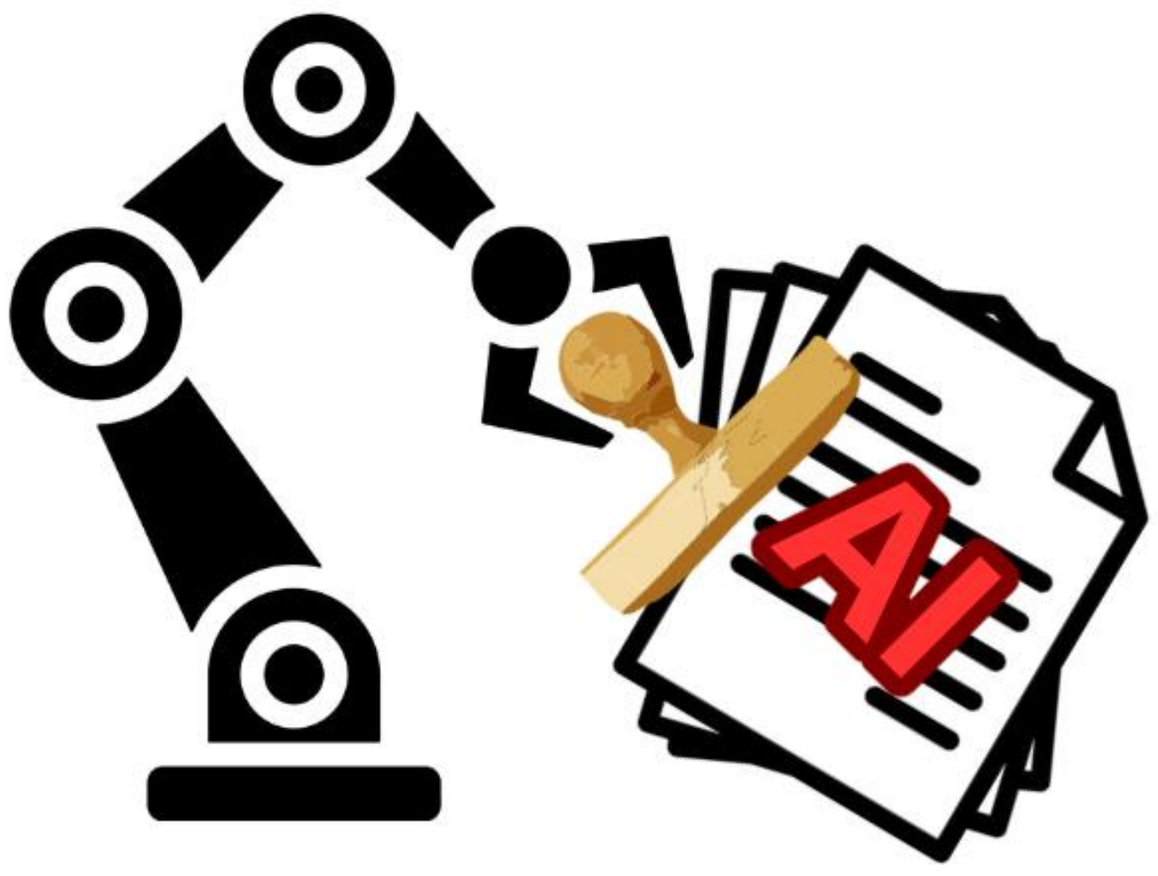
Detection of AI-generated Text

Bc. Matej Koreň

Supervisor: Ing. Radek Hranický, Ph.D.

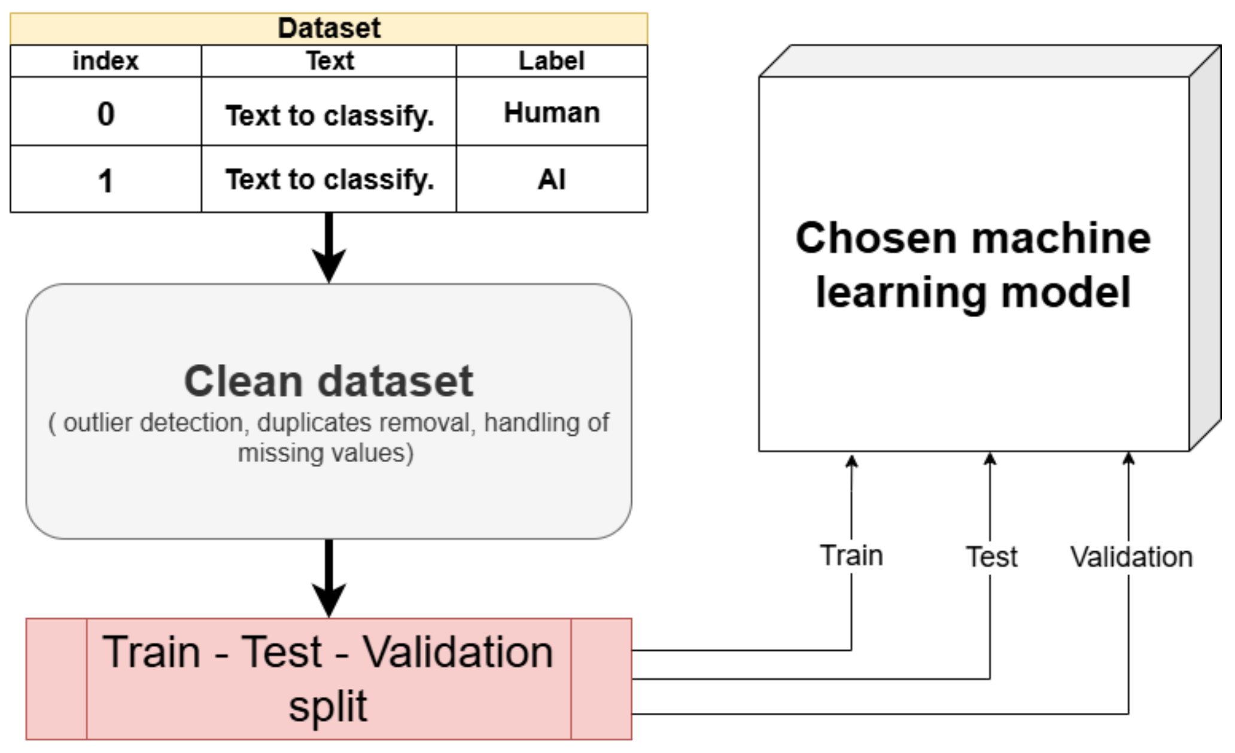
Motivation

The rise of AI-generated text presents new challenges in ensuring the authenticity of online content. Machine learning offers a powerful tool for detecting AI-generated text by analyzing linguistic patterns and inconsistencies that distinguish it from human writing. By identifying and flagging AI-generated content, we can maintain trust in digital communication and prevent the spread of misinformation. Using machine learning for this purpose helps protect the integrity of information and ensures a more transparent and reliable online space.



General idea

It is possible to find many online detectors that claim they are the most precise, but this is a very bold claim - in every machine learning task, the model is only as good as the training data. To tackle this problem, we found publicly available labeled datasets, as well as created one from the **Lakmoos AI** model responses, which we used to train, test and validate a variety of machine learning models, determining the most suitable one for this task.



Dataset		
index	Text	Label
0	Text to classify.	Human
1	Text to classify.	AI

Clean dataset
(outlier detection, duplicates removal, handling of missing values)
Train - Test - Validation split
Train
Test
Validation
Chosen machine learning model

Fig. 1.: Training pipeline visualisation.

Design

The project examines two machine learning techniques to text classification - a statistical (or **feature-based**) approach and the „modern“ way – using **transfromer based models**. To get the best of both worlds, a combined classifier was created, further improving the accuracy and confidence of these models. The application also has to be available online for enhancing the **Lakmoos AI** test, utilizing the solution for human-likeness scoring of model answers.

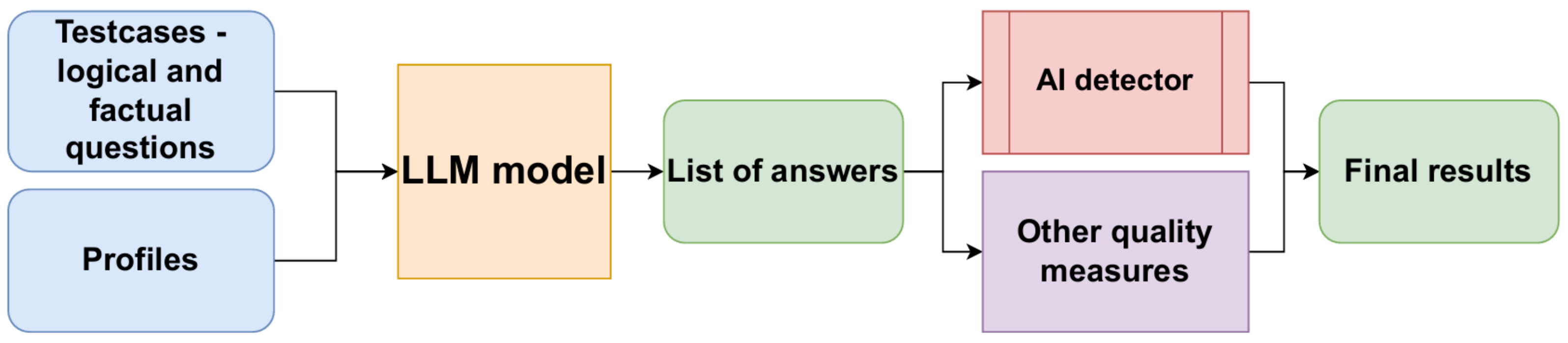


Fig. 2.: Integration to the existing test module.

Explanations

To improve interpretability of the classification decision, a **SHAP** explanation is available for both models. It shows the individual feature importance in the **XGBC** model as well as an interactive per-token influence visualisation of the **BERT** model decision on any given input text.

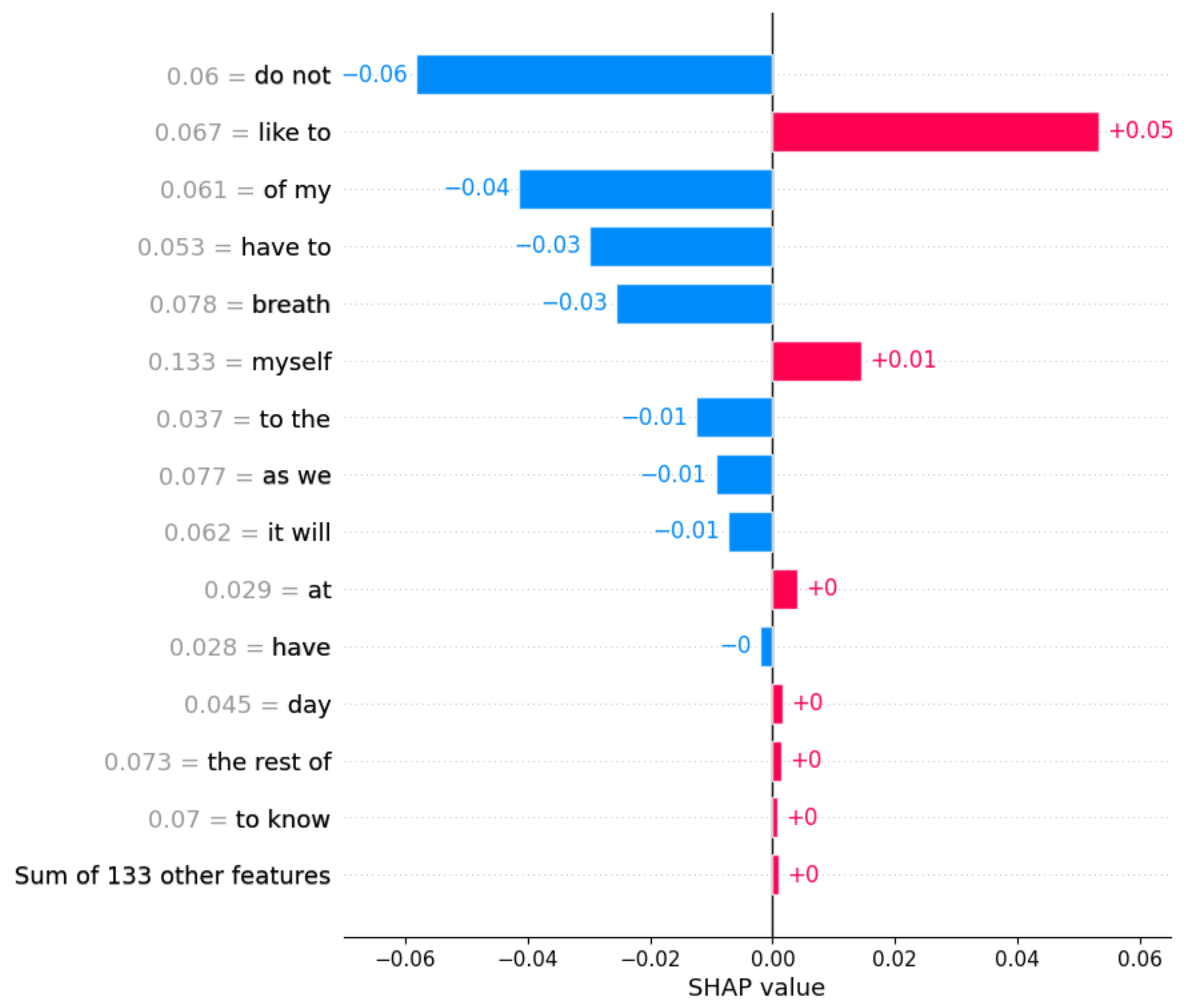
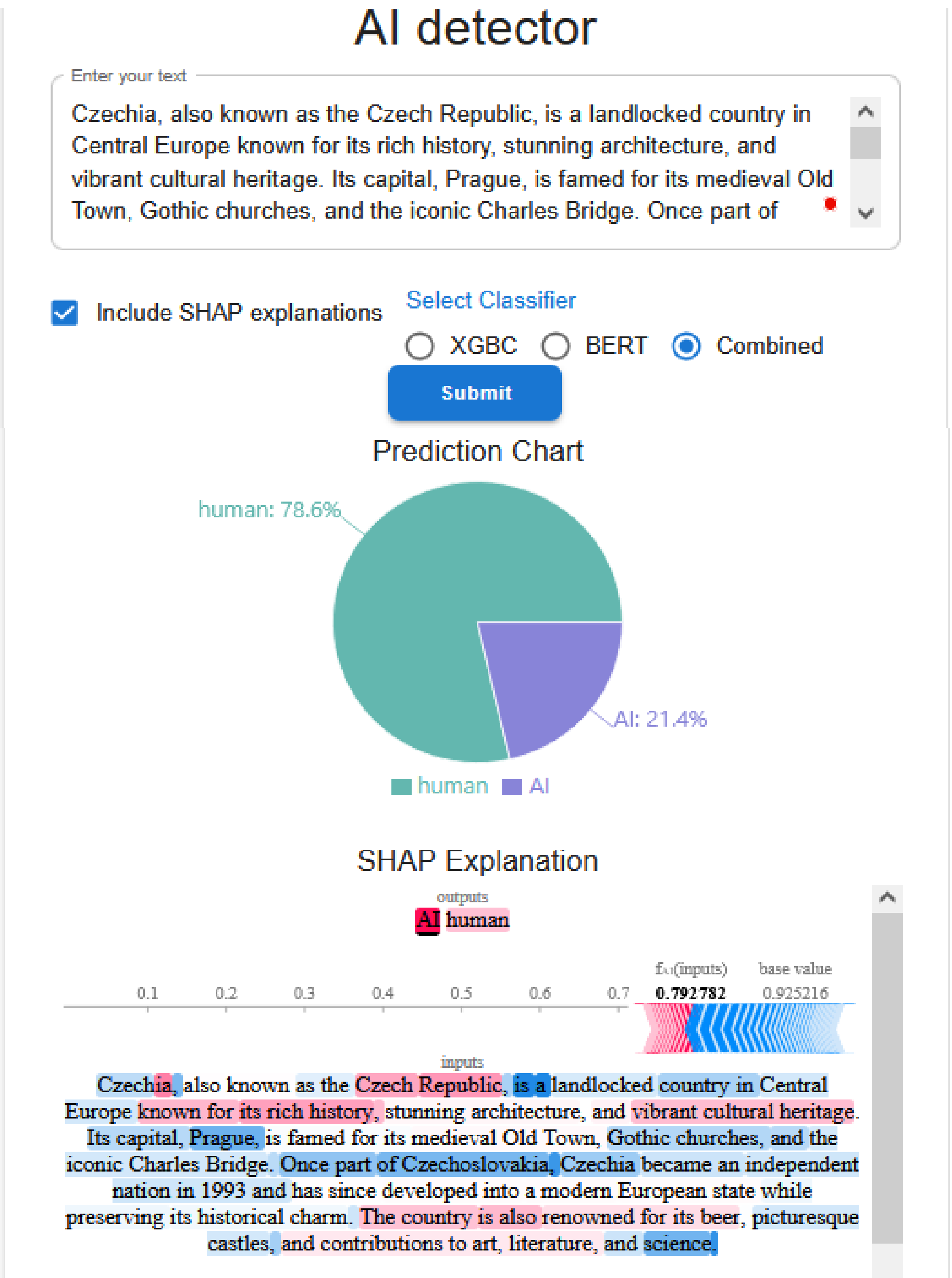


Fig. 3.: SHAP explanations

Implementation

With the use of a **Pycaret** library we have determined a few suitable feature based models such as **XGBoost** for the classification task. We also utilized a feature less **BERT** model for comparison of the feature-less approach. These models were then trained and put into a data pipeline, that was transformed into a **REST API** web application implemented in FastAPI and React for backend and frontend, respectively.



AI detector

Enter your text

Czechia, also known as the Czech Republic, is a landlocked country in Central Europe known for its rich history, stunning architecture, and vibrant cultural heritage. Its capital, Prague, is famed for its medieval Old Town, Gothic churches, and the iconic Charles Bridge. Once part of

☑ Include SHAP explanations

Select Classifier

○ XGBC ○ BERT ● Combined

Czechia, also known as the Czech Republic, is a landlocked country in Central Europe known for its rich history, stunning architecture, and vibrant cultural heritage. Its capital, Prague, is famed for its medieval Old Town, Gothic churches, and the iconic Charles Bridge. Once part of Czechoslovakia, Czechia became an independent nation in 1993 and has since developed into a modern European state while preserving its historical charm. The country is also renowned for its beer, picturesque castles, and contributions to art, literature, and science.

Fig. 4.: Web application screenshot

The final product is deployed online and available here:

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

lakmoos

Excel@FIT 2025