Data Classification and Summarize between AI and Human Text

by Fajar Athallah Yusuf

Source dataset:

https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text/data

Github:

https://github.com/fajarathallah/Capstone_Project_Model_IBM

Project Overview

- Penggunaan LLM seperti ChatGPT, Gemini, dll, telah menyulitkan identifikasi apakah sebuah teks ditulis oleh AI atau manusia. Permasalahan ini dapat menimbulkan resiko etika, penyebaran disinformasi, hingga kepercayaan terhadap keaslian konten.

- Dilakukan pengklasifikasi yang mampu membedakan teks buatan manusia dan AI hingga membuat kesimpulan berfokus apa yang membedakan antara teks buatan AI dengan manusia

- Pendekatan yang dilakukan adalah klasifikasi dan kesimpulan gaya penulisan teks dengan menggunakan model IBM Granite via Replicate di Google Collab melalui teknik prompting

Analysis Process

Dilakukan proses pengklasifikasian dan rangkuman untuk melihat apa yang membedakan penulisan / text antara AI dan Manusia

1.Statistika Deskriptif

- Melihat jumlah data, data missing dan null
- Melihat Distribusi data melalui visualisasi

2. Pre-Processing

- Mengatur API Token, environtment, dan model
- Mengatur Parameter dan prompting
- Labelling tulisan agar bisa dikomparasi dengan model

3. Processing

- Melakukan training data sebesar 1000 sample yang dilakukan secara *stratified sample*
- Mengambil sample sebesar 30 secara acak untuk melihat hasilnya

Insight & Findings

Labelling	granite_pred	style_analysis
Human	Human	AI. The text is.....
AI generated	Human	AI. The text is classified...
AI generated	Human	AI. The text is written...

Model belum baik dalam klasifikasi tulisan. Kecurigaan muncul diantara prompting yang kurang tepat, atau parameter yang belum Optimal

Insight & Findings

Style Penulisan:

1. AI :

- Menggunakan bahasa yang formal
- Sering memakai singkatan, ejaan, atau kode-kode tertentu
- Terkesan tidak natural dan tidak bernuansa

2. Human:

- Sering memuat pengalaman pribadi, dan sudut pandang subjektif.
- Argumen yang seimbang, disertai contoh

Conclusion & Recomendations

Conclusions:

- Hasil model yang kurang bagus dalam menentukan apakah tulisan dari AI atau Human
- Model dapat menyimpulkan style penulisan AI dan Human seperti apa

Recomendations

- Melakukan Refactor prompt kembali secara eksplist agar mendapatkan hasil yang optimal dalam pengklasfikasian
- Dapat melakukan plagiarisme AI setelah mengetahui gaya penulisan suatu text