LLM Detection

Practical Data Science: 3rd Project

Tsirmpas Dimitris

December 16, 2023

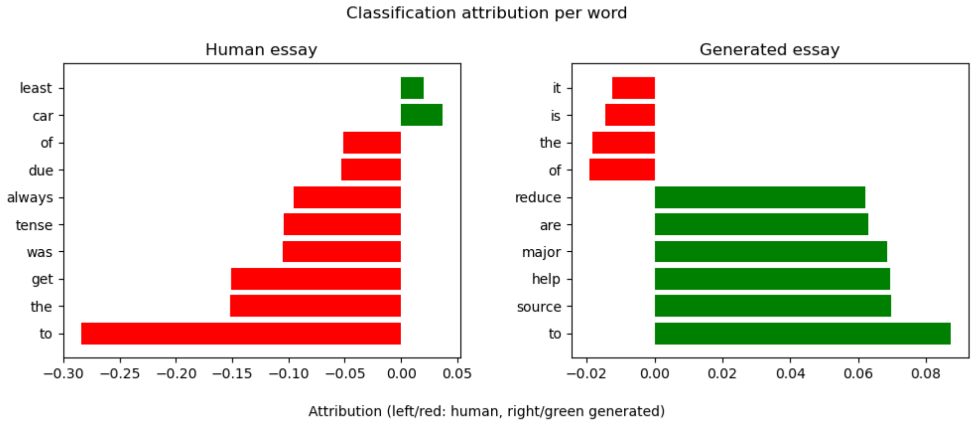
Athens University of Economics and Business
MSc in Data Science

1 Introduction

This report outlines results and conclusions drawn from the LLM detection project which can be found at `https://github.com/dimits-exe/practicaldatascience`. Implementation details, methodology and discussion can be found inside the relevant notebook and README file.

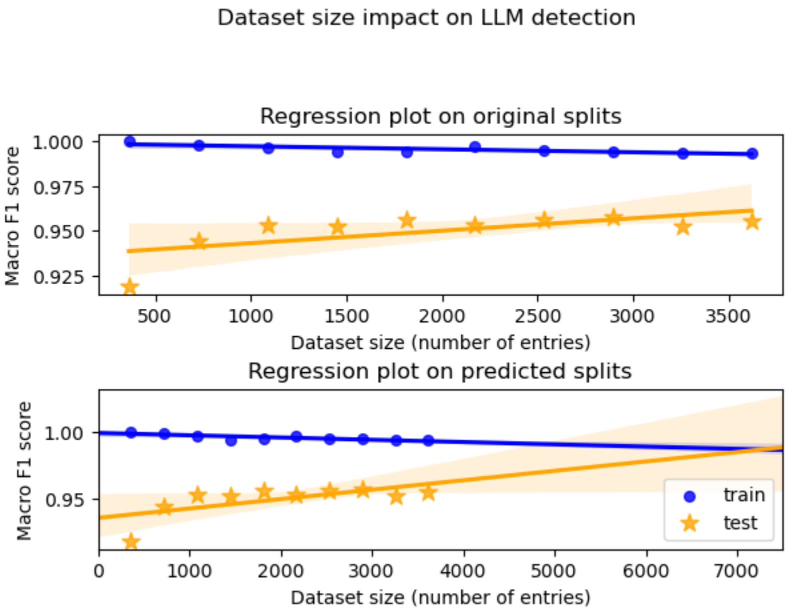
2 LLM Detection Results

Below we present graphs resulting from the LLM detection models presented in the original notebook.



Classifier attribution (LIME) per word for correctly classified human and generated essays. Red (left) lines indicate the classifier leans towards the text being a human essay because of the word's use, Green (right) lines indicate that it leans towards the text being LLM generated. The length of the lines indicates the certaintly of the classifier towards the classification. Note that this graph does not necessarily represent probability values for each word.

Figure 1: Classifier attribution (LIME) per word for correctly classified human and generated essays. Red (left) lines indicate the classifier leans towards the text being a human essay because of the word's use, Green (right) lines indicate that it leans towards the text being LLM generated. The length of the lines indicates the certainty of the classifier towards the classification. Note that this graph does not necessarily represent probability values for each word.



Impact of dataset size on LLM detection. The classifier used is the Random Forest Classifier with 15 simple decision trees. The green line indicates the expected classifier test score on unseen dataset sizes. This line follows the linear assumption which rarely applies to these kinds of problems. Shaded area represents the 95% Confidence Interval.

Figure 2: Impact of dataset size on LLM detection. The classifier used is the Random Forest Classifier with 15 simple decision trees. The green line indicates the expected classifier test score on unseen dataset sizes. This line follows the linear assumption which rarely applies to these kinds of problems. Shaded area represents the 95% Confidence Interval.

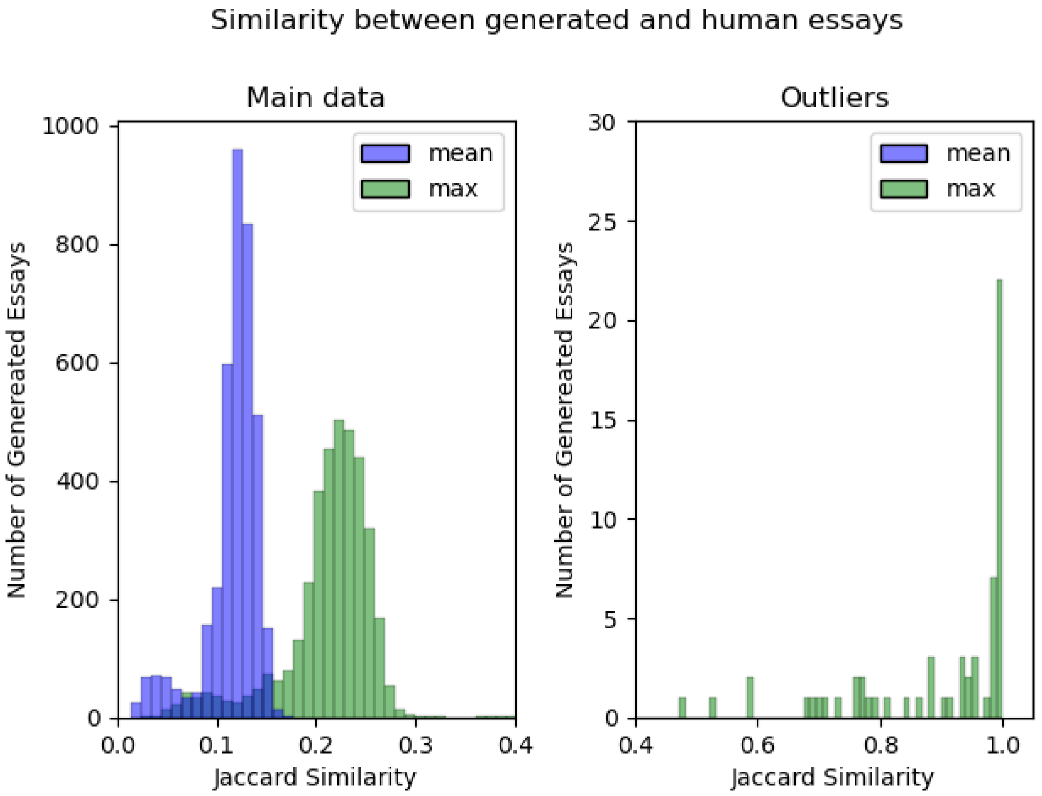


Figure 3: Mean and maximum Jaccard similarity between each generated essay compared to all human essays. This Figure demonstrates the very low similarity between the two kinds of essays in our dataset, largely explaining the exceptional performance of our classifiers.

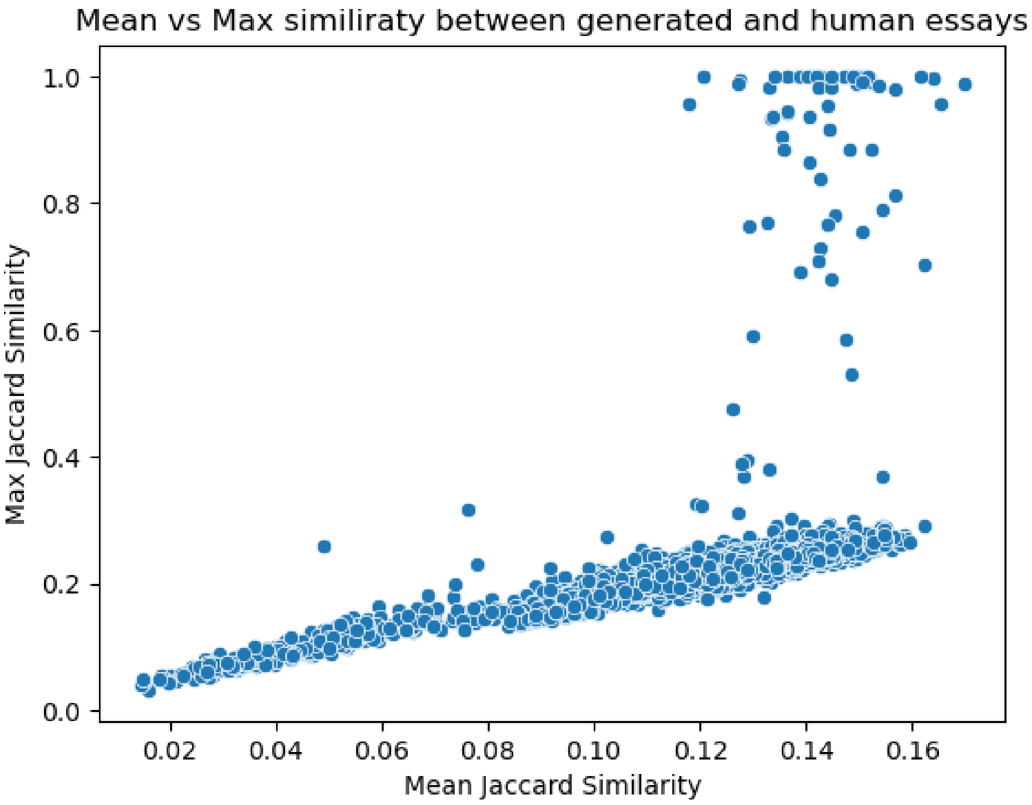


Figure 4: Relationship between mean and max similarity between each generated essay and all human essays. The relationship can be described as follows: From 0 to 0.12 mean similarity the relationship is clearly positive linear, while from 0.12 to 0.20 the relationship is still clearly positive and mostly linear but with significant outliers.

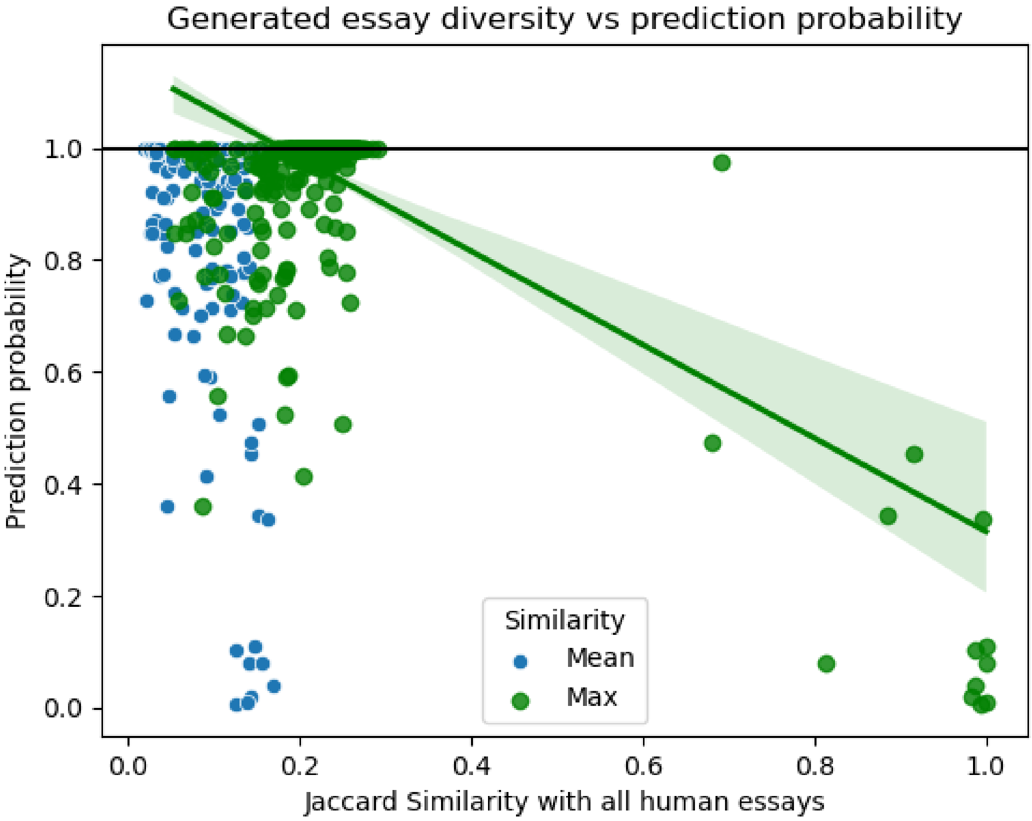


Figure 5: Relationship between essay similarity and prediction probability for that (generated) text. The distributions of mean and max similarity can also be seen in Figure 3. We notice a slight inverse trend between max similarity and prediction probability. When similarity becomes 1, or very close to 1, the classifier fails since the data point itself is mislabeled.

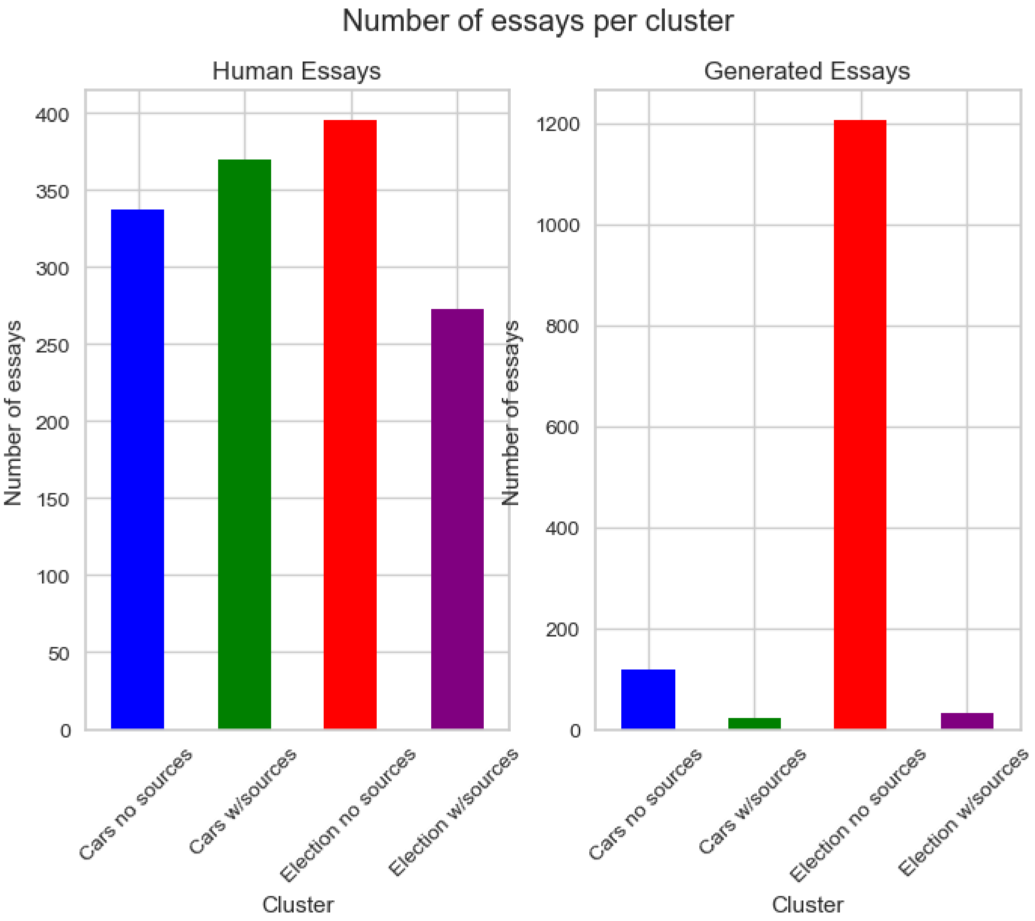


Figure 6: K-Means clustering of human and generated essays, labeled according to qualitative characteristics. The optimal generated

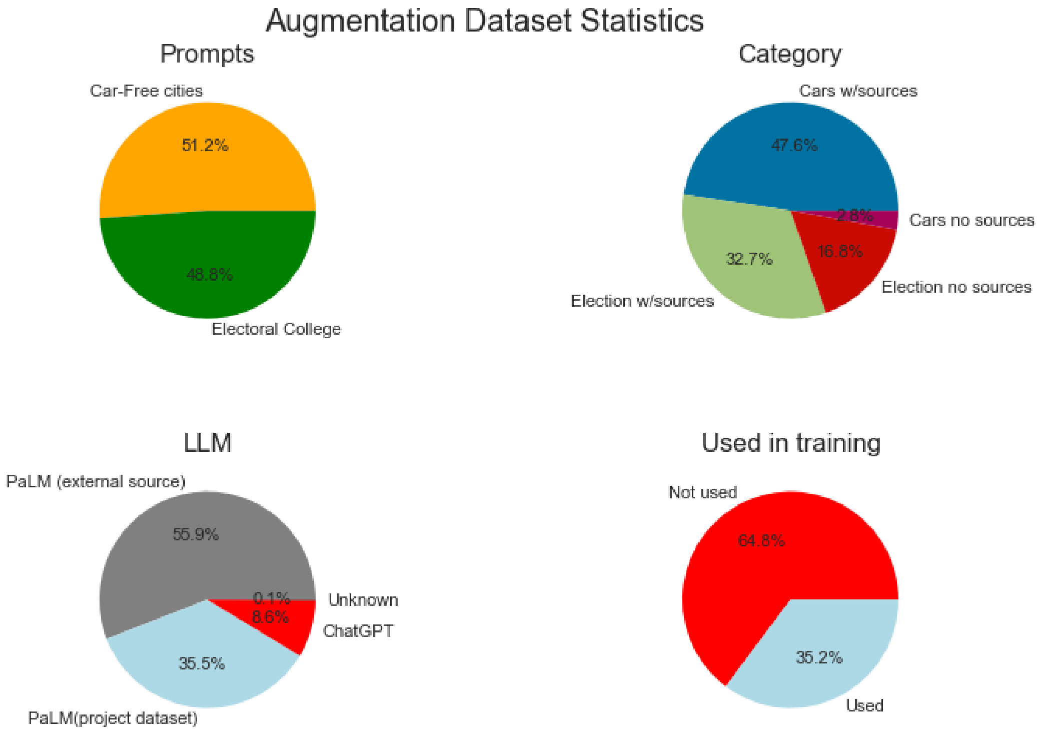


Figure 7: Statistics about the final, full augmented dataset comprising all the generated essays, even those not eventually used for the final model. The ”Category” pie-plot displays the relative counts of clusterings as defined by the K-means algorithm with $K = 4$ trained on the human texts.