Experimental Protocol for "TalkDown: A Corpus for Condescension Detection in Context"

Zijian Wang
Stanford University
zijwang@stanford.edu

1 Hypotheses

Our main hypothesis is context is decisive for condescension detection. Besides, we believe *slightly* oversampling the training set could yield good performance for very imbalanced task like condescension detection.

2 Data

We heuristically identified a collection of 66K comment-reply pairs, where 1) there is a condescending-related word in the reply, and 2) the reply contains a direct quotation from the comment (henceforcce, *quoted*), on Reddit from 2006 to 2018. We then crowdsourced a subset of the collection and obtained 4,992 valid labeled instances of such pairs using Amazon Mechanical Turk. Out of the valid instances, 65.2% labeled as *condescending* (henceforth, *positive*), and 34.8% as *non-condescending* (henceforth, *negative*).

To fully balance the dataset, we pulled out one random month's data for each year in 2011 to 2017. We extracted instances using the same methods as described above, but we filtered out comment-reply pairs in which a condescension-related word appeared.

Our final dataset thus consists of annotated positive and negative instances, with supplemental randomly-sampled negative instances. For our experiments, we partitioned the data into 80% train, 10% development, and 10% test splits. In addition, to simulate real world situations, we built a dataset with a 1:20 ratio of positive to negative instances.

3 Metrics

We use Macro-F1 score[1] on the test set as our evaluation metric. We evaluate our model on both the balanced setting and the 1:20 imbalanced setting.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

4 Models

We use the BERT model of Devlin et al. (2018), which uses a Transformer-based encoder architecture (Vaswani et al., 2017) to learn word representations by training against a masked language modeling task and a next-sentence prediction task. Our models are initialized with the pretrained representations released by the BERT team and a fully connected layer on the top (Figure 3 in Devlin et al. 2018), which are then fine-tuned to our dataset (Peters et al., 2019). We explore both BERT Base (BERT_B) and BERT Large (BERT_L), to determine whether the added expense of using BERT_L is justified.

5 General Reasoning

BERT has achieved state-of-the-art results for a variety of NLP tasks, including sentence classification (Devlin et al., 2018). It is suitable for our task because it could take one or two inputs without modifying the model architecture. Using BERT, we could easily test whether context matters and find the general strategy to deal with imbalanced dataset. The expected result would be 1) adding context would increase the performance compared to using the quoted part only, and 2) oversampling a small number of times, i.e., in our task $1 < n << 20$, would yield good performance when testing on the imbalanced dataset, which is crucial because it is close to the ratio of condescending in the real world. Besides these two main expectations, we also hope 1) increasing the capacity of the model could yield better performance, and 2) our model could operate well on unseen real world dataset, i.e., whether it could give reasonable predictions for comments in different subreddits/topics.

6 Summary of Progress

We have performed experiments on 1) the effect of including context as input, 2) the effect of increasing the capacity of the model, 3) the effect of applying different oversampling ratios, and 4) predicting condescension in unseen real world data. Our main findings are coherent to the expected results discussed in the reasoning part. More experiments and analysis could be done with 1) whether subreddit matters, i.e., could the same post be condescending in one subreddit but not the other, and 2) error analysis on the predictions of the real world experiments.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.