Text Classification

There are a total of 8 configurations done for the text classifier model, 4 of which utilizes Logistic Regression and 4 of which utilizes RNN.

Based on the precision-recall curve of all the models, the logistic regression model overall produced significantly better results, showcasing that for this particular dataset, a statistical classifier is better suited for the task in comparison to RNN. Classifiers that are trained using more data also produce significantly better results in comparison, with the statistical model that utilizes the full abstract dataset having the highest precision recall value.

From all of the RNN configurations, only the model that is trained on the title and using the full dataset resulted in a performance that is comparable to the rest of the statistical model. However, even then it is still slightly worse performance wise compared to its statistical model counterpart.

Meanwhile, RNN model trained on abstract even with the full dataset is not able to measure up to the Statistical Model trained on the same dataset, this may be caused due to overfitting, seeing as the RNN model trained on only 1000 rows on abstract displays a better precision-recall curve in comparison.

Based on the statistics gathered from the Accuracy, Precision, Recall and F1 Score:

Within the Statistical Models, the configuration utilizing abstract and all the rows within the dataset has the highest accuracy percentage amongst all of them, making it the configuration with the most amount of correct predictions among the statistical models. This may be due to the amount of token data that it would have gotten access to, allowing it to better perform predictions as compared to the configurations that utilizes titles, which has less data to work with in general.

In terms of other statistics, it also beats out the other configurations in this case, however, it has identified less true positives compared to the abstract with 1000 rows configuration and the title with 1000 rows configuration. However, these two may be a case of overfitting due to the recall percentage being very high compared to their overall accuracy and precision, so based on overall statistics, the abstract with all rows configuration remains the best performing model for now as the other two models have labeled too many things as positive, causing a decent portion of those positives to be false positives.

Within the RNN models, instead it is the Title with the full dataset configuration that has the best overall statistics, with it having the highest accuracy, precision and F1 score, though it does not have the highest overall recall score. Both abstract configurations have the highest amount of recall, however, they have quite low precision, indicating that similar to the statistical models, they also label the articles as positive very aggressively, causing a higher percentage of false positives, therefore lowering their overall accuracy. Therefore, the Title with full dataset configuration remains to be the most optimal choice for this dataset.

Overall, based on the statistics and the precision-recall curve of the models, the most ideal model for this case is the Statistical Model with Abstract and all rows configuration, and if there comes a situation where utilizing RNN may be more favourable, the model that was trained using the Title and all of the dataset is the best alternative option. If time is also a consideration, the RNN models also perform poorly in comparison to the statistical models, taking significantly longer to be trained on full datasets, therefore making the statistical models trained on the title and all rows the best option.

Topic Modelling

For both configurations, removal of stopwords was considered and attempted for the pre-processing stage, however this seems to have instead worsened topic coherency and therefore was deemed unnecessary for the model.

Configuration 1

- Pre-Processing: Removed Numbers, Rare Tokens and Common Tokens
- Did not utilize Bigrams
- Topics: 10

Using this configuration and running the model through 1000 rows of the training dataset, the most frequent token in most of the topics seems to be centered around model, algorithm, and learning. Based on the groupings on the results, the model seems to have understood some link between the words within the topic as the articles that are fetched by it displays similarities even across different topics, that being centered around large language models or algorithms to construct them, however, the applications of these models seems to differ between each of the articles, so it can be assumed that the LDA model mostly groups them due to their frequency and co-occurrence patterns rather than semantic context.

When this model is run through using 20,000 rows of the training dataset, the model token is no longer a part of the top words within each topic within the dataset, rather the more frequent tokens within each topic seems to be extracting, computational, language. The contents of the articles themselves however still seems to still mention models and machine learning concepts, even if it is not considered a prominent token within the topics themselves. The differing results may be caused due to more articles not having co-occurence of the token models specifically and the more popular topics within them, further showing that frequency is taken more into consideration over semantic context.

An overall advantage with this is that the LDA model is able to group the articles into broad thematic groupings without having to rely on labels, such as HumanComputerInteractions, etc that was in the original dataset.

The intertopic distance map for the 1000 row dataset seems to be less grouped up compared to the 20,000 row dataset one. Though with the latter, one of the topics is placed significantly further away and is smaller in frequency of tokens within it.

This configuration also resulted in quite poor topic coherence, with the topic coherence of the 20,000 rows results being significantly worse than the result of the 1000 rows one. The 1000 rows one's topic coherence however, remains at a decent number. This can also be seen through the results, as one of the topics of the 20,000 rows model was not able to find a matching article while the topics for the 1000 rows were able to find all matching articles.

Configuration 2

- Pre-Processing: Removed Numbers, Rare Tokens and Common Tokens
- Utilized Bigrams
- Topics: 25

This configuration’s pre-processing remains largely the same, with the exception of the inclusion of bigrams to the collection of tokens, and a higher topic count.

Running this model with 1000 rows results in frequent groupings of the tokens model, algorithm, layer, training, and generalization within most of the topics with those articles being centered around the fundamentals of machine learning and the architecture of models. There are a few other groupings as well that don't fully share most of the tokens with the most frequent group, those tokens being centered around large language models, optimization and mathematical terms related to it.

Meanwhile with 20,000 rows, the results are quite similar with the results of the 1000 rows, with the topic groupings similarly centered around LLM and learning. Though there are less groupings of tokens involving the direct mention of models.

The advantages of the model are the same as the advantages discussed in the previous iteration of the LDA model. The topic coherence is poorer for this configuration in comparison to the other configuration.