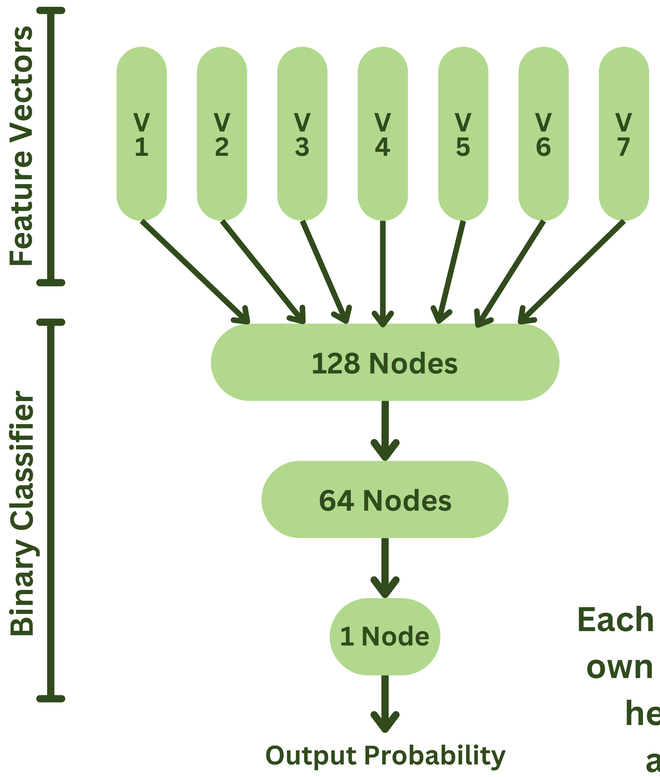
# Group 17 Authorship Verification

## Method 1: Traditional ML




### Model Structure

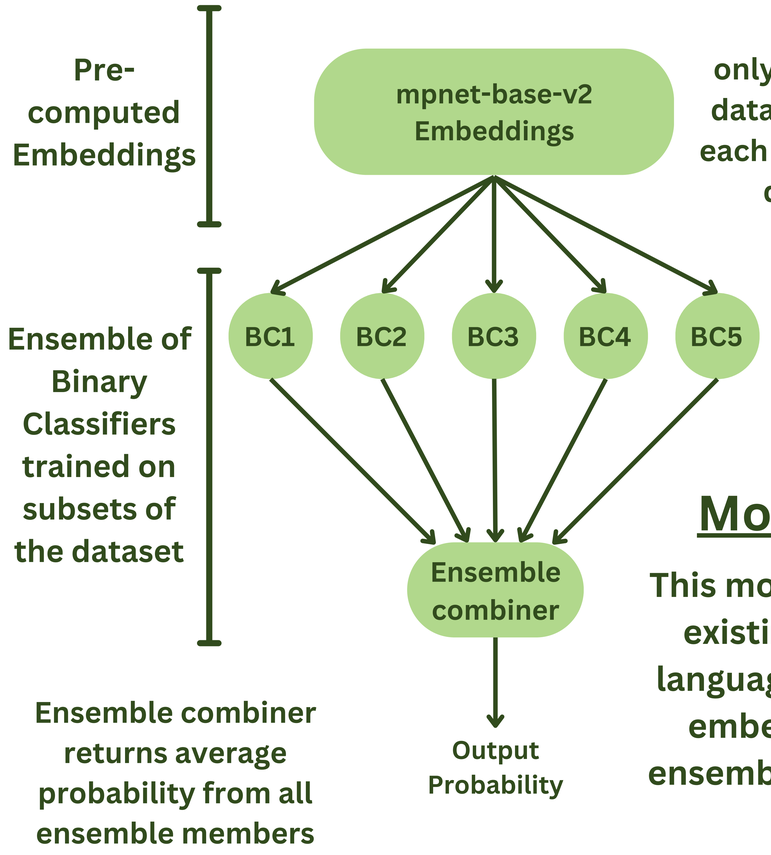
The main idea of this method is that we calculate multiple different feature vectors such as punctuation count, and feed these vectors into a binary classifier with trained weights.

Each feature vector measures its own informative stat that could help to identify if the same author wrote both texts.

### The Feature Vectors

| | |
|---|---|
| Punctuation | Measures count, complexity and distribution of punctuation. |
| Determinant/Noun ratio | Ratio of determinants to nouns. |
| Word/Sentence Lengths | Difference between the longest and shortest token, and the third quartile of length. |
| Capitals | Ratio of properly capitalised nouns, as well as how many sentences start with capitals. |
| Typos | How many typos are made compared to the total number of words. |
| Type-Token Ratio | Ratio of unique words to total words, shows if the vocabulary is diverse. |
| Readability | The Flesch-Kincaid Grade Level score measures how easy to read the text is. |

## Method 2: Deep Learning




Ensemble combiner returns average probability from all ensemble members

### Model Structure

This model leverages the pre-existing knowledge about language found in the MPnet embeddings and adds an ensemble of binary classifiers on top.

### Why an Ensemble?

Using an ensemble allows us to generalise better when averaged, as each member learns slightly different features from its own subset of data.

Bootstrap aggregating our training data like this can help reduce overfitting and improve robustness, as we average out any noise or weird outlier predictions from any single model.