

Eredet: Szövegek eredetiségének ellenőrzése mesterséges intelligencia használatával

eredet.app

Józsa Kristóf

Felkészítő: Szigeti-Pekár Emese

Zafféry Károly Szalézi Középiskola, 2536 Nyergesújfalu, Kossuth Lajos utca 63.

1. Bevezetés

2024-ben a mesterséges intelligenciák által írt beadandók, dolgozatok, házi feladatok egyre népszerűbbek lettek. Ennek az oka a könnyebben rendelkezésre álló MI programok (például: ChatGPT, Google Gemini, Meta LLaMA). Bár már sok ezeket felismerő program van, szinte az összes csak az angol nyelvet támogatja. Projektem készítése során egy könnyen elérhető eszközt próbáltam fejleszteni a magyar szövegek ellenőrzésére.

2. Probléma megoldásának menete

2.1. Adatok gyűjtése

A projekthez szükséges MI modellek készítéséhez sok-sok szövegre volt szükség. A modellek jó teljesítményéhez minél változatosabb szövegeket kell használni, hogy felismerje a különböző emberi írás stílusokat. Az emberek által írt szövegek a cseh Károly Egyetem magyar nyelvi korpuszából (mely az internetet használja forrásként), és további más forrásokból (könyvek, filmfeliratok, érettségi tételek, stb.) lett összeépítve. A nem emberi szövegek a piacon lévő MI szolgáltatások által generált szövegekből lett felépítve. A végső tanulási adatbázis kb. 4000 szöveget tartalmaz, amelynek 35%-a MI által generált és 65%-a ember által írt. Ez az arány általánosságban nehezítené a modellek tanulását (a legoptimálisabb esetben egy 50-50 arány lenne), de ebben az esetben ez az arány nem okozott problémát a modellek eredményeiben.

2.2. Modellek tanítása

A projekthez 3 MI modell lett készítve, ezek a modellek a tanulási fázis különböző szakaszaiban készültek. A modellek alapját a magyar BME egyetem által kiadott huBERT modell képezi. A hárommodelles szerkezet azért előnyösebb, mert vannak szituációk melyben a 3 modell egyidejű használatát véve sokkal pontosabb eredmények születnek. A modellek a transzformer architektúrát használják. A modellek a Google Colab rendszerben lettek

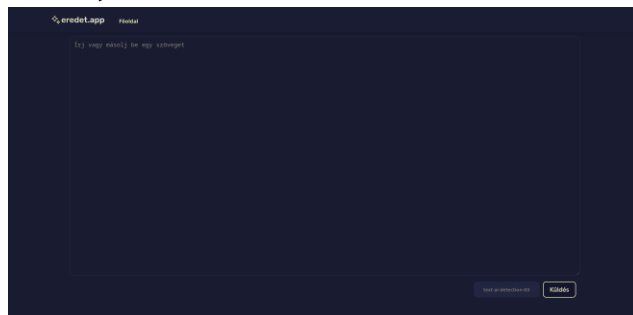
tanítva. A modellek tanítása 2 órát vett igénybe. A modellek 0 és 1 közötti értékeket adnak vissza (0 = emberi szöveg, 1 = MI szöveg), melyet egy 0-100 közötti tartományba átszámolva jelenít meg a felhasználónak.

2.3. Felhasználói felület létrehozása

A projekt egyik fő célja az, hogy a szövegek ellenőrzése minél elérhetőbb legyen. Ezzel a céllal készült el a projekt webes felülete, melyben egy felhasználó könnyedén igénybe veheti a szolgáltatást. A felület a SvelteKit keretrendszerben készült, TypeScript felhasználásával, mely egy gyors és olvasható kódbázist eredményezett.

A webes felületnek a másik fő célja az, hogy tanítsa a felhasználót a mesterséges intelligencia írásáról, és a felméréskor használt szempontok részletes leírásáról. Egy pozitív eredménynél a felhasználónak 4 gyors és egyszerű leírás található az eredmény alatt mely megpróbálja megértetni az elemzési szempontokat, és hogy miben ismerhetők fel az MI által írt szövegek. Ha a felhasználó további információt szeretne megtudni, akkor egy hosszabb formátumban meg tudja ismerni az MI eszközök használatáról, és a projekt modelljei által használt szempontokról. Ez a dokumentáció egy Markdown file-ba van, amelyet a webes kliensek automatikusan megformáznak egy könnyen olvasható honlapba.

A weboldal a felhasználó által begépett vagy másolt szöveget egy Flask-ban írt Python API-nak továbbítja, mely a felhasználó által kiválasztott modellen végigfuttatja az elemzést. Az API egy 1035 szavas mintaszöveget 564 milliszekundum alatt tud elemezni egy GTX 1050 Ti videokártyán (az MI modellek az NVIDIA CUDA technológia segítségével a videokártyán futnak). A projekt készítése során próbáltam követni a mikroszolgáltatásos szerkezetet, amelynek segítségével az alkalmazás hatékonyabban tudja fogadni a felhasználók gyors változását és a felhős szolgáltatások által biztosított előnyöket is fel tudja használni.



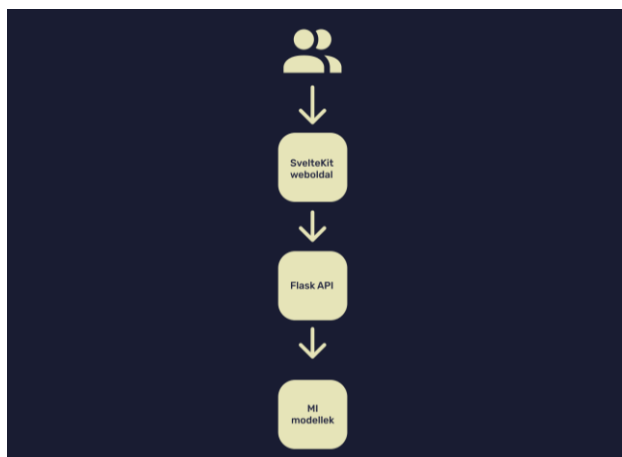
1. ábra: A felhasználói felület



2. ábra: A pozitív eredmény kijelzése



3. ábra: További információk a felismeréskor használt szempontokról



4. ábra: A projekt architektúrája

3. Elért eredmények

A modellek a vártnál sokkal jobb eredményeket hoztak. Ilyen kis modelleknél általában nem várható el a magas pontosság, ezek a modellek nem tanulták át minden ember stílusát. A legjobb modell a 650 szövegből álló validációs adatbázisból 98% pontossággal eltalálta a szövegek eredetiségét. A validációs adatbázis a magyar korpuszból tanuláshoz fel nem használt szövegeket használta, tehát lehet némi eltérés azon az alapon, hogy a szövegek ebben a korpuszban hasonló írásstílust alkalmaznak, de a valós használatban is remek eredményt adtak a modellek.

4. Összefoglalás

A projektem célja, hogy felhívjam a figyelmet a ChatGPT és további eszközök potenciális veszélyeire és előnyeire. A projekt nem azért készült, hogy a ellehetetlenítse az MI-k használatát szövegíráshoz, hanem azért hogy felhívja a figyelmet az MI-k túlzott használatára, és a felelősségteljes használatukra. A projekt a jövőben elérhető lesz a nyilvánosság számára az <https://eredet.app> linken, és bízom abban, hogy ezzel hozzájárulok a biztonságos MI fejlesztéshez.