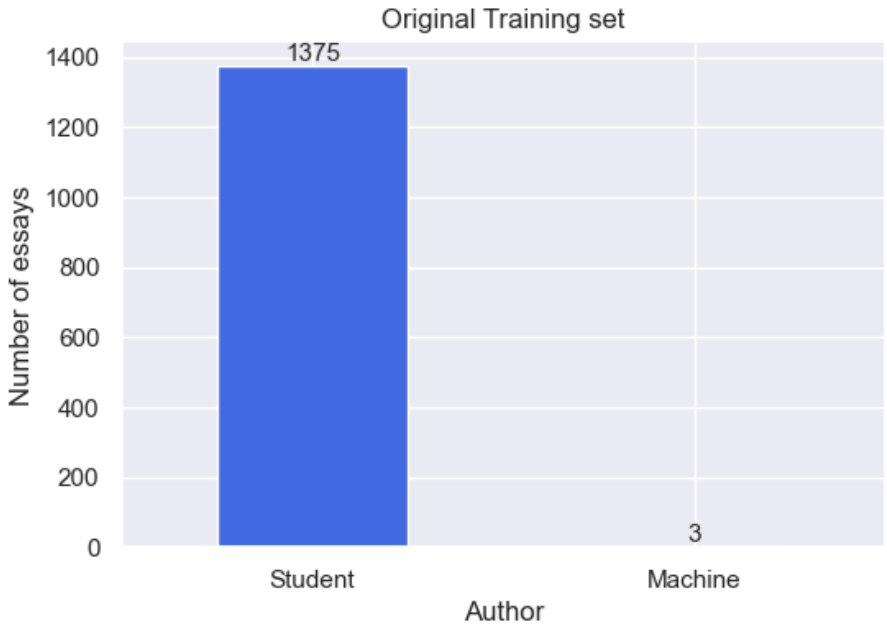
Kaggle Competition: Essay Classification Challenge

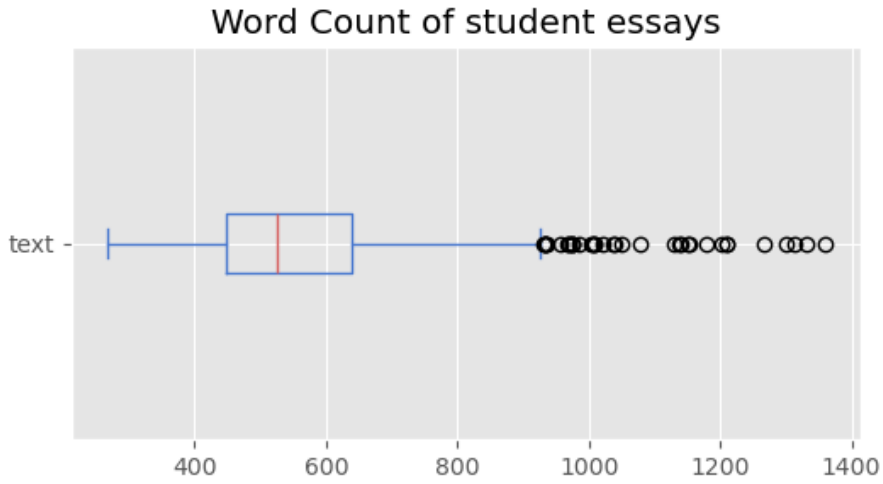
Practical Data Science - 2023

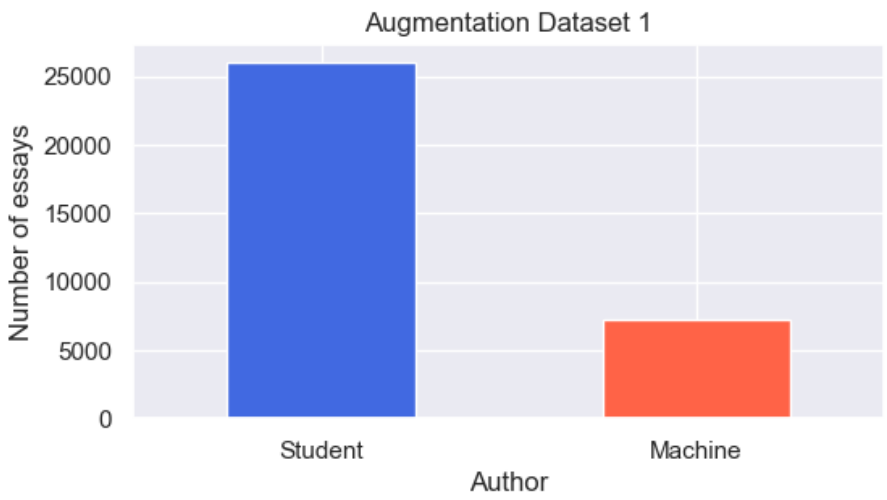
Phevos A. Margonis

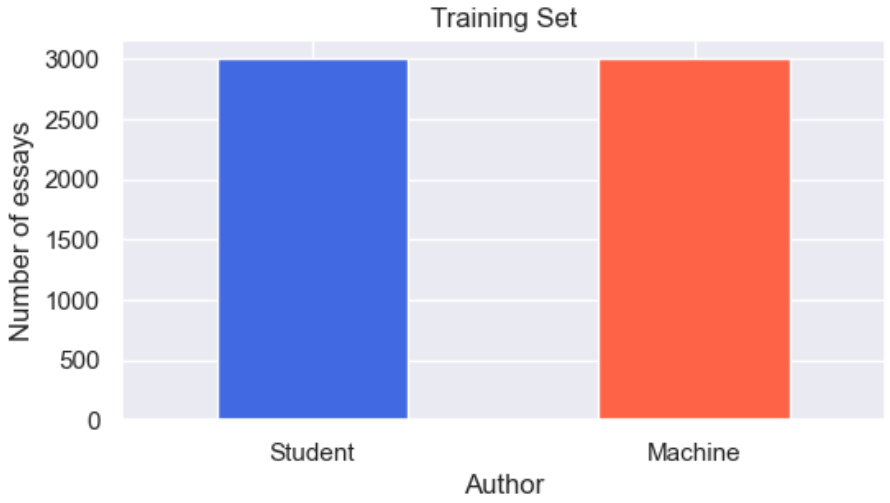
Athens University of Economics and Business

December 14, 2023

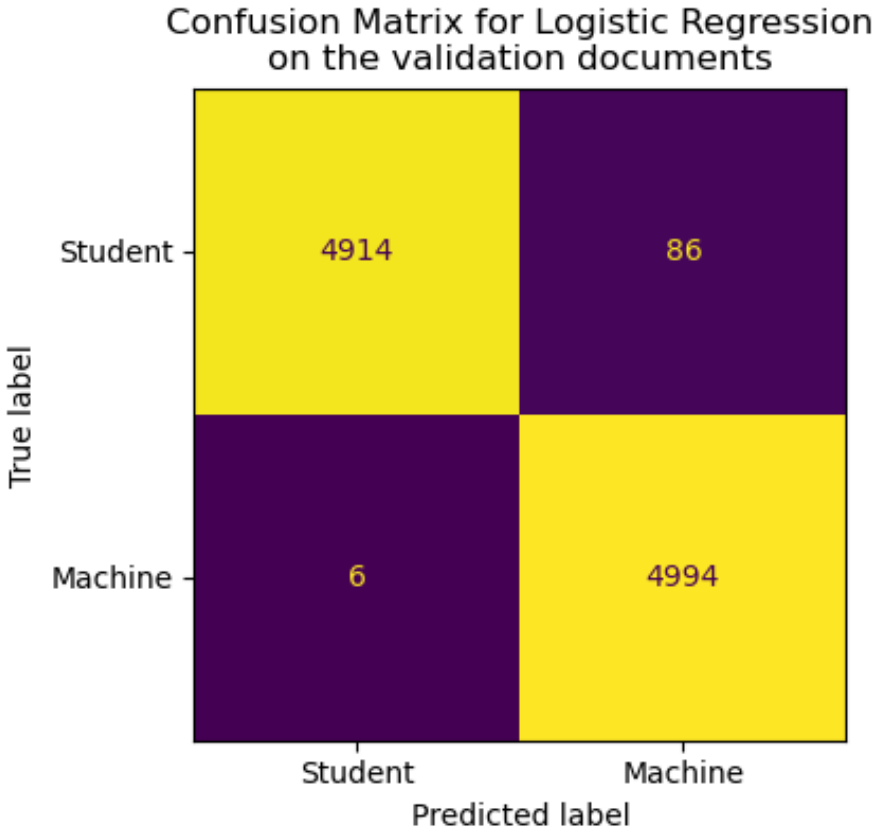

Original Training set
1375
3
Number of essays
Author
Student
Machine
0
200
400
600
800
1000
1200
1400

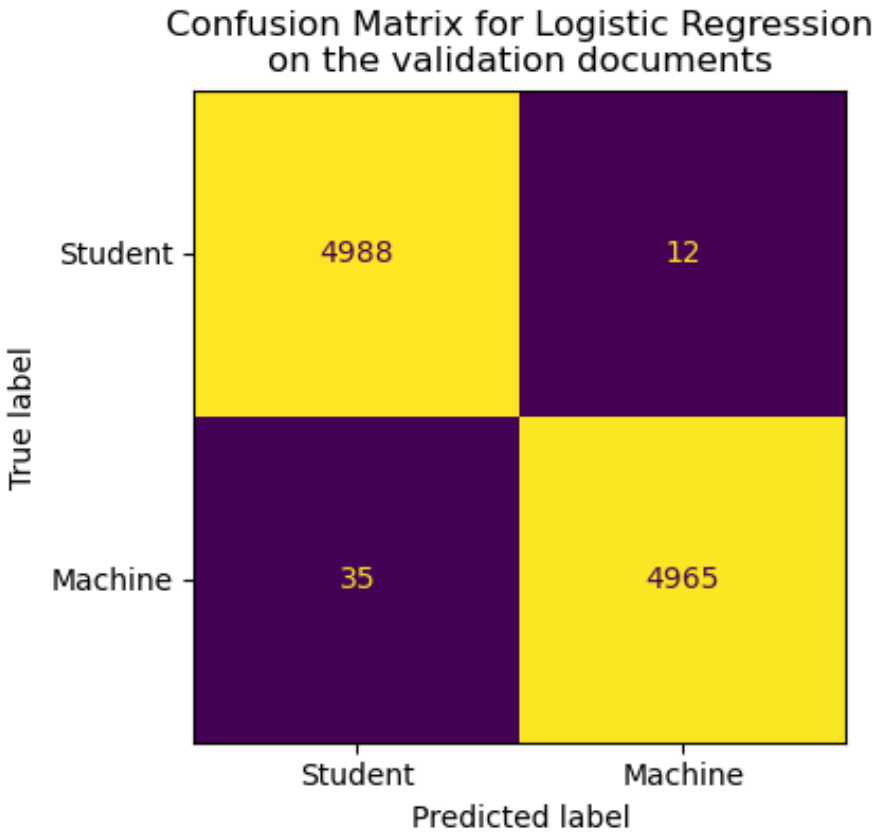

Word Count of student essays
text
400
600
800
1000
1200
1400

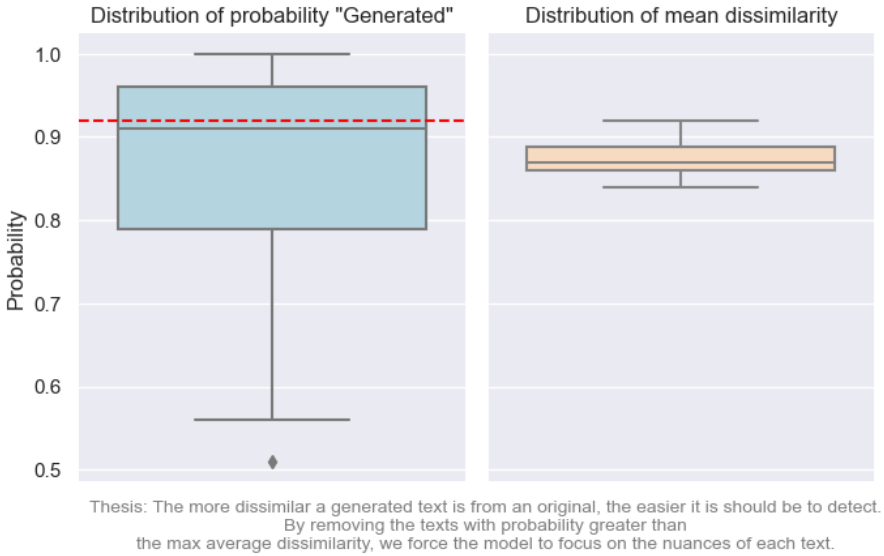

Augmentation Dataset 1
Number of essays
25000
20000
15000
10000
5000
0
Student
Machine
Author

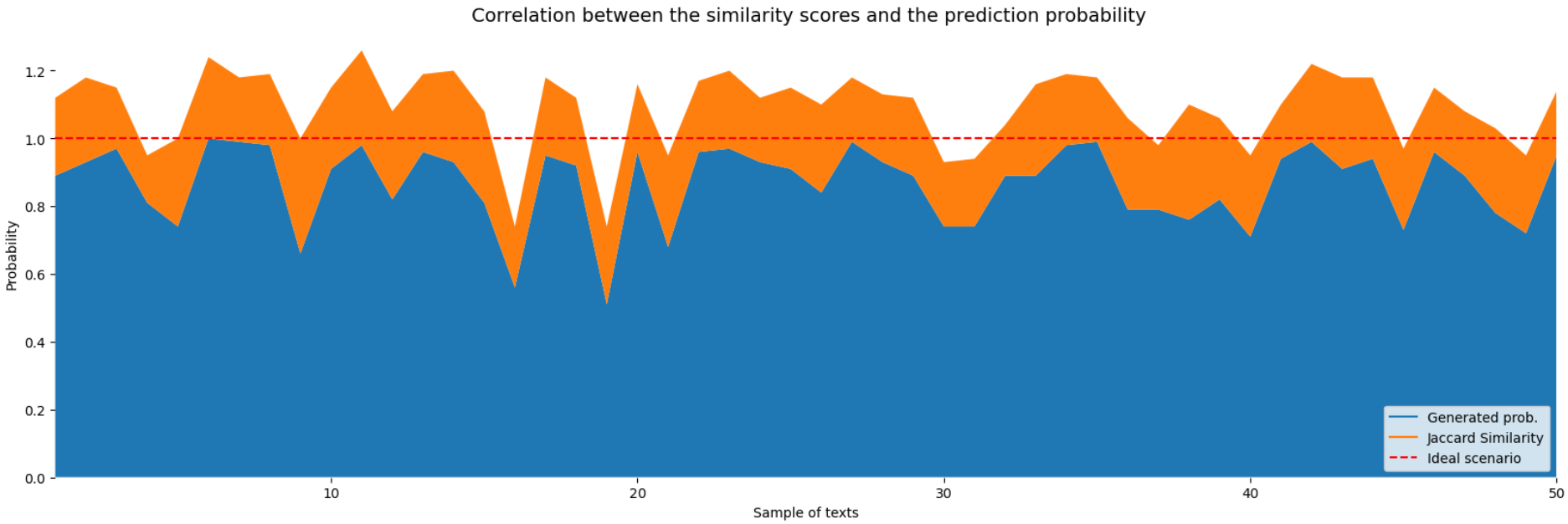

Training Set
Number of essays
3000
2500
2000
1500
1000
500
0
Student
Machine
Author


Validation Set
Number of essays
5000
4000
3000
2000
1000
0
Student
Machine
Author

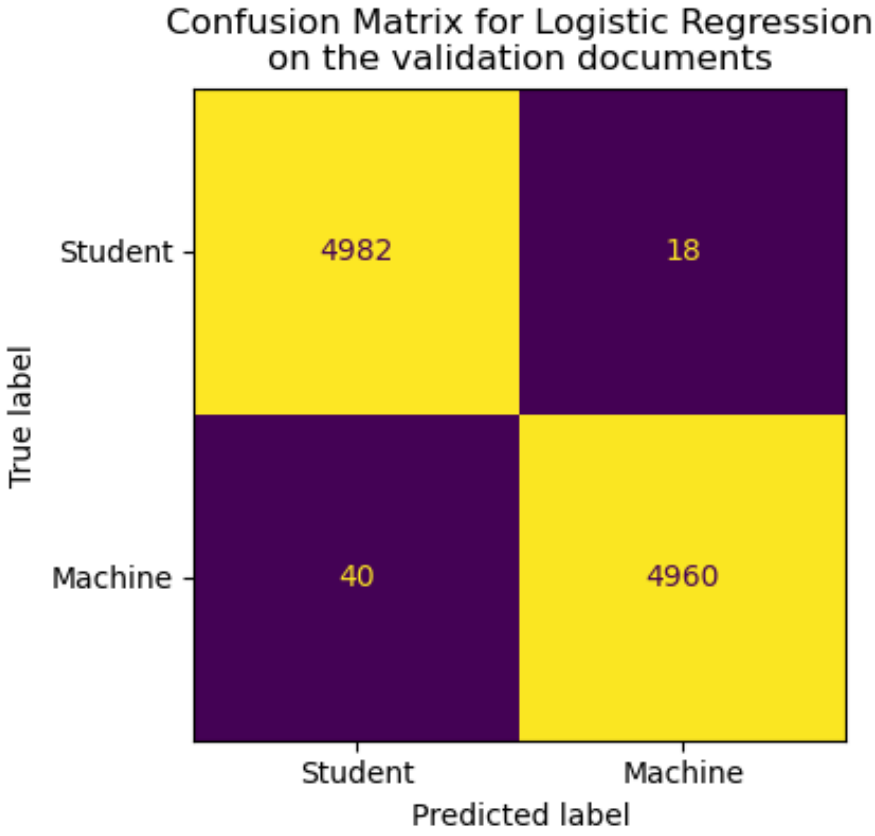

Confusion Matrix for Logistic Regression
on the validation documents
True label
Student
Machine
4914
86
6
4994
Student
Machine
Predicted label

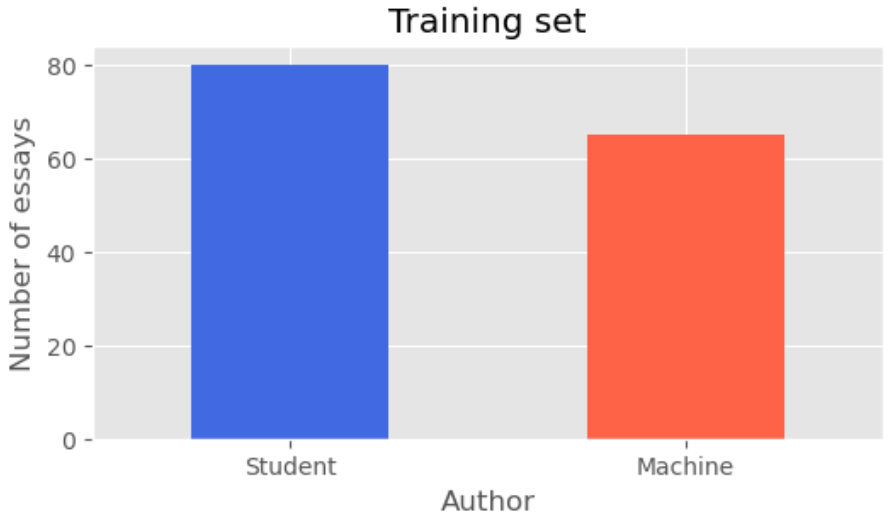

Confusion Matrix for Logistic Regression
on the validation documents
Student
Machine
True label
4988
12
35
4965
Student
Machine
Predicted label


Distribution of probability "Generated"
Distribution of mean dissimilarity
Probability
1.0
0.9
0.8
0.7
0.6
0.5
Thesis: The more dissimilar a generated text is from an original, the easier it is should be to detect.
By removing the texts with probability greater than
the max average dissimilarity, we force the model to focus on the nuances of each text.



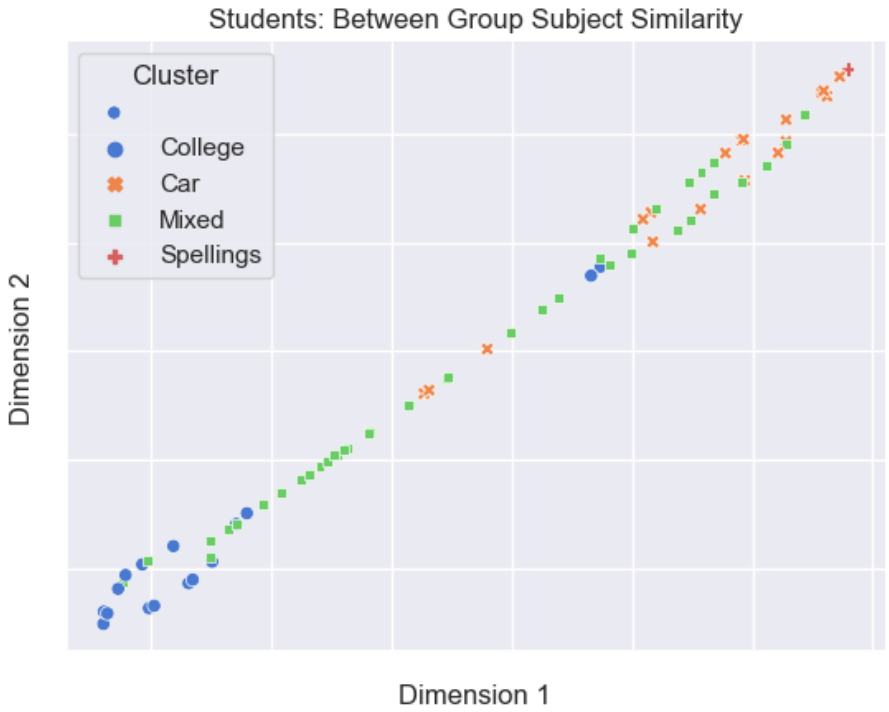
Thesis: The more similar pairs of student-LLM texts are, the harder it should be for the classifier to distinguish the True class.
Following that logic, the correlation between the probability that a text is generated, should be inversely proportional to the maximum similarity of that text to all the rest.
In the plot above, the ideal scenario is captured at points x=5, x=9, where the probabilities sum to one.

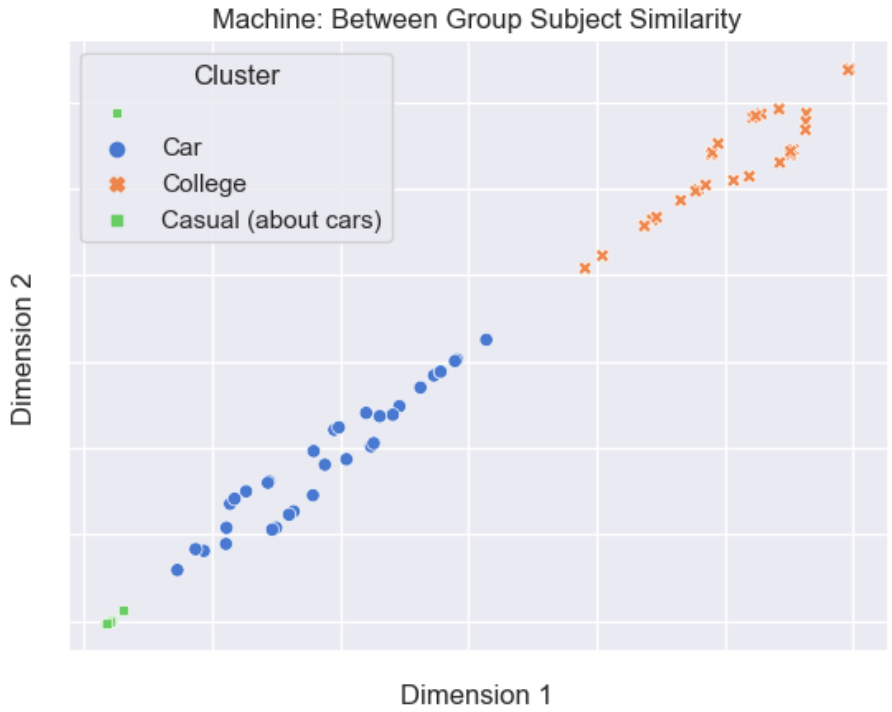

Confusion Matrix for Logistic Regression
on the validation documents
Student
Machine
True label
4982
18
40
4960
Student
Machine
Predicted label

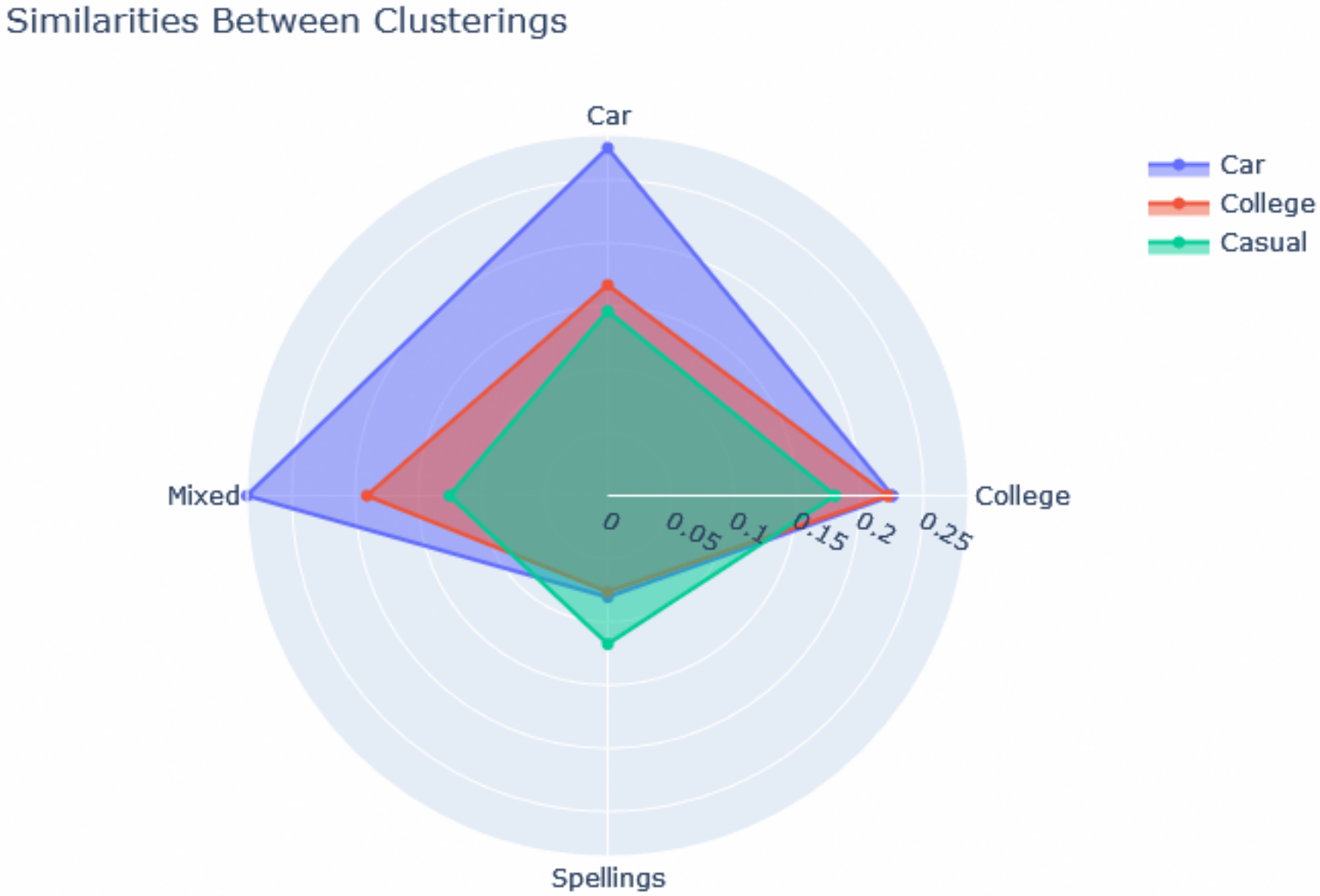

Training set
Number of essays
0
20
40
60
80
Student
Machine
Author


Learning curves
Training F1
Test F1
F1 Score
Size of training set
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
20
40
60
80
100
120
140


Cluster balance between clusterings
Legend Title
Student
Machine
Number of essays
Cluster
40
30
20
10
0
0
1
2
3


Students: Between Group Subject Similarity
Cluster
College
Car
Mixed
Spellings
Dimension 2
Dimension 1


Machine: Between Group Subject Similarity
Cluster
Car
College
Casual (about cars)
Dimension 2
Dimension 1


Similarities Between Clusterings
Car
College
Spellings
Mixed
0
0.05
0.1
0.15
0.2
0.25
Car
College
Casual