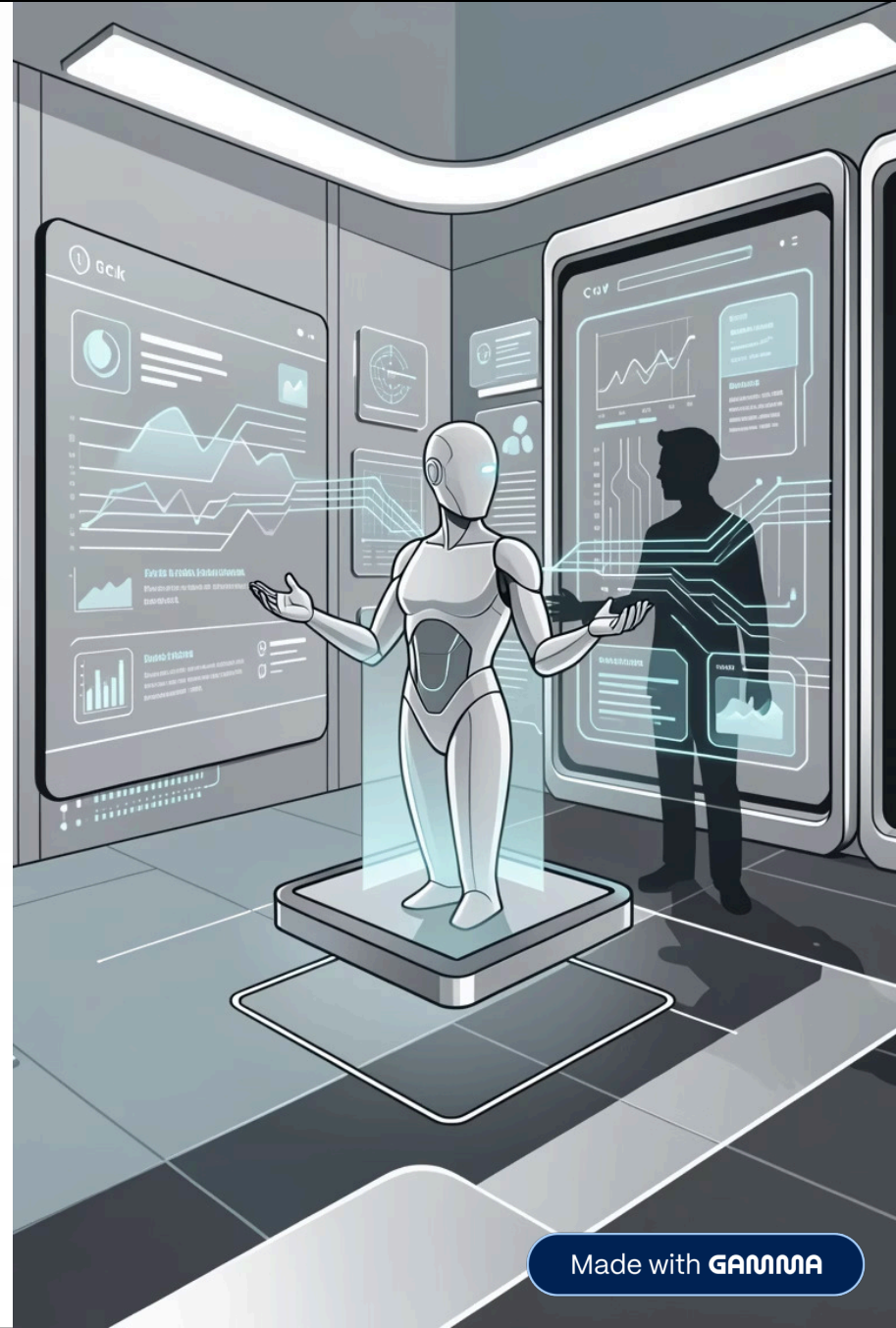
Wyjaśnialny klasyfikator toksyczności wypowiedzi

Maciej Andrzejewski

Krzysztof Osiński



Cel i zakres modelu

Automatyczna klasyfikacja treści internetowych w celu odciążenia zespołów moderacyjnych i zwiększenia bezpieczeństwa online.

1

Klasyfikacja trójstopniowa

- Toksyczne: Ataki, wulgaryzmy, mowa nienawiści.
- Nietoksyczne: Dialog neutralny i pozytywny.
- Graniczne (Borderline): Wymagające analizy kontekstowej.

2

Wyjaśnialność (XAI)

- Precyzja: Podświetlanie słów kluczowych.
- Transparentność: Uzasadnienia w języku naturalnym.
- Głębia: Identyfikacja niuansów (ironia, satyra).

Porównanie Architektur Systemu

Zaimplementowano dwa alternatywne potoki przetwarzania w celu optymalizacji kosztów i jakości.

Model(e)	Google Gemini 3.0 Flash	Hybryda: Toxic-BERT (Faza 1) + Llama 3.2 (Faza 2)
Mechanizm	Zero-shot prompting (Batch API)	BERT filtruje; Llama analizuje trudne przypadki (threshold > 0.4)
Zalety	Najwyższa jakość, świetne PL/EN, brak infrastruktury	Prywatność, 0 zł za API, szybkość
Wyzwania	Koszty API, zależność od dostawcy	Wymaga lokalnych zasobów (GPU/Ollama)

Implementacja i Interfejs

Architektura i Technologie

- **Backend:** Python + FastAPI (asynchroniczna obsługa zapytań).
- **Konteneryzacja:** Docker & Docker Compose (łatwe wdrożenie).
- **Logika biznesowa:** Pydantic (wymuszanie ustrukturyzowanych odpowiedzi JSON).
- **Silniki lokalne:** HuggingFace (transformers) oraz Ollama.

Interfejs i Komunikacja

- **REST API:** Endpoint /analyze-batch do masowego przetwarzania.
- **Dashboard (Streamlit):** Wizualizacja wyników i wykresów w czasie rzeczywistym.
- **Feedback Loop:** Mechanizm zgłaszania błędów (re-analiza).

Wyniki Ilościowe i Skuteczność

Ewaluacja wykazała wyraźny podział na wysoką precyzję semantyczną (Gemini) oraz efektywność kosztową (Hybrid).

Model	Dokładność	Obsługa języka	Koszt / Prywatność
Gemini 3.0 Flash (Cloud)	85%	Wybitna (PL/EN)	Płatny / Dane w chmurze
Hybrid (BERT + Llama)	72%	Dobra (Głównie EN)	Darmowy / Lokalnie

Metodologia i Zbiór Danych Testowych

Charakterystyka zbioru danych

- **Wolumen:** 400 zanonimizowanych komentarzy (PL i EN).
- **Specyfika:** Uwzględnienie **hard negatives** (trudne przypadki mylące proste filtry).
- **Różnice regionalne:** Wyższa tolerancja na toksyczność w zbiorze polskim; surowa moderacja w zbiorze angielskim.
- **Etykietowanie:** Ręczne (Ground Truth) z dodatkowymi metadanymi.

Analiza Jakościowa - Gdzie modele się mylą?

Nadwrażliwość (False Positives) - Problem Gemini

Model miewa problemy z wystąpieniem słów nacechowanych negatywnie w neutralnym lub medycznym kontekście.

- **Przykład:** "Jeżeli przyjąć za moment śmierci obumarcie mózgu to on nie żyje od lat"
- **Werdykt:** Toksyczny (0.65).
- **Błąd:** Model uznał to za atak na intelekt, nie rozpoznając kontekstu dyskusji medycznej.

Brak Nadwrażliwości (False Positives) - Przewaga Hybrydy

Toxic-BERT poprawnie odróżnił wszystkie nietoksyczne komentarze. Model jest konserwatywny – "niewinny, dopóki wina nie jest ewidentna".

- **Zaleta:** Brak niesłusznej cenzury i frustracji użytkowników z powodu blokady treści.
- **Korzyść biznesowa:** Brak fałszywych alarmów = zerowe marnotrawstwo zasobów obliczeniowych (Llama nie jest uruchamiana bez wyraźnej potrzeby).

Analiza Jakościowa - Gdzie modele się mylą?

Nadwrażliwość (False Positives) - Problem Gemini

- **Przykład:** "Trudno powiedzieć czy sarkazm czy skrajna głupota i odklejenie od realiów."
- **Werdykt:** Toksyczny (0.60).
- **Błąd:** Model uznał, że komentarz ma charakter obraźliwy, atakując intelekt rozmówcy, co nie zgadza się z oceną ludzi (moderatorów forum).

Brak Nadwrażliwości (False Positives) - Przewaga Hybrydy

Model doskonale radzi sobie z jawną, słownikową toksycznością, ale nie zauważa ’niuansów’.

- **Automatyzacja procesów:** W kontekście automatyzacji jest to cecha pożądana - lepiej przepuścić kilka toksycznych komentarzy do ręcznej moderacji, niż blokować niewinne wypowiedzi.

Analiza Jakościowa - Gdzie modele się mylą?

Niewykryte Zagrożenia (False Negatives) - Niewielki Problem Gemini

Model chmurowy czasem przepuszcza krytykę polityczną, którą ludzie uznali za zbyt ofensywną.

- **Przykład:** "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."
- **Werdykt:** Bezpieczny (0.25).
- **Błąd:** Model uznał, że wspomnienie o orientacji seksualnej w kontekście argumentu, że nie powinna ona stanowić tarczy przed odpowiedzialnością, nie stanowi bezpośredniego ataku nienawiści, co koliduje z osądem ludzi.

Niewykryte Zagrożenia (False Negatives) - Słabość Hybrydy

Jeśli pierwszy, prosty model (BERT) nie wykryje zagrożenia, zaawansowana Llama nie zostanie uruchomiona, co może prowadzić do przeoczeń.

- **Przykład:** "I hope all you Far Left whiners don't go out and commit suicide based on the Comey hearing today. […] 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."
- **Werdykt:** Bezpieczny (0.28).
- **Błąd:** BERT nie wykrył niestosownego odwołania do samobójstwa i ataku na grupę o innych poglądach politycznych.

Analiza Jakościowa - Gdzie modele się mylą?

Niewykryte Zagrożenia (False Negatives) - Gemini

Model chmurowy bywa "zbyt poprawny", nie uznając ostrej debaty za toksyczność.

- **Przykład:** "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)."
- **Werdykt:** Bezpieczny (0.30).
- **Błąd:** Model uznał to jedynie za krytykę osób publicznych bez bezpośredniej toksyczności/ataku wobec rozmówcy.

Niewykryte Zagrożenia (False Negatives) - Słabość kaskady Hybrid

W pojedynczych przypadkach model "przepuszcza" ewidentnie toksyczne wypowiedzi.

- **Przykład:** "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 then address a problem."
- **Werdykt:** Bezpieczny (0.13).
- **Błąd:** Bert nie wykrył rasizmu oraz nie wywołał wyjaśnienia od Llamy, która gdyby ją wymusić, poprawnie wskazałaby na mowę nienawiści i krzywdzące stereotypy.

Sukcesy - Ironia i Kontekst

Oba systemy (gdy LLM zostanie uruchomiony) świetnie radzą sobie z czytaniem między wierszami.

Żartobliwa Ironia (Gemini)

"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 chemikalia w sporych ilościach to wtedy owszem - tradycja okresu."

- **Werdykt:** Bezpieczny (0.10).
- **Zrozumienie:** Model zrozumiał ironiczny ton i cyniczne nawiązanie do podpaleń śmieci.

Sarkazm (Bert/Llama)

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

- **Werdykt:** Bezpieczny (0.001).
- **Zrozumienie:** Bert poprawnie zaklasyfikował niską toksyczność, a Llama odczytała sarkazm.

Sukcesy – Ironia i Kontekst

Modele stosunkowo dobrze radzą sobie z odróżnieniem żartobliwego i ofensywnego sarkazmu.

Ofensywna Ironia (Gemini)

"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."

- **Werdykt:** Toksyczny (0.65).
- **Zrozumienie:** Model zrozumiał kpiący ton wypowiedzi w stosunku do osoby prezydenta.

Slang i Eufemizmy (Bert/Llama)

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

- **Werdykt:** Toksyczny (0.91).
- **Zrozumienie:** Model poprawnie wychwycił ofensywny/ironiczny ton i zinterpretował eufemizm "okole" jako obraźliwy.

Umiarkowana Pewność - Graniczne Sukcesy

W niektórych niejednoznacznych przypadkach modele nie były pewne co do werdyktu i podejmowały decyzje wahające się na granicy thresholdu.

Gemini

"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."

- **Werdykt:** Bezpieczny (0.45).
- **Zrozumienie:** Model poprawnie zauważył atak słowny na osobę publiczną, ale uznał to za toksyczność "nie skrajną" w kontekście debat politycznych.

Hybryda - Bert/Llama

"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."

- **Werdykt:** Toksyczny (0.51).
- **Zrozumienie:** Bert poprawnie zaklasyfikował wypowiedź jako toksyczną, jednak z niską pewnością - na granicy decyzji. Llama wygenerowała uzasadnienie mówiące, że tekst zawiera słowne obelgi w kierunku prezydenta.

Umiarkowana Pewność - Graniczne Porażki

Gemini

"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."

- **Werdykt:** Bezpieczny (0.45).
- **Zrozumienie:** Model wyłapał pogardliwy ton i krytykę polityczną, ale uznał ją za niegroźną ze względu na brak gróźb i wulgaryzmów.

Hybryda - Bert/Llama

"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."

- **Werdykt:** Bezpieczny (0.49).
- **Zrozumienie:** Przypadek mocno na granicy - Bert źle określił wypowiedź jako nietoksyczną, jednak Llama poprawnie wykryła elementy toksyczne: dehumanizację, usprawiedliwianie przemocy wobec niewinnych.

Ograniczenia Systemu

Wariant Lokalny (Hybrid)

- **Bariera językowa:** BERT zoptymalizowany pod EN; słabiej radzi sobie z PL.
- **Efekt "Ślepej plamki":** Szybki filtr nie wykryje zagrożenia, zaawansowana Llama nie zostanie aktywowana.
- **Halucynacje:** Llama 3.2 (3B) może generować błędne uzasadnienia lub błędy w formacie danych.

Wariant Chmurowy (Gemini)

- **Ekonomia skali:** Koszty rosną liniowo z liczbą komentarzy (bariera dla dużych serwisów).
- **Vendor Lock-in:** Zależność od stabilności API Google i łącza internetowego.

Kluczowe Wnioski

- ### Wyjaśnialność (XAI) to fundament

  Uzasadnienia w języku naturalnym budują zaufanie i pozwalają na weryfikację decyzji AI.

- ### Strategia Hybrydowa = Efektywność

  Połączenie BERT + LLM pozwala oszczędzić **80-90% zasobów**, rezerwując moc obliczeniową dla trudnych przypadków.

- ### Pętla zwrotna

  Mechanizm "Skargi" to źródło cennych danych (Edge Cases) do przyszłego dotrenowania modeli.

Wnioski i Rekomendacje

1

Architektura Hybrydowa (BERT + Llama)

- **Zaleta:** Ochrona prywatności, brak kosztów.
- **Wada:** "Wąskie gardło" BERT – ryzyko przepuszczenia toksycznej treści.

2

Architektura Cloud (Gemini)

- **Zaleta:** Najwyższa jakość uzasadnień, głębokie rozumienie niuansów językowych.
- **Wada:** Wysoka cena, tendencja do nadmiernej cenzury (nadwrażliwość).

Rekomendacja końcowa:

- Dla serwisów o wysokim rygorze bezpieczeństwa – **Gemini**.
- Dla otwartych dyskusji politycznych i e-commerce – **Model Hybrydowy** z obniżonym progiem aktywacji modelu Llama.

Dziękujemy!