Evaluating NLP Systems

Group (Comp E Kids): Hamza Khan, Zeb Zimmer, Ben Moorlach
Github Repository: https://github.com/ZebZimmer/NLP

Step 1:

The model we chose was our Homework 1 fine-tuned model from RoBERTa base. The training was on the Stanford Natural Language Inference (SNLI) dataset. The task was to use the fine-tuned model as a text classifier on the SNLI test set.

Step 2:

The spreadsheet is linked in the submission. In the spreadsheet is an extra credit attempt in the form of Spacy analysis.

Step 3:

The spreadsheet has the desired columns. The main solution to fix errors was determined to be pre-processing. If the model is able to understand entities, and the relationship between them, it would help round out errors. Pre-processing is suggested as a counter strategy to simply requesting more data. We hypothesized pre-processing to be a reasonable solution due to the main error type being found to be a 'World Knowledge Gap'. Therefore, it would be beneficial to create connections within the data to bridge the gap. A possible way to implement this would be the creation of new tokens that represent a group of objects that mean roughly the same thing i.e. a group of people can be known as a crowd.

We attempted to create unique and specific error types and solutions for the bonus point and to better understand the issues the model faced. The below graphs show our understanding across the 49 incorrect prompts. It's interesting to note that the errors of the model follow a pattern where ~50% of the errors were on neutral prompts and the other ~50% of the errors were on the positive and negative prompts. This shows that the model model has some kind of linear output space where negative and positive are on opposite sides of the space.

NOTE: The spreadsheet has quasi-inverted confidence values as noted in the header: for only the spreadsheet a 1.0 is low confidence and 0.0 is high.

Potential solutions to fix errors	#	%
Pre-processing with entity recognition	19	38.77
Increase the size of the training data	11	22.45
Ground_Truth may be incorrect	7	14.29
Pre-processing with verb understanding	6	12.24
? Difficult Prompt	3	6.12
Pre-processing of ambiguous words	1	2.04
Pre-processing with verb tense understanding	1	2.04
Pre-processing with adjective recognition	1	2.04

Error type (distribution)	**#**	**%**
False neutral	24	48.98
False entailment	14	28.57
False contradiction	12	24.49

Error types for term prediction	**#**	%
World Knowledge Gap	12	24.49
Lexical Overlap Bias	8	16.33
Lack of Context	6	12.24
Unwarranted Assumption	6	12.24
Inadequate Spatial Reasoning	5	10.20
Ignoring Implicit Context	4	8.16
Ambiguity	3	6.12
Inadequate Emotional Intelligence	2	4.08
Inadequate Numerical Reasoning	2	4.08
Temporal Knowledge Gap	1	2.04

Step 4:

We spoke at length in the slack with Deb about the graphs below. Essentially we took the output of the last hidden layers, of which there were 12, and that layer is Nx768 where N is the number of input tokens. Figure 1 uses only the first token because it was discussed in class once that the first token (which is a special token like [START]) tends to contain a lot of information about the sentence. We then used the first 10 tokens to compare. It can be seen in figure 2 that the Resulting space is relatively the same with distinct groups for the 3 categories which we are happy with. The errors are all of the black and faded black (gray) dots which are somewhat favored to be nearer to the origin than correct predictions. Each dot's opacity is tied to its confidence value (dark is more confident). There are ~400 correct data points and 50 incorrect.

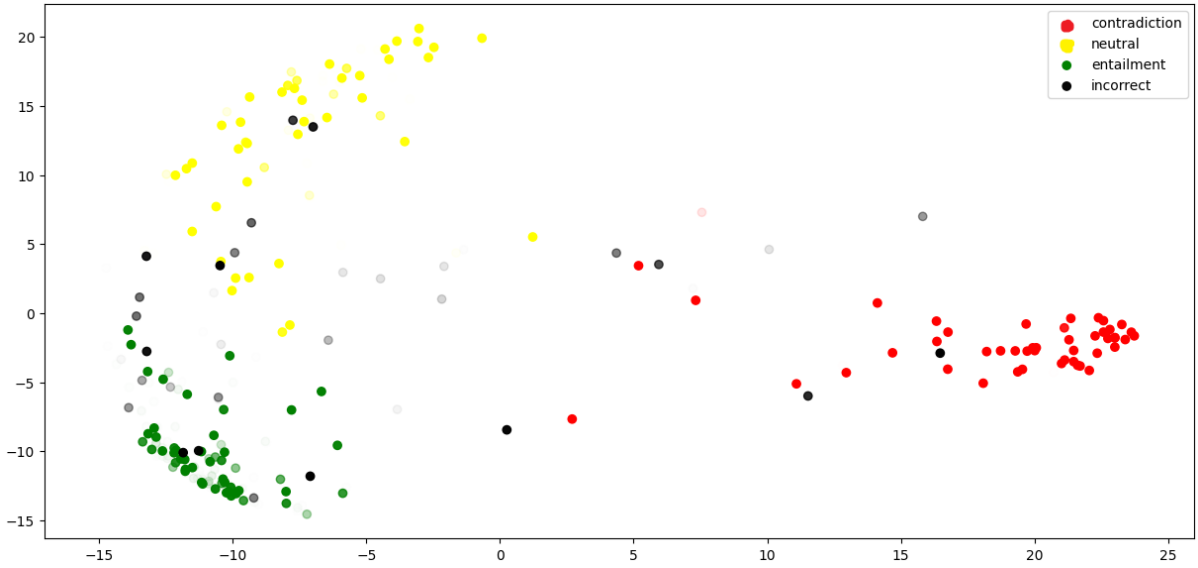

Figure 1 (First token)

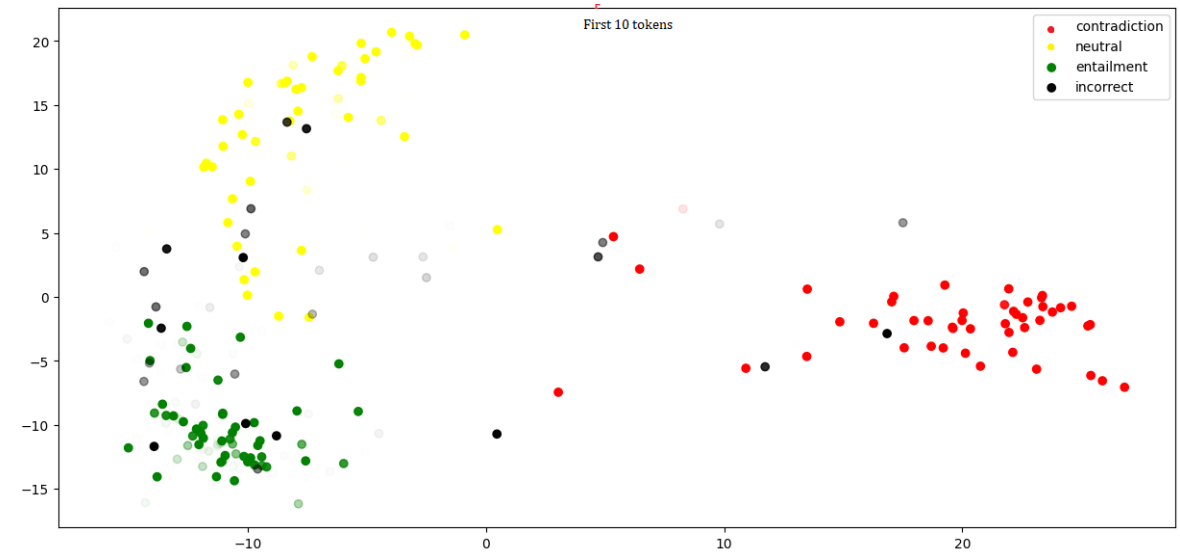

Figure 2 (First 10 tokens)

Step 5:

The work done in Homework 5 analyzed the errors found in the outputs from our fine-tuned RoBERTa model from Homework 1. The RoBERTa model was fine tuned using natural language inference (NLI) tasks. It took ~490 samples to gather the 49 error samples shown in the spreadsheet. The model's main limitation is that it struggles to understand entities and how an object can be described in different ways i.e. a group of people could be described as a crowd. A proposed future direction is to pre-process the text to create an entity object (maybe a token in the tokenizer) that learns how different nouns can be represented in new ways like the example just given. Another serious limitation is biases that became prevalent in the errors. Mainly underrepresented groups that are not being understood. A decent example of this is when the model assumed "little people" meant children. Biases are hard to root out as they tend to be a result of the training data. Increasing the size of the training data will always help but there's an extra caveat that the new samples need to be diverse. Data size increases generally always improve machine learning but are not always possible. That is why pre-processing the training data to learn even more context is likely the best option to improve a model such as ours.