

Raport Techniczny: Wyjaśnialny Klasyfikator Toksyczności Wypowiedzi (LLM-based)

Maciej Andrzejewski, Krzysztof Osiński

1 Definicja Zadania i Domena Problemu

1.1 Cel Projektu

Głównym celem projektu jest automatyzacja procesu moderacji treści internetowych poprzez wdrożenie modelu opartego na LLM (Large Language Model), który klasyfikuje komentarze do jednej z kategorii:

- **Toksyczny (Toxic):** Treści zawierające groźby, mowę nienawiści, wulgaryzmy lub ataki personalne.
- **Nietoksyczny (Non-toxic):** Treści neutralne lub pozytywne.
- **Graniczny/Niejednoznaczny (Borderline):** Treści trudne do oceny bez szerszego kontekstu, np. sarkazm, slang środowiskowy.

Kluczowym wyróżnikiem systemu jest **warstwa wyjaśnialności** (Explainability Layer), która ma na celu budowanie zaufania do decyzji algorytmu poprzez:

1. Podświetlanie konkretnych słów lub fraz decydujących o klasyfikacji.
2. Generowanie opisowego uzasadnienia decyzji (w języku naturalnym).
3. Identyfikację niuansów językowych, takich jak ironia czy satyra.

1.2 Grupa Docelowa i Zastosowanie

Rozwiążanie dedykowane jest dla:

- **Moderatorów treści:** Jako narzędzie wspomagające, przyspieszające weryfikację zgłoszeń poprzez wstępna analizę i wskazanie dowodów.
- **Administratorów małych i średnich serwisów:** Fora dyskusyjne, sekcje komentarzy w sklepach e-commerce, gdzie pełna ręczna moderacja jest nieefektywna kosztowo.
- **Zespołów Trust & Safety:** Do analizy trendów w mowie nienawiści i szybkiego reagowania na nowe formy toksyczności.

2 Zakres i Metodyka

2.1 Zbiór Danych

Na potrzeby projektu przygotowano dedykowany zbiór testowy składający się z 400 komentarzy.

- **Źródło:** Anonimizowane komentarze z sieci w języku polskim oraz angielskim.
- **Etykietowanie:** Każdy przykład posiada ręcznie nadaną przez twórców zbiorów (w roli moderatorów) etykietę (Ground Truth). Przykłady w języku angielskim posiadają również flagi metadanych (np. obecność ironii).

- **Balans klas:** Zadbano o reprezentację przypadków trudnych (tzw. *hard negatives*), które często mylą proste filtry słów kluczowych.
- **Różnice zbiorów:** Zbiór angielskich komentarzy zawiera wiele wypowiedzi dotyczących polityki, które były surowo moderowane (niska toleracja na toksyczność). Polski zbiór był zaś łagodniej oceniany (zauważalnie wyższa tolerancja moderatorów).

2.2 Architektura Systemu

W celu zbadania efektywności i kosztów, zaimplementowano i porównano dwa warianty potoku przetwarzania (pipeline):

1. Wariant A: Cloud-Native (Google Gemini 3.0 Flash)

- Wykorzystuje API Google do przetwarzania wsadowego (Batch API).
- Model realizuje zadanie *Zero-shot prompting*, zwracając ustrukturyzowany JSON z oceną i wyjaśnieniem.
- Zalety: Wysoka jakość rozumowania, obsługa wielu języków (PL/EN).

2. Wariant B: Local Hybrid (Toxic-BERT + Llama 3.2)

- Podejście dwuetapowe optymalizujące zasoby.
- **Faza 1:** Lekki model *unitary/toxic-bert* dokonuje szybkiej, wstępnej oceny statystycznej.
- **Faza 2:** Tylko w przypadku przekroczenia progu pewności (domyślny threshold > 0.4), uruchamiany jest model generatywny Llama 3.2 (poprzez aplikację Ollama) w celu analizy kontekstu i wygenerowania wyjaśnienia.
- Zalety: Prywatność danych, brak kosztów API, szybkość dla treści nietoksycznych.

3 Implementacja i Architektura Systemu

System został zaimplementowany w języku Python. Całość rozwiązania została skonteneryzowana przy użyciu platformy Docker oraz Docker Compose, co zapewnia powtarzalność środowiska uruchomieniowego i łatwość wdrożenia. Architektura została podzielona na warstwę serwerową (API) oraz warstwę prezentacji (UI).

3.1 Technologie i Biblioteki

W projekcie wykorzystano następujący stos technologiczny:

- **Backend / API:** FastAPI (asynchroniczna obsługa żądań HTTP).
- **Frontend / Dashboard:** Streamlit (wizualizacja danych i interakcja z modelem).
- **Modele Lokalne:** Biblioteka `transformers` (HuggingFace) dla modelu BERT oraz `ollama` jako serwer inferencyjny dla modelu Llama 3.2.
- **Walidacja Danych:** Pydantic (definiowanie schematów wejścia/wyjścia).

3.2 Wariant A: Cloud-Native (Google Gemini)

Implementacja opiera się na wykorzystaniu API modelu Gemini (wersja Flash). Kluczowym elementem implementacji jest wymuszenie ustrukturyzowanego formatu wyjściowego (*Structured Output*).

Zamiast standardowej odpowiedzi tekstowej, model otrzymuje instrukcję zwrócenia obiektu JSON ściśle zgodnego z definicją klasy Pydantic. Pozwala to na deterministyczne wyodrębnienie metadanych wyjaśnialności. Dzięki temu podejściu wyeliminowano konieczność stosowania wyrażeń regularnych do parsowania odpowiedzi modelu.

3.3 Wariant B: Local Hybrid Pipeline

W celu optymalizacji zasobów obliczeniowych zastosowano architekturę kaskadową. Proces decyzyjny przebiega dwuetapowo:

1. **Szybka filtracja:** Każdy tekst trafia do modelu `unitary/toxic-bert`. Jest to model dyskrymi- natywny, który w błyskawicznym tempie zwraca wektor prawdopodobieństwa dla 6 klas toksycz- ności.
2. **Warunkowa analiza:** System sprawdza maksymalny wynik z pierwszego etapu i w przypadku przekroczenia określonego progu odpytuje model LLama 3.2 odnośnie wyjaśnienia decyzji.
W projekcie przyjęto próg $\theta = 0.4$. Jeśli próg zostanie przekroczony, model generatywny otrzy- muje prompt z prośbą o pełnenie roli sędziego, wyjaśnienie decyzji oraz weryfikację pod kątem ironii (czego model BERT nie potrafi skutecznie wykryć).

3.4 Interfejs Użytkownika i API

Aplikacja w obu wariantach udostępnia dwie opcje pracy:

- **REST API (FastAPI):** Endpoint `/analyze-batch` przyjmuje listę tekstów i zwraca zagre- gowane wyniki. Obsługuje przetwarzanie wsadowe, co jest kluczowe dla wydajności przy dużej liczbie zapytań.
- **Dashboard (Streamlit):** Interfejs graficzny umożliwia użytkownikowi ręczne wprowadzanie tekstu, wizualizację wyników na wykresach oraz odczyt wygenerowanego przez AI uzasadnienia. Dodano również mechanizm pętli zwrotnej pozwalający użytkownikowi zgłosić niezgodność z werdyktem, co wymusza ponowną, głębszą analizę przez model.

4 Uruchomienie i Wdrożenie

W celu zapewnienia przenośności i łatwości reprodukcji wyników, całe środowisko zostało skontenery- zowane. Projekt nie wymaga ręcznej instalacji bibliotek Python ani modeli językowych na maszynie hosta - wszystkie zależności są zarządzane automatycznie przez platformę Docker.

4.1 Wymagania Wstępne

Jednym wymogiem systemowym jest obecność zainstalowanego silnika **Docker** oraz narzędzia **Docker Compose**.

4.2 Konfiguracja i Uruchomienie

W zależności od wybranego wariantu architektonicznego, przygotowano dwa niezależne pliki konfigu- racyjne orkiestracji kontenerów.

4.2.1 Wariant A: Cloud (Gemini)

Przed uruchomieniem wymagane jest ustawienie klucza API jako zmiennej środowiskowej.

```
export GOOGLE_API_KEY="Twój_Klucz_API"  
docker-compose -f docker-compose.gemini.yml up --build
```

System uruchamia lekki kontener z aplikacją Streamlit oraz FastAPI, komunikujący się z chmurą Google.

4.2.2 Wariant B: Local Hybrid

Ten wariant uruchamia zestaw dwóch usług: serwera inferencyjnego Ollama oraz aplikacji analitycznej.

```
docker-compose -f docker-compose.local.yml up --build
```

Uwaga: Przy pierwszym uruchomieniu system automatycznie pobiera niezbędne wagi modeli:

- Model `unitary/toxic-bert` (ok. 500 MB) jest pobierany podczas budowania obrazu Docker.
- Model `llama3.2` (ok. 2.0 GB) jest pobierany przez skrypt startowy (`entrypoint`) kontenera Ollama.

Dzięki zastosowaniu wolumenów dyskowych (*Docker Volumes*), pobrane modele są trwałe i nie wymagają ponownego pobierania przy kolejnych restartach aplikacji.

4.3 Dostęp do Aplikacji

Po poprawnym uruchomieniu kontenerów, usługi są dostępne pod następującymi portami lokalnymi:

4.3.1 Wariant chmurowy:

- **Interfejs Graficzny (GUI):** <http://localhost:8501>
- **Dokumentacja API (Swagger UI):** <http://localhost:8000/docs>

4.3.2 Wariant lokalny:

- **Interfejs Graficzny (GUI):** <http://localhost:8502>
- **Dokumentacja API (Swagger UI):** <http://localhost:8001/docs>

5 Ewaluacja i Analiza Wyników

Ewaluację przeprowadzono na zbiorze testowym składającym się z 400 komentarzy, zawierających zbalansowaną liczbę przykładów toksycznych, bezpiecznych oraz trudnych przypadków granicznych (sarkazm, slang). Jako punkt odniesienia (Ground Truth) przyjęto etykiety nadane ręcznie przez człowieka.

5.1 Wyniki Ilościowe (Metryki)

Poniższa tabela przedstawia porównanie skuteczności obu wariantów architektonicznych. Wariant chmurowy (Gemini) osiągnął wyższą ogólną dokładność, szczególnie w rozumieniu kontekstu, natomiast wariant lokalny wykazał się bardzo wysoką precyzją w detekcji wulgaryzmów dzięki modelowi BERT.

Model	Accuracy	Obsługa Języka	Koszt/Prywatność
Gemini 3.0 Flash (Cloud)	~85%	Wybitna (PL/EN)	Płatny / Dane w chmurze
Toxic-BERT + Llama 3.2 (Hybrid)	~72%	Dobra (Głównie EN)	Darmowy / Dane lokalne

Tabela 1: Porównanie wydajności modeli (szacunkowe).

5.2 Analiza Jakościowa (Case Studies)

Poniżej przedstawiono szczegółową analizę wybranych przypadków, które sprawiły modelom trudność lub zademonstrowały ich unikalne możliwości.

5.2.1 Kategoria 1: False Positives (Nadwrażliwość)

Sytuacje, w których model błędnie oznaczył bezpieczną treść jako toksyczną. Często wynika to z użycia słów nacechowanych negatywnie w neutralnym lub medycznym kontekście lub łagodnej oceny ludzi, do której porównano wyniki LLM.

- Cloud (Gemini 3.0)

- Przykład 1:

"Ten chory typ jest całkowicie oderwany od rzeczywistości."

Decyzja modelu: Toksyczny (Score: 0.65)

Analiza: Model uzasadnił: *Wypowiedź ma charakter obraźliwy, używa stigmatyzującego określenia 'chory typ' w celu zdyskredytowania osoby i sugeruje jej brak kontaktu z rzeczywistością.*

- Przykład 2:

"Wystarczyłoby się z sąsiadami zgadać. Jak kogoś okradli to krew się gotuje i każdy ma ochotę na rewanż. Tylko, żeby patyków nie połamali. Niech te kiepy złodziejskie mają jak wyjść z klatki po akcji. Tak bym to widział."

Decyzja modelu: Toksyczny (Score: 0.75)

Analiza: Model uznał, że wypowiedź promuje samosąd i zawiera groźby użycia przemocy fizycznej, co nie zgadza się z etyktą nadaną przez ludzi (moderatorów).

- Przykład 3:

"trudno powiedzieć czy sarkazm czy skrajna głupota i odklejenie od realiów"

Decyzja modelu: Toksyczny (Score: 0.60)

Analiza: Model uznał, że komentarz ma charakter obraźliwy, atakując intelekt rozmówcy poprzez określenia takie jak 'skrajna głupota' i 'odklejenie od realiów'.

- Przykład 4:

"Jeżeli przyjąć za moment śmierci obumarcie mózgu to on nie żyje od lat"

Decyzja modelu: Toksyczny (Score: 0.65)

Analiza: Model uznał, że wypowiedź ma charakter obraźliwy, ponieważ sugeruje brak sprawności intelektualnej lub „śmierć mózgową” adresata i jest to atak mający na celu poniżenie osoby. W rzeczywistości komentarz dotyczy dyskusji medycznej/swiatopoglądowej czy za śmierć człowieka należy uznać moment obumarcia mózgu, często poruszanej w kontekście eutanazji.

- Local (Toxic-BERT + Llama 3.2)

- Brak przypadków FP:

- * Model Toxic-BERT poprawnie odróżnił wszystkie nietoksyczne komentarze. Świadczy to o jego wysokim konserwatyzmie - model działa zgodnie z zasadą 'niewinny, dopóki wina nie zostanie udowodniona'.
- * Z perspektywy biznesowej jest to korzystne zjawisko - eliminuje ryzyko niesłusznego cenzurowania użytkowników i frustracji wywołanej blokadą bezpiecznych treści. Dodatkowo brak fałszywych alarmów oznacza zerowe marnotrawstwo zasobów obliczeniowych (Llama nie była uruchamiana bez wyraźnej potrzeby).
- * Model doskonale radzi sobie z jawną, słownikową toksycznością, ale nie zauważa 'niuanów'. W kontekście automatyzacji jest to cecha pożądana - lepiej przepuścić kilka toksycznych komentarzy do ręcznej moderacji, niż blokować niewinne wypowiedzi.

5.2.2 Kategoria 2: False Negatives (Niewykryte Zagrożenia)

Najbardziej ryzykowne przypadki, gdzie toksyczna treść została przepuszczona. Zazwyczaj dotyczy to hejtu "w białych rękawiczach" lub błędów ortograficznych (obfuscation). Modele głównie niedoszacowały toksyczności komentarzy ze zbioru w j. angielskim, które były surowo oceniane przez ludzi.

- Cloud (Gemini 3.0)

- Przykład 1:

"The Trudeau cabinet members and staff appear to have no sense of ethical responsibility. Brison's response is cavalier and shows he is completely unaware that this is a serious allegation of impropriety. Having a gay husband doesn't give him a free pass."

Decyzja modelu: Bezpieczny (Score: 0.25)

Analiza: Moderatorzy oznaczyli treść jako toksyczną, zaś model uznał, że: *Tekst zawiera krytykę polityczną i etyczną skierowaną pod adresem rządu oraz konkretnej osoby publicznej. Choć wspomina o orientacji seksualnej, robi to w kontekście argumentu, że nie powinna ona stanowić tarczy przed odpowiedzialnością, co nie stanowi bezpośredniego ataku nienawiści, ale porusza kwestie tożsamościowe w sposób krytyczny.*

- Przykład 2:

"I believe your whole post is contained in your last sentence, but I can't personally appreciate the jolly remarks about Trump and Spicer. They are both LYING INTENTIONALLY to the American people, and even with my twisted sense of humor I can't chuckle or even snort. I applaud almost everything you write, so this isn't meant as a criticism...what those two buffoons are doing should be halted, and damned fast. Blessings to you (anyway)."

Decyzja modelu: Bezpieczny (Score: 0.30)

Analiza: Gemini uzasadnił decyzję: *Tekst zawiera krytykę osób publicznych, używając określeń takich jak 'buffoons' (pajace) oraz oskarżeń o kłamstwo. Choć wypowiedź jest naczchowana emocjonalnie i zawiera lekkie obelgi pod adresem polityków, nie jest bezpośrednio toksyczna wobec rozmówcy.*

- Local (Toxic-BERT + Llama 3.2)

- Przykład 1:

"I hope all you Far Left whiners don't go out and commit suicide based on the Comey hearing today. Your leadership and the news media has led you over the cliff with their fictitious articles and speeches. You need to go back to the election and not accept the false narrative that Clinton lost the election because of the Russians. Clinton lost because she was a poor candidate under investigation by the FBI. Get over it."

Decyzja modelu: Bezpieczny (Score: 0.28)

Analiza: Model BERT nie wychwycił niestosownego odwołania do samobójstwa i ataku na grupę o innych poglądach politycznych. Llama nie została ztriggerowana, jednak po sprawdzeniu wychwyciła toksyczność: *The text contains hate speech and incitement towards suicide, targeting a specific political group. It also promotes misinformation about the Clinton election loss.*

- Przykład 2:

"Indians are already started a race war, it is just that white people aren't biting. Indians have already killed one person with their night hunting. Indians would rather complain than address a problem."

Decyzja modelu: Bezpieczny (Score: 0.13)

Analiza: BERT nie wychwycił rasistowskiego charakteru wypowiedzi, a przez nadanie bardzo niskiej toksyczności Llama nie została wywołana. Odpowiedź Llamy na prompt z komentarzem: *The text contains racist and xenophobic language, making assumptions about an entire group of people based on their ethnicity. It also promotes violence and stereotypes.*

– **Przykład 3:**

"The article says he confessed and they caught him doing the killing. How in the hell can he plead not guilty??? Why spent time and money screwing with this person. Sent him to the gas chamber in what ever state still has one and uses it!!!"

Decyzja modelu: Bezpieczny (Score: 0.39)

Analiza: Model BERT nie uznał nawoływanego do wykonania wyroku kary śmierci za wystarczająco agresywne. Po dodatkowej weryfikacji Llama wychwyciła agresywny język, ale nie wskazała jawnie na toksyczność wypowiedzi.

– **Przykład 4:**

"Crooked lieberals. Any surprise ?"

Decyzja modelu: Bezpieczny (Score: 0.002)

Analiza: BERT zupełnie nie dostrzegł toksyczności w tekście, przez co Llama nie została pobudzona, jednak LLM podał bliższe prawdzie spostrzeżenia: *The term contains derogatory language towards a specific political group, which can be perceived as discriminatory and hurtful.*

5.2.3 Kategoria 3: Sukcesy w Detekcji Ironii i Kontekstu

Złożone kontekstowo przypadki, w których proste modele zawodzą, ale zastosowanie LLM (Gemini lub Llama 3.2) pozwoliło na poprawną klasyfikację nawet niejednoznacznych i sarkastycznych treści.

- **Cloud (Gemini 3.0)**

– **Przykład 1 (żartobliwa ironia):**

"Nie ma co przesadzać - nawet na zdjęciach widać, że to jakaś góra karton-śmieci a słonko dzisiaj nieźle dawało. Wystarczy soczewka i pożar gotowy. Co innego, gdyby to były opony albo jakieś chemikalkia w sporych ilościach to wtedy owszem - tradycja okresu."

Decyzja modelu: Bezpieczny (Score: 0.10)

Analiza: Gemini tak uzasadnił swoją decyzję: *Wypowiedź ma charakter sceptyczny i ironiczny, ale nie zawiera treści toksycznych, obraźliwych ani nienawiistnych. Autor spekuluje na temat przyczyn pożaru, używając cynicznego tonu w odniesieniu do częstych pożarów składowisk odpadów.*

– **Przykład 2 (obraźliwa ironia):**

"Workaholic ... you mean the Pres who takes of every weekend and goes golfing. Lol - what a sick joke that is. So far his biggest job has been throwing rolls of paper towels at Puerto Ricans. And when he's not on the golf course he spends most of his time on the porcelain throne tweeting out idiot comments."

Decyzja modelu: Toksyczny (Score: 0.65)

Analiza: Gemini uzasadnił, że "tekst zawiera obraźliwe sformułowania skierowane przeciwko osobie publicznej, używając kpin i wyzwisk takich jak 'idiot comments' oraz 'sick joke'. Choć jest to krytyka polityczna, jej forma jest lekceważąca i konfrontacyjna".

- **Local (Toxic-BERT + Llama 3.2)**

– **Przykład 1 (nieszkodliwy sarkazm):**

"Let me guess, you've never been an NFL player right? That's a rhetorical question BTW."

Decyzja modelu: Bezpieczny (Score: 0.001)

Analiza: BERT poprawnie zaklasyfikował wypowiedź jako nisko toksyczną, a Llama odczytała sarkazm: *The tone implies that the speaker believes the person has never been an NFL player and is using sarcasm to express this belief.*

– **Przykład 2 (obraźliwa ironia):**

"Stick your head between your knees and kiss your okole goodbye."

Decyzja modelu: Toksyczny (Score: 0.91)

Analiza: BERT słusznie uznał wypowiedź za toksyczną, a Llama poprawnie wykryła ironię/sarkazm: *The phrase is a euphemistic way of saying goodbye to one's buttocks, which may be considered rude or off-color.*

5.2.4 Kategoria 4: Przypadki Umiarkowanej Pewności Modelu (Graniczne)

Niejednoznaczne przypadki w których modele nie były pewne co do werdyktu i podejmowały decyzje na granicy thresholdu (przewidziana toksyczność w granicach 0.4 – 0.5).

- **Cloud (Gemini 3.0)**

– **Przykład 1:**

"Alaska's great because of "this man." Alaska's been spiraling into the toilet under Republican majority for decades and all we need is one more flush and we're done."

Decyzja modelu: Bezpieczny (Score: 0.45)

Analiza: Gemini podjął poprawną decyzję, uzasadniając: *Tekst zawiera silną krytykę polityczną skierowaną przeciwko konkretnej partii, używając przy tym pogardliwych metafor sugerujących upadek i potrzebę usunięcia danej grupy. Choć jest to wypowiedź agresywna i obraźliwa, mieści się w ramach ostrej retoryki politycznej i nie zawiera bezpośrednich gróźb przemocy ani wulgarności.*

– **Przykład 2:**

"I will never give it up. You may not mind that the president has broken tradition but I sure do. He has lied and so I will always remind people like you who forget that we have a big fat liar for president."

Decyzja modelu: Bezpieczny (Score: 0.45)

Analiza: Gemini niepoprawnie zakwalifikował komentarz jako nieszkodliwy uzasadniając, że *"Tekst zawiera bezpośrednią zniewagę skierowaną do osoby publicznej oraz konfrontacyjny ton wobec rozmówcy. Użycie sformułowania 'big fat liar' jest formą ataku personalnego, co podnosi wynik w kategorii obrazy, jednak w kontekście debaty politycznej nie jest uznawane za skrajną toksyczność.*

- **Local (Toxic-BERT + Llama 3.2)**

– **Przykład 1:**

"I will never give it up. You may not mind that the president has broken tradition but I sure do. He has lied and so I will always remind people like you who forget that we have a big fat liar for president."

Decyzja modelu: Toksyczny (Score: 0.51)

Analiza: BERT poprawnie zaklasyfikował wypowiedź jako toksyczną, jednak z niską pewnością na granicy decyzji. Llama wygenerowała uzasadnienie: *The text contains a threat and name-calling towards the president, which can be perceived as toxic.*

– **Przykład 2:**

"Some of these comments are obviously from Syria's useful idiots. Look, even if it is the rebels who used the gas, what choice did they have? Assad is a monster, and with Putin's backing he is winning. What are they supposed to do, just let him? Not everyone gets to survive war. Sometimes innocents are sacrificed. It's ugly but the rebels need our support and showing the world the risks is justified."

Decyzja modelu: Bezpieczny (Score: 0.49)

Analiza: BERT źle określił wypowiedź jako nietoksyczną, jednak Llama poprawnie wykryła elementy toksykczne: *The text contains a morally ambiguous and potentially dehumanizing justification for supporting rebels who may have used chemical weapons, downplaying the suffering of innocents.*

5.3 Wnioski z analizy błędów

Analiza wykazała, że rozwiązywanie hybrydowe (BERT + Llama) jest skutecznym kompromisem, który zapewnia prywatność danych oraz tanie i szybkie działanie, jednak posiada też pewne wady. Jeśli model BERT uzna treść za całkowicie bezpieczną (bardzo niski wynik przewidywanej toksykczności), zaawansowany model Llama nigdy nie zostanie uruchomiony, nawet jeśli treść zawiera subtelną toksykczność kontekstową. Rozwiązywanie oparte na Gemini, analizujące każdy tekst pełnym modelem LLM, eliminuje ten problem kosztem wyższego czasu przetwarzania i ceny, zapewniając w zamian lepszą dokładność predykcji oraz wsparcie decyzji użytkowników systemu, dzięki wyższej jakości uzasadnieniom.

6 Ograniczenia Rozwiązań

Mimo obiecujących wyników, zaimplementowany system posiada pewne ograniczenia techniczne i merytoryczne, które należy uwzględnić przy ewentualnym wdrożeniu produkcyjnym.

6.1 Ograniczenia Wariantu Lokalnego (Hybrid)

- **Bariera językowa modelu Toxic-BERT:** Model `unitary/toxic-bert` został wytrenowany głównie na korpusach anglojęzycznych. Choć radzi sobie z uniwersalnymi wulgaryzmami, jego skuteczność w detekcji polskiego slangu, odmian gramatycznych czy literówek jest niższa niż w przypadku języka angielskiego.
- **Ślepa plamka (Blind Spot):** Architektura kaskadowa zakłada, że jeśli Toxic-BERT uzna treść za bezpieczną (wynik < 0.4), to Llama 3.2 nie jest uruchamiana. Oznacza to, że subtelna toksykczność bez użycia wulgaryzmów (np. wyrafinowana ironia) może zostać przepuszczona, ponieważ "szybki filtr" jej nie wychwyci.
- **Halucynacje małych modeli:** Model Llama 3.2 (3B) jest modelem relatywnie małym. Zdarza mu się generować uzasadnienia, które brzmią logicznie, ale nie mają pokrycia w tekście, lub sporadycznie łamać format JSON mimo instrukcji systemowych.

6.2 Ograniczenia Wariantu Chmurowego (Gemini)

- **Koszt przy dużej skali:** Wykorzystanie modelu Gemini Flash dla każdego komentarza generuje koszty liniowe. Przy milionach komentarzy miesięcznie może to być bariera nie do przejścia dla małych serwisów.
- **Zależność od API:** Działanie systemu jest uzależnione od dostępności usług Google oraz stabilności łącza internetowego.

7 Wnioski i Kierunki Rozwoju

Realizacja projektu pozwoliła na sformułowanie następujących wniosków:

1. **Wyjaśnialność buduje zaufanie:** Dodanie warstwy generującej uzasadnienie zmienia postrzeżenie narzędzia przez moderatora. System przestaje być "czarną skrzynką", a staje się intelligentnym asystentem. Wskazywanie konkretnych fragmentów tekstu pozwala na błyskawiczną weryfikację decyzji.
2. **Hybryda jako optymalny kompromis:** Połączenie lekkiego klasyfikatora (BERT) z LLM (Llama) to najefektywniejsza strategia dla systemów lokalnych. Pozwala zaoszczędzić około 80-90% zasobów obliczeniowych, uruchamiając kosztowniejsze obliczeniowo narzędzia tylko wtedy, gdy jest to niezbędne.
3. **Pętla zwrotna:** Zaimplementowany mechanizm "Skargi" jest kluczowy dla długoterminowego utrzymania jakości. Pozwala on na zbieranie trudnych przypadków (Edge Cases), które mogą posłużyć do dotrenowania (Fine-tuning) mniejszych modeli w przyszłości.

8 Podsumowanie

W ramach projektu udało się stworzyć funkcjonalny prototyp wyjaśnialnego klasyfikatora tokuszczności, działający w dwóch architekturach. Przygotowane rozwiązanie jest kompletnie - posiada warstwę serwerową (FastAPI), interfejs wizualny (Streamlit) oraz pełną konteneryzację (Docker), co czyni je gotowym do łatwego wdrożenia i testowania.

Wyniki ewaluacji potwierdzają, że współczesne modele LLM, nawet w wersjach skompresowanych (Llama 3.2), są w stanie skutecznie wspierać ludzi w walce z hejtem w sieci, oferując nie tylko binarną klasyfikację, ale także głębsze zrozumienie kontekstu wypowiedzi.