An analysis of productivity on Reddit

Methods

We leverage the Huggingface Datasets and Transformers Trainer natural language model training pipeline. We choose a single model to leverage, `microsoft/deberta-v3-large` because it is the best performing large language model (LLM) on the SuperGLUE benchmark for natural language understanding and because it's the largest model that we can feasibly use on 12GB-16GB GPU RAM Google Colab GPUs (unlike massive language models like Gopher or Google's recently released PaLM).

We use a training batch size of 16, 16-bit floating point precision leveraging PyTorch's native AMP, a learning rate warmup ratio of 0.1 of the total number of optimization steps, and a total of 25 training passes (epochs) through the data. We use exact match accuracy as our model selection metric instead of loss to avoid the model overfitting and becoming more confident with lower loss.

Our dataset was heavily imbalanced with the majority label consisting of around 50% of all of our datapoints and the minority label consisteing of less than 1%. As a result, we leveraged SKLearn's `compute_class_weights` function to inversely weight the loss of samples in the training phase. Intuitively, the larger a class, the less it's samples' training losses would be weighted, and vice versa for smaller classes. We reformulated our multiclass classification problem into a multiple binary classification problems leveraging crossentropy loss.

Our data consisted of three separate sequences: the title of the Reddit post, the parent comment, and the child comment. Since the model tokenizer is only able to receive at most pairs of sequences, we concatenate the title and parent comment into a single sequence. This aligns with our notion of "parent context" exactly, where the title and parent comment provide the context to evaluate the authorial response intent of the child comment.

For hyperparameter configurations, we use a weight decay value of `1e-1` and tune over the learning rate values of `1e-5`, `2e-5`, and `3e-5` similar to the pattern in Liu et al. 2019.

Results

The learning rate value of `3e-5` performed the best on the dev set with a dev set accuracy of `0.606`. The best model yielded at test set exact match accuracy of `0.588`, with a 95% confidence interval of `[0.575, 0.703]`, which is significant compared to the baseline BOW L2-regularization logistic regression classifier exact match accuracy of `0.53` (and the majority class baseline of 0.46).

Analysis

Overall, the model is able to correctly identify authorial intents, but struggles to leverage the post title and parent comment to identify the correct context to frame the intention. In other words, the model lacks the common sense to accurately determine the authorial intent.

The confusion matrix for the best model on the dev set by exact match accuracy evaluated on the test set is shown in the code output above.

Arbitrate

There were a total of 9 (4.17%) datapoints in the test set with a true authorial intent label of "Arbitrate", with a model accuracy of 4/9 (44.44%). An interesting failure case to analyze when an "Arbitrate" authorial intent datapoint was misclassified as the "Inform" authorial intent, signaling that the model found some commonality between the two intent types. Indeed, given the following failure case:

Feature name	**Text**
Article title	Ukraine joins European power grid, ending its dependence on Russia
Parent comment	Ukraine joins European power grid, ending its dependence on Russia
Child comment	Hey /u/Picture-unrelated, This is now the top post on reddit. It will be recorded at /r/topofreddit with all the other top posts.

The child comment is providing information, but the information isn't related to the content of the post; rather, it's related to the metadata about the post i.e. that it will be recorded in "/r/topofreddit".

The model is able to ascertain that intent of the child comment is partially informative, but it isn't able to distinguish between whether the information is related to the content of the post or the metadata of the post.

Argue

There were a total of 34 (15.74%) datapoints in the test set with a true authorial intent label of "Argue", with a model accuracy of 23/34 (67.65%). 4/34 (12.5%) of "Argue" authorial intent datapoints were misclassified as "Rhetorical". An example of a failure case is as follows:

Feature name	**Text**
Article title	'Price gouging from Covid': student ebooks costing up to 500% more than in print - Call for inquiry into academic publishers as locked-down students unable to access study material online
Parent comment	'Price gouging from Covid': student ebooks costing up to 500% more than in print - Call for inquiry into academic publishers as locked-down students unable to access study material online
Child comment	And that’s why I don’t feel bad pirating as many books as I can. Fuck the required online workbook apps connected to some textbooks though. Scams within

Feature name	Text
	scams.

The child comment cites ("that's why...") the parent comment as evidence for their point that the author "[doesn't] feel bad pirating as many books as I can." Although to human annotators, we can recognize how the phrase "And that's why" means the author is using the parent context as the basis for their argument, perhaps the model focuses on the strong language, "Fuck the ..." (or perhaps perceives the author's comment to be overwhelmingly blunt/direct in its argument).

Connect

There were a total of 22 (10.19%) datapoints in the test set with a true authorial intent label of "Connect", with a model accuracy of 8/22 (31.82%). 8/22 (36.36%) of "Connect" authorial intent datapoints were misclassified as "Rhetorical". An example of a failure case is as follows:

Feature name	Text
Article title	"Churchill's grandson slams Trump for skipping cemetery visit because of weather: ""They died with their face to the foe and that pathetic inadequate @realDonaldTrump couldn’t even defy the weather to pay his respects to The Fallen,"" Soames tweeted
Parent comment	Churchill's grandson slams Trump for skipping cemetery visit because of weather: ""They died with their face to the foe and that pathetic inadequate @realDonaldTrump couldn’t even defy the weather to pay his respects to The Fallen,"" Soames tweeted
Child comment	Again, Imagine Obama. Not just what he would have done in this situation, but what would have happened if he'd acted like *this*."

The child comment draws parallels between Trump's behavior and Obama's behavior. While the comment does make a rhetorical comment about behavior i.e. "... he'd acted like *this*.", it pertains only to Obama's behavior, not Trump's behavior.

The model is able to recognize that the intent given the connection reference to Obama is rhetorical (based on the exaggerated reference), but it's not able to recognize that the comment is topical towards Obama specifically, either because it was not able to coresolve the referent of "he'd" or because the model wasn't able to understand that a connection to new information was made. (The lack of context regarding the characteristics, policies, etc. of Obama's presidency versus Trump's presidency may mean the model can't recognize the weight of this connecting statement.)

Inform

There were a total of 34 (15.74%) datapoints in the test set with a true authorial intent label of "Inform", with a model accuracy of 13/34 (38.24%). The model wrongly predicted "Argue" instead of the correct label of "Inform" 9/34 times. Looking through the datapoints, reflecting on our own experiences as annotators, and reading over our peers' feedback on AP2, we notice that sometimes there are "Inform" comments where the author's "voice" may be more strongly reflected in their comment, interweaving their opinions. Due to the nature of the Reddit platform and how discussions on the r/worldnews subreddit may lean more towards a casual (i.e. non-academic) tone, comments with the "Inform" intent may have been more difficult to correctly identify (and thus get confused or mislabeled with "Argue" and other labels) due to this mix of a casual tone with new information.

Inquire

There were a total of 15 (6.94%) datapoints in the test set with a true authorial intent label of "Inquire", with a model accuracy of 3/15 (20.00%). There was no majority classification label of the datapoints in the test set with true authorial intent label of "Inquire". As such, the model did not learn this class well, if at all, as performance is close to chance.

Rhetorical

The majority of datapoints were in this class, accounting for 100/216 (46.29%) of the test set datapoints. The model largely classified these datapoints correctly, with an exact match accuracy on the test set of 81/100 (81%). We guess that due to the nature of r/worldnews (and Reddit's social media and public forum platform), the space possibly leans more towards casual conversations, which overshadow and affect possibly the majority of comments. With this in mind, this category may be creating a "noisy" bias that prevents the model from recognizing/labeling other labels.

Summarize

There were a total of 2 (0.93%) datapoints in the test set with a true authorial intent label of "Summarize". The model had a 0% accuracy rate on the test set "Summarize" datapoints. As that is a very small proportion of data in the overall datapoint distribution, the model did not learn this class well, if at all. There are not enough datapoints in the test set to tell whether the model performed better than chance.

Reflection, Combined with Analysis

Given that the model we used does not have clear feature weights, it is difficult to see what exact features of our data our model was learning.

Looking at the confusion matrix, we saw that Argue and Connect are both confused with Rhetorical, and vice versa. This could be a result of annotation guidelines that are not clear enough; annotators noted that our annotation guidelines were at times ambiguous when distinguishing between these labels: one annotator noted that “if it was a general comment or emotion, the boundaries of ‘connect’ vs ‘rhetorical’ often blended, or ‘connect’ vs ‘inform’ was another that blended based on the circumstance.” Another annotator noted that “one difficulty I found in annotating the data was deciding between the category of ‘Argue’ and ‘Rhetorical’ because of the difficulty in determining whether something was opinionated enough to be an argument rather than just a statement.”

We can see that the model struggles when considering comments with sarcasm, as these comments are often found in parent context/comment pairs with the label ‘Rhetorical’; however, sarcasm can also be found in other labels, and in this regard the model struggles to look beyond sarcasm to see authorial intent.

Our dataset overwhelmingly consisted of comment/parent pairs with Rhetorical labels, reflective of the sarcastic tone most often used on Reddit that is meant to be funny rather than provide meaningful discourse. Thus, our dataset would be a good candidate for oversampling (or undersampling the majority class of ‘Rhetorical’), as doing so would likely help our model learn more features of comment/parent pairs with true labels that are not ‘Rhetorical.’ We addressed this problem in our model by changing class weights to assign less weight to this majority class while training the model.