Towards Explainable AI-Generated Text Detection Using Ensemble and Combined Model Training

Hadi Mohammadi[1,*], Anastasia Giachanou[1] , and Ayoub Bagheri[1].

1. Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands

Introduction

Language model evolution has brought a new era of AI-generated text, reducing the distinctions between human and machine writing. While impressive, this advancement brings important challenges, especially in protecting the authenticity and integrity of digital content. Our research addresses this growing problem by developing a novel way of recognizing AI-generated material. The goal of this research is to develop an ensemble and mixed-model training strategy that improves not only detection accuracy but also explainability. This methodology is unique because it includes a broad range of advanced text classification algorithms and applies them to two languages, English and Dutch, and several genres such as newspaper text, tweets, reviews, poetry, and a mystery genre.

The significance of our work lies in its dual goal: to distinguish between human and AI-generated texts while also clarifying the decision-making process of the models involved. In an era when **explainable AI (XAI)** decisions are as important as their accuracy, our technique differs by incorporating *SHapley Additive exPlanations (SHAP)* into the model. This integration provides clear and understandable insights into the factors influencing its forecasts. As the digital landscape evolves with the development of AI-generated material, our study contributes significantly to guaranteeing the reliability and validity of digital information.

key words: AI-generated Text Detection, Ensemble Model, Explainable AI (XAI), Natural Language Processing (NLP), Transformer-Based Models.

Methodology

Data Sources: The CLIN33 shared dataset and the AuTexTification dataset, which contain over 160,000 texts in English and Dutch across five domains, were used [1].

Data Preprocessing: Steps include converting texts to lowercase, removing non-informative elements, and tokenization and lemmatization.

Data Augmentation: Techniques such as substitution, deletion, introducing spelling variations, back translation (English⇔Dutch) , and paraphrasing using AI models (*GPT-2*).

Addressing Class Imbalance: Using RandomOverSampler, SMOTE, and computing class weights for balanced training.

Experimental Setup: Use of the Adam optimizer, learning rate scheduler, early stopping mechanism, mixed-precision training, and a suite of transformers for text processing.

Model Training and Optimization: Description of the training process, including hyperparameter optimization, model architecture (BERT-based models), and evaluation metrics.

The ensemble model architecture: A sophisticated system that combines the outputs of multiple transformer-based models, including *bert-base-multilingual-uncased*, *xlm-roberta-base*, and *distilbert-base-multilingual-cased*. These individual models, each with unique language understanding strengths, are first trained on relevant datasets. Their outputs are then concatenated and fed into a dense layer for final binary classification.

A notable aspect of this architecture is the use of both frozen and fresh models. The frozen models retain their learned weights, preserving their specialized knowledge, while the fresh models are trained again, allowing adaptability to the current dataset. This combination enhances the model's accuracy and generalization ability.

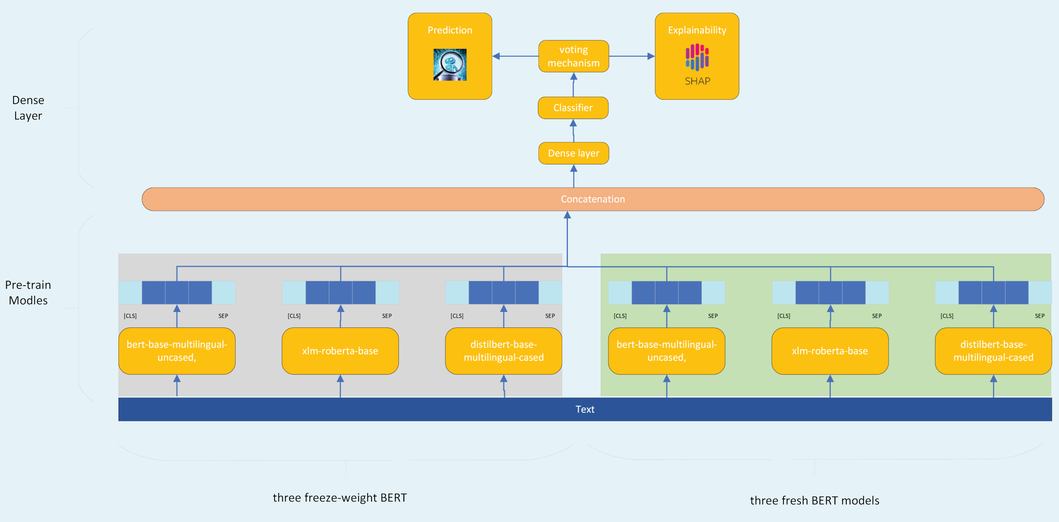
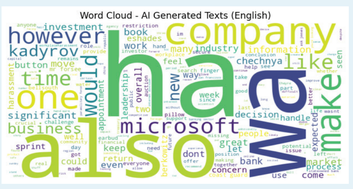


Table 1: summary of model hyperparameters

Parameter	Description
Tokenization Max Length	256 tokens
Learning Rate Range	1e-5 to 1e-4 (Default: 3e-5)
Batch Sizes	16, 32, 64
Learning Rate Scheduler	Cosine decay schedule
Warmup Steps	200 steps
Early Stopping Patience	3 epochs
Loss Function	Binary cross-entropy
Optimizer	Adam
Precision Training Policy	Mixed float16



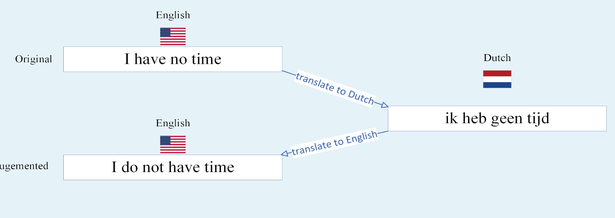


Figure 1: Model Architecture Visualization (left), Word Cloud and Back transation Example (right)

Results

- **Research Contributions and Results:**
  - Developed a custom model combining various *BERT* versions with both frozen and fresh models.
  - Captures relationships between pretrained model outputs using Dence layer.
  - Enhances robustness, especially for multilingual challenges.
  - Used *SHAP* for result transparency.
  - Better capture AI genetated text (TN) on Dutch
- **Limitiotion:**
  - limition on disticnt human form AI text from EDA analysis
  - limition on disticnt human form AI text on new genre
- **Future Research Direction:**
  - Plan for in-depth analysis of model components.
  - Objective: Improve explainability of large language models.
  - Aim for models that excel across datasets and align with human thought.

Table 1: Performance in different genre / language

Genre	newspaper	tweets	reviews	New (poetry, and mystery)
Lnaguage	**Englsih**			
Accuracy	0.9750	0.9600	0.8150	0.7667
F1 Score	0.9750	0.9600	0.8121	0.7604
Lnaguage	**Dutch**			
Accuracy	0.9250	0.9600	0.8400	0.7500
F1 Score	0.9247	0.9600	0.8400	0.7350

A) word, unique word and character count , average word length distribution B) Boxplot by language and domain

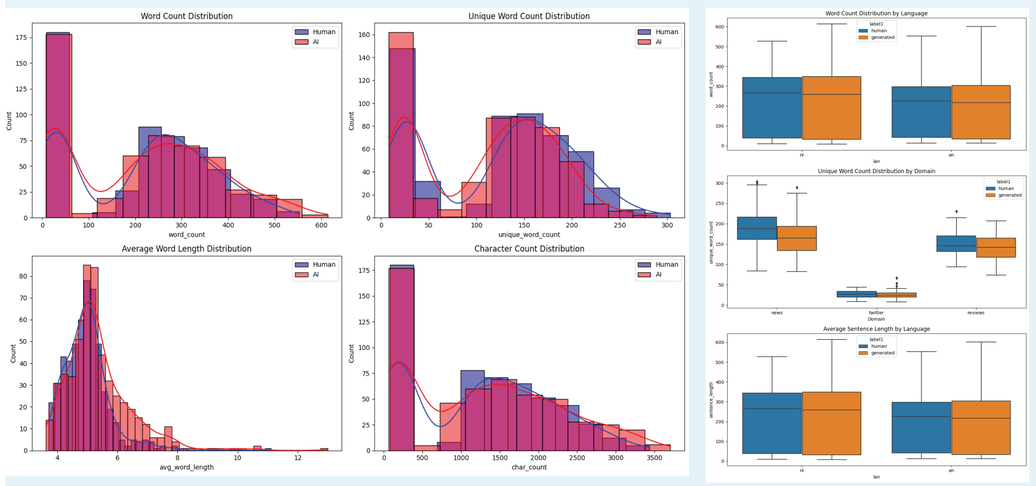


Figure 2: Exploratory Data Analysis for Human / AI generated texts

B) Confusion Matrix for different languages and domains

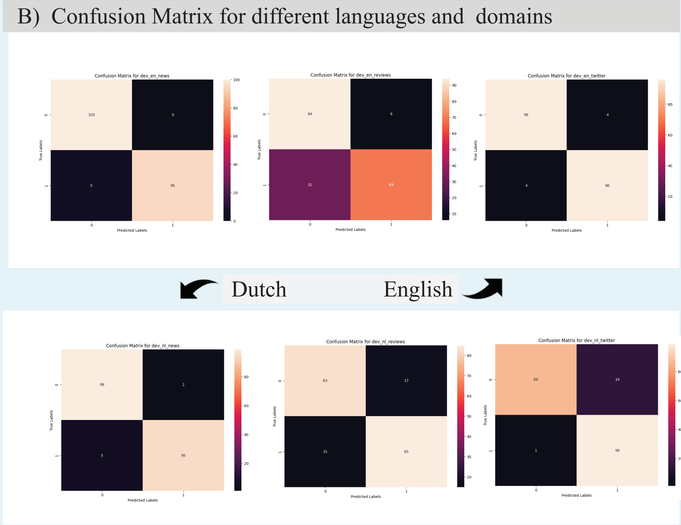
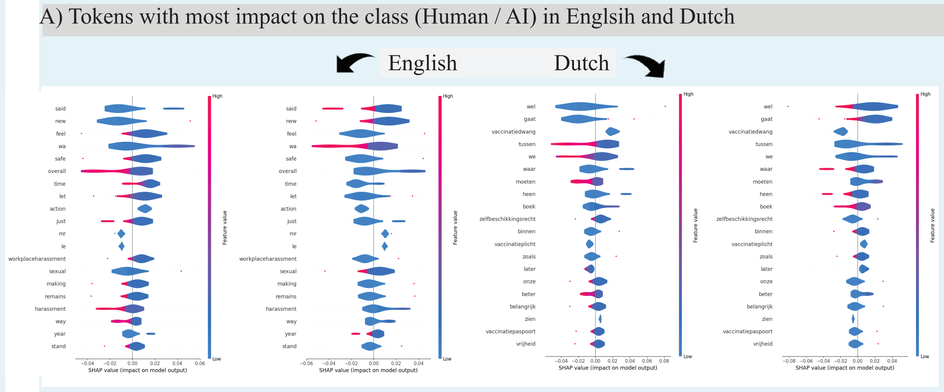


Figure 3: Performance for the model on test data-set

A) Tokens with most impact on the class (Human / AI) in Englsih and Dutch

B) Examples of effects of each tokens on the class (top: human / down: AI)



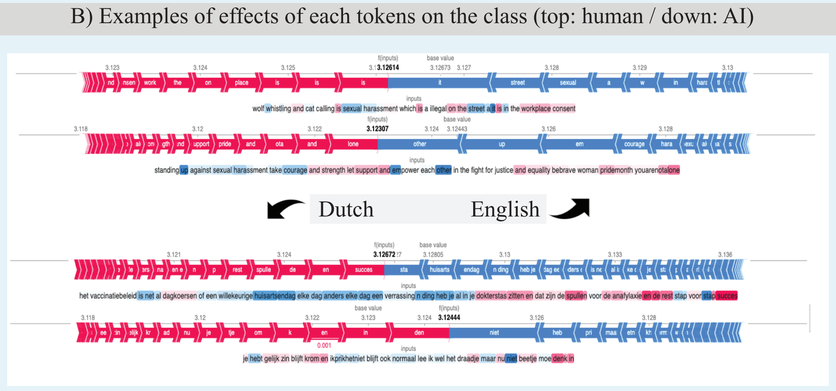


Figure 4: Using Explaniable AI (SHAP) to find out the Most Effective Factors (Left) and Two Examples (Right)

Conclusion

Our research advances AI-generated text detection by showing that an ensemble model architecture that mixes various transformer-based models works. Compared textual features demonstrated patterns and traits that identify human and AI-generated literature. Merging datasets from different languages and areas helps researchers understand text generation's complexities. This strategy improves detection accuracy and raises questions about AI transparency and trustworthiness in digital content verification. This study advances AI-generated text detection algorithms to a higher level of sophistication, accuracy, and explainability, opening the way for digital content authenticity research and applications.

*Corresponding Author: h.mohammadi@uu.nl

Universiteit Utrecht

[ΨΣ] interuniversity graduate school of iops Psychometrics and Sociometrics

[1] the dataset utilized is specifically associated with AuTexTification: Automated Text Identification competition and CLIN33 shared task .

AUTEXTIFICATION

CLIN 33