IMPROVING FAKE NEWS AI DETECTION ALGORITHMS THROUGH XAI TEHCNIQUES.

PROJECT AUTHOR: KACPER LIKUS

FIRST SUPERVISOR: DR DHUHA AL-SHAIKHLI

SECOND SUPERVISOR: DR MARYAM SHAHPASAND

Introduction

Fake news is a problem deeply rooted in todays highly interconnected digital landscape. To combat this problem, many automated systems have begun to arise, most notably, AI powered detection systems.

Problem

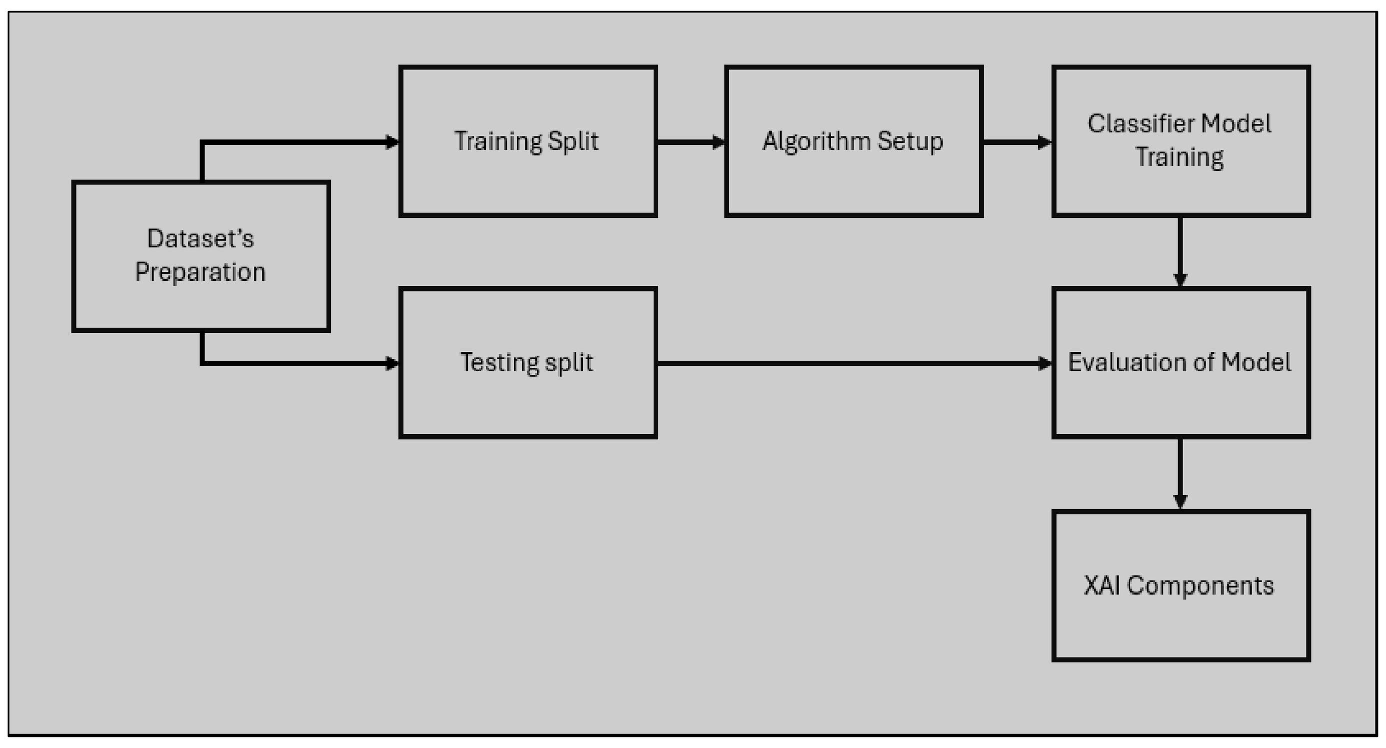
While AI deep learning classifiers often reach high 90% accuracies on different datasets of fake news, there is a large concern with a lack of explainability within these systems. In domains such as healthcare and politics, explained decisions are vital for trust.

Literature

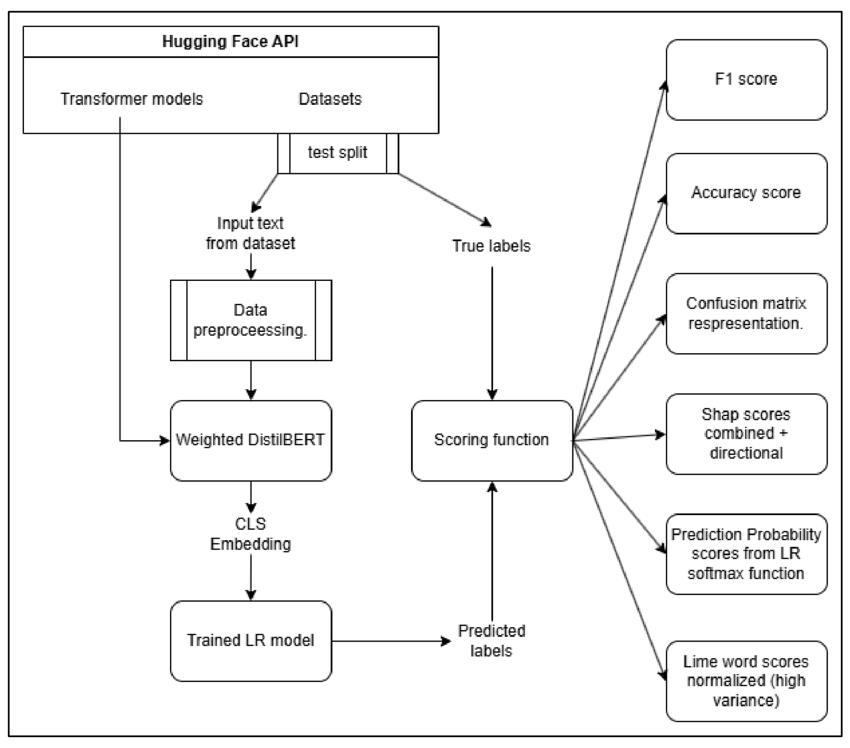
There are a variety of fake news detection approaches, from mixes of machine learning and deep-learning models, to fully functioning transformer models. While accuracy has improved, most studies focus heavily on classification performance, often overlooking explainability. Explainable AI (XAI) frameworks like SHAP and LIME which can shed light on model decisions.

Methodology

Based on the literature, a transformer-logistic classifier was developed to prioritise lightweight, explainable results. Numerous factors such as f1-score were used to get quantitative scoring out of the model in testing.



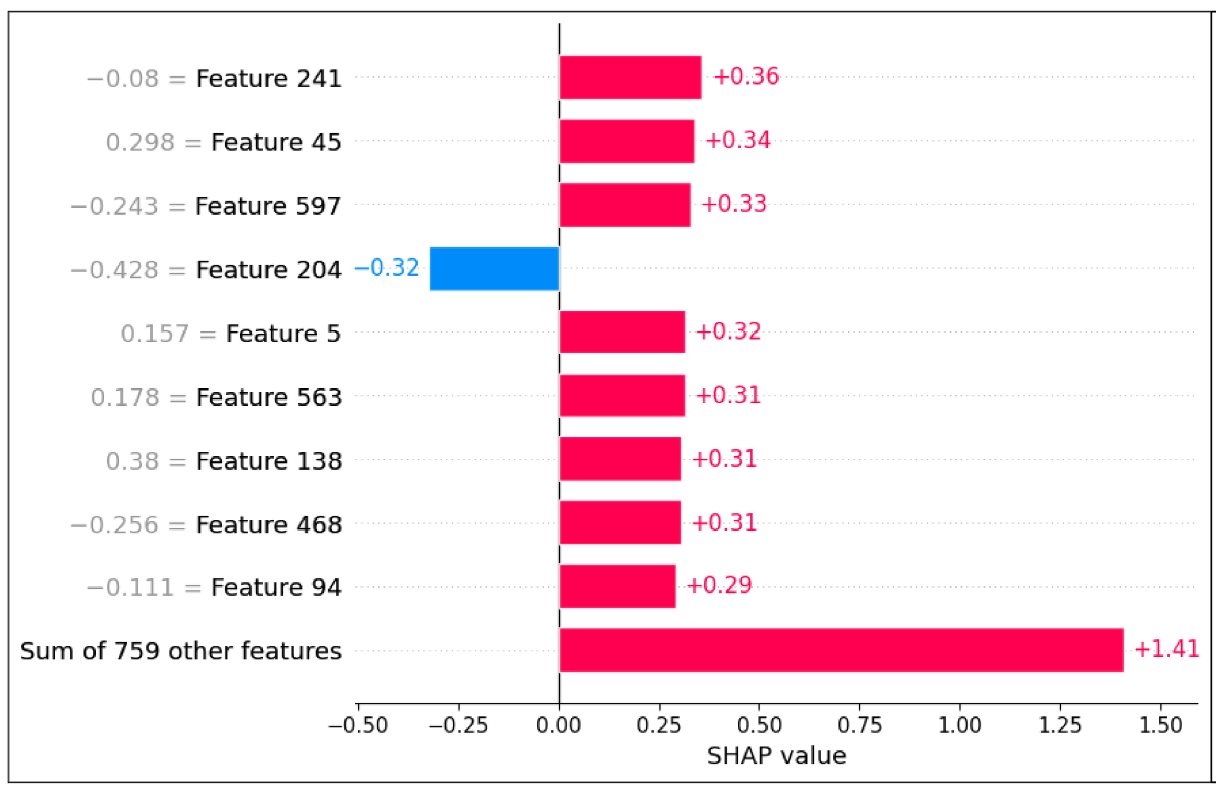
General Artefact Framework



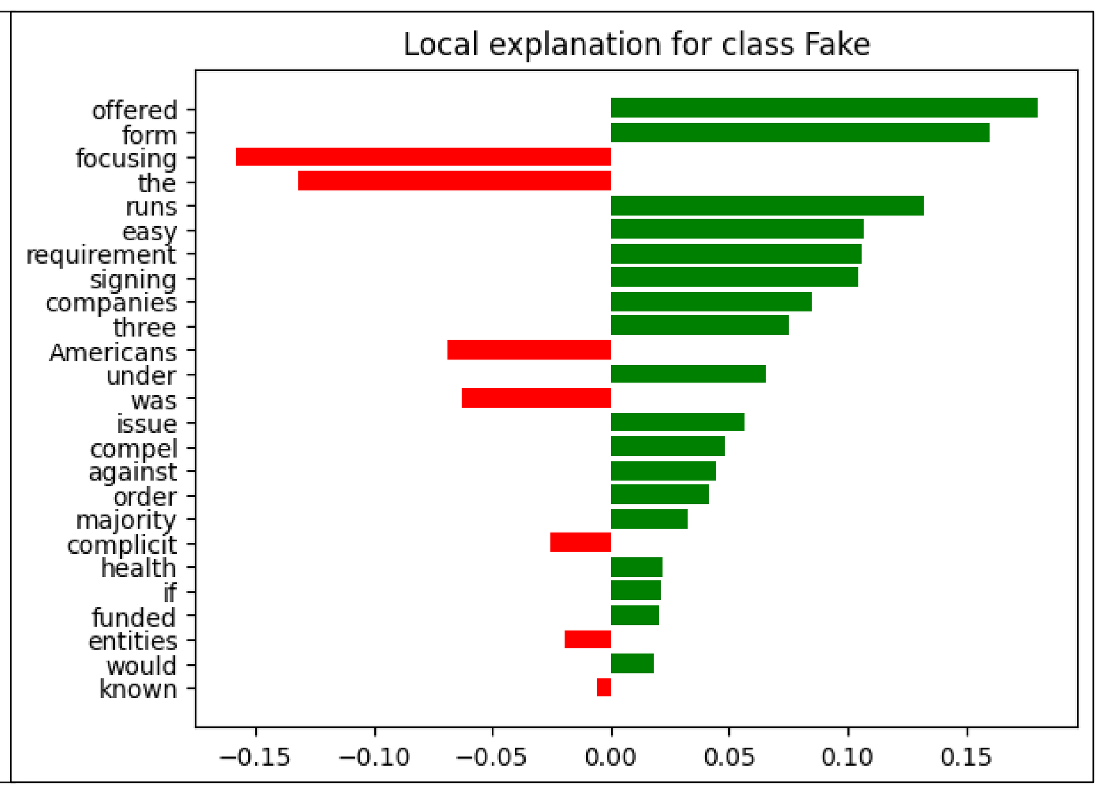
Technical Framework

Objectives

- Understand fake news detection landscape.
- Note most capable AI models for the task.
- Discover XAI techniques for classifiers.
- Build a simple prototype classifier to use on public datasets.
- Implement XAI techniques to gain visual interpretation of predictions.



SHAP output visual



LIME output visual

Analysis

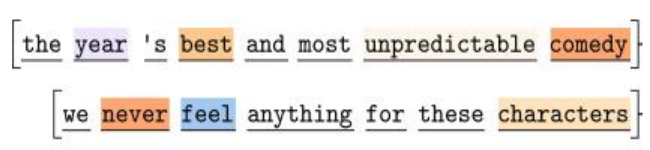
The lightweight model uses binary-labelled data and DistilBERT embeddings to classify fake news with strong results on clean input. While performance drops on complex datasets, SHAP and LIME successfully provide global and local interpretability. The prototype balances simplicity with insight, though deeper models could improve generalisation.

Component	Justification
DistilBERT	Fast, lightweight, strong embeddings
Logistic Regression	Interpretable, simple, works with SHAP
SHAP + LIME	Complements each other (global + local explanation)
Minimal cleaning	Avoid losing contextual meaning

Model Justifications

Model	Accuracy	Description
DistilBERT + LR	60%~	Deep embeddings to Logistic classifier.
TriFN	88%~	Composite result of Article, User and Publishers.
GCN	90%~	Leverages graph representation to link users, articles, publishers, shares, retweets, social connections etc…

Comparison against benchmark dataset



Example user facing approach

Conclusion

Fake news poses serious risks by spreading misinformation rapidly and eroding public trust. This project demonstrates how modern AI, combined with explainable tools like SHAP and LIME, can help classify and interpret fake news effectively. While limitations remain, the system offers a transparent foundation for safer and more trustworthy information filtering.

University of Staffordshire London