Readings

This week's weekly reading, "Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations" (Bao et al. 2021) has a lot of overlap with our group's research goals. That paper examines prosocial outcomes and behaviours generally, with the expression of gratitude being an example of a prosocial behaviour. I've taken on the view that the expression of gratitude can itself be a prosocial/generous act, which presumably often arises from the feeling of gratitude, but not necessarily, and that a person feels gratitude does not diminish the moral virtue of expressing that gratitude. In our research we have been aiming to measure the expression of gratitude, primarily as a proxy for the feeling of gratitude, but I now realise that the act of expressing gratitude is also interesting for our study.

From Bao et al.'s (2021) paper I discovered that the BERT model has a special token, CLS, whose embedding is useful for classifying entire sentences. I was previously considering a way to construct meta-embeddings, by averaging the embeddings, but I think we should use the CLS embedding instead. If time permits I will check the performance using CLS against the mean meta-embeddings.

We should adopt the TLC (Top Level Comment) term used by Bao et al. (2021), for consistency with the literature. What we have been calling a submission, they call a post, what we have been calling a comment at depth 1, they call a TLC.

Bao et al. (2021), like our group, study a large corpus of Reddit discussions, but their corpus is isolated to a single year. They account for the presence of automated posts by bots, which we should also account for in our analysis. Their approach is to exclude TLCs with more than 3500 words, because they found by manual inspection that thit criterion captures a large number of bot posts. We must also exclude comments and posts that have been deleted by the user or removed by moderators. It's also interesting to consider the potential bias introduced by deletion and removal of posts and comments. Perhaps comments and posts expressing more severe animosity are more likely to be removed or deleted.

From the above paper I came across the paper "It's going to be okay: Measuring Access to Support in Online Communities" (Wang and Jurgens 2018). The methodology for measuring supportiveness in this work is similar to our approach to measuring thankfulness and animosity. Wang and Jurgens (2018) collect manual annotation of supportiveness, but to isolate supportiveness from say politeness, they also ask annotators for judgements of politeness. Wang and Jurgens (2018) use a five-point Likert scale, whereas we use binary annotation.

We should report the Krippendorf's α of our annotations.

Processing and Analysis

I have now finished contriving 20 example sentences and annotating 100 sentences manually. When the rest of the sentences are ready for annotation I will annotate them.

I have generated a list of unique submission names from the raw dataset. I sampled 10200 names from this and extracted all submissions and comments related to these 10200 names (about 1 hour of compute). Many of these will be either deleted or removed. Pranav has selected 106 TLCs to be used for manual annotation.

We can now iterate through each submission and quickly load associated comments. We need to decide on an algorithm for sampling the comments.

References

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring Access to Support in Online Communities. In Proceedings of EMNLP.

Bao, Jiajun, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. "Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations." In Proceedings of the Web Conference 2021, pp. 1134-1145. 2021.