جامعة قطر
QATAR UNIVERSITY

IntegrityAI at GenAI Detection Task 2

Detecting Machine-Generated Academic Essays in English and Arabic Using ELECTRA and Stylometry

Mohammad AL-Smadi

January 19, 2025

Shared Task: GenAI Detection Task 2

Workshop on Detecting AI-Generated Content at COLING 2025.

GenAI Detection Task 2 Overview

- **Objective:** Evaluate and rank models based on their ability to detect AI-generated academic essays.

- **Languages:** Arabic and English.

- **Phases:**
  - **Training & Validation:** Teams develop models using provided datasets.
  - **Evaluation Phase:** Performance assessed on controlled datasets.
  - **Testing Phase:** Final ranking based on model accuracy on unseen data.

Dataset Details

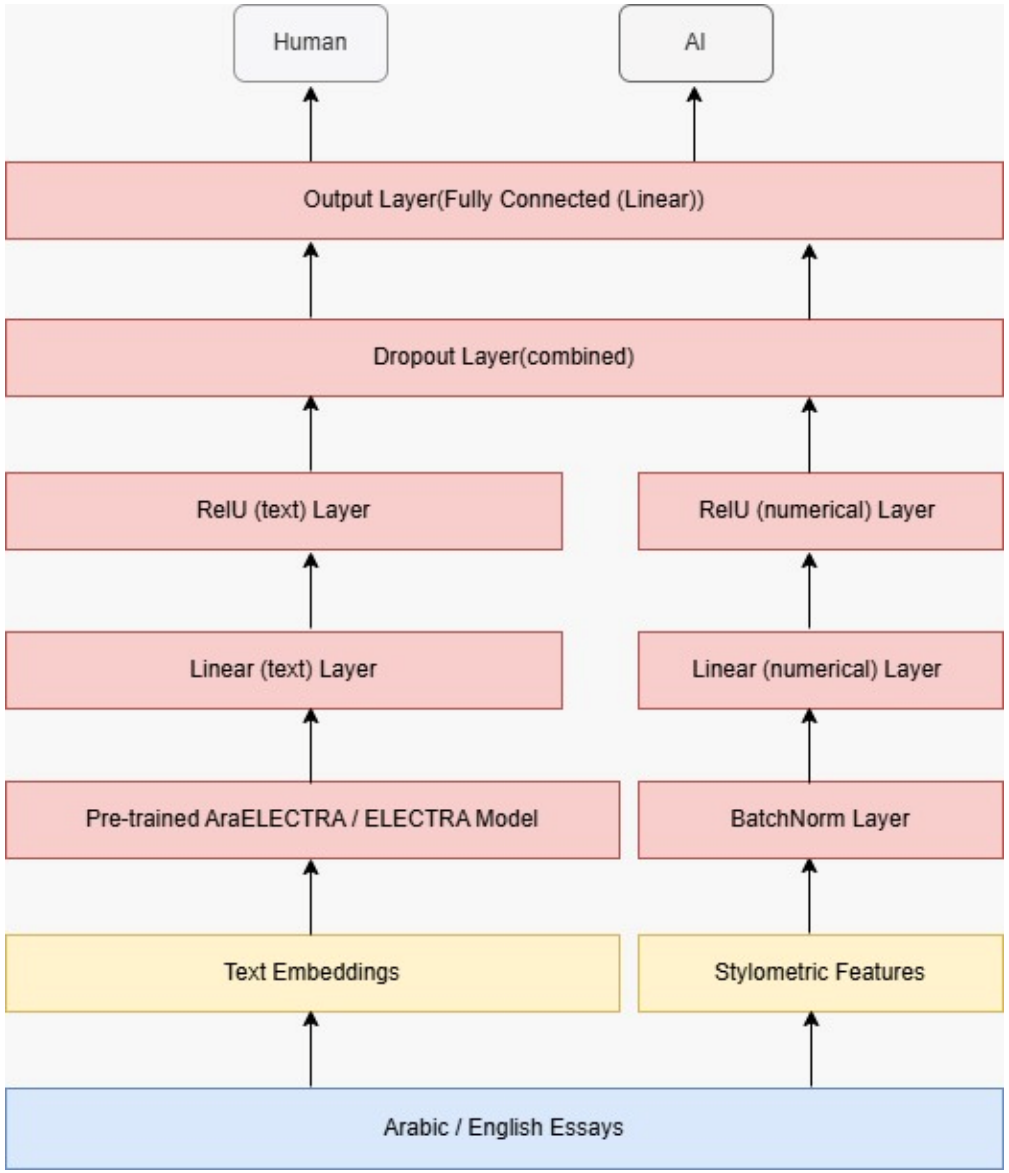
- **Dataset Structure:**

Language	Train Size	Dev. Size	Eval. Size	Test Size
Arabic	2070 (AI: 925, Human: 1145)	481 (AI: 299, Human: 182)	886	293
English	2096 (AI: 1467, Human: 629)	1626 (AI: 391, Human: 1235)	869	1130

- **Sources:**
  - **Human-written:** ETS Corpus of Non-Native Written English.
  - **AI-generated:** Outputs from GPT-3.5, GPT-4, Gemini-1.5, Llama-3.1, and more.
- **Languages:**
  - Arabic and English.

Methodology

- **Model Architecture:** ELECTRA and AraELECTRA with stylometric feature integration.
- **Stylometric Features:**
  - Word Count, Sentence Count, Vocabulary Richness, Average Word Length, Commas, Periods.
- **Training Enhancements:**
  - Dropout Layers, Batch Normalization, Fully Connected Layers, and ReLU Activation.
- **Evaluation Metrics:** macro F1-score.



Results

	Model	Eval. Phase F1 (%)	Testing Phase F1 (%)
Arabic	AraELECTRA_base_discriminator	99.8	**98.4**
	AraELECTRA_base_discriminator without features	-	96.9
	Baseline-Arabic Model	57.5	46.1
English	ELECTRA_small_discriminator	100.0	98.5
	ELECTRA_small_discriminator without features	-	96.1
	ELECTRA_large_discriminator	100.0	**99.7**
	Baseline-English Model	29.8	47.8

Key Insights:

- Stylometric features improved F1-scores by 1.5% (Arabic) to 2.4% (English).
- ELECTRA-Large achieved superior results but required more computational resources.
- Ranked **2nd** among 26 teams in English subtask (F1-score: **99.7%**).
- Ranked **1st** among 23 teams in Arabic subtask (F1-score: **98.4%**).

Conclusion & Future Work

- **Key Takeaways:**
  - High detection accuracy proves the efficiency of ELECTRA-based models and Stylometric features in AI text detection.
  - Maintaining models high performance while being computationally efficient, making it suitable for deployment on GPUs with moderate memory capacity.
  - Using larger models, such as ELECTRA-Large, achieves even higher F1-score of 99.7% on the English dataset, highlighting the potential for further accuracy gains when using more computationally intensive models.

- **Future directions:**
  - **Enhancing real-time detection** capabilities.
  - **Expanding the model's language support** beyond Arabic and English.
  - **Adapting the approach** for diverse academic fields and writing styles.

جامعة قطر
QATAR UNIVERSITY

Thank You!

Mohammad AL-Smadi
malsmadi@qu.edu.qa