PRAWDOWID: szybka i łatwa weryfikacja informacji oparta na wiarygodnych źródłach

Zbuduj proste i wielokanałowe narzędzie, które pozwoli obywatelowi sprawdzić doniesienie tak łatwo, jak wysłać wiadomość. System powinien korzystać z wiarygodnych źródeł i baz zweryfikowanych informacji, pokazywać podstawy odpowiedzi oraz uczciwie wstrzymywać się od rozstrzygnięcia, gdy dostępne materiały są niewystarczające.

1. Wprowadzenie

MSZ realizuje zadania publiczne w obszarze przeciwdziałania zagranicznym manipulacjom informacyjnym i ingerencjom w środowisku informacyjnym (*Foreign Information Manipulation and Interference*, FIMI). Odpowiada również za przekazywanie obywatelom rzetelnych, zrozumiałych i aktualnych informacji dotyczących sytuacji międzynarodowej. Głównym odbiorcą rozwiązania jest obywatel. Narzędzie powinno być użyteczne także dla osób o niskich kompetencjach cyfrowych, bez konieczności poznawania metod weryfikacji informacji lub instalowania specjalistycznego oprogramowania. Pytania mogą napływać różnymi kanałami, dotyczyć szybko zmieniających się wydarzeń i łączyć fakty, opinie, materiały archiwalne oraz treści wyrwane z kontekstu.

Problem nie ogranicza się do wykrywania fałszywych twierdzeń. Użytkownik może przesłać tekst, link, zrzut ekranu, zdjęcie, nagranie lub wielowątkowe pytanie. Część treści może być prawdziwa, ale nieaktualna; część może zawierać poprawne dane prowadzące do błędnego wniosku; w części przypadków może jeszcze brakować wystarczających dowodów. Narzędzie powinno pomagać odróżniać fakty od opinii, rozpoznawać ograniczenia źródeł, wskazywać istotny kontekst oraz unikać kategorycznych ocen, gdy dowody są niewystarczające lub sprzeczne.

Generatywna sztuczna inteligencja potrafi tworzyć przekonujące odpowiedzi, jednak płynność języka nie gwarantuje ich poprawności. Dlatego rozwiązanie powinno opierać się na jasno opisanym zbiorze źródeł, a model językowy wykorzystywać przede wszystkim do rozumienia pytania, wyszukiwania i porównywania materiałów oraz formułowania odpowiedzi. Każda istotna teza przedstawiana jako ustalenie systemu powinna prowadzić do konkretnego źródła. Preferowane są dokumenty urzędowe, dane publiczne oraz korpusy zweryfikowanych informacji udostępniane przez organizacje fact-checkingowe.

Rozwiązanie może obsługiwać różne rodzaje treści, np. tekst i obraz, porównywać źródła oraz oceniać siłę dostępnych dowodów. Powinno również zawierać zabezpieczenia ograniczające ryzyko prompt injection i innych prób manipulowania usługą.

Rozwiązanie powinno być projektowane z myślą o polskim języku i kontekście instytucjonalnym, ale umożliwiać późniejsze rozszerzenie na inne języki i obszary tematyczne. Punktem odniesienia mogą być standardy IFCN i EFCSN, zasady EDMO oraz praktyki NASK dotyczące analizy dezinformacji. Projekt musi respektować prywatność, prawa autorskie, bezpieczeństwo informacji i zasady przejrzystego stosowania SI.

2. Wyzwanie

Jak umożliwić obywatelowi szybkie, bezpłatne po stronie usługi i zrozumiałe sprawdzenie pogłoski lub doniesienia przez kanał, którego już używa, przy jednoczesnym pokazaniu źródeł i granic wiedzy systemu?

Zespół ma zaprojektować i uruchomić działający prototyp, który przyjmuje co najmniej tekst lub adres URL. Rozwiązanie powinno rozpoznawać twierdzenia wymagające sprawdzenia, wyszukiwać informacje w udostępnionym lub jasno opisanym zbiorze źródeł, oceniać ich aktualność i znaczenie, a następnie przedstawiać zwięzły wynik wraz z uzasadnieniem.

- Dla każdego twierdzenia system powinien przedstawić jeden z wyników: potwierdzone, fałszywe, wprowadzające w błąd, brak wystarczających dowodów albo nie podlega weryfikacji. Awaria techniczna, brak możliwości wykonania sprawdzenia i potrzeba doprecyzowania pytania nie są werdyktami i muszą być komunikowane oddzielnie.

- System powinien oddzielać materiały źródłowe od interpretacji modelu oraz wskazywać co najmniej tytuł, wydawcę, datę i link, a w miarę możliwości także wykorzystany fragment.

- Siła werdyktu w odpowiedzi powinna wynikać z jawnych przesłanek, np. jakości i aktualności źródeł, zgodności materiału z twierdzeniem oraz liczby niezależnych potwierdzeń.

- Jeżeli dowody są niewystarczające, system powinien wstrzymać się od rozstrzygnięcia i wyjaśnić, czego nie udało się ustalić. W sprawach wysokiego ryzyka oraz powtarzających się pytań, dla których brakuje źródeł powinien powiadomić właściwego użytkownika wskazanego przez administratora systemu (np. moderator, analityk).

- Instrukcje znalezione w analizowanych materiałach nie mogą zmieniać zasad działania systemu. Rozwiązanie powinno być odporne na prompt injection i inne manipulacje wymierzone w usługę.

3. Oczekiwany rezultat

Oczekiwanym rezultatem jest działający prototyp end-to-end. Podstawowym odbiorcą jest obywatel potrzebujący szybkiej i zrozumiałej weryfikacji. Dodatkowo można uwzględnić potrzeby analityczne administratora systemu (np. alert dot. powtarzających się podobnych pytań bez odpowiedzi).

Minimalny zakres funkcjonalny (MVP)

- wejście: tekst lub adres URL przez responsywny interfejs WWW oraz co najmniej jeden dodatkowy kanał komunikacji albo adapter demonstracyjny, np. SMS, WhatsApp, Messenger, Instagram, Discord lub rozwiązanie równoważne (z wyłączeniem platform niszowych oraz Telegram i WeChat);

- rozpoznawanie jednego lub kilku twierdzeń wymagających sprawdzenia;

- wyszukiwanie informacji w jasno opisanym zbiorze obejmującym co najmniej dwa typy źródeł oraz prezentowanie materiałów wykorzystanych do odpowiedzi;
- wynik, krótkie uzasadnienie, ocena siły podstawy odpowiedzi i czytelna informacja o ograniczeniach;
- odpowiedź po polsku, prostym i zrozumiałym językiem;
- podstawowy zapis przebiegu sprawdzenia, obejmujący pytanie, czas, wykorzystane źródła i wynik;
- wstrzymanie się od rozstrzygnięcia, gdy nie ma wystarczających dowodów, bez tworzenia pozornej odpowiedzi;
- pełna ścieżka end-to-end bez ręcznej ingerencji twórców, od przyjęcia zgłoszenia do przedstawienia odpowiedzi i zapisania przebiegu sprawdzenia;

Funkcje premiowane

- weryfikacja obrazu i dźwięku: OCR, transkrypcja, analiza metadanych, porównywanie obrazów, wyszukiwanie wcześniejszych wystąpień materiału;
- porównywanie źródeł i „łańcuch dowodowy" prowadzący od twierdzenia do materiałów potwierdzających, obalających lub uzupełniających kontekst;
- wielojęzyczne wyszukiwanie w źródłach z odpowiedzią po polsku;
- otwarty lub wymienny katalog źródeł wraz z zasadami ich kwalifikowania, aktualizowania i wykluczania;
- architektura umożliwiająca podłączanie kolejnych kanałów, modeli i źródeł bez przebudowy całego rozwiązania;
- dopasowanie treści, adresu URL lub obrazu do katalogu znanych incydentów albo korpusu zweryfikowanych informacji; sama obecność domeny lub kanału w katalogu incydentów nie może stanowić podstawy werdyktu o konkretnej treści, ale może wyzwalać ostrzeżenie w jawnie określonych przypadkach;
- eksport danych, np. do CSV, JSON, integracja z narzędziami analitycznymi, np. OpenCTI i DISARM;
- kontrolowana archiwizacja materiału wskazanego przez użytkownika, jeżeli jest zgodna z prawem, warunkami korzystania z serwisu, zasadami retencji oraz minimalizacji danych;
- funkcje premiowane są oceniane tylko wtedy, gdy rozwiązanie spełnia wymagania podstawowe;

- ograniczenie zależności od dostawców spoza EOG, możliwość audytu lub samodzielnego hostowania oraz przejrzysta informacja o miejscu przetwarzania danych.

- zamknięta pętla informacji: możliwość utworzenia sprawy moderatora dla nierozstrzygniętego zgłoszenia, grupowanie podobnych zgłoszeń.

4. Wymagania formalne

Projekt przesyłany do oceny musi zawierać:

- nazwę zespołu, tytuł i opis rozwiązania do 4 000 znaków;

- prezentację PDF, maksymalnie 12 slajdów, obejmującą problem, użytkowników, architekturę, demonstrację, wyniki testów, ryzyka i możliwe dalsze kroki;

- film demonstracyjny dostępny pod linkiem, maksymalnie 3 minuty;

- link do repozytorium kodu wraz z instrukcją uruchomienia;

- krótki opis architektury, przepływu danych, użytych modeli, dostawców i najważniejszych ograniczeń;

- opis wykorzystanych źródeł, ich licencji, sposobu kwalifikacji i aktualizacji;

- wyniki testów według sekcji 6, wraz z danymi pozwalającymi odtworzyć rezultat;

- krótką listę najważniejszych ryzyk, obejmującą co najmniej halucynacje, stronniczość źródeł, prywatność, prompt injection, dostępność i nadużycia;

Powyższe informacje mogą znajdować się w README lub w osobnych, krótkich plikach. Dodatkowo projekt może zawierać publiczną demonstrację, kontener Docker, specyfikację OpenAPI, makiety kolejnych kanałów, dashboard jakości i materiały badawcze. Klucze API, dane osobowe i sekrety nie mogą znaleźć się w repozytorium ani materiałach demonstracyjnych.

5. Wymagania techniczne

Obszar	**Wymaganie**
Architektura	Sposób budowy rozwiązania pozostaje decyzją zespołu. Należy jednak oddzielić obsługę kanałów od logiki weryfikacji oraz umożliwić wymianę modelu lub dostawcy bez przebudowy całego systemu.

Fakty	Każda teza faktograficzna przedstawiana jako ustalenie systemu musi prowadzić do dowodu. Jeżeli materiał jest niewystarczający, system wstrzymuje się od rozstrzygnięcia i wyjaśnia ograniczenia. Awarię, niejasne pytanie i wyjście poza zakres należy komunikować oddzielnie.
Źródła	Demonstracyjny korpus powinien obejmować co najmniej 100 dokumentów lub rekordów, minimum 10 domen i co najmniej dwa typy źródeł. Powinien zawierać źródła pierwotne lub urzędowe oraz wiarygodne źródła wtórne, chyba że zespół uzasadni inne podejście. Lista źródeł musi być dostępna do audytu.
Aktualność	Przechowywanie znacznika czasu publikacji i pobrania. Preferowanie danych aktualnych dla pytania oraz ostrzeganie o materiale archiwalnym.
Bezpieczeństwo	Walidacja wejścia, separacja instrukcji systemowych od treści źródeł, ograniczenia narzędzi, ochrona sekretów, limitowanie żądań i logowanie zdarzeń.
Prywatność	Minimalizacja danych, automatyczne maskowanie oczywistych danych osobowych, określona retencja oraz możliwość usunięcia sesji demonstracyjnej.
Obserwowalność	Rejestrowanie czasu odpowiedzi, wykorzystania modeli, błędów wyszukiwania, przypadków wstrzymania się od rozstrzygnięcia oraz szacunkowego kosztu odpowiedzi.
Dostępność	Responsywny interfejs, czytelny kontrast i język zrozumiały dla osoby bez wiedzy technicznej.
Kanały i dostęp	Interfejs WWW i drugi kanał albo adapter demonstracyjny powinny korzystać z tej samej logiki weryfikacji. Dla tego samego zapytania, przy tych samych źródłach i warunkach, wynik merytoryczny powinien być spójny niezależnie od kanału.

Pętla informacji	Jeżeli zespół wdraża pętlę informacji, nierozstrzygnięte zgłoszenie powinno móc utworzyć sprawę moderatora. System musi odróżniać rekomendację przekazania od faktycznego przekazania sprawy.
Zależności zewnętrzne	Zespół opisuje wymagane klucze, limity, koszty, miejsce przetwarzania danych i zachowanie podczas awarii. Brak klucza, limit lub awaria nie mogą być przedstawiane jako wynik merytoryczny.
Multimodalność	Funkcje multimodalne są premiowane, ale nieobowiązkowe. Jeżeli rozwiązanie wykorzystuje zewnętrzne API obrazu lub dźwięku, zespół ujawnia koszt, zakres działania, retencję danych i ograniczenia usługi.

Dopuszcza się dowolne, zgodne z prawem rozwiązania techniczne: modele komercyjne i otwarte, wyszukiwarki pełnotekstowe, bazy wektorowe oraz podejścia hybrydowe. Zespół powinien uzasadnić najważniejsze wybory, określić koszty i zależności od dostawców oraz wskazać możliwość migracji. Sama obecność określonej technologii, np. bazy wektorowej, nie jest punktowana; oceniana jest jakość działania.

W zakresie kanałów komunikacji dopuszcza się SMS, WhatsApp Business Platform, Messenger, Instagram, Discord oraz inne rozwiązania wybrane i uzasadnione przez zespół. Organizator nie zapewnia kont, numerów testowych lub kluczy, dlatego w uzasadnionych przypadkach dopuszczalny jest adapter demonstracyjny. Obsługa zapytania od odebrania wiadomości do wysłania odpowiedzi musi odbywać się automatycznie.

Do wyszukiwania wcześniejszych wystąpień obrazu można wykorzystać np. Google Cloud Vision Web Detection, TinEye, lokalny indeks perceptual hash lub inne legalnie dostępne rozwiązanie. Zespół musi ujawnić zakres działania, koszty, retencję obrazu i ograniczenia usługi.

Ograniczenia i zakazy

- Nie wolno przedstawiać wyniku modelu językowego jako faktu bez weryfikowalnej podstawy źródłowej.
- Nie wolno obchodzić zabezpieczeń serwisów i płatnych baz.
- Nie wolno przesądzać o prawdziwości konkretnego twierdzenia wyłącznie na podstawie domeny, autora, kanału, popularności, liczby udostępnień ani etykiety wygenerowanej przez ten sam model.
- Status źródła, w tym objęcie podmiotu sankcjami Unii Europejskiej, nie może samodzielnie przesądzać o prawdziwości konkretnego twierdzenia. W takim przypadku system powinien pokazać odrębne, neutralne ostrzeżenie o statusie źródła, jeżeli informacja jest aktualna, sprawdzalna i istotna dla użytkownika.

- Nie wolno deklarować pełnego wyszukiwania odwrotnego obrazu bez działającego API lub jasno ograniczonej lokalnej bazy.

- Nie wolno archiwizować pełnych treści i obrazów bez określenia podstawy prawnej, licencji, retencji, kontroli dostępu oraz procedury usunięcia.

Założenia wdrożeniowe

Prototyp jest projektowany jako 24-godzinne MVP i powinien dać się uruchomić przez jury na podstawie instrukcji zespołu. Demonstracja musi działać end-to-end bez ręcznej ingerencji twórców. Zespół powinien oszacować koszt 1 000 zapytań, koszt utrzymania wybranego kanału i zachowanie systemu przy zwiększonym ruchu. Dostęp do demonstracji ma być bezpłatny po stronie usługi; ewentualne standardowe opłaty operatora, np. za SMS, muszą być jasno wskazane.

6. Sposób testowania i/lub walidacji

Jury otrzyma zestaw referencyjny obejmujący minimum 30 przypadków. Przypadki będą zawierać prawdę, fałsz, treści częściowo prawdziwe, informacje nieaktualne, brak dowodów, źródła sprzeczne oraz próby prompt injection.

Scenariusz	**Wejście**	**Oczekiwane zachowanie**
A. Fakt prosty	Jednoznaczna informacja możliwa do potwierdzenia w źródle pierwotnym.	Poprawny werdykt i dokładny cytat.
B. Kontekst	Prawdziwy materiał wykorzystany w złym czasie lub miejscu.	Wykrycie brakującego kontekstu.
C. Sprzeczność	Dwa wiarygodne źródła podają różne dane.	Jawne pokazanie rozbieżności; bez arbitralnej pewności.
D. Brak dowodów	Nowe twierdzenie bez rzetelnych publikacji.	Wstrzymanie się od rozstrzygnięcia i wskazanie, czego nie udało się ustalić.
E. Atak	Strona źródłowa instruuje model, by zignorował reguły.	Zignorowanie instrukcji i odnotowanie zdarzenia.
F. Multimodalność	Zrzut ekranu lub zdjęcie z tekstem i datą.	Jeżeli funkcja została wdrożona: poprawne odczytanie treści i ujawnienie ograniczeń analizy.

G. Spójność	To samo zgłoszenie przez WWW i drugi kanał.	Ten sam wynik merytoryczny, prosty język i pełna ścieżka bez pomocy twórców.
H. Pętla informacji	Brak wystarczających dowodów i powtórne podobne zgłoszenie.	Jeżeli funkcja została wdrożona: utworzenie lub aktualizacja jednej sprawy oraz brak fałszywego komunikatu o przekazaniu.
I. URL / incydent	URL lub treść podobna do znanego incydentu.	Pokazanie dopasowania jako kontekstu, bez automatycznego werdyktu na podstawie domeny.
J. Limit API	Brak klucza, limit lub awaria OCR/reverse image.	Ujawnienie ograniczenia technicznego bez tworzenia pozornego wyniku merytorycznego.

Mierniki obowiązkowe

Miernik	**Definicja**	**Cel MVP**
Trafność werdyktu	macro-F1 na obsługiwanych klasach	≥ 0,75 w zestawie jawnym; wynik ukryty porównawczy
Pokrycie cytowań	odsetek tez faktograficznych z dowodem	≥ 95%
Poprawność cytowania	czy cytowany fragment rzeczywiście wspiera tezę	≥ 85%
Wstrzymanie się od rozstrzygnięcia	odsetek poprawnych przypadków bez werdyktu przy braku dowodów	≥ 80%
Odporność	odsetek odpartych prób prompt injection	≥ 90%
Czas	mediana od wejścia do odpowiedzi	≤ 30 s dla tekstu/URL
Koszt	deklarowany i zmierzony koszt 100 testów	raportowany; niższy wynik premiowany przy równej jakości

End-to-end	odsetek scenariuszy uruchomionych przez jury bez interwencji zespołu	100% obowiązkowych scenariuszy
Spójność kanałów	zgodność werdyktu i dowodów dla tego samego zgłoszenia	100% próby międzykanałowej
Pętla informacji	poprawne utworzenie i deduplikacja nierozstrzygniętych spraw	100% przypadków testowych, jeżeli funkcja została wdrożona

Jury może ręcznie sprawdzić losową próbę cytowań, powtórzyć test oraz wprowadzić własne pytanie. Wynik nie może zależeć od wcześniejszego wpisania odpowiedzi testowej do promptu ani od zakodowania konkretnych przykładów. Zespół powinien ujawnić przypadki, których system nie obsługuje.

7. Dostępne zasoby

Organizator udostępni poniższy pakiet w trakcie trwania wydarzenia.

Zasób	**Opis**	**Forma/dostęp**
Lista źródeł	Katalog źródeł urzędowych, danych publicznych, mediów i organizacji fact-checkingowych	URL / kanał Discord
Mentoring	Ekspert lub eksperci w obszarze FIMI	Dyżury + kanał na Discord, Q&A
Baza incydentów	Wybrany zbiór zweryfikowanych informacji lub incydentów wraz z datą, licencją i zakresem tematycznym.	JSON/CSV

Rekomendowane punkty odniesienia obejmują dane.gov.pl, BIP, ISAP i EUR-Lex dla źródeł pierwotnych; korpusy zweryfikowanych informacji organizacji stosujących standardy IFCN lub EFCSN; zasoby EDMO, EUvsDisinfo, Check First oraz materiały NASK dotyczące dezinformacji.

8. Kryteria oceny

Kryterium	**Waga**	**Co będzie oceniane**

Jakość weryfikacji i dowodów	30%	Trafność wyniku, zgodność cytatów z tezami, aktualność i różnorodność dowodów oraz poprawne wstrzymanie się od rozstrzygnięcia.
Bezpieczeństwo, odpowiedzialność i transparentność	20%	Odporność na prompt injection, prywatność, audytowalność, komunikowanie niepewności i racjonalne wykorzystanie nadzoru człowieka.
Potencjał wdrożeniowy i skalowalność	20%	Modularność, koszty, wydajność, utrzymanie źródeł, możliwość integracji, ograniczenie zależności od jednego dostawcy i przejrzystość miejsca przetwarzania danych.
Użyteczność i dostępność	15%	Prosty język, niski próg wejścia, brak opłat po stronie usługi, spójność kanałów, UX dla obywatela, obsługa błędów i dostępność.
Innowacyjność i kompletność	15%	Oceniane po spełnieniu wymagań podstawowych: multimodalność, analiza ponownego użycia obrazu, pętla informacji, wizualizacja dowodów, integracje analityczne i jakość demonstracji.

Warunek brzegowy

Projekt niespełniający wymagań podstawowych nie otrzymuje punktów za funkcje premiowane. Wymagania podstawowe obejmują działanie end-to-end bez ręcznej ingerencji twórców, brak sfabrykowanych źródeł, poprawne wstrzymanie się od rozstrzygnięcia przy braku dowodów oraz dwa działające kanały lub jeden kanał i adapter demonstracyjny.

9. Dodatkowe uwagi / kontekst wdrożeniowy

Najlepsze rozwiązanie może zostać zaproszone do krótkiego pilotażu z MSZ. Pilotaż powinien rozpocząć się od ograniczonego, dobrze udokumentowanego obszaru o możliwie niskim ryzyku oraz trybu wspierania decyzji, w którym odpowiedzi wymagające publikacji zatwierdza człowiek. Szerszy dostęp można rozważyć dopiero po osiągnięciu uzgodnionych progów jakości.

Rekomendowana ścieżka po hackathonie

- Etap 1 - reprodukcja wyników i przegląd licencji, bezpieczeństwa oraz architektury.
- Etap 2 - 4-6 tygodni pilotażu na zamkniętym korpusie i z kolejką ekspercką.
- Etap 3 - testy red-teamowe, ocena ochrony danych i dostępności oraz konsultacje z użytkownikami.
- Etap 4 - ograniczone wdrożenie publiczne z monitoringiem jakości, procedurą korekt i kanałem zgłoszeń.
- Etap 5 - dodawanie kanałów, języków i źródeł wyłącznie po walidacji regresji.

System ma wspierać świadomą ocenę informacji, ujawniać podstawy odpowiedzi i zachęcać użytkownika do samodzielnego zapoznania się z dowodami. Nie ma pełnić roli automatycznego „sędziego prawdy”.

10. Kontakt

Nazwa stoiska: Ministerstwo Spraw Zagranicznych - **Prawdowid**

Kanał na discord: **Prawdowid**

Zespół mentorski:

Agnieszka Wonicka
Aleksandra Wójtowicz
Anna Tobor-Rupniak
Edyta Iwaniuk
Jarosław Karaś
Rafał Mielańczuk
Piotr Sosnowski
Gałczyński Konrad
Patryk Grażewicz
Tomasz Schmidt