Detection of AI-Generated Arabic Text

A Data Mining Approach

Moutzz Ahmed Abokhashabh
Student Number : 4714244
MSIS-822 Advanced Data Analytic Techniques

December 14, 2025

Abstract

This project investigates the problem of detecting AI-generated Arabic text using data mining and machine learning techniques. Using the KFUPM-JRCAI Arabic Generated Abstracts dataset, multiple preprocessing, feature engineering, and modeling approaches were applied. Traditional machine learning models and a deep learning model based on BERT embeddings were evaluated. Experimental results show that traditional classifiers combined with TF-IDF and stylometric features achieve high performance, with Random Forest and SVM outperforming the deep learning approach in this task.

1 Introduction

The rapid advancement of large language models has enabled the large-scale generation of human-like text. While beneficial, this also introduces challenges related to authorship verification, academic integrity, and misinformation. Detecting AI-generated Arabic text is especially challenging due to Arabic's rich morphology and linguistic diversity. This project aims to develop and evaluate machine learning models capable of distinguishing between human-written and AI-generated Arabic abstracts.

2 Related Work

Prior research on AI-generated text detection relies on stylometric analysis, lexical statistics, and neural language models. Recent approaches utilize transformer-based embeddings such as BERT. However, Arabic-focused detection studies remain limited, motivating this work.

3 Dataset Description

The dataset used is the **KFUPM-JRCAI Arabic Generated Abstracts** dataset from Hugging Face. It consists of human-written abstracts and AI-generated versions produced by multiple language models.

3.1 Dataset Statistics

Category	Samples
Human-written abstracts	8,388
AI-generated abstracts	33,552
Total	41,940

Table 1: Overall dataset distribution

4 Methodology

4.1 Preprocessing

Arabic-specific preprocessing was applied, including normalization, removal of diacritics, stop-word removal, and stemming using the ISRI stemmer.

4.2 Feature Engineering

Two main feature types were extracted:

- TF-IDF features using unigrams and bigrams.
- Stylometric and linguistic features such as average syllables per word, adverb counts, and emotional valence.

4.3 Data Splitting

The dataset was split into training (70%), validation (15%), and test (15%) sets to avoid data leakage.

5 Models

The following models were implemented:

- Logistic Regression (baseline)
- Support Vector Machine (SVM)
- Random Forest
- XGBoost
- Feedforward Neural Network using BERT embeddings

6 Results

Model	Accuracy	F1-score
Logistic Regression	96.3%	0.96
SVM	97.4%	0.97
Random Forest	97.8%	0.98
XGBoost	97.1%	0.97
FFNN (BERT)	85.9%	0.84

Table 2: Test set performance comparison

7 Conclusion

Results show that traditional machine learning models combined with TF-IDF and stylometric features outperform the deep learning approach for this dataset. Random Forest achieved the best overall performance. Future work may explore larger Arabic transformer models and cross-domain evaluations.