Deliverable 2: Progress and Preliminary Results

Project:
A program that can identify the school(s) of philosophy to a given statement belongs. We want to create a model that can classify text as belonging to one of 9 possible schools of philosophy.

Data preprocessing:
https://www.kaggle.com/datasets/kouroshalizadeh/history-of-philosophy
We are using the dataset listed above from kaggle, the same one stated in deliverable 1. To process the data we extracted the columns we decided to use, the school of philosophy (the classes) and the string of tokenized words that was already preprocessed in the dataset. We split the data into testing and training data sets and separated out the classes from the text samples.

Machine Learning Model:
In order to ensure the data kept some of its meaning we used the Word2vec to vectorize words rather than a bag of words model. In order to implement this we used the Gensim library to create a dictionary of vectorized words from our training data. WE then used this to vectorize the sample sentences word by word and take the average of all of them in order to have a standard length for all for each sample. With our testing and training data vectorized we used the random forest model from sci-learn to classify our data. The model is currently very overfit.

Preliminary Results:
As stated above, the model is very overfit, the training data has a very high accuracy and the test data has a very low accuracy. Some of the hyper parameters need to be tuned as this should give more accurate results.

Next Steps:
To improve our model we will continue to tune hyperparameters to see what can be improved. There are several parameter we have yet to experiment with, but we will look at some graphs of our data in particular the vectors to determine how easy it will be to classify our data we may need to change our classification method or choose a different model from word2vec depending on the result.