitive results compared to CASCADE which uses personality embeddings.
- Moreover, we have shown that BERT broadly performs as well as the baseline models with a quarter of the amount of training data. Additionally, as we reduce the number of training examples, the performance of baseline models decreases at a quicker rate than BERT.

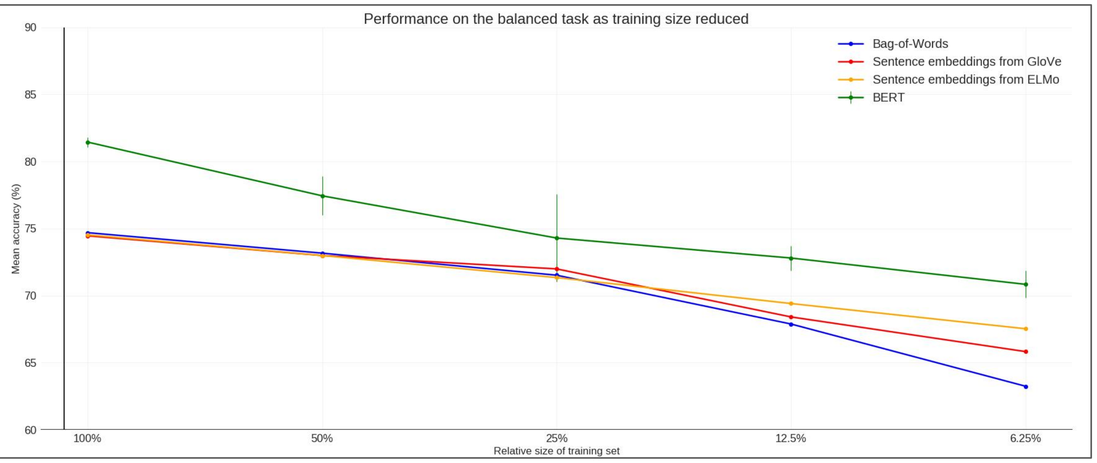
Comparison to previous work





Model Comparisons

- BERT based on Devlin et al. (2018)
- Logistic Regression applied to:
  - Bag-of-Words
  - Sentence embeddings from GLoVe
  - Sentence embeddings from ELMo
- BERT on average performs better than the other models but shows significant variability on 25% reduction of the dataset.



Future work

- Investigate whether BERT performance can be improved by treating the task as a sentence-pair task and testing whether the response appears to be a sarcastic response to the ancestor.
- Investigate the effect of variation in the content of the training sets by drawing different sized samples from the much larger, main SARC dataset while using models with the same initialized weights on each repeat.
- Perform a similar analysis on other tasks in NLP where getting good quality labelled examples is challenging.

Key: (a) – Khodak et al. (2017) (b) – Ilic et al. (2018) (c) – Hazarika et al. (2018) NP – No personality embedding

Reddit source from comments: https://www.reddit.com/r/politics/comments/b0be47/trump_said_he_wouldnt_cut_medicaid_social/