KINGDOM OF SAUDI ARABIA
Ministry of Higher Education
Taibah University
College of Computer Science and Engineering (Boys Section)

المملكة العربية السعودية
وزارة التعليم العالي
كلية علوم و هندسة الحاسب الآلي
(قسم الطلاب)

Detection of AI-Generated Arabic Text: A Data Mining Approach

Final Report

Mohammed Alnajjar – 4714245

MSIS822-M4

Supervised by

Dr. Mohammed Alsarem

1.Abstract

The increasing use of large language models has led to a growing presence of AI-generated text, raising concerns about authorship authenticity, particularly for Arabic. This study addresses the detection of AI-generated Arabic text using interpretable linguistic features and traditional machine learning models. A dataset of academic abstracts was compiled and labeled as human-written or AI-generated. Four linguistic features were extracted: Honoré's R measure, noun count, genitive construction frequency, and entity density.

Logistic Regression, Support Vector Machine, and Random Forest classifiers were evaluated using a stratified data split. Results show that Random Forest outperforms the other models, achieving the most balanced performance. Feature importance analysis identifies entity density as the most influential feature, highlighting differences in named entity usage between human and AI-generated text. The findings demonstrate that shallow linguistic features provide an effective and interpretable baseline for Arabic AI-text detection.

Link githup for code:

2.Introduction

The rapid advancement of large language models (LLMs) has significantly increased the availability of AI-generated text across various online platforms. While these models offer substantial benefits in automation, content generation, and accessibility, they also introduce challenges related to academic integrity, misinformation, authorship verification, and digital trust. As AI-generated text becomes increasingly indistinguishable from human writing, the need for reliable detection methods has grown more urgent—especially for under-studied languages such as Arabic, where limited annotated datasets and fewer linguistic tools hinder progress.

This project aims to address this gap by developing a machine-learning–based framework for detecting AI-generated Arabic text using a set of handcrafted linguistic and statistical features. Instead of relying on deep neural models or large-scale embeddings, the approach focuses on extracting interpretable text features, including Honoré's R measure, noun frequency, genitive constructions, and entity density. These features capture stylistic and syntactic properties that differ between human and AI-generated writing.

The primary objectives of this project are:

1. **To design and implement a feature extraction pipeline for Arabic text** using Stanza NLP and custom linguistic functions.
2. **To build and compare traditional machine-learning classifiers**—including Logistic Regression, Support Vector Machine (SVM), and Random Forest—trained on the extracted feature set.
3. **To evaluate these models using comprehensive metrics**, such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.
4. **To identify which linguistic features contribute most** to distinguishing AI-generated text from human-authored text.

Through this research, we demonstrate that interpretable linguistic features, combined with traditional machine-learning algorithms, can achieve competitive performance in AI-text detection for Arabic—offering a lightweight and transparent alternative to more complex deep-learning approaches.

3.Related Work

The rapid rise of large language models (LLMs) such as GPT, LLaMA, and PaLM has motivated extensive research on detecting AI-generated text. Existing detection approaches generally fall into three categories: **statistical feature–based methods**, **neural network–based detectors**, and **watermarking techniques**.

1. Statistical and Linguistic Feature–Based Methods.
Early work on text detection relied on handcrafted statistical features such as perplexity, burstiness, part-of-speech (POS) distributions, and vocabulary richness. Measures like Honoré's R, type–token ratio (TTR), and lexical diversity have been used to differentiate human writing from machine-generated content. Classical machine-learning classifiers—including Logistic Regression, Naïve Bayes, and Random Forest—have shown strong performance when combined with interpretable linguistic features. However, most studies focus on English, with very limited work on Arabic linguistic cues such as genitive constructions (,(إضافة noun usage patterns, or named-entity sparsity.

2. Neural Network–Based Detectors.
More recent approaches employ deep learning, often based on transformers.

Models like GPT-Detect, RoBERTa-based classifiers, and fine-tuned BERT variants attempt to distinguish AI-generated text using contextual embeddings. While these models often achieve high accuracy, they have important drawbacks: high computational cost, sensitivity to paraphrasing attacks, reduced interpretability, and poor cross-model generalization. Moreover, pretrained Arabic transformer detectors are scarce, which limits their applicability to the Arabic language.

3. Watermarking and Generation-Time Techniques.
Another direction focuses on embedding statistical watermarks into the output of LLMs during generation. These methods modify token sampling distributions to leave detectable patterns in the generated text. Although watermarking can be effective, it is not always implemented by model providers and can be defeated through paraphrasing, translation, or text corruption. Additionally, watermarking is not applicable to existing unlabeled text where the generation process is unknown.

Despite these advances, **research on AI-text detection for Arabic remains limited**. Existing work is constrained by the scarcity of labeled datasets and the complexity of Arabic morphology and syntax. This motivates the need for lightweight, interpretable, and language-aware approaches. The present project builds on the feature-based detection paradigm but adapts it to Arabic by incorporating noun frequency, genitive constructions, Honoré's R, and entity density—features that capture stylistic differences between AI-generated and human-written Arabic.

4.Dataset Description

This project uses the **Arabic Generated Abstracts** dataset published by **KFUPM–JRCAI Lab**, available publicly on HuggingFace. The dataset is designed to support research on distinguishing human-written Arabic text from AI-generated text, particularly in academic writing domains.

Source and Composition

The original HuggingFace dataset consists of five separate datafiled, each containing abstracts generated from different sources:

1. **Human-written Arabic abstracts**
2. **AI-generated abstracts using GPT models**
3. **AI-generated abstracts using LLaMA-based models**
4. **AI-generated abstracts using other ALLAM**
5. **AI-generated abstracts using JIAS**

For this project, all four subsets were **combined into a single unified dataset**, and a binary classification label was created:

- **"ai_generated"** for all machine-generated abstracts
- **"human"** for all human-written abstracts

This restructuring allows for a clear **binary classification task**: Human vs. AI-generated text.

Dataset Size

After merging the datasets and applying preprocessing steps, the final dataset contains:

- **Total samples: 41,940**
- **AI-generated abstracts: 33,552 samples**
- **Human-written abstracts: 8,388 samples**

This shows that the dataset is **imbalanced**, with AI-generated text representing nearly **80%** of the data. This imbalance is important to consider when evaluating model performance.

Data Fields

The final dataset contains two main columns:

- **norm_text** – The normalized textual content of each abstract
- **human vs. AI_generated** – A binary label indicating whether the text is human-written or AI-generated

(Additional features such as Honoré's R, noun count, entity density, and genitive count were extracted later as part of the feature engineering process.)

5.Methodology

The proposed methodology focuses on interpretable linguistic features and traditional machine-learning models for detecting AI-generated Arabic text.

1. Text Preprocessing

All text samples were normalized prior to feature extraction to reduce orthographic variation and noise common in Arabic text. The preprocessing pipeline included:

- **Normalization of Arabic characters**, unifying different forms of Alif (إ، أ، آ، ا) into a single form (ا).
- **Conversion of Alif Maqsura (ى) to Ya (ي)** to maintain lexical consistency.
- **Removal of Tatweel (-)** and Arabic diacritics (ً ٌ ٍ َ ُ ِ ّ ْ).
- **Filtering non-Arabic characters**, retaining only Arabic letters, digits, and whitespace.
- **Whitespace normalization**, collapsing multiple spaces into a single space.

These steps ensure consistent tokenization and reliable downstream linguistic analysis.

2. Feature Engineering

Instead of relying on high-dimensional text embeddings, this work extracts a compact set of linguistically motivated features designed to capture stylistic differences between human-written and AI-generated Arabic text.

2.1 Honoré's R Measure

Honoré's R is a lexical richness metric that quantifies vocabulary sophistication based on word frequency distributions. It is computed using the total number of words, the number of unique words, and the number of words appearing only once. This feature helps capture repetitive or uniform vocabulary patterns commonly observed in AI-generated text.

2.2 Noun Count

The total number of nouns in each text is computed using part-of-speech tagging via the Stanza Arabic pipeline. Both common nouns (NOUN) and proper nouns (PROPN) are counted. This feature reflects differences in informational density and syntactic structure.

2.3 Genitive Construction Count (Idafa)

Arabic genitive constructions (إضافة) are approximated by identifying consecutive noun–noun sequences within a sentence. These constructions are frequent in formal and academic Arabic writing and may appear with different patterns in AI-generated text.

2.4 Entity Density

Entity density is defined as the ratio of named entities to the total number of words in a text. Named entities are extracted using Stanza's Arabic Named Entity Recognition (NER) module. This feature captures how often specific entities (e.g., organizations, locations, persons) are referenced relative to text length.

All extracted features are numeric and form a low-dimensional feature vector for each text sample.

3. Model Selection and Training

Three traditional machine-learning classifiers were trained and compared:

- **Logistic Regression**, serving as a linear baseline model.
- **Support Vector Machine (SVM)** with a radial basis function (RBF) kernel to model non-linear decision boundaries.
- **Random Forest**, an ensemble of decision trees capable of capturing complex feature interactions.

The dataset was split into **training (70%)**, **validation (15%)**, and **test (15%)** sets. Hyperparameters were tuned using the validation set. Feature scaling was applied where required (e.g., for SVM), while tree-based models operated on raw feature values.

4. Evaluation Metrics

All final models were evaluated on a held-out test set using multiple metrics to provide a comprehensive assessment:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **ROC–AUC**

Additionally, **confusion matrices** were generated to visualize classification errors and class-wise performance.

5. Feature Importance Analysis

For the best-performing model (Random Forest), feature importance scores were extracted to analyze which linguistic features contributed most to distinguishing AI-generated text from human-written text. This analysis enhances the interpretability of the model and provides insight into stylistic differences between the two classes.

6.Results and Analysis

1. Overall Model Performance

Three classifiers—Logistic Regression, Support Vector Machine (SVM), and Random Forest were evaluated on a held-out test set. Table 1 summarizes the performance of each model across standard evaluation metrics.

Table 1: Model Performance on the Test Set

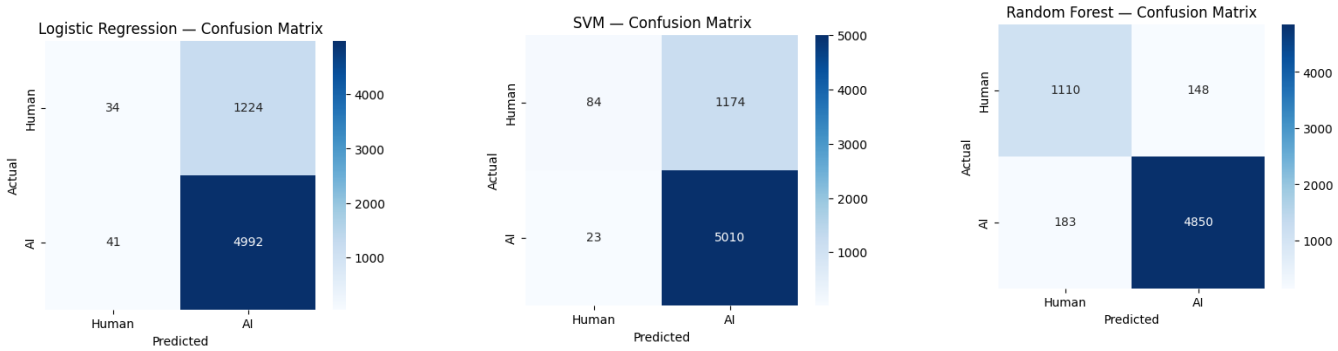
Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
Logistic Regression	0.79	0.80	0.99	0.88	0.67
SVM	0.8	0.81	0.99	0.89	N/A
Random Forest	0.94	0.97	0.96	0.96	0.95

Note: ROC-AUC is not compatible with SVM model.

The results indicate that the **Random Forest model consistently outperforms both Logistic Regression and SVM** across all evaluation metrics. This suggests that non-linear decision boundaries and feature interactions play a significant role in distinguishing AI-generated text from human-written text.

2. Confusion Matrix Analysis

Figures 1–3 present the confusion matrices for the three evaluated models.



- Logistic Regression exhibits a high false positive rate, misclassifying a substantial number of human-written texts as AI-generated, while maintaining a low false negative rate for AI-generated content.
- **SVM** shows improved balance between precision and recall but still struggles with minority-class (human) predictions.
- **Random Forest** achieves the most balanced performance, with fewer misclassifications across both classes.

These results demonstrate Random Forest's robustness to class imbalance and its ability to model complex linguistic patterns.

3. Feature Importance Analysis

To interpret the decision-making behavior of the best-performing model, feature importance scores were extracted from the Random Forest classifier. Figure 4 presents the relative contribution of each linguistic feature to the final classification.

Key observations:

- **Entity density** emerges as the **most influential feature**, indicating that the proportion of named entities relative to total words is a strong discriminator between human-written and AI-generated Arabic text. This suggests that AI-generated abstracts tend to exhibit distinct patterns in referencing named entities compared to human-authored content.
- **Honoré's R measure** also plays a significant role, reflecting differences in vocabulary sophistication and lexical diversity between the two classes.
- **Noun count** contributes moderately to the model's decisions, capturing variations in informational and descriptive density across texts.
- **Genitive construction count** provides additional, though comparatively smaller, discriminative power by modelling syntactic structures common in formal Arabic writing.

Overall, this analysis highlights that **entity-level information is a critical signal** for AI-text detection in Arabic and confirms that linguistically motivated features offer both strong predictive performance and interpretability.

4. Error Analysis

Despite strong overall performance, certain types of errors were observed:

- **Human-written abstracts with highly formal or templated language** were occasionally misclassified as AI-generated.
- **AI-generated texts with high lexical diversity and realistic entity usage** were sometimes predicted as human-written.
- Short abstracts with limited content often lacked sufficient linguistic cues, leading to ambiguous predictions.

These errors highlight the limitations of relying solely on surface-level linguistic features and suggest potential benefits from incorporating semantic or contextual representations.

5. Discussion

The experimental results demonstrate that **traditional machine-learning models, particularly Random Forest, can effectively detect AI-generated Arabic text using a small set of handcrafted linguistic features**. Compared to deep-learning approaches, this method offers advantages in interpretability, computational efficiency, and ease of deployment.

However, the observed errors indicate that as AI-generated text becomes more sophisticated, feature-based approaches may require augmentation with contextual or hybrid models to maintain performance.

7.Conclusion and Future Work

This project investigated the detection of AI-generated Arabic text using a set of linguistically motivated features and traditional machine learning models. Four interpretable features were extracted—Honoré's R measure, noun count, genitive construction count, and entity density—and evaluated using Logistic Regression, Support Vector Machine (SVM), and Random Forest classifiers.

Experimental results demonstrated that **Random Forest significantly outperformed both Logistic Regression and SVM**, achieving the most balanced performance across classes. While Logistic Regression and SVM exhibited a strong bias toward predicting AI-generated text—resulting in high false positive rates for human-written samples—Random Forest effectively captured nonlinear relationships among features and handled class imbalance more robustly.

Feature importance analysis revealed that **entity density is the most influential feature**, highlighting the critical role of named entity usage patterns in distinguishing human-authored text from AI-generated content. This finding suggests that AI-generated abstracts exhibit systematic differences in how entities are introduced and referenced compared to human writing. Overall, the study confirms that shallow linguistic features, when carefully selected, can provide both strong performance and interpretability for Arabic AI-text detection.

Several directions can be explored to extend this work:

- **Feature Expansion**: Incorporating additional syntactic, semantic, and discourse-level features, such as dependency tree depth, cohesion measures, or semantic role patterns.
- **Hybrid Models**: Combining handcrafted linguistic features with contextual embeddings from Arabic pretrained language models to balance interpretability and performance.
- **Cross-Domain Evaluation**: Testing the proposed framework on multiple Arabic text domains to assess robustness and generalizability.
- **Multilingual Extension**: Adapting the methodology to other low-resource languages to examine cross-linguistic patterns in AI-generated text.
- **Efficiency Optimization**: Exploring lightweight NLP pipelines or approximation techniques to reduce computational cost and enable real-time applications.