Martyna Florkiewicz
Uniwersytet Ekonomiczny w Katowicach,
Politechnika Śląska

DEZINFORMACJA W PRZESTRZENI CYFROWEJ I WYKORZYSTANIE SZTUCZNEJ INTELIGENCJI W PROCESIE JEJ WERYFIKACJI

Streszczenie (abstrakt): W dzisiejszych czasach, przez rozwój Internetu i mediów społecznościowych, dezinformacja stała się poważnym problemem, który może wpływać na decyzje ludzi. Celem artykułu jest analiza tego zjawiska oraz przedstawienie projektu aplikacji, która pomaga użytkownikom weryfikować treści publikowane w sieci. W części teoretycznej opisano różnice między misinformacją, dezinformacją a malinformacją oraz techniki manipulacji, takie jak *clickbait* czy *cherry picking*. W części praktycznej przedstawiono prototyp narzędzia, które wykorzystuje język Python, bibliotekę spaCy oraz modele Sentence Transformers. Aplikacja automatycznie analizuje tekst, wyszukuje informacje w Internecie i porównuje je, co pozwala szybciej ocenić, czy dany artykuł jest wiarygodny.

Słowa kluczowe: dezinformacja, fake news, weryfikacja informacji, sztuczna inteligencja, NLP, Python

DISINFORMATION IN THE DIGITAL SPACE AND THE USE OF ARTIFICIAL INTELLIGENCE IN ITS VERIFICATION PROCESS

Abstract: Nowadays, due to the development of the Internet and social media, disinformation has become a serious problem that can influence people's decisions. The aim of the article is to analyze this phenomenon and present a project of an application that helps users verify content published online. The theoretical part describes the differences between misinformation, disinformation and malinformation, as well as manipulation techniques such as clickbait or cherry picking. The practical part presents a prototype tool that uses the Python language, the spaCy library and Sentence Transformers models. The application automatically analyzes the text, searches for information on the Internet and compares it, which allows for a faster assessment of whether a given article is credible.

Keywords: disinformation, fake news, information verification, artificial intelligence, NLP, Python

Wstęp

Rozwój Internetu i mediów społecznościowych w znaczący sposób wpłynął na pozyskiwanie i rozpowszechnianie przez ludzi informacji. Za pośrednictwem Internetu użytkownicy mają szybki dostęp do informacji z całego świata, co z jednej strony zwiększa dostępność wiedzy, a z drugiej ułatwia szybkie rozpowszechnianie się niezweryfikowanych lub nie-

prawdziwych informacji. Łatwy dostęp zarówno do tworzenia, jak i udostępniania informacji sprzyja szybkiemu szerzeniu się dezinformacji.

Nieprawdziwe informacje mogą wpłynąć na decyzje ludzi, dlatego są szczególnie niebezpieczne, kiedy dotyczą sytuacji kryzysowych, tematów politycznych czy zdrowotnych. Weryfikowanie informacji stało się kluczową potrzebą nie tylko dla pojedynczych ludzi, ale także dla całego społeczeństwa.

Problemem badawczym podjętym w niniejszym artykule jest określenie, czy i w jakim stopniu możliwe jest zautomatyzowane wspieranie procesu wykrywania nieprawdziwych informacji w Internecie. Pomimo istnienia serwisów *fact-checkingowych* oraz inicjatyw edukacyjnych, wiele osób nie dysponuje czasem ani odpowiednimi narzędziami do samodzielnej weryfikacji treści w Internecie.

Celem głównym artykułu jest analiza zjawiska dezinformacji oraz prezentacja koncepcji aplikacji wspierającej użytkowników w procesie weryfikacji informacji publikowanych w Internecie.

1. Charakterystyka zjawiska dezinformacji

W dzisiejszych czasach, kiedy dostęp do informacji jest większy i łatwiejszy niż kiedykolwiek wcześniej, coraz trudniej jest rozróżnić, co jest prawdą, a co fałszem. Zarówno dostęp do przeglądania różnych treści, jak i do ich tworzenia i publikowania jest tak duży, że weryfikacja wiarygodności treści staje się poważnym wyzwaniem. Dezinformacja jest zjawiskiem znanym od dawna, ale to głównie za sprawą rozwoju Internetu i mediów społecznościowych jej skala oraz tempo rozprzestrzeniania stanowią realne zagrożenie dla społeczeństwa.

1.1. Definicja i podział nieprawdziwych informacji:

Według Komisji Europejskiej dezinformacja to „możliwe do zweryfikowania nieprawdziwe lub wprowadzające w błąd informacje, tworzone, przedstawiane i rozpowszechniane w celu uzyskania korzyści gospodarczych lub wprowadzenia w błąd opinii publicznej, które mogą wyrządzić szkodę publiczną”[1]. W literaturze przedmiotu można spotkać się z różnymi definicjami, chociaż większość z nich zawiera podobne elementy: intencje nadawcy, fałszywość treści oraz ich negatywny wpływ. Istotne jest jednak precyzyjne rozróżnienie pojęć, które często są mylone. Jak wskazują Karina Stasiuk-Krajewska i Michał Wenzel, należy wyróżnić trzy kategorie:

Misinformacja – nieumyślne udostępnianie fałszywych informacji (np. wynikające z błędu).

Dezinformacja – celowe tworzenie i udostępnianie fałszywych lub zmanipulowanych informacji, które mają na celu oszukanie odbiorców, wyrządzenie szkody lub osiągnięcie korzyści (politycznych, finansowych).

[1] Europejski Trybunał Obrachunkowy, Dezinformacja: wspólne wyzwanie UE (Sprawozdanie specjalne nr 09/2021), Luksemburg: Urząd Publikacji Unii Europejskiej, Luksemburg 2021, s. 8-11.

Malinformacja – sytuacja, gdy prywatne, prawdziwe informacje są przenoszone do sfery publicznej i udostępniane w celu wyrządzenia szkody (np. mowa nienawiści)[2].

1.2 Techniki manipulacji i mechanizmy psychologiczne

Twórcy dezinformacji wykorzystują szereg technik mających na celu wpłynięcie na odbiorcę. Zestaw takich technik został opracowany przez Naukową i Akademicką Sieć Komputerową (NASK)[3]. Do najczęściej stosowanych w środowisku internetowym należą:

- ***Clickbait*** – sensacyjny lub emocjonalny nagłówek, który nie odpowiada treści, a służy przyciągnięciu uwagi.
- **Język emocjonalny** – sformułowania wywołujące silne emocje, co utrudnia racjonalną ocenę.
- **Podszywanie się** – wykorzystanie wizerunku znanej osoby lub instytucji do uwiarygodnienia fałszu.
- ***Cherry picking*** – wybór tylko tych informacji, które pasują do tezy, z pominięciem kontekstu.
- **Fałszywa przyczyna** – sugerowanie nieuzasadnionego związku przyczynowo-skutkowego.
- **Dowód anegdotyczny** – powoływanie się na pojedyncze doświadczenia w celu podważenia danych ogólnych.
- ***Deepfake*** – materiał audio lub wideo wygenerowany przez sztuczną inteligencję.
- ***Cheapfake*** – manipulowanie materiałem audiowizualnym.

Dezinformacja nie jest zjawiskiem chaotycznym – twórcy fałszywych treści bardzo często wykorzystują mechanizmy psychologiczne. Jak zauważa Krzysztof Kaczmarek, naturalne środowisko działań dezinformacyjnych to chaos informacyjny i emocjonalny charakter przekazów medialnych[4]. Media społecznościowe sprzyjają temu zjawisku poprzez algorytmy promujące kontrowersyjne treści oraz tworzenie tzw. baniek filtrujących, w których użytkownik otrzymuje jedynie informacje zgodne z jego przekonaniami.

2. Metody weryfikacji informacji i rola technologii

Weryfikacja informacji (ang. *fact-checking*) to proces, którego celem jest sprawdzenie prawdziwości twierdzeń zawartych w przekazach medialnych. W obliczu rosnącej fali dezinformacji, metody weryfikacji ewoluują, przechodząc od analizy w pełni manualnej do rozwiązań wspieranych przez zaawansowane algorytmy.

---

[2] K. Stasiuk-Krajewska, T. Wenzel, *Dezinformacja w czasach kryzysu,* Wydawnictwo Adam Marszałek, Toruń 2024, s. 5.

[3] NASK, *Dezinformacja – informacje i przeciwdziałanie*, https://www.nask.pl/dezinfo [dostęp: 30.04.2025].

[4] K. Kaczmarek K., *Konsekwencje dezinformacji: przegląd wybranych narzędzi i technik manipulacji*, „Bezpieczeństwo Narodowe" 2024, nr 2, s. 15.

2.1. Od metod ręcznych do automatyzacji

Tradycyjny *fact-checking* opiera się na pracy ekspertów i dziennikarzy. Wykorzystują oni metody ręczne, polegające na:

- Analizie źródła (sprawdzenie wiarygodności domeny, autora, daty publikacji).
- Weryfikacji u źródła pierwotnego (dotarcie do oryginalnych dokumentów, nagrań lub danych statystycznych).
- Analizie logiczno-semantycznej (poszukiwanie wewnętrznych sprzeczności w tekście).

Przykładem takich działań są organizacje takie jak polski Demagog czy międzynarodowe agencje prasowe. Choć metody te charakteryzują się najwyższą precyzją i uwzględniają kontekst kulturowy, to napotykają na barierę skalowalności. Weryfikacja pojedynczego artykułu może zająć człowiekowi od kilkunastu minut do kilku godzin. W obliczu tysięcy treści generowanych każdego dnia w mediach społecznościowych, weryfikacja manualna staje się niewystarczająca. Odpowiedzią na to wyzwanie jest automatyzacja procesu. Jak zauważają Marek Troszyński i Aleksander Wawer, rozwiązaniem jest połączenie jakościowego kodowania z narzędziami udostępnionymi przez lingwistykę komputerową[5].

2.2. Zastosowanie Przetwarzania Języka Naturalnego (NLP)

Podstawą automatycznej analizy treści jest dziedzina sztucznej inteligencji zwana Przetwarzaniem Języka Naturalnego (NLP – Natural Language Processing). Umożliwia ona systemom informatycznym nie tylko odczytywanie tekstu, ale także jego „zrozumienie" na poziomie strukturalnym. W kontekście walki z dezinformacją kluczowe są następujące procesy NLP:

- Tokenizacja i lematyzacja: Procesy te polegają na podziale tekstu na mniejsze jednostki i sprowadzaniu słów do ich formy podstawowej. W języku polskim, który charakteryzuje się skomplikowaną fleksją, jest to niezbędne, aby algorytm zrozumiał, że różne formy tego samego słowa dotyczą jednego pojęcia.
- Rozpoznawanie Nazwanych Encji (NER – Named Entity Recognition): Algorytmy potrafią automatycznie wyłowić z tekstu kluczowe informacje, takie jak nazwiska osób publicznych, nazwy organizacji, lokalizacje geograficzne czy daty. Pozwala to na zamianę nieustrukturyzowanego tekstu w zbiór faktów, które można łatwiej zweryfikować w zewnętrznych bazach danych.

2.3. Analiza semantyczna i modele sztucznej inteligencji

Tradycyjne wyszukiwarki internetowe działają głównie w oparciu o słowa kluczowe. Jest to podejście niewystarczające w walce z dezinformacją, ponieważ twórcy fake newsów często używają synonimów lub manipulują szykiem zdań. Rozwiązaniem tego problemu są nowoczesne modele uczenia maszynowego, takie jak Sentence Transformers.

---

[5] M. Troszyński, A. Wawer, *Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych*, „Przegląd Socjologii Jakościowej" 2017, t. 13, nr 2, s. 62-80.

Technologia ta pozwala na analizę semantyczną, czyli badanie znaczenia tekstu. Modele te przekształcają całe zdania w wektory liczbowe w wielowymiarowej przestrzeni. Działa to na zasadzie porównywania „odległości" między znaczeniami. Jeśli teza zawarta w badanym artykule jest semantycznie bliska informacjom pochodzącym z wiarygodnych źródeł, wektory te będą znajdować się blisko siebie, nawet jeśli użyto zupełnie innych słów.

Połączenie analizy semantycznej z mechanizmami wyszukiwania informacji (np. poprzez API wyszukiwarek) pozwala na stworzenie narzędzi, które pełnią rolę „cyfrowego asystenta". System taki może w czasie rzeczywistym przeszukiwać Internet, porównywać fakty i wskazywać użytkownikowi potencjalne manipulacje, znacznie przyspieszając proces weryfikacji.

3. Projekt i implementacja systemu wspomagającego weryfikację informacji – studium przypadków

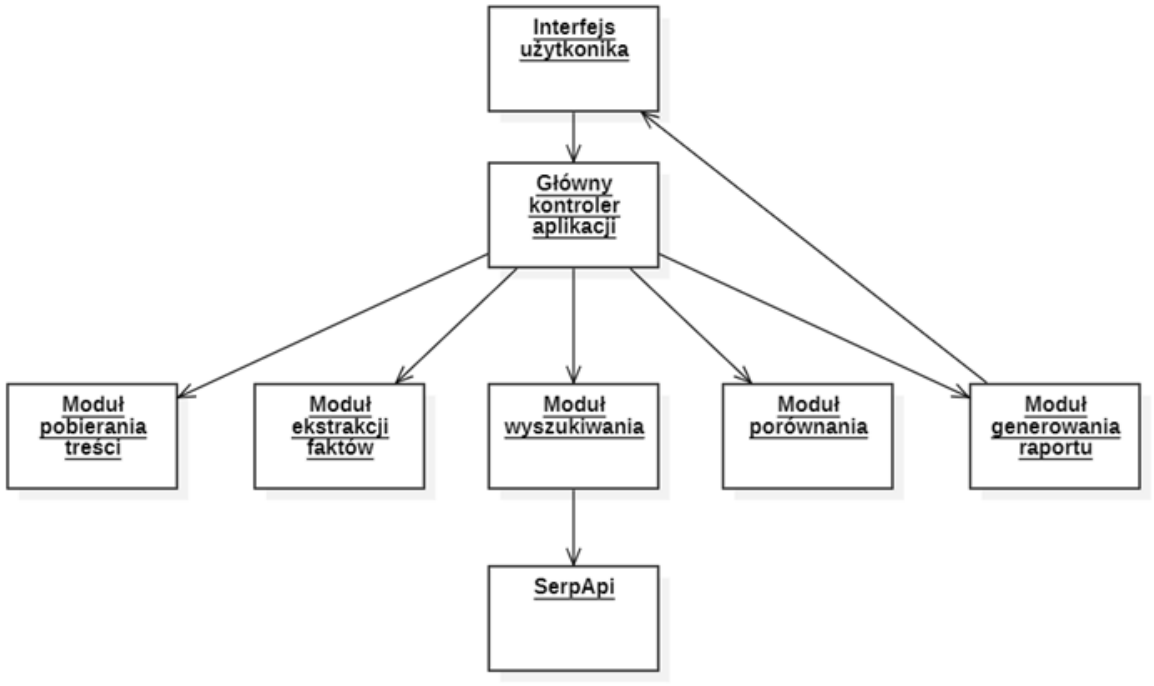
W celu weryfikacji teoretycznych założeń dotyczących możliwości automatyzacji *fact-checkingu*, opracowano projekt i zaimplementowano prototyp systemu informatycznego. Głównym założeniem było stworzenie narzędzia, które w czasie rzeczywistym analizuje treść artykułu i konfrontuje ją z dostępnymi w sieci źródłami, pełniąc rolę asystenta użytkownika.

3.1. Architektura i środowisko technologiczne

System został zaprojektowany w architekturze klient-serwer. Warstwa logiczna (*backend*) odpowiada za przetwarzanie danych, natomiast warstwa prezentacji (*frontend*) umożliwia interakcję z użytkownikiem poprzez przeglądarkę internetową. Do realizacji projektu wybrano język Python, ze względu na dostępność zaawansowanych bibliotek wspierających sztuczną inteligencję. Jako szkielet aplikacji wykorzystano *framework* Flask. Jest to rozwiązanie lekkie i elastyczne, które nie narzuca sztywnej struktury, co pozwala na łatwe prototypowanie rozwiązań badawczych. Flask odpowiada za *routing* (kierowanie ruchem w aplikacji) oraz komunikację między interfejsem a silnikiem analitycznym.

W warstwie przetwarzania danych zaimplementowano następujące technologie:

- spaCy: Biblioteka ta służy do wstępnej obróbki tekstu. Odpowiada za tokenizację oraz rozpoznawanie jednostek nazwanych w języku polskim.
- Sentence Transformers: Wykorzystany do tworzenia reprezentacji wektorowej zdań i obliczania podobieństwa semantycznego.
- SerpApi: Interfejs programistyczny umożliwiający pobieranie wyników wyszukiwania z Google w czasie rzeczywistym, co zapewnia dostęp do aktualnych informacji.
- Beautiful Soup 4: Biblioteka służąca do tzw. *web scrapingu*, czyli pobierania i oczyszczania treści ze stron internetowych (usuwanie reklam, skryptów i elementów nawigacyjnych).



Rysunek 1. Schemat architektury logicznej systemu

Źródło: Opracowanie własne

3.2. Algorytm przetwarzania danych

Działanie prototypu opiera się na sekwencyjnym przetwarzaniu informacji. Proces ten został zautomatyzowany i przebiega w kilku etapach, niewidocznych dla użytkownika końcowego.

Etap 1: Ekstrakcja i czyszczenie danych – Proces rozpoczyna się od wprowadzenia przez użytkownika adresu URL artykułu. System pobiera kod źródłowy strony, a następnie separuje właściwą treść artykułu od elementów technicznych strony.

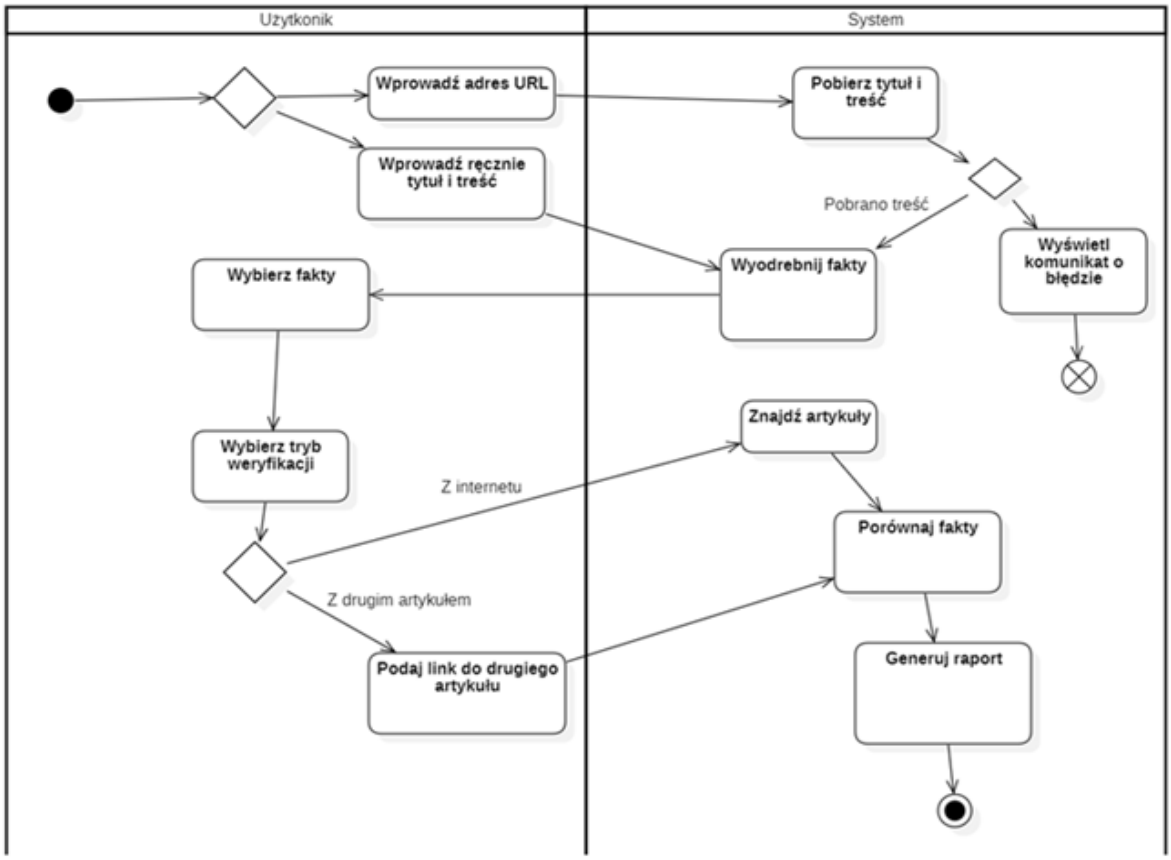
Etap 2: Analiza lingwistyczna (NLP) – Oczyszczony tekst trafia do modułu NLP. Przy użyciu biblioteki spaCy tekst jest dzielony na zdania. Algorytm identyfikuje w nich kluczowe podmioty (osoby, organizacje, kraje) oraz orzeczenia. Na tym etapie system dokonuje również wstępnej selekcji, odróżniając zdania zawierające fakty (weryfikowalne) od zdań będących opiniami autora.

Etap 3: Wyszukiwanie kontekstowe – Dla każdego wyodrębnionego twierdzenia system generuje zapytanie do wyszukiwarki. Dzięki integracji z API aplikacja pobiera listę artykułów powiązanych tematycznie z wiarygodnych domen.

Etap 4: Weryfikacja semantyczna – Pobrane fragmenty tekstów są porównywane z twierdzeniem z badanego artykułu. W tym celu model Sentence Transformers zamienia oba teksty na wektory liczbowe. Następnie obliczane jest podobieństwo.

- Wartość bliska 1 oznacza, że znalezione źródła potwierdzają tezę.

- Wartość niska lub ujemna sugeruje sprzeczność lub brak potwierdzenia w dostępnych źródłach.



Rysunek 2. Diagram sekwencji procesu weryfikacji

Źródło: Opracowanie własne

3.3. Prezentacja wyników

Ostatnim etapem jest agregacja wyników i ich prezentacja. System nie podejmuje ostatecznej decyzji za człowieka, lecz dostarcza mu dowodów. W interfejsie użytkownika wyświetlany jest raport zawierający listę przeanalizowanych tez wraz z linkami do źródeł, które je potwierdzają lub podważają. Pozwala to użytkownikowi na szybką ocenę wiarygodności materiału bez konieczności ręcznego przeszukiwania wielu stron.

Podsumowanie

Zrealizowana analiza teoretyczna oraz prace badawcze nad prototypem systemu weryfikacyjnego pozwoliły na realizację głównego celu artykułu, jakim była ocena możliwości wykorzystania technologii informatycznych w walce z dezinformacją.

Przeprowadzona charakterystyka zjawiska wykazała, że współczesna dezinformacja nie jest przypadkowym błędem, lecz zorganizowanym działaniem wykorzystującym mechanizmy psychologiczne (np. błędy poznawcze) oraz technologiczne (algorytmy mediów społecznościowych). Zrozumienie różnic między misinformacją a dezinformacją jest klu-

czowe dla doboru odpowiednich metod przeciwdziałania. Jak wykazano, metody ręczne, choć precyzyjne są niewystarczające w obliczu skali problemu, co rodzi konieczność wdrażania rozwiązań zautomatyzowanych.

Studium przypadku, polegające na opracowaniu modelu systemu opartego na języku Python i metodach Przetwarzania Języka Naturalnego (NLP), doprowadziło do sformułowania następujących wniosków:

1. **Możliwość automatyzacji:** Wykorzystanie bibliotek takich jak spaCy oraz modeli Sentence Transformers pozwala na skuteczne wyodrębnianie faktów z tekstu i ich automatyczne porównywanie z zewnętrznymi bazami wiedzy.
2. **Skuteczność weryfikacji:** Zastosowane algorytmy wykazują wysoką skuteczność w weryfikacji danych twardych (liczby, daty, nazwiska). System poprawnie identyfikuje sprzeczności między badanym artykułem a źródłami referencyjnymi.
3. **Ograniczenia technologii:** Modele sztucznej inteligencji wciąż napotykają trudności w interpretacji kontekstu kulturowego, ironii oraz tekstów publicystycznych, gdzie granica między faktem a opinią jest nieostra.
4. **Rola człowieka:** Technologia nie może w pełni zastąpić krytycznego myślenia użytkownika. Zaprojektowane narzędzie pełni rolę „inteligentnego asystenta", który przyspiesza proces docierania do źródeł, ale ostateczna ocena wiarygodności musi pozostać po stronie człowieka.

Podsumowując, integracja narzędzi informatycznych z wiedzą o mechanizmach dezinformacji stanowi obiecujący kierunek rozwoju systemów bezpieczeństwa informacyjnego. Dalsze prace powinny koncentrować się na udoskonalaniu modeli semantycznych oraz rozszerzaniu analizy o materiały wizualne (wykrywanie *deepfake*), które stanowią coraz większe zagrożenie w przestrzeni cyfrowej.

Bibliografia

1. Europejski Trybunał Obrachunkowy. Dezinformacja: wspólne wyzwanie UE (Sprawozdanie specjalne nr 09/2021), Luksemburg: Urząd Publikacji Unii Europejskiej, Luksemburg 2021, s. 8-11.
2. Kaczmarek K., *Konsekwencje dezinformacji: przegląd wybranych narzędzi i technik manipulacji,* „Bezpieczeństwo Narodowe" 2024, nr 2.
3. NASK, Dezinformacja – informacje i przeciwdziałanie, https://www.nask.pl/dezinfo [dostęp: 30.04.2025].
4. Stasiuk-Krajewska K., Wenzel T., *Dezinformacja w czasach kryzysu,* Wydawnictwo Adam Marszałek, Toruń 2024.
5. Troszyński M., Wawer, A., *Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych,* „Przegląd Socjologii Jakościowej" 2017, t. 13, nr 2.

Dane kontaktowe

Martyna Florkiewicz, mf319892@student.polsl.pl