Streszczenie

Rozprawa doktorska

Tytuł: Classifying and Mining Disinformation Narratives With Large Language Models

Autor: mgr inż. Witold Sosnowski

Promotor: prof. dr hab. Adam Wierzbicki

Promotor pomocniczy: dr Kinga Skorupska

Niniejsza rozprawa bada, w jaki sposób narracje dezinformacyjne mogą być definiowane, klasyfikowane oraz automatycznie wydobywane z wykorzystaniem dużych modeli językowych. Podczas gdy większość podejść obliczeniowych analizuje dezinformację na poziomie pojedynczych treści (twierdzenia, posty, artykuły), w tej pracy podstawową jednostką analizy staje się narracja dezinformacyjna. Takie ujęcie pozwala stabilniej opisywać i mierzyć, jak dezinformacja rozprzestrzenia się w czasie i pomiędzy różnymi źródłami. Rozprawa wprowadza trzy komponenty wspierające ewaluację modeli w klasyfikacji narracji dezinformacyjnych oraz ich wydobywanie zarówno globalnie, jak i na poziomie pojedynczych mediów.

Po pierwsze, rozprawa przedstawia EUDisinfoTest, benchmark do oceny zdolności dużych modeli językowych do rozróżniania narracji dezinformacyjnych od wiarygodnych w wielu domenach. Narzędzie dodatkowo testuje wpływ, jaki strategie retoryczne (Logos, Etos, Patos) wywierają na decyzje klasyfikacyjne modeli.

Po drugie, praca proponuje DiNaM, metodę, która rekonstruuje globalny zbiór narracji dezinformacyjnych na podstawie dużych korpusów artykułów fact-checkingowych. Metoda identyfikuje zweryfikowane przypadki fałszywych informacji, grupuje treści o zbliżonym znaczeniu w spójne zbiory, a następnie generuje narracje na podstawie wspólnego rdzenia każdej grupy, tworząc tematyczną mapę narracji wraz z analizą trendów w czasie.

Po trzecie, rozprawa prezentuje DiNO, metodę, która skupia się na analizie pojedynczych źródeł medialnych. W przeciwieństwie do metody DiNaM, która rekonstruuje narracje na podstawie artykułów weryfikujących fakty, DiNO wykorzystuje katalog fałszywych twierdzeń jako punkt odniesienia do identyfikacji podobnych treści w artykułach danego źródła, a następnie grupuje je w spójne klastry i na tej podstawie generuje narracje dezinformacyjne.

Wyniki badań pokazują, że zaproponowane metody wydobywania narracji przy wykorzystaniu dużych modeli językowych skutecznie odtwarzają zbiory narracji dezinformacyjnych zarówno w skali globalnej, jak i na poziomie pojedynczych źródeł medialnych. Ponadto, modele tego typu osiągają wysoką skuteczność w klasyfikacji narracji jako dezinformacyjne lub wiarygodne, jednak ich oceny pozostają podatne na wpływ strategii retorycznych.