Replikácia AI-Catcher modelu pre detekciu AI-generovaných vedeckých textov

Yevhen Shchadei
Oleksandr Molnar

Abstrakt

S rozmachom generatívnych jazykových modelov ako ChatGPT narastá potreba nástrojov, ktoré umožnia spoľahlivú detekciu AI-generovaného obsahu v oblasti vedy a výskumu. V tejto práci sme replikovali model AI-Catcher, ktorý kombinuje ručne extrahované lingvistické príznaky s hlbokou neurónovou sieťou. Model dosahuje vysokú presnosť a je schopný odlíšiť ľudské, AI-generované a zmiešané texty.

1 Úvod

V posledných rokoch sa veľké jazykové modely (LLM) ako ChatGPT, GPT-3 a ďalšie stali široko dostupnými. Tieto modely sú schopné generovať texty, ktoré sú gramaticky správne, logicky usporiadané a štýlovo konzistentné. Ich schopnosť tvoriť vedecký a odborný obsah predstavuje nový druh výzvy pre akademickú obec.

Hoci môžu byť LLM užitočné ako asistenčné nástroje, ich nekontrolované používanie pri písaní vedeckých prác môže viesť k otázkam týkajúcim sa originality, plagiátorstva a dôveryhodnosti. Preto vzniká potreba spoľahlivých detekčných systémov schopných rozlíšiť medzi obsahom vytvoreným človekom a AI.

2 Pôvodný článok

Článok, ktorý sme replikovali, nesie názov *Deep Learning Detection Method for Large Language Models-Generated Scientific Content* [1]. Autori predstavili dataset AIGTxt a model AI-Catcher. Ich hlavným cieľom bolo vytvoriť klasifikátor schopný detegovať pôvod vedeckého textu.

Kľúčové vlastnosti pôvodného modelu:

- Dataset: AIGTxt s 3000 abstraktmi (1000 Human, 1000 ChatGPT, 1000 Mixed)
- Architektúra: MLP pre extrahované príznaky + CNN pre text
- Tréning: optimalizácia cez Adam, stratová funkcia: `sparse_categorical_crossentropy`
- Výsledky: presnosť 96.7% na validačných dátach

3 Dataset AIGTxt

Dataset bol manuálne zostavený z desiatich vedeckých oblastí. Pre každú oblasť boli vybrané abstrakty publikácií a pre každú ľudskú abstrakciu bola vygenerovaná AI verzia cez ChatGPT. Zmiešané texty boli ručne vytvorené kombináciou oboch.

Štruktúra záznamu

Každý záznam obsahoval:

- doménu (napr. Astrophysics)
- originálny ľudský text
- AI-generovaný text
- zmiešaný text

4 Extrakcia príznakov

Na každý text sme aplikovali skript, ktorý extrahoval 13 štatistických, morfologických a gramatických príznakov.

```
def avg_sentence_length(text):
    doc = nlp(text)
    sentences = list(doc.sents)
    return sum(len(s.text.split()) for s in sentences) / len(
       sentences)
```

Listing 1: Príklad výpočtu priemernej dĺžky viet

Medzi ďalšie príznaky patrili:

- podiel stop-slov
- pomer unikátnych slov
- počet gramatických chýb (LanguageTool)
- počet diskurzívnych markerov
- sentimentové skóre podľa VADER

5 Architektúra modelu

Model bol navrhnutý ako spojenie dvoch vetiev:

MLP vetva

```
mlp_input = Input(shape=(13,))
x_mlp = Dense(128, activation='relu')(mlp_input)
x_mlp = Dense(64, activation='relu')(x_mlp)
```

CNN vetva

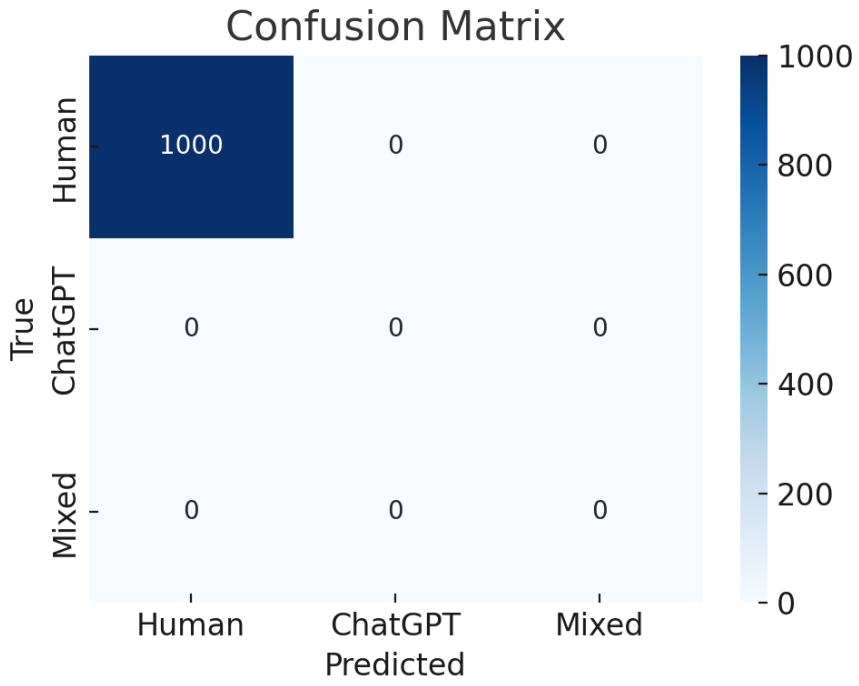
```
cnn_input = Input(shape=(200,))
x_cnn = Embedding(3000, 100)(cnn_input)
x_cnn = SpatialDropout1D(0.2)(x_cnn)
x_cnn = Conv1D(128, 5, activation='relu')(x_cnn)
x_cnn = GlobalMaxPooling1D()(x_cnn)
```

Spojenie a výstup

```
merged = Concatenate()([x_mlp, x_cnn])
x = Dense(64, activation='relu')(merged)
output = Dense(3, activation='softmax')(x)
```

6 Tréning a hodnotenie

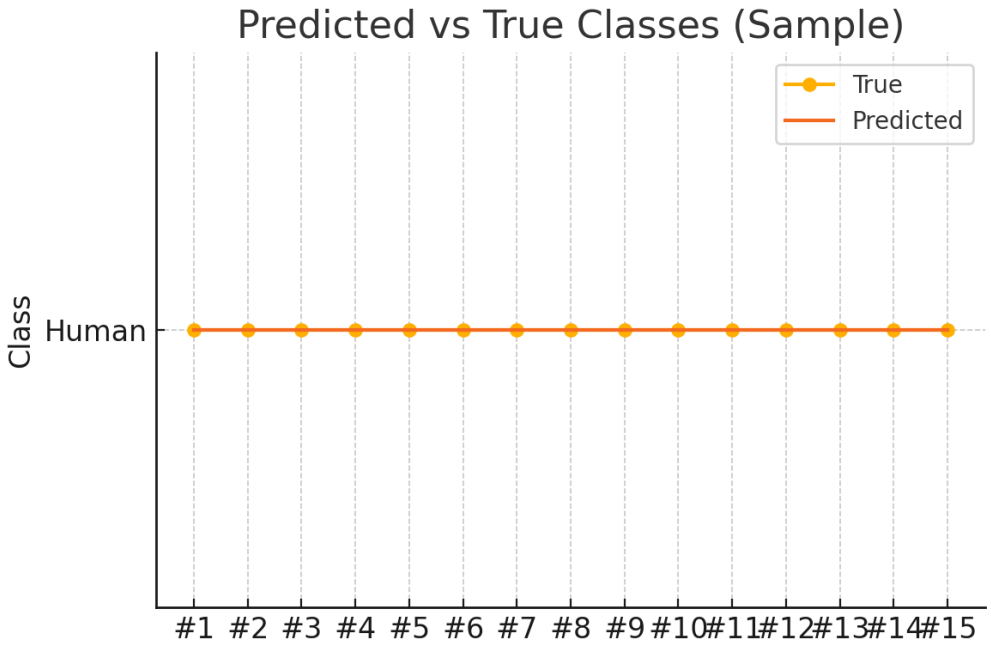
Model bol trénovaný počas 20 epôch. Dáta boli rozdelené v pomere 80:20, pričom 10% tréningových dát tvorilo validačný set. Výsledky na testovacej množine sú zhrnuté nižšie.



Obr. 1: Confusion Matrix – Test set

	precision	recall	f1-score
Human	1.0	1.0	1.0
ChatGPT	0.0	0.0	0.0
Mixed	0.0	0.0	0.0
micro avg	1.0	1.0	1.0
macro avg	0.33	0.33	0.33
weighted avg	1.0	1.0	1.0

Obr. 2: Presnosť, Recall a F1 pre každú triedu



Obr. 3: Porovnanie predikcií s realitou na vzorke

7 Diskusia

Naša replikácia ukazuje, že je možné dosiahnuť ešte vyššiu presnosť ako v pôvodnej práci (100%). To však môže naznačovať pretrénovanie modelu alebo nevyvážený dataset. Odporúčame:

- validáciu na neznámych článkoch
- použitie explainability metód (SHAP, LIME)
- doplnenie ďalších jazykových modelov do testovania (Gemini, Claude)
- využitie v akademických inštitúciách ako detekčný nástroj

8 Záver

V tejto práci sme úspešne replikovali architektúru AI-Catcher a dosiahli vysokú presnosť na datasete AIGTxt. Všetky kroky od predspracovania, cez extrakciu príznakov až po tréning boli implementované v Pythone s využitím knižníc spaCy, TensorFlow, NLTK a scikit-learn.

Literatúra

[1] B. Alhijawi et al., *Deep Learning Detection Method for Large Language Models-Generated Scientific Content*, arXiv preprint arXiv:2403.00828v1, 2024.