Bryan Mangosing

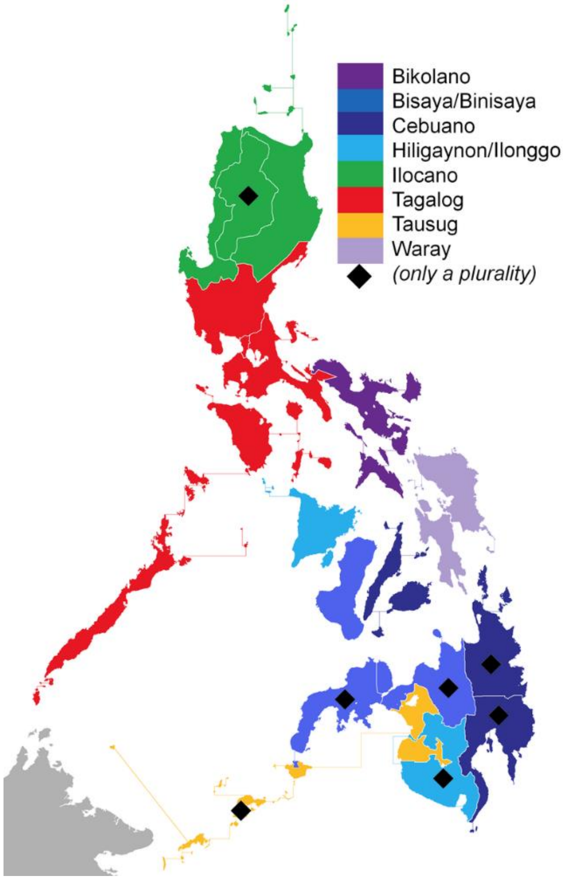
Ling 583

Prof. Malouf

Applying Transfer Learning to Low-Resource Philippine Languages

Introduction:

The Philippines is a very linguistically diverse country said to contain over 120 languages. Filipino, the national language, is based on the Tagalog language, but even then, other languages are more prevalent depending on the region of the Philippines a person comes from. Despite how numerous the languages are in the Philippines, on the scale of the languages of the world, many of the languages does not compare to the size of speakers of the world's largest languages. All of the native languages of the Philippines are considered low-resource languages since there is not as much text data available as there is to English for example. Even languages such as Tagalog do not have as many resources despite it and Filipino being the more dominant languages in the country.



Methods:

For my project, I decided to try using a pre-trained model and then using it on different on a dataset containing different languages of the Philippines. The six languages that are included in this data set are Tagalog, Cebuano, Ilocano/Ilokano, Bikol/Bilokano, Hiligaynon/Ilonggo, and Waray. The text data itself comes from the PALITO Corpus, which unfortunately is no longer active, but one of the researchers uploaded the text data onto GitHub and can be downloaded from there (https://github.com/imperialite/Philippine-Languages-Online-Corpora/tree/master/PALITO%20Corpus). The PALITO Corpus contains two sets of text data, literary text and religious text, for each of the 8 languages that are in there. For my project, I took 6 of the languages and both types of text to create a dataset. With each language, the text entry is either marked as "literature" or "religious" and then also considering the language, giving us 12 features. I decided not to use two of the languages for two main reasons. The first one being that the accuracy diminished each time another language was added. Another reason was because it would have been 8 languages to look at and the confusion matrix would have been difficult to read since there would have been 16 language variants.

After I downloaded the data from the corpus, I had to do a bit of text processing as well as creating the actual dataset. When I looked through the contents of the different text in the corpus, I noticed that there were a bunch of HTML tags in it, so I had to make sure I removed them. Although I am not sure what effect it would have had, I thought it would just be better to remove it and keep the text data a bit cleaner. After that, I combined the 12 different datasets into one which I would use the model on. In addition, for the main dataset, they were split into a training and test set, but I only used a portion of the training data. When I was playing around with the size of how much of the training set to include, the more I added, the higher the accuracy and macro average became. I ended up settling on just 5000 entries of the training set, which in total contains 40446 entries.

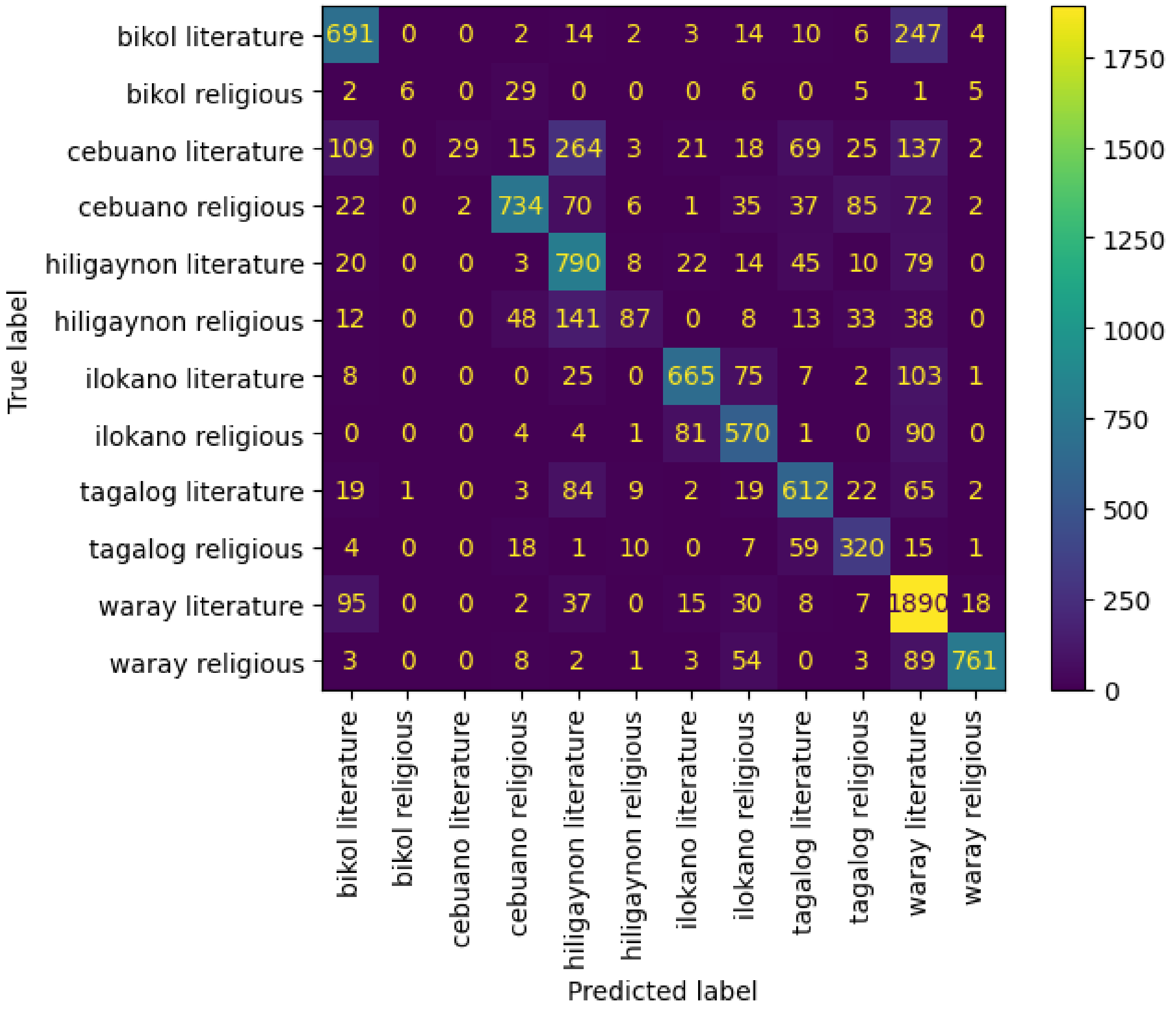
For the actual model, I decided to go with DistilBERT mainly because we had used it in class previously and since it would work with our limited access to large computing power. With DistilBERT, I also used an SGD Classifier on vectors since there are two different types of text, and it seemed that certain words could be more likely to appear next to each depending on the genre. I ran the code to establish a baseline and received the following results:

```
                      precision    recall  f1-score   support

    bikol literature       0.70      0.70      0.70       993
     bikol religious       0.86      0.11      0.20        54
  cebuano literature       0.94      0.04      0.08       692
   cebuano religious       0.85      0.69      0.76      1066
hiligaynon literature       0.55      0.80      0.65       991
 hiligaynon religious       0.69      0.23      0.34       380
  ilokano literature       0.82      0.75      0.78       886
   ilokano religious       0.67      0.76      0.71       751
  tagalog literature       0.71      0.73      0.72       838
   tagalog religious       0.62      0.74      0.67       435
    waray literature       0.67      0.90      0.77      2102
     waray religious       0.96      0.82      0.88       924

            accuracy                           0.71     10112
           macro avg       0.75      0.61      0.61     10112
        weighted avg       0.74      0.71      0.68     10112
```

Here you can see that the accuracy us 71% and the macro average is 61%, which is not too bad. When looking at the confusion matrix from the baseline, I did see some interesting patterns. Looking at the confusion matric, it looks like the Bikol Literature and the Waray Literature seemed to have a significant portion being mistaken for the other. Geographically the areas that Waray and Bikol are spoken are relatively close, but Waray would actually be more related to Cebuano since both are in the Bisayan language branch. I thought it was interesting since it could hint at the possibility of them borrowing more from each other, but it is also just possible that the model itself calculated that the two were somehow similar. There is nothing really conclusive we can say, but it is intriguing.



Graph 1) Confusion matrix of baseline

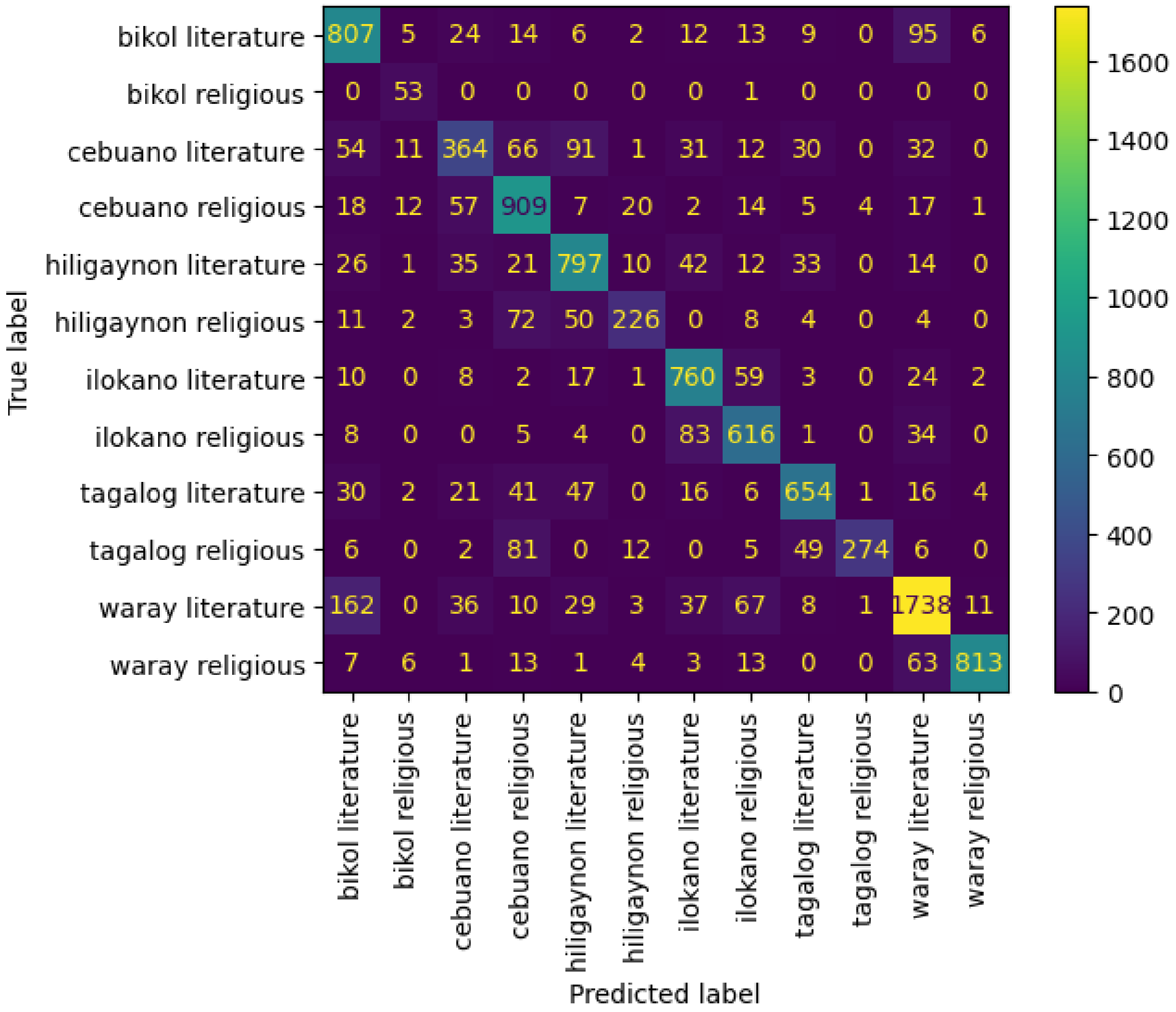
After that, I had made the model find the optimal value for alpha and then use what value that was to get these results:

```
                      precision    recall  f1-score   support

     bikol literature       0.71      0.81      0.76       993
      bikol religious       0.58      0.98      0.73        54
   cebuano literature       0.66      0.53      0.59       692
    cebuano religious       0.74      0.85      0.79      1066
hiligaynon literature       0.76      0.80      0.78       991
 hiligaynon religious       0.81      0.59      0.69       380
   ilokano literature       0.77      0.86      0.81       886
    ilokano religious       0.75      0.82      0.78       751
   tagalog literature       0.82      0.78      0.80       838
    tagalog religious       0.98      0.63      0.77       435
     waray literature       0.85      0.83      0.84      2102
      waray religious       0.97      0.88      0.92       924

             accuracy                           0.79     10112
            macro avg       0.78      0.78      0.77     10112
         weighted avg       0.80      0.79      0.79     10112
```

Here we can see that with the optimization that both the accuracy and the macro average increased to 79% and 77% respectively. Interestingly, the optimization improved how much certain categories were correctly labeled, such as the Cebuano Literature. In the baseline the number was extremely low, and most of the results were placed into Bikol, Hiligaynon, and Waray. Hiligaynon, Waray, and Cebuano are all Bisayan languages, so it is not surprising to see the model confuse those three. It is interesting that Cebuano is being mixed with Bikol, like it did with Waray. In general, the model was improved significantly after the optimization.



	bikol literature	bikol religious	cebuano literature	cebuano religious	hiligaynon literature	hiligaynon religious	ilokano literature	ilokano religious	tagalog literature	tagalog religious	waray literature	waray religious
bikol literature	807	5	24	14	6	2	12	13	9	0	95	6
bikol religious	0	53	0	0	0	0	0	1	0	0	0	0
cebuano literature	54	11	364	66	91	1	31	12	30	0	32	0
cebuano religious	18	12	57	909	7	20	2	14	5	4	17	1
hiligaynon literature	26	1	35	21	797	10	42	12	33	0	14	0
hiligaynon religious	11	2	3	72	50	226	0	8	4	0	4	0
ilokano literature	10	0	8	2	17	1	760	59	3	0	24	2
ilokano religious	8	0	0	5	4	0	83	616	1	0	34	0
tagalog literature	30	2	21	41	47	0	16	6	654	1	16	4
tagalog religious	6	0	2	81	0	12	0	5	49	274	6	0
waray literature	162	0	36	10	29	3	37	67	8	1	1738	11
waray religious	7	6	1	13	1	4	3	13	0	0	63	813

Graph 2) Confusion matrix after optimizations

Conclusion:

In conclusion, what I have learned from this project is that using transfer learning on different languages does seem to work, and that it can actually distinguish between genres of those languages to a degree. We also saw that there may be an effect of the model confusing more closely related languages. Although this was just a very basic use of transfer learning, if one were given more computing power, someone may be able use a larger amount of training data. Furthermore, I think it would be interesting to use data that is sourced elsewhere that is not in the PALITO Corpus and see how effectively the model would work on that after being trained on this data. Originally, I had wanted to train the model on Ilokano data and the try and see if I can get it to parse through spoken Ilokano, but my main issues my main issue was just finding data for that. At least with transfer learning, I know that it is possible to train a language model on limited data and still get some effectiveness. For now, I think this shows that in essence, transfer learning is really useful for those who want to do research on low-resource languages.