LLM Detection

Practical Data Science: 3rd Project

Tsirmpas Dimitris

December 9, 2023

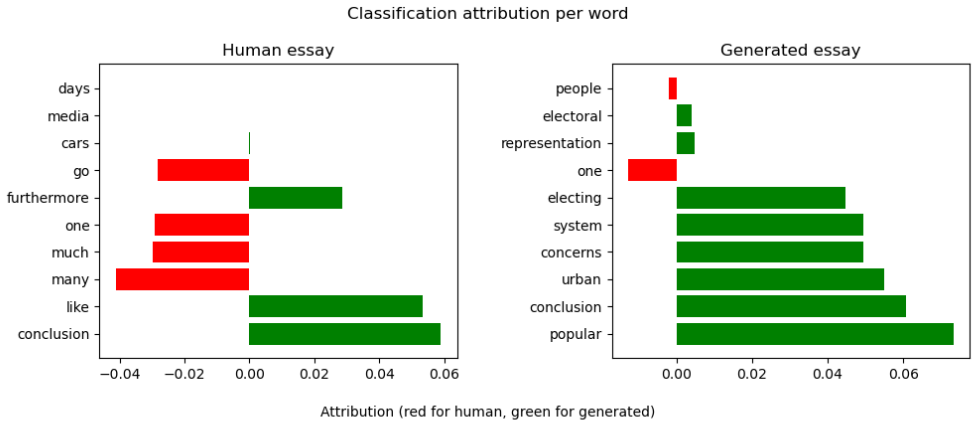
Athens University of Economics and Business
MSc in Data Science

1 Introduction

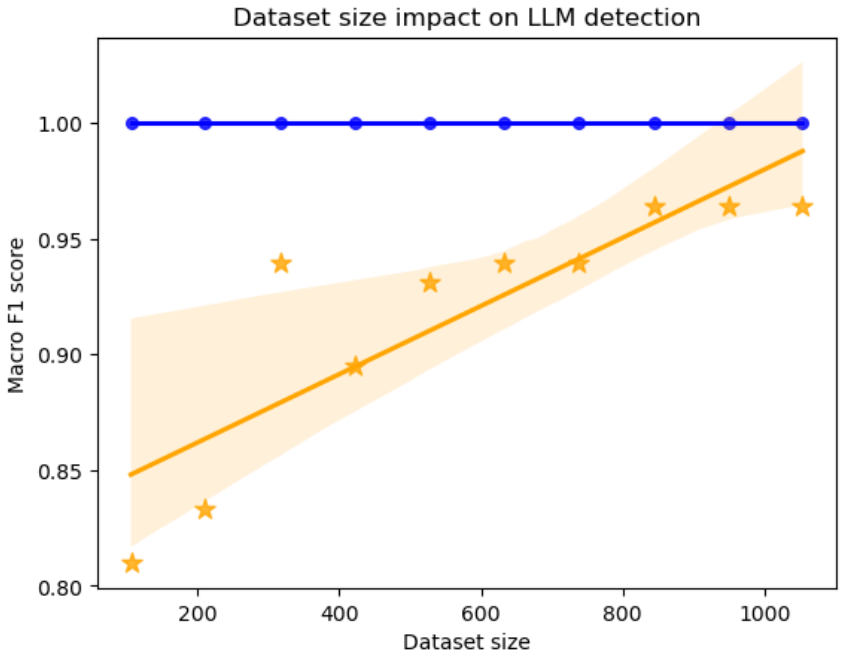
This report outlines results and conclusions drawn from the LLM detection project which can be found at `https://github.com/dimits-exe/practicaldatascience`. Implementation details, methodology and discussion can be found inside the relevant notebook and README file.

2 LLM Detection Results

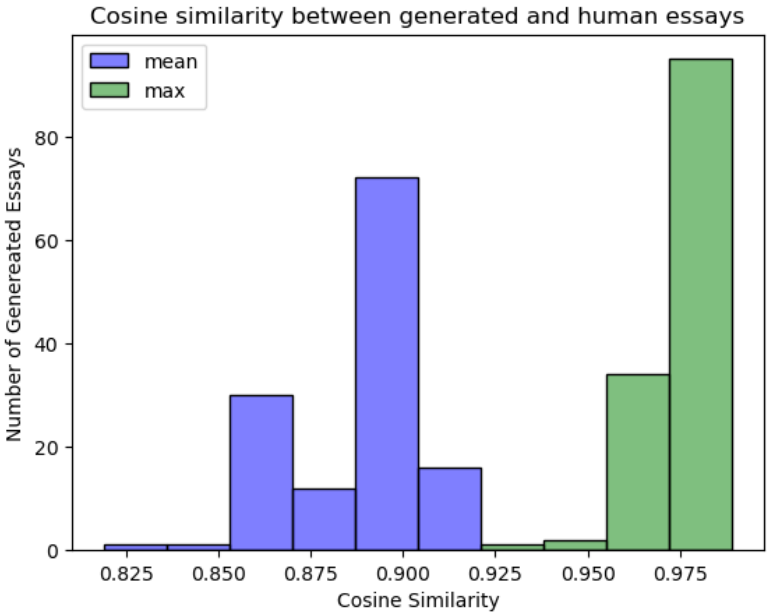
Below we present graphs resulting from the LLM detection models presented in the original notebook.



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 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.



Mean and maximum cosine similarity between each generated essay compared to all human essays. The similarities were computed using Word2Vec. Note that most generated texts are very similar on average with human essays (mean similarity), and almost all have at least one almost identical human counterpart (maximum similarity).