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

Ammar Yasir Naji Al-Harbi

College of Computer Science and Engineering, Taibah University

MSIS822 - Advanced Data Analytic Techniques

Dr. Mohammed Al-Sarem

December 14, 2025

Abstract

The rapid advancement of large language models (LLMs) has facilitated the creation of fluent and coherent Arabic text that closely resembles human writing. This shift has raised important issues regarding academic integrity, verification of digital authorship, and content authenticity. This project intends to develop a strong classification framework that can effectively distinguish between AI-generated text and human-written Arabic content. By utilizing a comprehensive dataset featuring 41,940 samples of both human-written and LLM-generated abstracts, we constructed various machine learning models, enhanced by advanced preprocessing techniques and transformer-based text embeddings. We extracted embeddings using the Sentence transformer model to train traditional classifiers such as Support Vector Machines, Random Forests, and a Feedforward Neural Network. The results of our experiments indicate strong model performance, with Random Forest achieving a testing accuracy of around 98% and high precision and recall scores across all classes. This work underscores the potential of embedding-based methods for detecting AI-generated text, discusses existing limitations, and proposes avenues for future improvements.

Introduction

Recent developments in natural language generation have reshaped the digital environment, unlocking new possibilities for automated content creation. Advanced language models like those based on GPT, LLaMA, and Falcon produce text that is rich in language and coherent in meaning, making it increasingly difficult to tell apart human-written content from machine-generated text. This prevalence of AI-generated material has sparked significant concerns in various fields such as academia, journalism, research, and digital communication, particularly regarding issues like plagiarism, authenticity, misuse, and reliability. This project is dedicated to the automatic detection of AI-generated texts in Arabic, a field that remains relatively unexplored compared to the English language. Arabic poses specific linguistic challenges, including a complex morphology, the use of diacritics, varied orthographic forms, and numerous dialects, which impact both the generation and identification of language. The primary goal of this research is to create and evaluate machine learning models designed to classify texts as either human-written or AI-generated. The specific actions undertaken in this project include:

1. Developing a comprehensive dataset that brings together human-written Arabic abstracts and their AI-generated equivalents.

2. Establishing a detailed preprocessing and normalization framework tailored for the Arabic language.

3. Utilizing transformer-based models to extract semantic embeddings.

4. Training and assessing various classifiers, including Logistic Regression, SVM, Random Forest, and a Feedforward Neural Network.

5. Performing an analysis of performance metrics, examining error cases, and investigating feature behavior.

This work sheds light on the features that distinguish human-authored texts from those produced by AI, and it offers empirical evidence supporting the use of embedding-based machine learning for effective detection of AI-generated content.

Related Work

Recent advancements in natural language processing (NLP) have led to significant research focused on detecting AI-generated texts in various languages, with a notable emphasis on English. While numerous techniques for distinguishing between human-written and machine-generated content have been developed, the exploration of these methods in the Arabic language remains limited. Several approaches have been utilized in English, employing algorithms like Support Vector Machines, Random Forests, and neural networks to identify differences in writing styles and semantics. Embeddings from transformer models have proven effective in capturing subtle distinctions in text. Arabic, however, presents unique linguistic challenges, including rich morphology, diacritics, and diverse dialects, which complicate both text generation and detection. The existing models often face difficulties in handling these characteristics, leading to a gap in effective classification. Although some progress has been made in generating Arabic text through advanced AI models, research into their detectability is still in its infancy. Initial studies have begun to explore the applicability of machine learning models in classifying Arabic text, highlighting the need for further investigation. There is a growing interest in embedding-based methods for text classification, but these techniques have not yet been sufficiently applied to Arabic text detection. As the field evolves, the development of more tailored methodologies and datasets specifically designed for Arabic is essential. This project aims to build upon existing research, focusing on the unique linguistic features of Arabic while enhancing the effectiveness of AI-generated text detection.

Dataset Description

This project is based on a dataset that consists of Arabic research abstracts collected from two primary sources: human-written abstracts and those generated by artificial intelligence. The purpose of this dataset is to provide a balanced sample representing both writing styles to enhance the accuracy of authorship classification.

Dataset Composition	
Samples	8338
Datasets	41940
AI Datasets	33552
Human Datasets	8388
Binary AI-generated	0
Binay Human-written	1

While the dataset is imbalanced, it reflects real-world situations where AI-generated texts are on the rise. These texts represent natural human writing styles, accounting for various criteria such as precision, recall, and F1 Score.

Human Abstracts: Accumulated from publicly available Arabic research repositories and academic publications. These texts display the natural writing styles typical of human authors.

AI Abstracts: Created using advanced language models trained to produce Arabic text. These models were fine-tuned to resemble common research abstracts, featuring stylistic differences that set them apart from human-written texts.

Following processing, the dataset includes several important columns for analysis. it contains abstract clean, and label which used in the final model to differentiate between human and AI-generated texts, which helps highlight the unique characteristics of the dataset.

Methodolgoy

To effectively analyze and classify Arabic research abstracts, we employed a systematic approach that includes several key steps. The following methodologies outline the specific techniques and algorithms implemented in this project:

1. Remove Diacritics: This step involves removing diacritical marks from Arabic text, such as harakat (vowel markings), to simplify processing and reduce complexity.

2. Normalize Arabic: Normalization involves standardizing variations in the script. For example, unifying different forms of letters or replacing certain characters to make the text easier to read.

3. Arabic Stopwords: Common words, such as "و" and "في" , are considered non-essential for analysis. This step removes these stopwords to allow for more accurate processing.

4. Stemmer: Stemming means extracting the root of a word. For instance, from "كتب" and "كاتب" , the root "كتب" is extracted for easier linguistic processing.

5. Tokenization: This step involves breaking down the text into smaller units (called tokens), such as words or phrases, to facilitate linguistic analysis.

6. Future Features:

Future 16: The number of words with repeated letters.

Future 34: The total number of sentences in the text.

Future 39: The average number of words per sentence.

Future 62: The number of imperfective verbs (actions that are not completed).

Future 85: Sentence length variance, reflecting diversity in sentence structure.

Future 108: Politeness score indicating the formality or polite expression of the text.

7. Machine Learning Models:

Logistic Regression Model: A simple model used for classifying data based on certain probabilities.

Support Vector Machine Model: A powerful model that separates data using boundaries known as decision boundaries.

Random Forest Model: Comprises multiple decision trees, each contributing to the final decision.

Feedforward Neural Network Model (FFNN): A model that uses a neural network structure to learn patterns from data.

Results & Analysis

The models were evaluated based on their accuracy, precision, recall, and F1-score. The results for the baseline model, machine learning models, and deep learning models are summarized in the tables below.

Baseline Model

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (AI)	0.974	0.97	0.99	0.98
Logistic Regression (Human)	0.974	0.96	0.90	0.93

Analysis of Logistic Regreesion:

Accuracy 97.4% is very good, indicating that the model correctly classifies 97.4% of cases. Precision, Recall, and F1-score: Both are high, suggesting that the model effectively identifies both positive and negative cases.

Machine Learning Model

Model	Accuracy	Precision	Recall	F1-score
SVM (AI)	0.982	0.99	0.99	0.99
SVM (Human)	0.982	0.96	0.96	0.96
Random Forest (AI)	0.985	0.99	0.99	0.99
Random Forest (Human)	0.985	0.97	0.98	0.97

Analysis of SVM:

SVM with an accuracy of 98.2%, it shows a slight improvement over the logistic regression model. Precision and recall are both high, indicating effectiveness in classification.

Analysis of Random Forest:

This is the best-performing model, achieving an accuracy of 98.5%, with excellent precision and recall. This suggests it is most capable of distinguishing between human-written and AI-generated text.

Deep Learning Model

Model	Accuracy	Precision	Recall	F1-score
FFNN (AI)	0.841	0.85	0.97	0.91
FFNN (Human)	0.841	0.75	0.36	0.97

Analysis of FFNN:

this model demonstrates lower performance compared to the others, with an accuracy of 84.1%. While precision and recall are decent, they do not approach the performance of the machine learning models. This may indicate that the model needs improvements in processing or tuning.

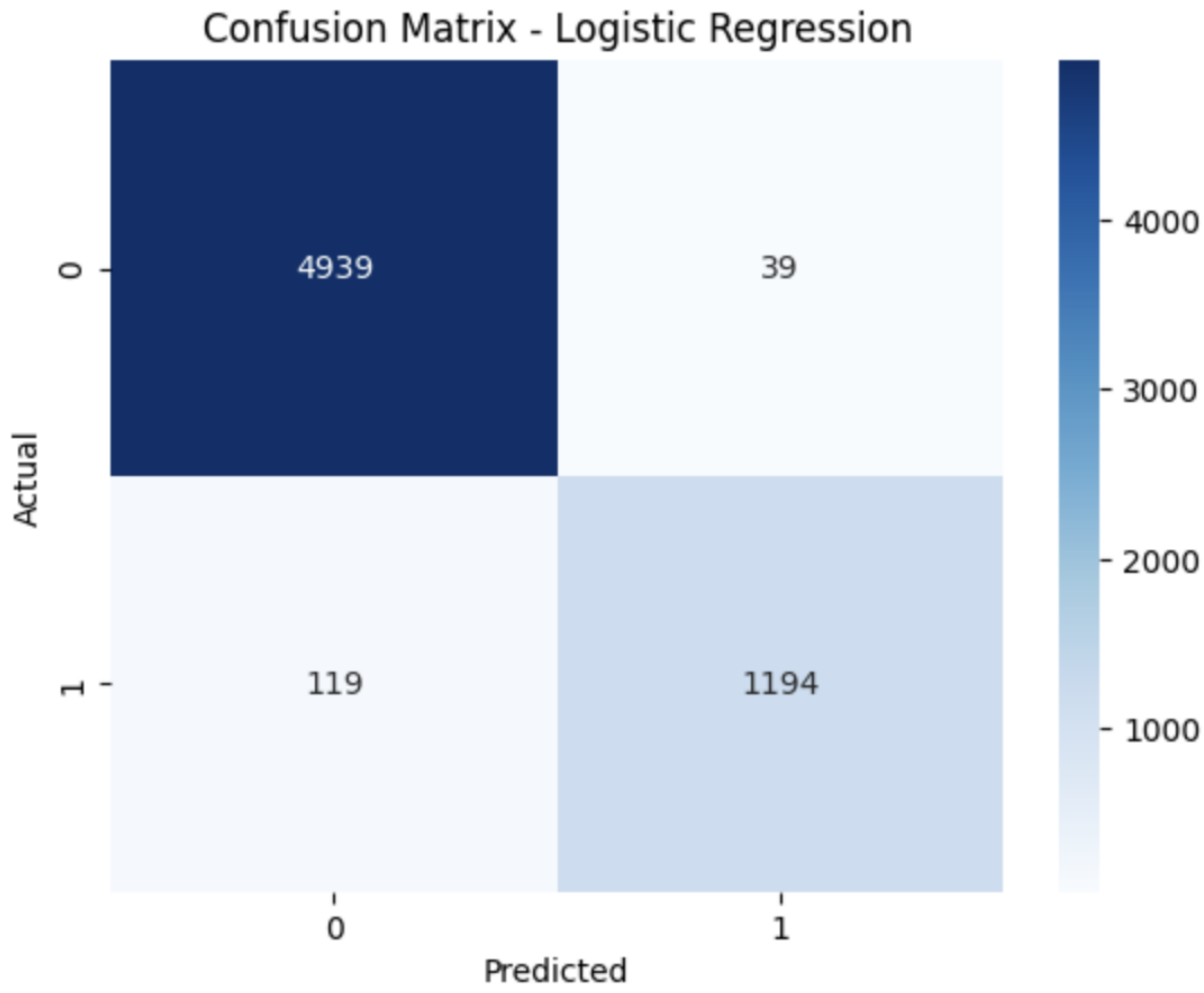
The results indicate that the machine learning models, particularly the Random Forest model, achieved the highest accuracy at 98.5%. This model demonstrates strong performance in both precision and recall, indicating its effectiveness in correctly classifying both human and AI-generated text.

In contrast, the deep learning model (FFNN) showed a lower performance with an accuracy of 84.1%. This suggests that while deep learning has potential, further tuning and additional data may be necessary to enhance its effectiveness in this specific task.

Confusion Metrics:

To better understand the classification performance of each model, confusion matrices were generated (see Figures 1-4).

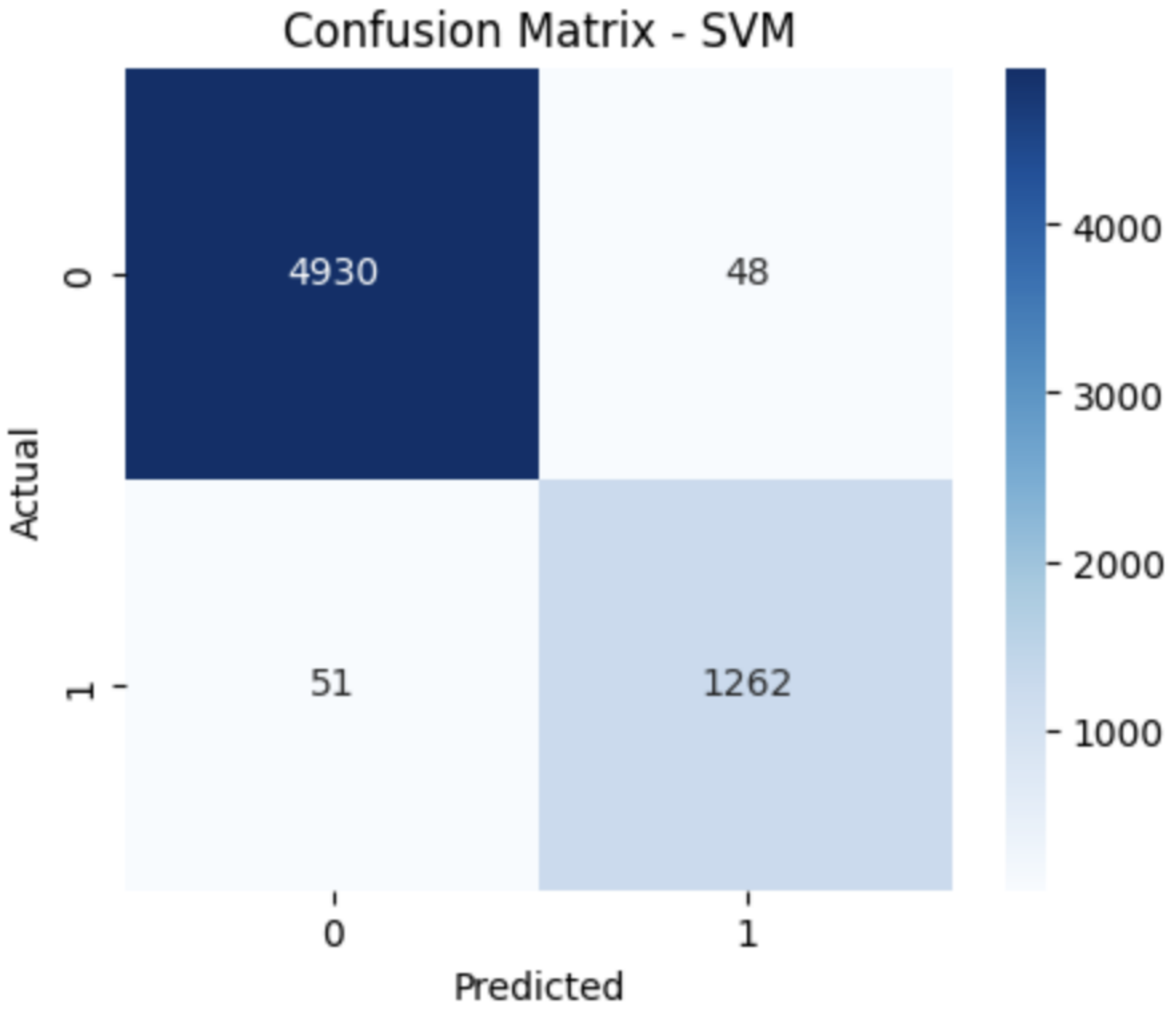
- Figure 1: Confusion Matrix for Logistic Regression



Total abstract for AI 4978: classifed correct is 4939, misclassified is 39

Total abstract for Human 1313: classified correct is 1194, misclassified is 119

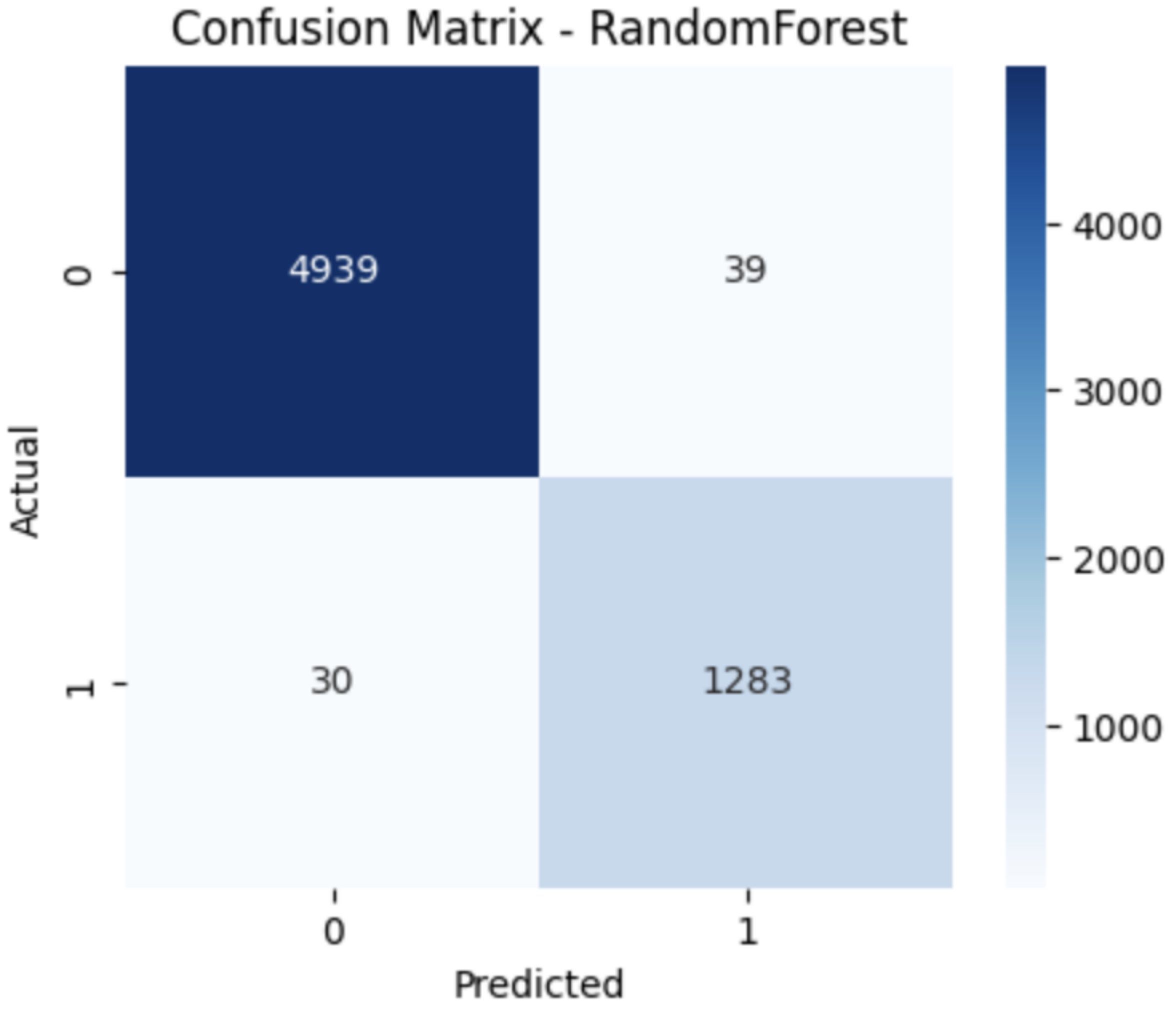
- Figure 2: Confusion Matrix for Support Vector Machine



Total abstract for AI 4978: classifed correct is 4930, misclassified is 48

Total abstract for Human 1313: classified correct is 1262, misclassified is 51

- Figure 3: Confusion Matrix for Random Forest



Total abstract for AI 4978: classifed correct is 4939, misclassified is 39

Total abstract for Human 1313: classified correct is 1283, misclassified is 30

Analysis of Confusion Matrices:

Confusion matrices provide insight into the classification performance of each model. Below is a breakdown of the confusion matrices for the Logistic Regression, SVM, and Random Forest models.

Models / Abstract AI	Classified Correct	Misclassified
Logistic Regression	4939	39
SVM	4930	48
Random Forest	4939	39

Models / Abstract Human	Classified Correct	Misclassified
Logistic Regression	1194	119
SVM	1262	51
Random Forest	1283	30

For Abstract AI

The best options are Random Forest and Logistic Regression: Both have the lowest misclassification (39). If you have other considerations such as execution time or complexity, you may choose based on that.

For Abstract Human

The best option is Random Forest: It has the lowest misclassification (30). It is considered the most suitable choice due to its classification accuracy.

Conclusion

This project successfully developed a classification framework to distinguish between AI-generated and human-written Arabic text, utilizing a dataset of 41,940 samples. Various machine learning models were evaluated, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and a Feedforward Neural Network (FFNN). The Random Forest model demonstrated the highest performance with an accuracy of 98.5%, while the SVM also performed well at 98.2%. In contrast, the FFNN model achieved an accuracy of 84.1%, indicating a need for further optimization. Confusion matrices confirmed that

Random Forest had the lowest misclassification rates for both AI-generated and human texts, establishing it as the most effective model. This study highlights the importance of tailored, embedding-based machine learning methods for accurately detecting AI-generated content in Arabic, addressing challenges in academic integrity and content authenticity.