Artificial thesis detection

Aleksandr Makarov, Danila Kuklianov

UNIVERSITY OF TARTU
Institute of Computer Science

Introduction

It is common these days to rely on assistance of chat bots like Chat-GPT. They are helpful and convenient tools to simplify everyday life. However, this help can go hand in hand with abuse, especially when it comes to Bachelor's thesis work. Therefore, tools that can tell apart text written by humans and text written by a language model are needed. This project aims to address this problem by creating a dataset of thesis works from previous students in a specific curriculum. These theses were rephrased using the GPT-3.5 turbo model, and the transformer-based model was fine-tuned to determine whether the text was written by humans or by ChatGPT.

Dataset preparation

For dataset creation, the DSpace[1] portal was used. Works from Science and Technology Bachelor's Curriculum were collected, as we believe there is something "special" in the mindset of people that have chosen to enroll in this curriculum.

The files were automatically downloaded using Selenium and parsed using PYPDF2. The region of interest was chosen from the **Introduction** header and up to the **References** or **Bibliography** header.

The parsed text was then cut into 150 word fragments and send to the GPT-3.5 turbo using API[2] and a prompt "Rephrase to avoid plagiarism: {{ 150 words text }}".The final dataset contains a total of 2472 original text fragments and 2472 rephrased by GPT-3.5 turbo. The mean length of text samples is ≈150 for both human and GPT. The metrics on the number of words per sample are in Figure 1.

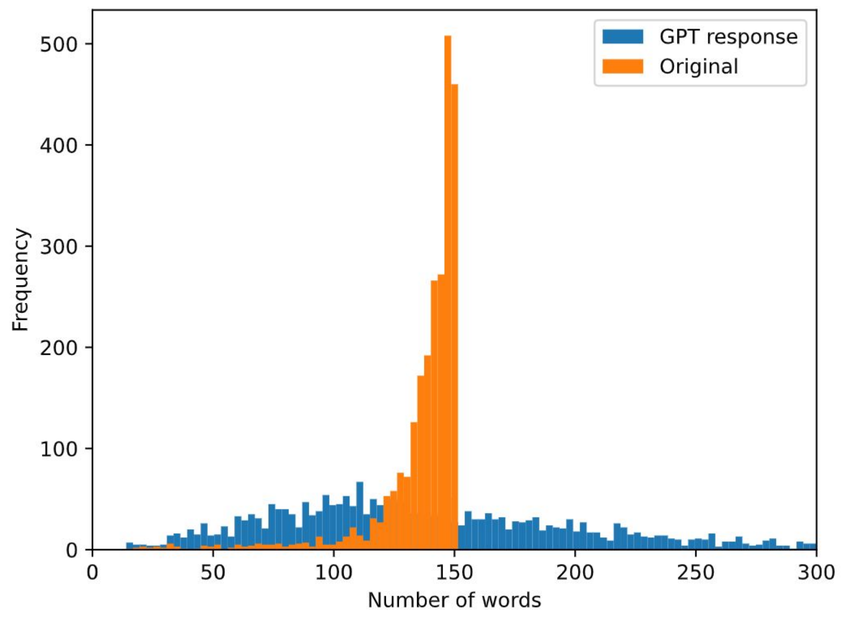

Figure 1. *Dataset distribution*

Training

The model in question is DistillBERT uncased model, which is perfectly balanced in terms of training time and performance. The model was rigged and trained to output a binary value - whether the text is human- or GPT-written. It was tested on the authors, as well as a few of their coursemates, even though everyone was encouraged to try. The model generally performs well, although the dataset required tweaking after SHAP pointed out that the model overused certain text features. The total workflow is depicted in Figure 2 and a web-demo is accessible via QR-code.

Explainability

In order to get insights into model's decision-making SHAP[4] explainability technique was used. It is perturbation based and assesses importance of each feature and is found incredibly important for model debugging. However, if the model is not running on GPU SHAP values take a lot of time to compute.

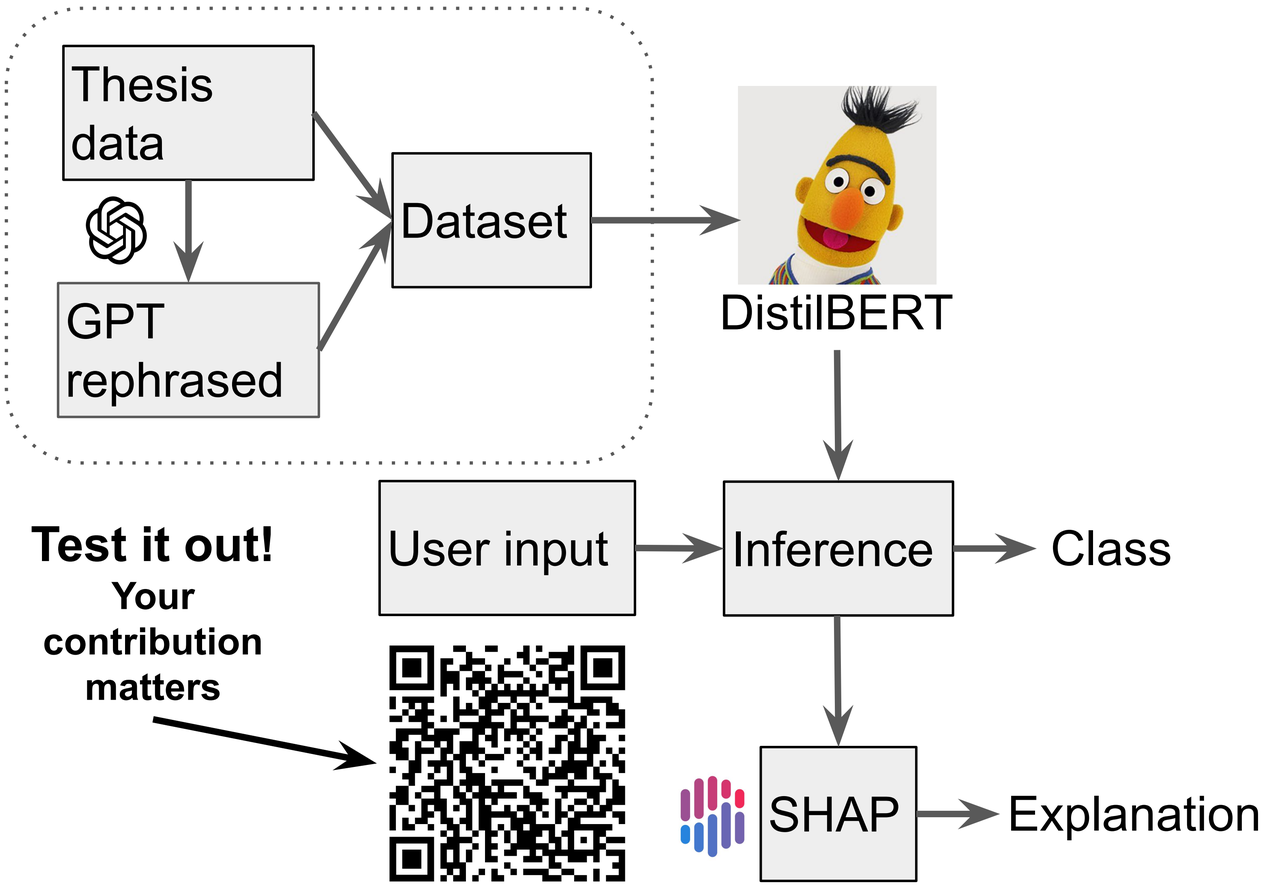

Figure 2. *The total workflow*

Results

All-in-all a valuable tool was developed for detection of GPT-written text fine tuned on Bachelor's work of a specific curriculum. The loss graph signifies that there might be an overfit (Fig. 3), while validation accuracy graph shows that there was not drop in accuracy(Fig. 4), hence overfitting is unlikely. Additionally, feedback from our coursemates was collected, although it was not very conclusive: some people claimed that it really worked, while for others it did not perform very well. As a result we arrived at the conclusion, that if it is a well-written, elaborate text without grammar issues model will most likely consider it GPT-written.

Most of the conclusions as well as project improvements were derived using SHAP method, and further directions would include improving the dataset and using other explainability techniques, like gradient-base explanations, which are claimed to be more accurate in NLP related tasks,

Figure 3. *Train & validation loss*

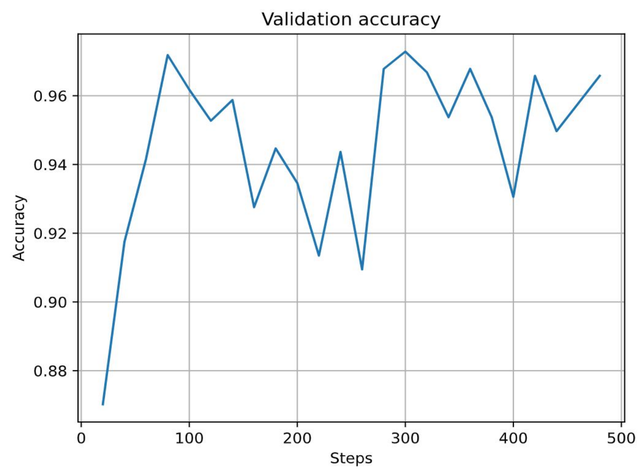

Figure 4. *Validation accuracy*

References:

[1] https://dspace.ut.ee/handle/10062/63912

[2] https://platform.openai.com/docs/guides/chat/introduction

[3] Sandra Mitrovi´c, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain.explaining decisions of machine learning model for detecting short chatgpt-generated text.

[4] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.