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 Nethertlands

Introduction

Our research tackles the challenge of distinguishing between human and AI-generated text, a crucial issue in the era of advanced language models. We propose a unique approach using an ensemble and mixed-model strategy, focusing on accuracy and explainability. This method involves a variety of advanced text classification algorithms, applied to both English and Dutch texts across multiple genres. Notably, our work integrates *SHapley Additive exPlanations (SHAP)* for clearer insights into model decisions, emphasizing the importance of **explainable AI (XAI)**. Our study is significant in ensuring the authenticity and integrity of digital content in an increasingly AI-driven world.

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

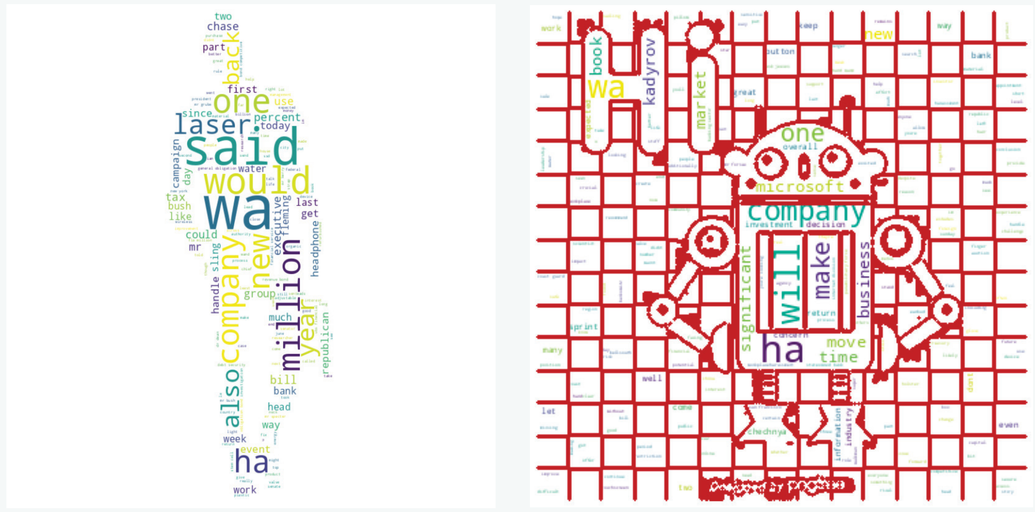


Figure 1: Word Clouds with text generated by human (left) and AI (right)

Methodology

Data Sources: The the AuTexTification dataset and CLIN33 shared dataset, which contain over 160,000 texts in English and Dutch across five domains, were used [1, 2].

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.

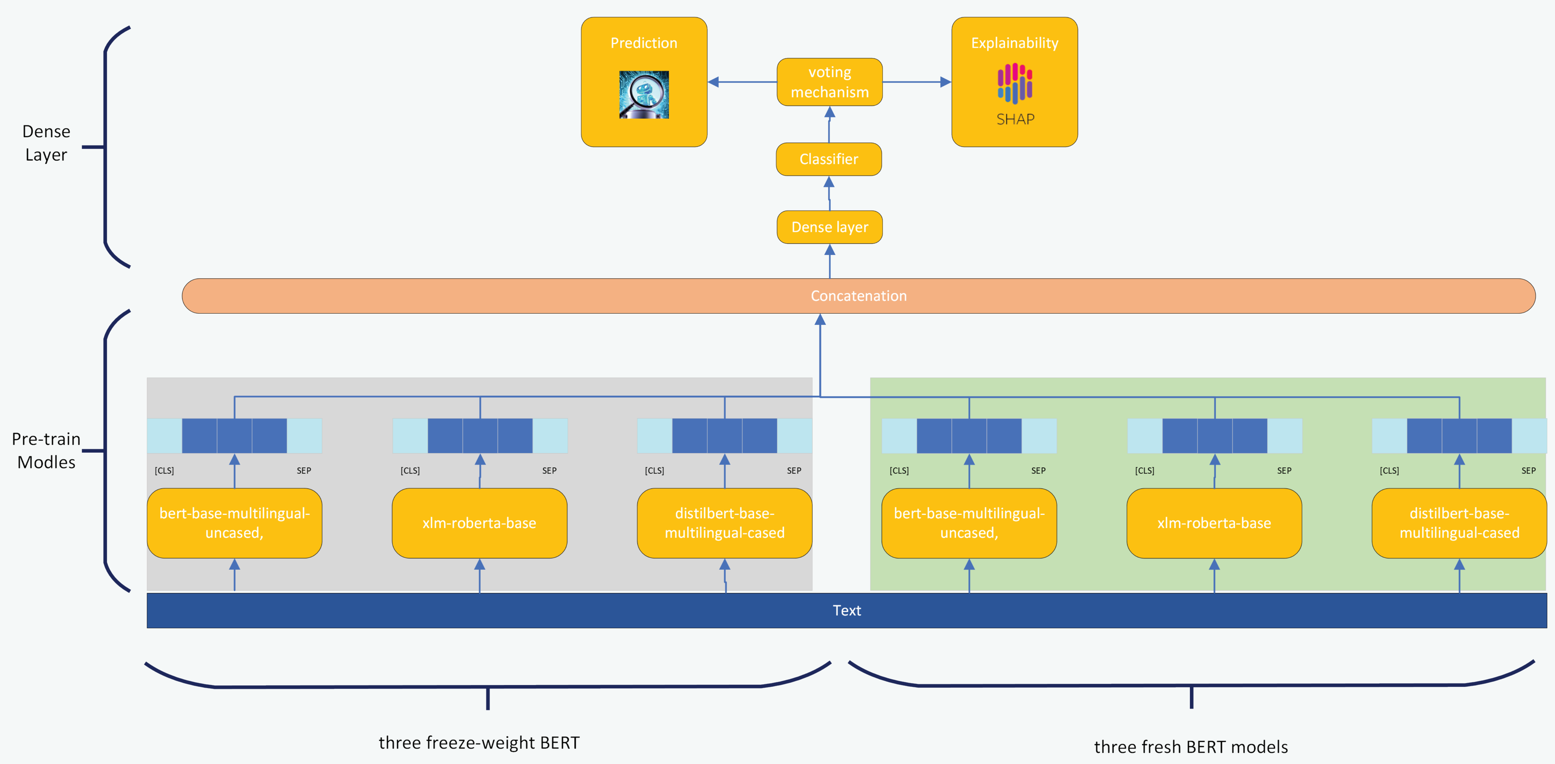
Experimental Setup: Use of the Adam optimizer, learning rate scheduler, early stopping mechanism, mixed-precision training, and a suite of transformers for text processing.

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

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

Ensemble Model Architecture: ombines outputs from several transformer-based models like *bert-base-multilingual-uncased*, *xlm-roberta-base*, and *distilbert-base-multilingual-cased*. Initially, these models are fine-tuned on the AuTexTification dataset. After this, their weights are frozen, and they are combined with freshly fine-tuned models on a training task dataset. The outputs of all models are merged and passed through a dense layer for binary classification. The architecture uses a mix of frozen models (to retain specialized knowledge) and freshly trained models (for adaptability), enhancing accuracy and generalization. A voting mechanism aggregates the predictions from each model to ensure robust and balanced detection, particularly effective in distinguishing between human and AI-generated text.



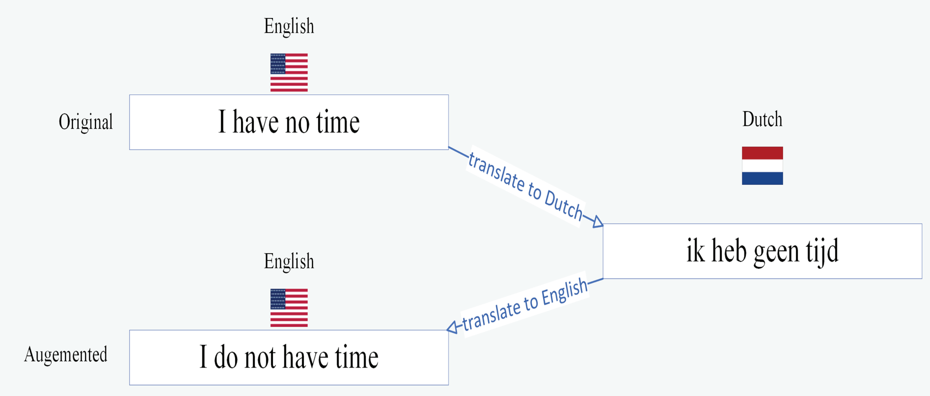


Figure 2: Model Architecture visualization (left), and Back transation Example (right)

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.
  - *SHAP* was used to improve result explainability and transparency.
  - The model is better at capturing AI-generated text (**TN**) in Dutch
- **Limitiotions:**
  - Limitation on distinguishing human from AI text from EDA analysis.
  - Limitation on distinguishing human from AI text in new genres.
  - Limitation on using the Dutch BERT version (like *BERTje*).
- **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 2: 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

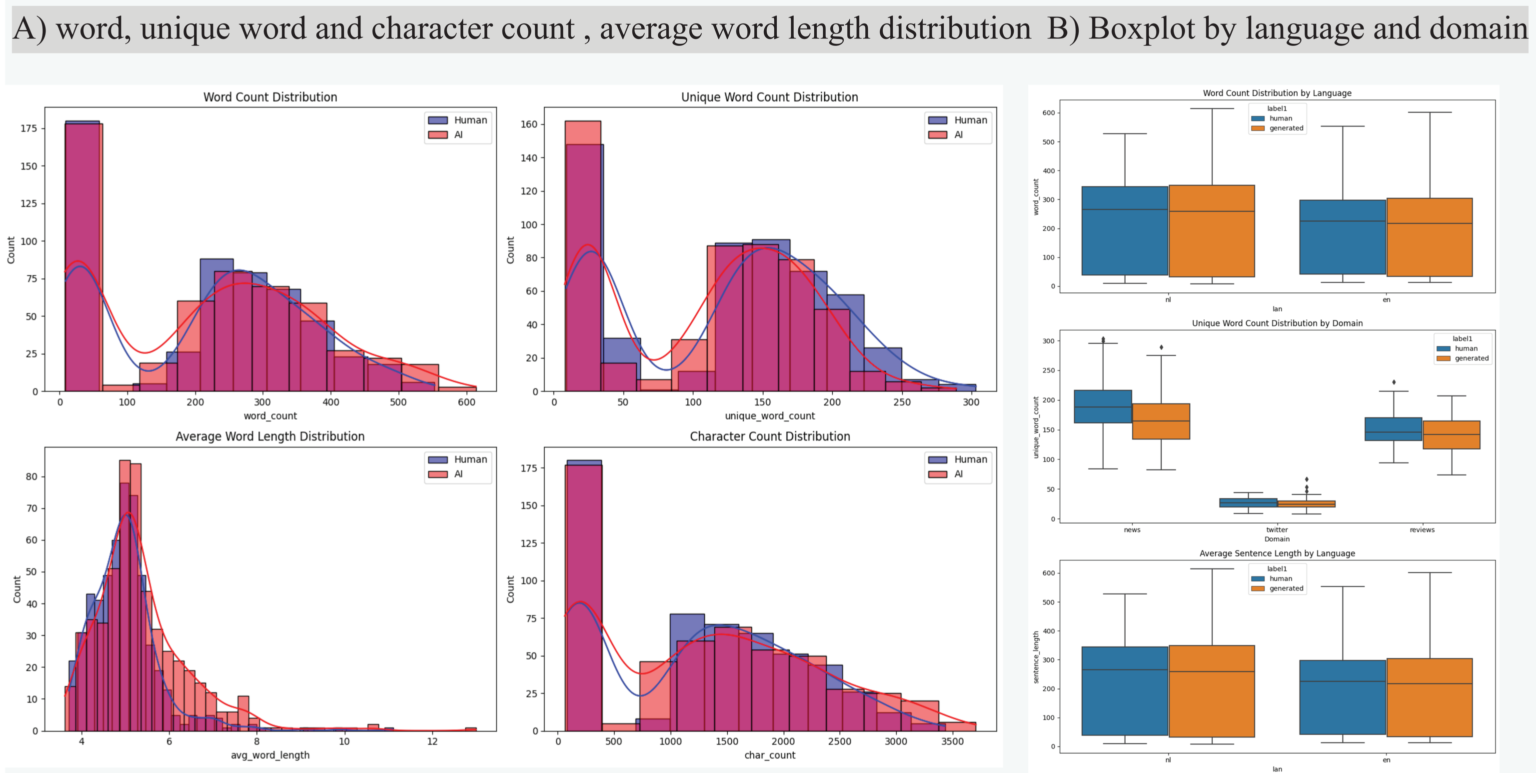


Figure 3: Exploratory Data Analysis for Human / AI generated texts

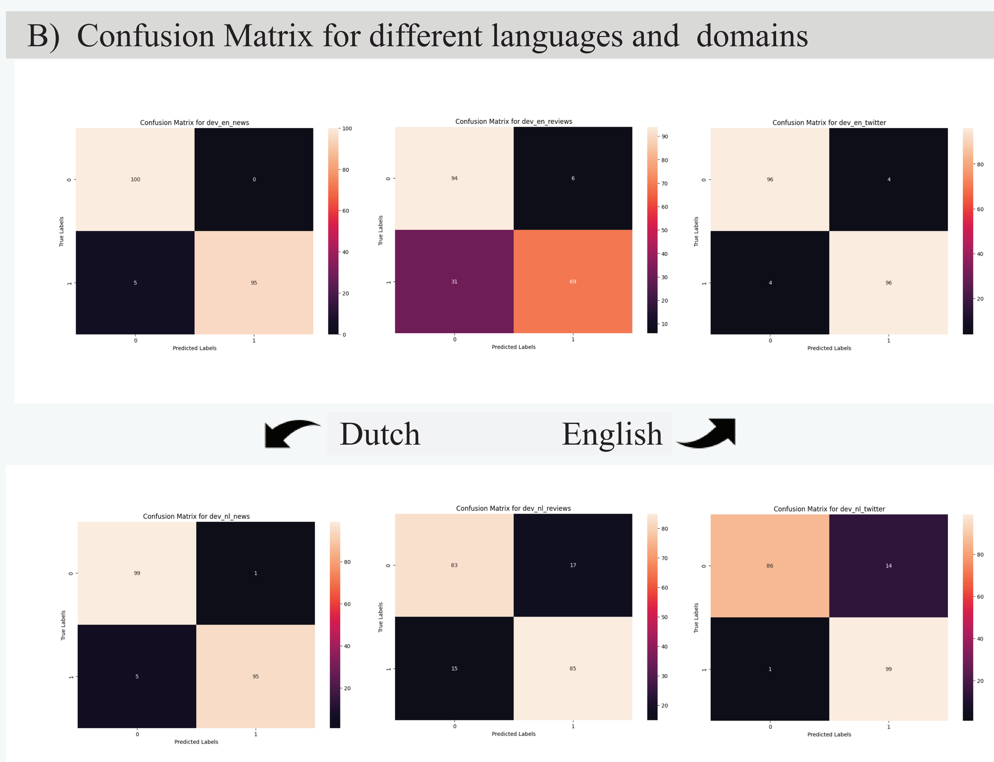


Figure 4: Performance for the model on test data-set

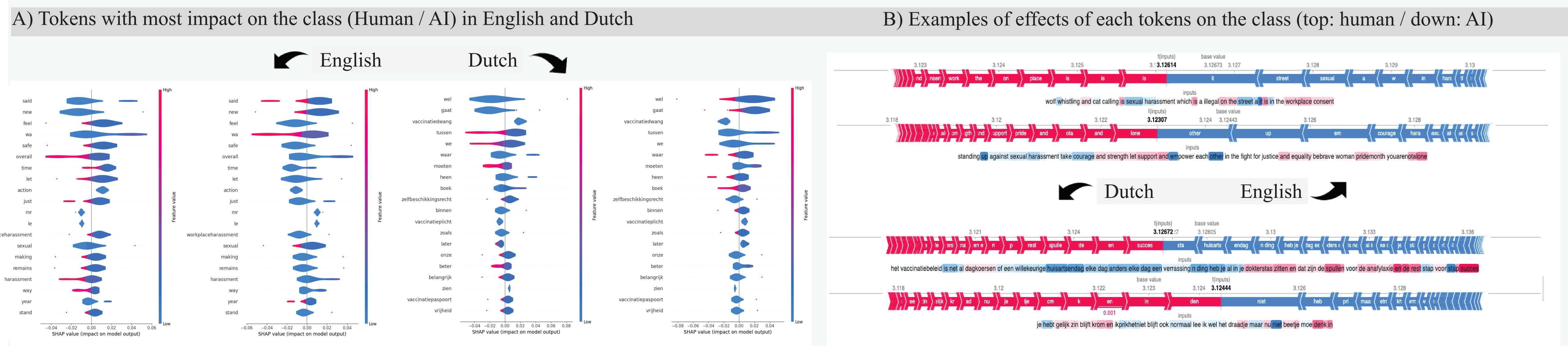


Figure 5: Using Explaniable AI (SHAP) to find out the Most Effective Factors (Left) and Two Examples (Right)

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

[1] https://sites.google.com/view/autextification
[2] https://sites.google.com/view/shared-task-clin33/home