



Milena Królikowska <mj.krolikowska@uw.edu.pl>

ZAPROSZENIE DO UDZIAŁU WE WSTĘPNYCH KONSULTACJACH RYNKOWYCH //

Nauka.Sprawdza

Nauka.Sprawdza <nauka.sprawdza@cwid.uw.edu.pl>

7 sierpnia 2025 14:59

Do: Milena Królikowska <mj.krolikowska@uw.edu.pl>

Demagog

----- Forwarded message -----

From: Marcel Kiełtyka <marcel.kieltyka@demagog.org.pl>

Date: Thu, Jul 24, 2025 at 5:58 PM

Subject: Re: ZAPROSZENIE DO UDZIAŁU WE WSTĘPNYCH KONSULTACJACH RYNKOWYCH // Nauka.Sprawdza

To: Agnieszka Tobijasiewicz <agnieszka.tobijasiewicz@demagog.org.pl>, Nauka.Sprawdza <nauka.sprawdza@cwid.uw.edu.pl>

Cc: Zarząd <zarzad@demagog.org.pl>

Dzień dobry,

Pani Tatiano, poniżej przesyłam odpowiedź:

Nie istnieją obecnie żadne narzędzia automatyczne o wystarczającej dokładności, które same oceniałyby treści jako „prawdziwe” lub „fałszywe” bez znaczącego ryzyka błędu. Przykładowo, badania podkreślają, że algorytmy wykorzystujące NLP i ML do wykrywania dezinformacji w praktyce napotykają na trudności: weryfikacja dużych ilości danych, brak reprezentatywnych zbiorów treningowych i zjawisko tzw. deepfake wymagają stałego nadzoru człowieka. Nowe formy dezinformacji („deep fakes”) pozostają obecnie trudne do wykrycia algorytmicznie i często **wymagają interwencji ludzkiej**. W konsekwencji rolę klasyfikatora prawdziwości dalej pełni w praktyce **zweryfikowany fact-checker**, działający według ustalonej metodyki weryfikacji źródeł i dowodów.

Dostępne narzędzia potrafią mierzyć np. **sentiment publikowanych treści** lub wykrywać potencjalne **nieautentyczne konta** (np. boty). Jednakże należy podkreślić, że takie metody są zawodne. Badania pokazują, że algorytmy automatycznego wykrywania botów (np. Botometer) generują znaczące odsetki **fałszywych alarmów i pominieć** – ludzie są oznaczani jako boty i odwrotnie. Analiza sentymentu bazuje zwykle na uproszczonych modelach, które mogą mylić ironię czy dwuznaczne wypowiedzi. Z tego powodu zarówno wyniki analizy nastroju społecznego, jak i oceny wiarygodności kont należy traktować z dystansem i poddawać weryfikacji eksperckiej.

W praktyce AI może pomóc **przyspieszyć przetwarzanie danych** (np. wstępne filtrowanie lub grupowanie treści), ale dziś pełni raczej funkcję pomocniczą. Nawet zaawansowane modele bazujące na sztucznej inteligencji są „trenowane” na prostych zbiorach binarnych (prawda/fałsz), co powoduje utratę niuansów, charakterystyczną dla wieloznacznych informacji prasowych. NiemanLab zauważa, że AI może pomóc m.in. w szybkiej detekcji gwałtownie rozprzestrzeniających się relacji (zwłaszcza multimedialnych), ale **istotny kontekst i ocena prawdziwości** wciąż wymaga eksperckiego spojrzenia człowieka. W miarę rozwoju technologii możliwa jest coraz szersza automatyzacja rutynowych etapów (np. identyfikacja rozgłosu czy korelacja informacji z bazą faktów), lecz obecnie jakiegokolwiek decyzje o prawdziwości treści powinien podejmować człowiek-fact-checker na podstawie metodyki weryfikacji.

Demagog, w ramach programu weryfikacji na Facebooku (Meta), stosuje sześć kategorii oceny treści: **Falsz, Przeróbka, Częściowy fałsz, Brak kontekstu, Satyra** oraz **Prawda**. Definicje tych kategorii są następujące:

- **Falsz:** Informacja nie zgadza się z żadnym wiarygodnym źródłem, bazuje na nieaktualnych danych sprzecznych z nowszymi faktami lub zawiera jedynie fragmenty prawdziwych informacji, których brak kluczowego kontekstu fałszywie zmienia przekaz. Demagog uznaje wtedy twierdzenie za fałszywe.
- **Przeróbka:** Materiały audiowizualne (zdjęcia, wideo, dźwięk) poddane zostały obróbce przekłamującej pierwotny kontekst lub treść. Na przykład zmanipulowany obraz lub wideo, sklejone fragmenty czy dodane efekty audio, które wprowadzają odbiorcę w błąd.
- **Częściowy fałsz:** Wypowiedź lub informacja zawiera zarówno elementy prawdziwe, jak i błędne. Części treści mogą być zgodne z rzeczywistością, ale miesza się to z przeinaczonymi danymi, błędnymi cytatami bądź wyciągniętymi wnioskami, co prowadzi do wprowadzenia w błąd.
- **Brak kontekstu:** Prezentowane dane lub cytaty zostały oderwane od istotnego tła lub uzupełniających informacji. Samo wystąpienie fragmentu może sugerować fałszywe wnioski, mimo że sformułowanie oryginalne nie zawiera bezpośredniego kłamstwa. Przykładem jest wycięcie wypowiedzi z komentarzem sugerującym nierzetelność, którą należy rozumieć inaczej przy poznaniu całego kontekstu.
- **Satyra:** Treść ma charakter ironiczny lub humorystyczny i nie jest jednoznacznie oznakowana jako żart. W związku z tym osoby bez wskazówek co do konwencji mogą uznać ją za poważną informację. Demagog klasyfikuje materiał jako satyryczny, gdy brak jest oznaczenia gatunku satyrycznego, a mimo to przekaz ewidentnie kpi z pewnych faktów lub postaci.
- **Prawda:** Informacja nie zawiera w istotny sposób żadnych błędnych informacji ani manipulacji. Wszystkie cytowane dane i fakty są zgodne z dostępnymi, wiarygodnymi źródłami. W tym wypadku nie stwierdza się żadnej dezinformacji ani wypaczeń.

Kategorie te stosuje Meta na potrzeby programu fact-checking (tzw. „Content Ratings”). Oceny: **False, Partly False, Altered Photo/Video, Missing Context, Satire** (oraz zasadniczo „True” jako brak dezinformacji). Meta opisuje „False” jako wypowiedzi bez oparcia w faktach, „Partly False” – zawierające nieścisłości, „Altered Photo/Video” – zmodyfikowane multimedialnie materiały, „Missing Context” – przekaz wymagający dopowiedzenia, a „Satire” – celową karykaturę faktów. Te kategorie są analogiczne do powyższych ocen Demagoga i funkcjonują globalnie (w tym również na polskim rynku) w programach weryfikacji wspieranych przez Meta.

Inne podejścia – mis-/dezinformacja, DISARM

Ponieważ dezinformacja to złożone zjawisko, w literaturze proponuje się też alternatywne podziały. Często rozróżnia się **misinformację** (błędną informację rozprzestrzeganą bez świadomości jej fałszu) od **dezinformacji** (świadomego wprowadzania w błąd). Ramy te pomagają rozumieć intencje nadawców. Dodatkowo istnieją klasyfikacje TTP (tactics, techniques, procedures) opisujące całe kampanie. Przykładem jest model **DISARM** (Disinformation Analysis and Resilience Mapping) – metoda opisująca fazy i metody dezinformacji. DISARM pozwala analitykom mapować „łańcuch dezinformacyjny” (planowanie, produkcję, dystrybucję itp.) i wskazuje punkty w łańcuchu, gdzie można przeciwdziałać. Na przykład **amplifikację** przez sieci botów (etap rozpowszechniania) można ograniczać poprzez polityki platform wykrywające skoordynowane działania i blokujące nieautentyczne konta. Model DISARM zwraca uwagę, że skuteczne przeciwdziałanie wymaga opisu nie tylko treści (co zostało powiedziane), ale też **technik manipulacji** (jak powstała kampania, jaka jest rola skoordynowanych kont).

„Fingerprints of misinformation” – badania i krytyka

W literaturze naukowej pojawił się termin „fingerprints of misinformation”, np. artykuł w czasopiśmie Nature („The fingerprints of misinformation”). Autorzy sugerują, że teksty dezinformacyjne mają statystycznie inne cechy językowe niż rzetelne informacje – są prostsze („łatwiejsze do czytania”) i silniej nacechowane emocjonalnie. Jednak podejście to budzi kontrowersje. Krytycy argumentują, że powierzchowne cechy językowe nie są wystarczającym wskaźnikiem prawdziwości wypowiedzi. Wskazują, że nawet jeśli pewne cechy (jak emocjonalny język) występują częściej w fałszywych informacjach, niekoniecznie predysponują do uznania treści za fałszywą z uwagi na tzw. efekt bazowy (większość treści w sieci jest prawdziwa). Williams podkreśla, że prawdziwość wypowiedzi zależy od zgodności ze światem, nie od stylu wypowiedzi i że brak „odcisków palców” umożliwia dezinformatorom unikać wykrycia. W praktyce Demagog traktuje więc wszelkie takie algorytmy tylko jako pomocne wskazówki – ostateczny werdykt należy do fact-checkera, a nie do automatu.

Zależność od metodyki w projekcie

Zwracamy uwagę, że **wybrana metodyka weryfikacji** będzie kluczowa. Skuteczność narzędzi i procesów oceny treści zależy od określonych procedur – przykładowo ustalenie, jakie źródła uznajemy za wiarygodne, kto zatwierdza oceny i jak informacja jest dokumentowana. Demagog dysponuje doświadczeniem opracowywania takich metodyk (w tym zgodnie ze standardami Meta, DISARM czy wypowiedzi polityków), jednak ostateczne efekty monitoringu będą zależały od wspólnie przyjętych kryteriów na etapie realizacji projektu.

Monitorowanie aktywności nieautentycznej

Ważnym elementem monitoringu jest też wykrywanie **nieautentycznej aktywności społecznej**. Przykładowo, skoordynowane działania to technika manipulacji, gdzie sieć fałszywych i „wspomaganych” kont działa synchronicznie w mediach społecznościowych. Badacze opisują to jako użycie mieszanki prawdziwych, fałszywych i powielanych kont do realizacji sieciowej kampanii. Detekcja takich wzorców (np. powtarzalne wiadomości, współdzielenie treści przez wiele kont, podobne sygnały językowe) pozwala wcześniej wskazać organizowane operacje dezinformacyjne. W projektowaniu monitoringu należy uwzględnić również algorytmy wykrywające koordynację (np. nagły, zorganizowany wzrost publikacji na określony temat) czy filtrujące content pod kątem cech botów.

Podsumowując: skuteczne wykrywanie dezinformacji w nauce wymaga połączenia automatycznych narzędzi monitoringu z profesjonalną metodyką fact-checkingu. Automatyczne systemy mogą wspierać pracę ekspertów (np. wykrywając potencjalne ogniska dezinformacji), lecz to człowiek-fact-checker decyduje o ostatecznej weryfikacji zgodnie z przyjętymi kategoriami i procedurami. Wszelkie narzędzia (np. analizy sentymentu czy wykrywanie botów) należy stosować ostrożnie i weryfikować wyniki. W efekcie pełna wiarygodność monitoringu zależy zarówno od użytych technologii, jak i od przyjętej metodologii kontroli i interpretacji danych.



Marcel Kiełtyka

Członek zarządu, Dyrektor ds.
Komunikacji i PR.

tel. +48 507 099 993

marcel.kieltyka@demagog.org.pl

 **DEMAGOG**
demagog.org.pl



Kontrolujemy polityków. **Wesprzyj nas na Patronite!**

[Ukryto cytowany tekst]