Detection of AI-Generated Arabic Text

Abstract:

This project is specifically designed to detect forgery in Arabic abstracts that were faked by artificial intelligence. Using classification models to determine whether Arabic text was written by a human or generated by artificial intelligence. Traditional machine learning models (Logistic Regression, SVM, Random Forest) were compared with a deep learning model (FFNN). The results were presented using standard evaluation metrics (Accuracy, Precision, Recall, F1 score), as well as confusion matrices and training curves (Sebastiani, 2002).

Introduction:

The rapid development and improvement of artificial intelligence (AI) technology have led to the wide use of advanced language models and their impact on our daily lives (Gehrmann et al., 2019). The advancement in generative AI has enabled AI tools such as LLMs to generate text, generate images, and more. Although this technology is attractive, it poses security and social threats (Uchendu et al., 2020).

Using generative AI and its impact on various aspects of life and work introduced significant risks and challenges. This includes its misuse through the generation of misleading content, such as fake comments (Gehrmann et al., 2019). This development made it more difficult to distinguish between texts written by humans and those produced by LLM models; therefore, analyzing and detecting this output is a crucial issue that cannot be ignored (Uchendu et al., 2020).

Detecting AI-generated Arabic text has become a major concern with advancements in generative text tools. As a result, many studies have attempted to address this problem using Machine learning, Deep Learning (DL), and Baseline methods. These techniques include Machine learning algorithms such as support vector machines (SVM) and random forest (RF), as well as deep learning architectures, including feedforward networks (FFN) (Khorsheed & Al-Thubaity, 2013; Sebastiani, 2002).

Related work:

The classification of Arabic texts has witnessed a significant leap in recent years, especially with the introduction of rapid learning technologies (Khorsheed & Al-Thubaity, 2013). The story began with Al-Furaih's 2019 study, which focused on using learning algorithms to classify Arabic texts. The key point here? Arabic is a widely used language, and it requires specific techniques normalization, cleaning, and book naming, resulting in excellent results (Khorsheed & Al-Thubaity, 2013). Then came Al-Rifai and his team's 2021 study, which presented a real challenge: Arabic content doesn't always fall into a single category; sometimes it's coined, with exceptions from other categories simultaneously (El-Rifai et al., 2021).

With the proliferation of AI linguistic models, the crucial question became how to differentiate between a text written by a human and one edited by a machine. Al-Shaibani and Ahmed (2025) proposed the idea of a beautiful girl called "AI fingerprinting" meaning that by analyzing writing style, you can determine whether a text is AI-generated or not (Gehrmann et al., 2019; Uchendu et al., 2020). Barhoum (2025) went a bit further by defining the AraBERT model with LSTM, and the results show that Transformer models are indeed powerful when supported by deniability learning layers (Antoun et al., 2020).

Key Applications and Existing Problems

Najjar (2025) specializes in the topic of sensitivity: detecting AI-generated text in educational content. This topic is crucial for academic integrity and the quality of education (Uchendu et al., 2020).

Darsa and his team (2025) proudly undertake the AraGenEval task and compare several models. This type of comparative analysis is useful because it allows the community to see exactly where it stands (Zain et al., 2023).

The Big Picture

Hamzawi and his team (2025) present a comprehensive study on classifying Arabic texts, highlighting the challenges: data pressure, dialectal diversity, and the complexity of Arabic (Oueslati et al., 2020). These problems affect overall classification, but also the detection of machine-generated text.

One final, very important point: ethics. The World Health Organization (2022) has emphasized the importance of ethics and integration in AI applications (World Health Organization, 2022).

Dataset Description:

The dataset Source is Sadiya collaborating with KFUPM, with a total size of 22.4 MB and 8388 rows downloaded from Hugging Face. The dataset contains Arabic abstracts written by humans as well as AI-generated abstracts using multiple LLMs such as allam_generated_abstract, jais_generated_abstract, llama_generated_abstract, and openai_generated_abstract (Zain et al., 2023).

The project addresses a binary classification for the Arabic text, where the models predict whether the text is a human-written written, its label will be 0, and if it is an AI-generated text, its label will be 1. Each model is mapped with a consistent label to ensure fairness between the models' results (Sebastiani, 2002).

Sample examples:

كثيرا ما ارتبطت المصادر التاريخية في الأندلس خاصة منها كتب التراجم والفهرست والبرامج وغيرها بدراسة حياة العلماء والرواة والقضاة والساسة ؛ وقد تطورت هذه المادة حتى ترك لنا المؤلفون الأندلسيون سلسلة متواصلة الحلقات من كتب التراجم كالصلة لابن بشكول ، وصلة الصلة لابن الزبير، والتكملة لكتاب الصلة لابن الآبار، والذيل والتكملة لكتابي الموصول والصلة لابن عبد الملك المراكشي إضافة إلى الإحاطة في أخبار غرناطة لابن الخطيب ، إلا أنها لم تنس أن تشير في ثنايا أو بالأحرى في خواتم هذه المؤلفات إلى فئة المرأة العالمة التي ساهمت في الإنتاج الفكري والحضاري الأندلسي. ومن خلالها سنسعى إلى الوقوف على حالة التعليم عند المرأة الأندلسية، وكيف كانت تأخذ فنون العلم. وما مدى إسهامها في الفكر التربوي والإنتاج الفكري الأندلسيين؟

Methodology:

In the beginning, I downloaded the data from Hugging Face, then I converted each split into a single data frame, and then merged it into a data frame named df_all. Then it is taken all " original abstract" naming it as text and labeling it as human. After that, collect all the columns that were generated from AI and label them as AI. The result will be two columns of text and labels, after that, saving it into a CSV file named as raw.csv. Then check the quality of the data by counting missing, duplicated values. And calculating the Lengths of human texts.

Then, for applying Word length frequency distribution, first extract Arabic words by deleting any non-Arabic symbols, adjusting space, and return a list that contains only Arabic words. Then, for each word, count its length with count its frequency. for instance, {3: 10, 4: 7, 5:2}.

Divide paragraphs and sentences + calculate S/. Firstly, for each text paragraph and then into sentences (using appropriate punctuation), and then the following was calculated.

To apply the Number of words found in the 50 positions within word embedding, I used a pre-trained Arabic **fast Text** embedding model. starting with building a corpus from the tokenized Arabic words that were extracted from all texts.

Perplexity: I used A GPT 2 model in such a way that calculates Perplexity for each batch and counts loss for each text.

GPT-2 output probability: computing average negative log, from loss

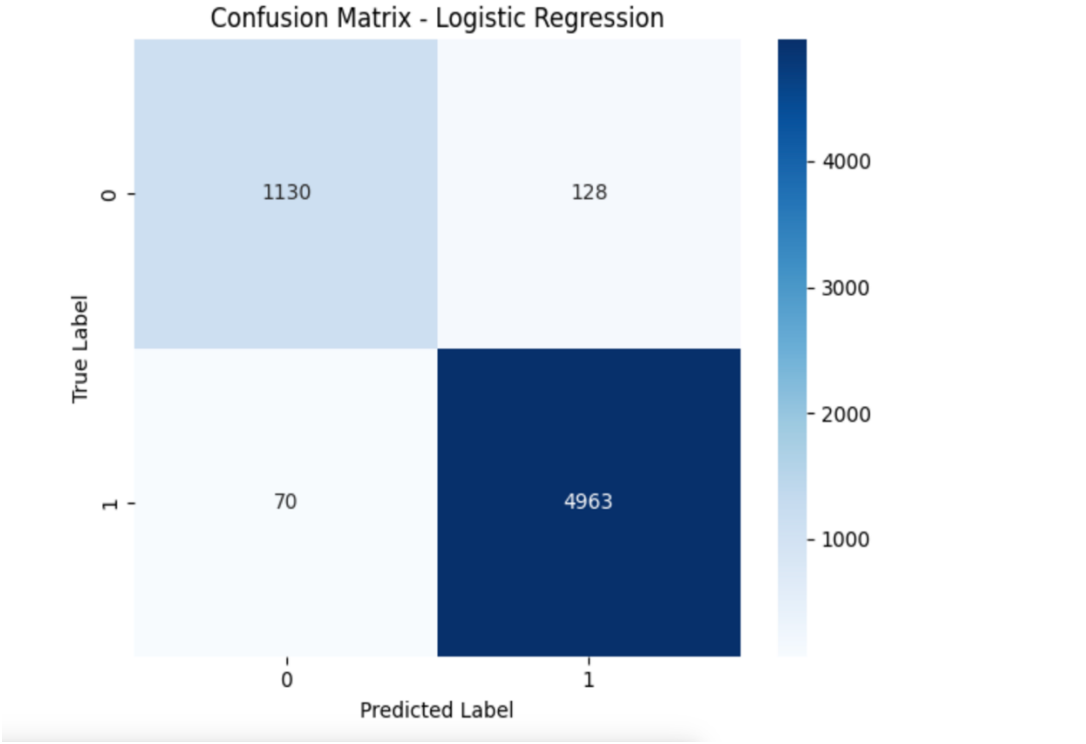
Data cleaning: downloading stop words from GitHub, then removing the diacritics and standardizing the letters by replacing the alif with a hamza with an alif without a hamza, the alif Mansurah with a ya, and the ta marbuta with a ha. After that, anything outside the scope of Arabic was deleted. Finally, the average word length and average sentence length (Type Token Ratio) were calculated for each category.

After that there were three traditional methods were trained: logistic regression, support vector machine, and Random Forest with 200 trees. These models were trained on the training set and evaluated using both validation, test sets.

On the other hand, Deep learning is implemented by using sentence-level embeddings generated by which are used as inputs to a feedforward neural network, normalizing the batches for regularization. (Khorsheed & Al-Thubaity, 2013; Sebastiani, 2002).

Logistic Regression:

```
                precision    recall  f1-score   support

...         human       0.94      0.90      0.92      1258
               ai       0.97      0.99      0.98      5033

         accuracy                           0.97      6291
        macro avg       0.96      0.94      0.95      6291
     weighted avg       0.97      0.97      0.97      6291

[Saved Figure] /content/Detection-AI-Generated-Arabic-Text/reports/figures/lr_cm.png
```

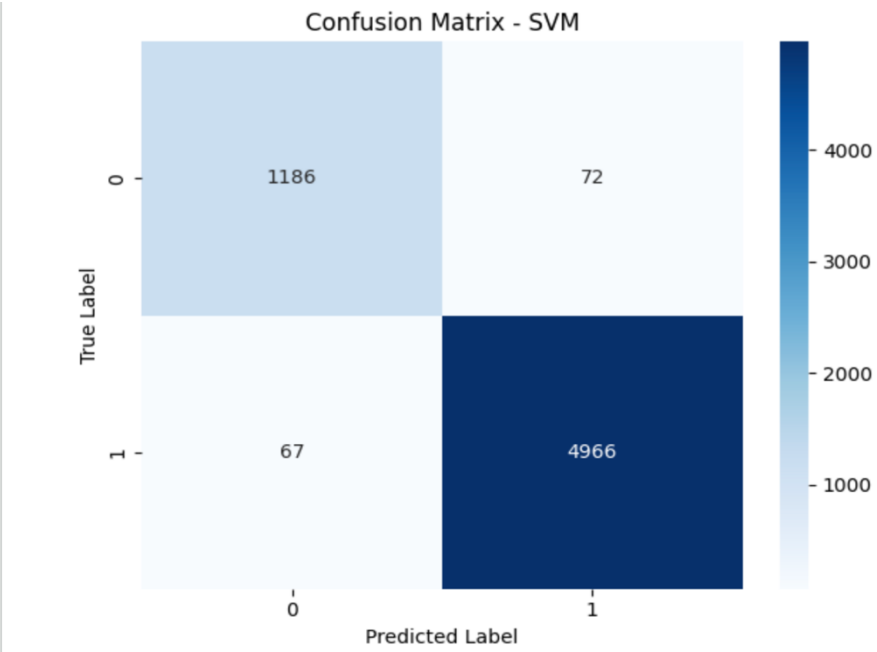


The Logistic Regression model performed exceptionally well in classifying Arabic texts, achieving an overall accuracy of approximately 96.85%.

The confusion matrix shows that the model excels at detecting AI-generated texts (correctly classifying 4,963 AI texts), with a small number of errors, including misclassifying 70 AI texts as human.

Regarding human texts, 1,130 were correctly classified, while 128 were incorrectly classified as AI, indicating a slight bias in classifying some human texts as AI due to a data imbalance (a larger number of AI texts).

SVM:

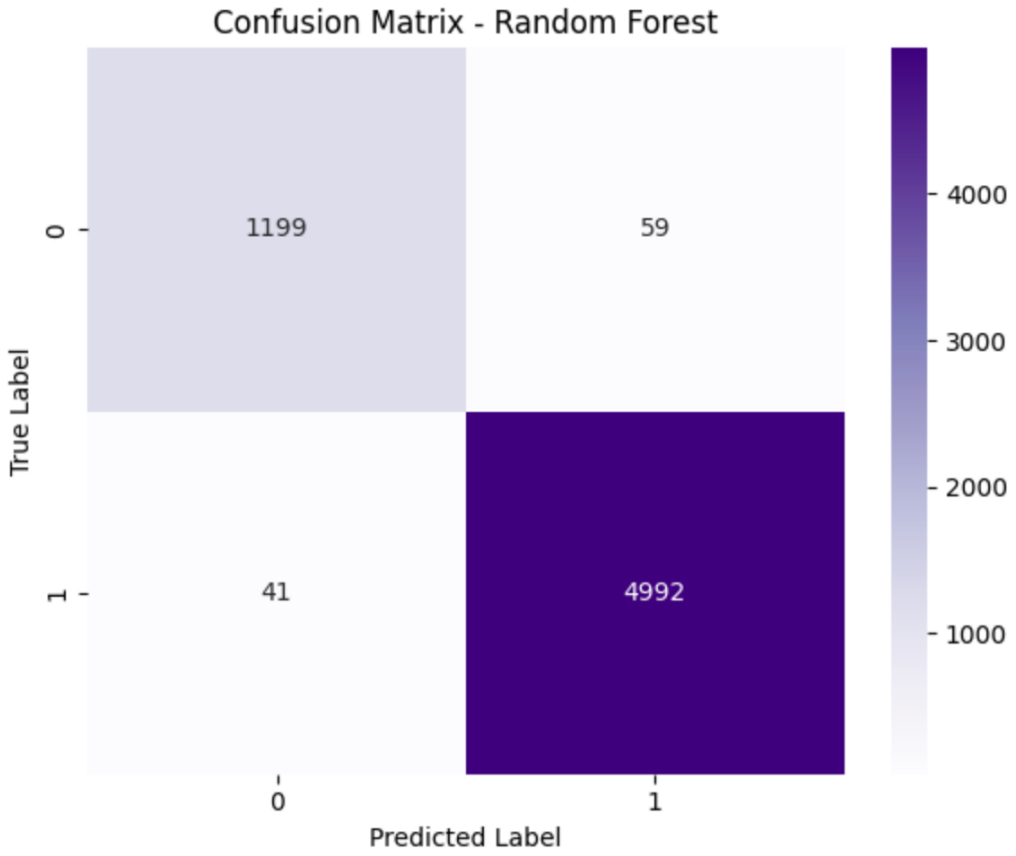



```
Training SVM
SVM Test Accuracy: 0.9779049435701797
Confusion Matrix:
 [[1186   72]
 [  67 4966]]

[SVM] Val Accuracy: 0.9737678855325914
              precision    recall  f1-score   support

       human       0.95      0.94      0.94      1258
          ai       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.97      0.96      0.97      6291
weighted avg       0.98      0.98      0.98      6291

[SVM] Test Accuracy: 0.9779049435701797
              precision    recall  f1-score   support

           0       0.95      0.94      0.94      1258
           1       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.97      0.96      0.97      6291
weighted avg       0.98      0.98      0.98      6291
```

The SVM model performed exceptionally well in classifying Arabic texts, achieving a test accuracy of approximately 97.79%. The confusion matrix shows that the model correctly classified 4966 AI-generated texts, with only a few errors, misclassifying 67 AI texts as human. It also correctly classified 1186 human texts, while misclassifying 72 human texts as AI. The evaluation metrics demonstrate high Precision/Recall values, particularly for the AI category (≈0.99), reflecting the model's ability to accurately detect generated texts with minimal errors.

Random Forest

```
Training Random Forest
Confusion Matrix:
 [[1199   59]
 [  41 4992]]

[RF] Val Accuracy: 0.9813990461049285
              precision    recall  f1-score   support

       human       0.96      0.95      0.95      1258
          ai       0.99      0.99      0.99      5032

    accuracy                           0.98      6290
   macro avg       0.97      0.97      0.97      6290
weighted avg       0.98      0.98      0.98      6290

[RF] Test Accuracy: 0.9841042759497695
              precision    recall  f1-score   support

       human       0.97      0.95      0.96      1258
          ai       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.98      0.97      0.98      6291
weighted avg       0.98      0.98      0.98      6291
```

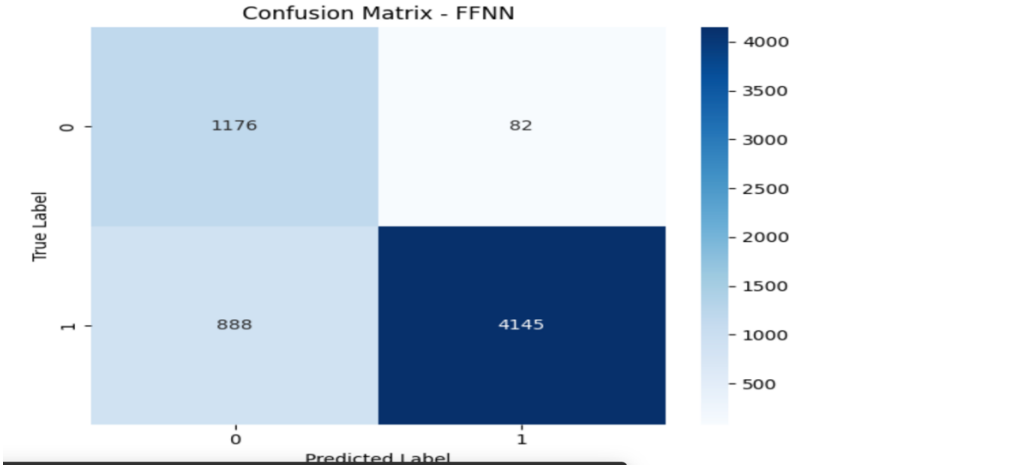


The Random Forest model achieved the highest performance among traditional models, with a test accuracy of approximately 98.41%. The confusion matrix shows that the model correctly classified 4992 AI-generated texts with a very limited number of errors (41 AI texts were misclassified as Human). It also correctly classified 1199 human texts, while 59 human texts were incorrectly classified as AI. The evaluation metrics confirm high Precision and Recall for the AI category (≈0.99), as well as

strong performance for the Human category (F1 ≈0.96), indicating the model's ability to minimize errors in both categories.

FFNN:

```
[FFNN] Test Accuracy: 0.8458114767127642
              precision    recall  f1-score   support

       human     0.5698    0.9348    0.7080      1258
          ai     0.9806    0.8236    0.8952      5033

    accuracy                         0.8458      6291
   macro avg     0.7752    0.8792    0.8016      6291
weighted avg     0.8984    0.8458    0.8578      6291

[Saved Figure] /content/Detection-AI-Generated-Arabic-Text/reports/figures/ffnn_cm.png
```



The FFNN model achieved a test accuracy of approximately 84.58%, which is lower than the performance of traditional models in this experiment. The confusion matrix shows that the model correctly classified 4145 AI-generated texts, but misclassified a significant number of AI texts as human (888 cases), indicating a lower recall for the AI category compared to other models. Conversely, the model achieved a high recall for the human category (≈0.93) but a low precision for the human category (≈0.57), meaning that a considerable portion of the texts it predicted as human were actually AI. Overall, the results suggest that FFNN needs further refinement (such as improving the text representation, threshold, or network structure) to reach the performance level of traditional models.

Conclusion:

This academic study aims and focuses on distinguish between human-written and AI-generated Arabic text by developing and evaluating classification models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Feedforward Network. Findings indicate that Random Forest achieved a high accuracy at 98.41%, followed by SVM at 97.79% and Logistic Regression at 96.85% while Feedforward Network performed less effectively, with an accuracy of 84.58%. (Zain et al., 2023).

References:

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the International Conference on Language Resources and Evaluation (LREC). https://aclanthology.org/2020.lrec-1.747/

El-Rifai, H., Al-Qadi, R., & Elnagar, O. (2021). Arabic text classification: The need for multi-labeling systems. Journal of King Saud University – Computer and Information Sciences, 33(8), 991–1002. https://doi.org/10.1016/j.jksuci.2019.03.002

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). https://aclanthology.org/P19-3019/

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large Arabic corpus. Journal of King Saud University – Computer and Information Sciences, 25(1), 13–24. https://doi.org/10.1016/j.jksuci.2012.09.004

Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, 408–430. https://doi.org/10.1016/j.future.2020.06.034

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

Uchendu, A., Le, T., Shu, K., & Liu, H. (2020). Authorship attribution for neural text generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). https://aclanthology.org/2020.emnlp-main.673/

World Health Organization. (2022). Ethics and governance of artificial intelligence for health. https://www.who.int/publications/i/item/9789240029200

Zain, A., Farooqui, S., & Rafi, M. (2023). BUSTED at AraGenEval: Detection of AI-generated Arabic text. arXiv preprint arXiv:2309.01940. https://arxiv.org/abs/2309.01940