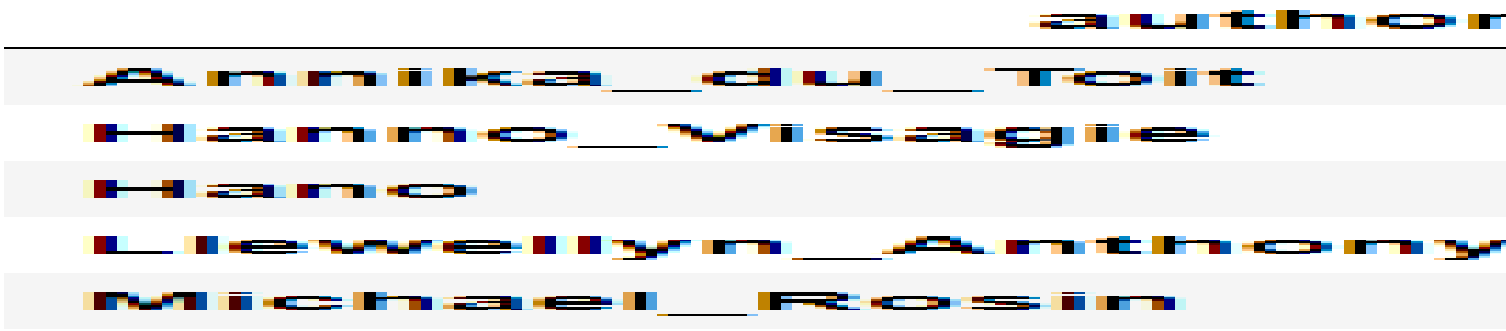
Student number that was investigated: 31597793

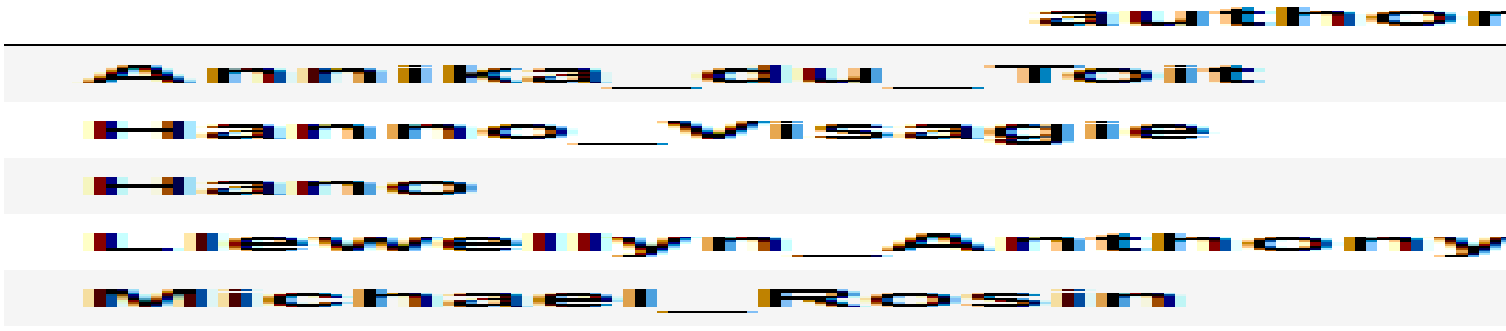
Date the report was created: 2022-11-14 19:34:03

Burrows delta score



The Burrows delta score is calculated from the z-scores of every common word in the vocabulary, which represents the fingerprint of the document and can be across multiple documents. The z-scores are numerical measuremnts that descibes a value's relationship to the mean of a group of values. The author (student number) of the submitted document is the x-axis of the matrix and the authors (student number) of all the documents in the corpus is the y-axis of the matrix. If the value is very low it means that the author in the corpus is the author of the submitted document.If the value is very high it means that the author in the corpus is not the author of the submitted document. The closer the value is to 0 the more likely the author in the corpus is the author of the submitted document.

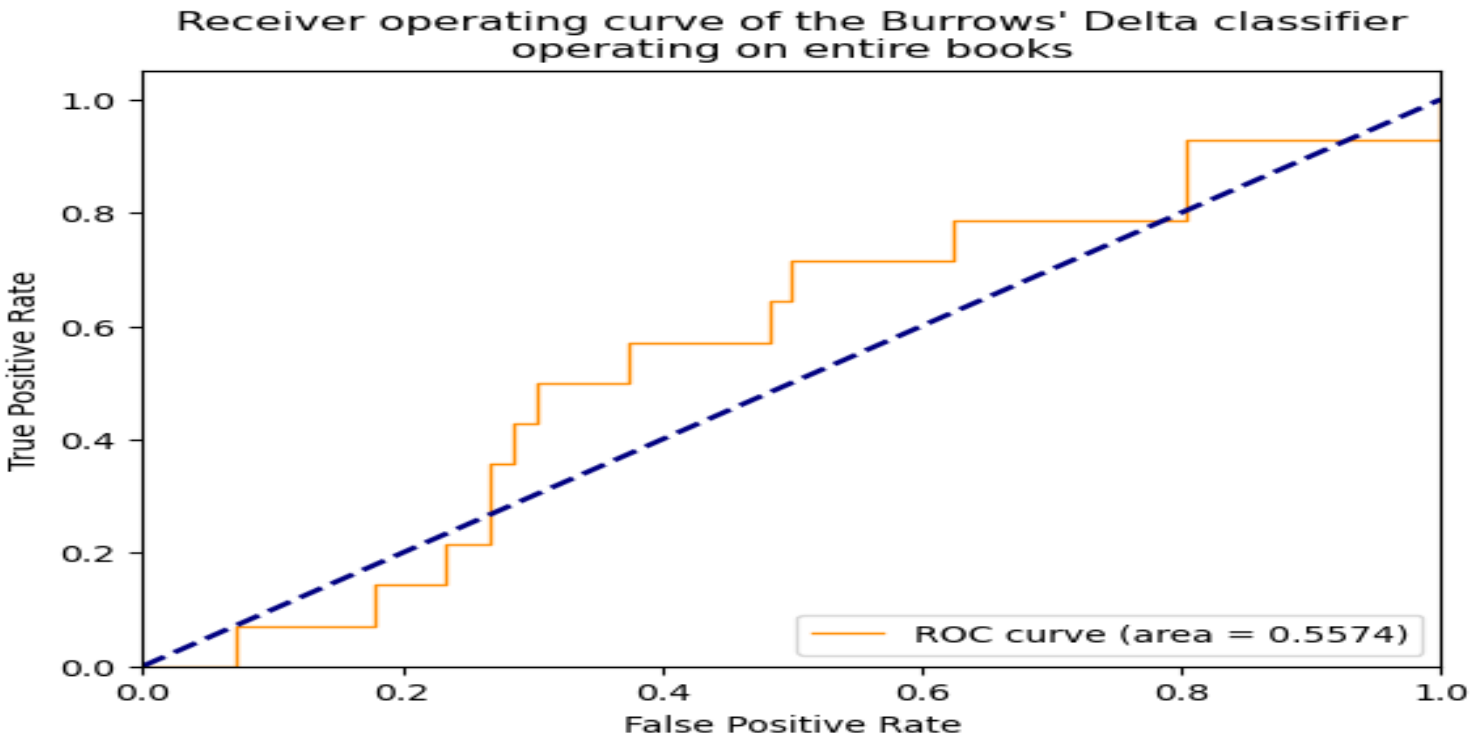
Authorship score



The authorship score is a probability corresponding to the delta values from the Burrows delta matrix. A high value indicates a high probability thatthe author in the corpus is the author of the submitted document. The author (student number) of the submitted document is the x-axis of the matrix and the authors (student number) of

all the documents in the corpus is the y-axis of the matrix. If the value is very low it means that the author in the corpus is not the authorof the submitted document. If the value is very high it means that the author in the corpus is the author of the submitted document. The closer the value is to 1 the more likely the author in the corpus is the author of the submitted document.

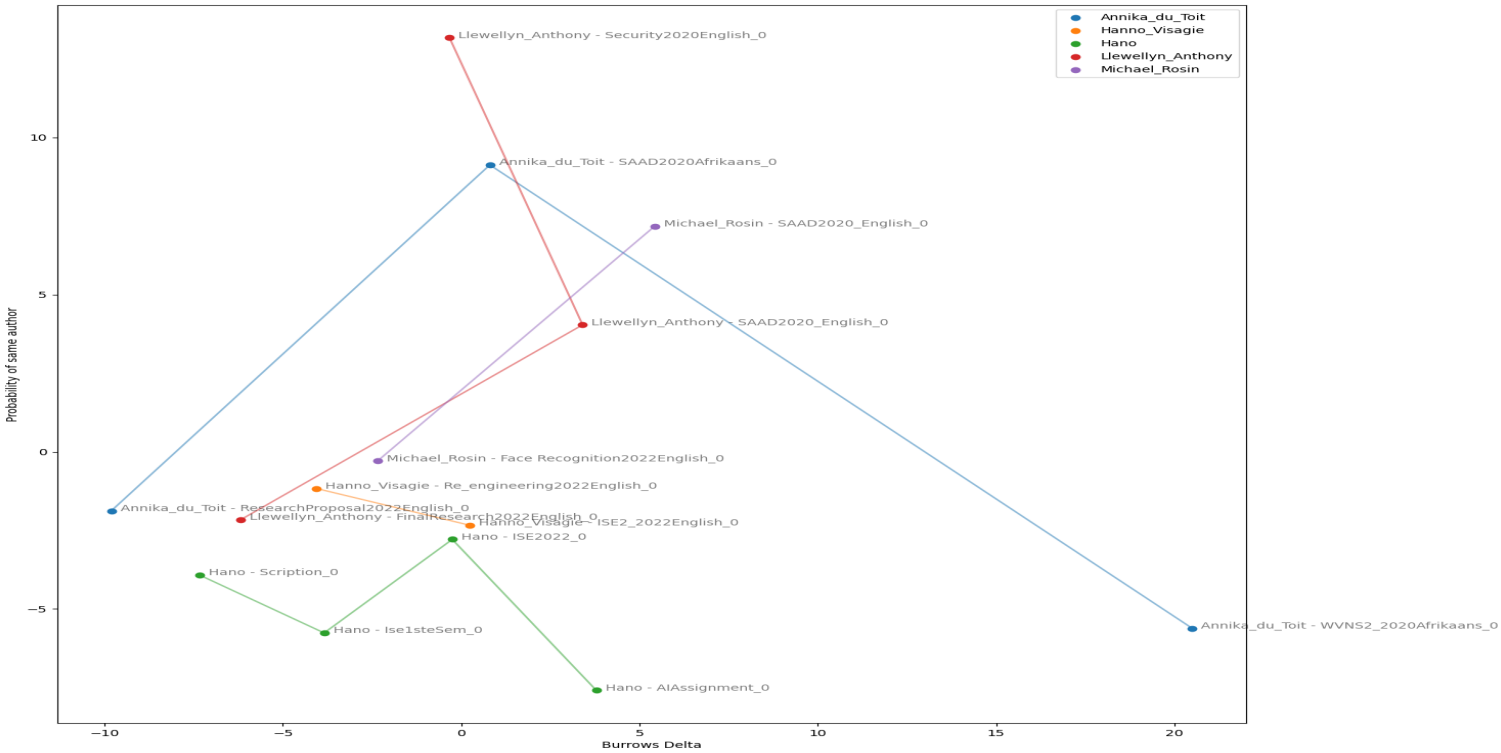
Receiver Operator Characteristic (ROC) curve



The Receiver Operator Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings to determine the performance of the Machine Learning Model which is used to calculate the Burrows delta score and the authorship score. Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve is to the 45 degree diagonal of the ROC space, the less accurate the test is. An AUC score of 0.5 means that a classifier is performing badly, and a 1.0 score is a perfect score. The AUC score is

on the bottom right corner of the graph. The model will be more accurate if the corpus consists of more documents and if each document contain more words.

Principal Component Analysis (PCA) curve



The Principal Component Analysis (PCA) curve is a cluster of all the authors and their documents. It is used to visualise the sylistic similarities between the documents in the corpus. The closer the documents are to each other the more similar they are. The documents are clustered based on their Burrows delta score.