Automatic AI Detection for Romanian Language

Cristina Andreea Tomoiu, Paul Ștefan Popescu, Cristian Mihăescu

University of Craiova, Romania

tomoiu.cristina.g4i@student.ucv.ro, stefan.popescu@edu.ucv.ro, cristian.mihaescu@edu.ucv.ro

This study presents a model that is designed to detect AI-generated text in Romanian, a dataset created specifically for training the model to classify sentences, and an Android application that enables the model's usage on a large scale.

We have built a data analysis pipeline that creates a high-quality dataset used to train a model that classifies sentences. The dataset has 10600 instances and 3 attributes: text (the sentences), label (human or AI) and AI_used (version of the AI used in order to generate the text). Out of 10600 instances, 5774 (54,5%) were written by humans, and 4826 (45.5%) are AI-generated, ensuring a near-balanced distribution. The starting point of the dataset was collecting human-written content, which was extracted from Romanian articles published at RRIOC [1] between 2008-2015, before AI text generation was possible. The content was proofed, processed to remove irrelevant sections, chunked, then stored in a JSONL file. This JSONL file was then used to rephrase the obtained chunks using a randomly chosen prompt with the help of GPT (gpt-4-turbo, gpt-3.5-turbo), Claude (claude-sonnet-4-20250514) and Gemini (gemini-1.5-flash-latest), which are among the most widely used AI models. The chunks were rephrased in order to keep any name or surname used in the original text, thus ensuring the classification relies on linguistic patterns rather than named entity cues. After the rephrasing was complete, the results were merged into a single JSONL file, creating the dataset in its final form.

Although several architectures were explored during experimentation, the final and most effective classifier is TFBertForSequenceClassification. It consists of a pretrained BERT model with an additional classification layer on top of BERT's neural network. The BERT model used is bert-base-romanian-cased-v1 [2], which, as the name implies, is case sensitive.

Our model achieved an accuracy of 99.15% on the 2120 entries verified. 1154 entries were correctly classified as written by a human, and 948 entries were classified as AI-generated. Only 18 samples were misclassified, 17 of those being false negatives (misclassfied as human written), and 1 a false positive. Beyond accuracy, the other classification metrics considered are precision, recall and F1-score. The recall for human sentences is 99.91% and F1-score is 99.23%. The recall for AI-generated sentences is 98.24%, and the F1-score is 99.06%.

This model has been integrated in an Android application that enables detection of AI-generated content in Romanian or English. The app has multiple input types: plain text, PDF files and images (with OCR extraction). That input is encoded in a desirable format for the model, chunked, then reconstructed. After the prediction is obtained, a content analysis is displayed. The results are stored in cloud and can be accessed anytime from the user's history.

We have successfully built and integrated a high-quality model that can detect AI generated text in Romanian language. This model has been trained using a dataset tailored specifically to the model's needs, thus promoting authenticity in Romanian texts.

The application, model, and dataset are available on Google Drive **here**

References

[1] RRIOC: Available at: https://rochi.utcluj.ro/rrioc/en/.

[2] Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online, November 2020. Association for Computational Linguistics.