Streszczenie

Szybkie rozpowszechnianie dezinformacji w przestrzeni internetowej stanowi poważne zagrożenie dla procesów demokratycznych, zdrowia publicznego oraz stabilności społecznej. Chociaż dezinformacja jest definiowana jako perswazyjna komunikacja celowo wprowadzająca w błąd, większość podejść redukuje jej automatyczne wykrywanie do problemu binarnej klasyfikacji, pomijając kluczowe mechanizmy jej skuteczności, takie jak manipulacyjne techniki perswazyjne oraz złośliwe intencje nadawcy. Ograniczenie to wpływa negatywnie na wyjaśnialność systemów, zwłaszcza w scenariuszach obejmujących różne kategorie tematyczne oraz gatunki tekstów.

Niniejsza rozprawa doktorska odpowiada na tę lukę badawczą poprzez zaproponowanie obliczeniowego podejścia do wykrywania dezinformacji z wykorzystaniem modeli językowych, wzbogaconego o rozumowanie oparte na perswazji i intencji. Praca rozwija stan wiedzy poprzez cztery powiązane ze sobą badania. Po pierwsze, wprowadza nowe zbiory danych, które wykraczają poza binarne etykiety dezinformacji, uwzględniając techniki manipulacji oraz złośliwe intencje, w tym pierwszy korpus w języku polskim oraz pierwszy anglojęzyczny zbiór anotowany pod kątem złośliwych intencji wraz z adnotacjami pochodzącymi z kolejnych etapów procesu anotacji. Zasoby te umożliwiają szczegółowe badanie dezinformacji jako zjawiska komunikacyjnego ukierunkowanego na realizację określonych celów oraz ustanawiają nowe punkty odniesienia dla wielowymiarowej analizy dezinformacji w języku polskim i angielskim.

Po drugie, czerpiąc z dorobku psychologii oraz badań nad komunikacją, rozprawa proponuje wzbogacone rozumowanie dla dużych modeli językowych, wykazując, że analiza perswazyji jako pośredniego kroku rozumowania istotnie poprawia skuteczność wykrywania dezinformacji oraz dobrze generalizuje się pomiędzy różnymi tematami, gatunkami tekstów i zmianami temporalnymi. Po trzecie, rozprawa przedstawia rozumowanie wzbogacone o analizę intencji, pokazując, że jawne wnioskowanie o złośliwych intencjach zwiększa zdolność modeli do wykrywania dezinformacji, również w kontekście wielojęzycznym.

Ponadto, rozprawa analizuje teksty perswazyjne generowane przez modele sztucznej inteligencji, wprowadzając wielojęzyczny zbiór danych umożliwiający porównanie trudności detekcji oraz charakterystyk językowych treści tworzonych przez ludzi i modele językowe. Na podstawie szeroko zakrojonych analiz lingwistycznych i empirycznych wykazano, że perswazja generowana przez AI charakteryzuje się systematycznymi różnicami językowymi i stanowi nowe wyzwanie dla istniejących systemów wykrywania dezinformacji.

Podsumowując, praca łączy teoretyczne definicje dezinformacji z ich automatyczną detekcją, dostarczając nowych zbiorów danych, ram rozumowania dużych modeli językowych oraz wniosków empirycznych, które wspierają rozwój przejrzystych, uogólnialnych i odpornych na przyszłe wyzwania systemów wykrywania dezinformacji w erze generatywnej sztucznej inteligencji.