Imię i nazwisko: Kamil Kanclerz
Tytuł rozprawy: Metody spersonalizowanego rozpoznawania tekstów obraźliwych
Tytuł rozprawy w języku angielskim: Personalized Offensiveness Detection Methods

Streszczenie

Niniejsza rozprawa dotyczy problematyki spersonalizowanego rozpoznawania tekstów obraźliwych z wykorzystaniem technik przetwarzania języka naturalnego (ang. *natural language processing*, NLP) oraz metod głębokiego uczenia. Główne wyzwania związane z rozpoznawaniem treści obraźliwych wynikają z różnorodności percepcji użytkowników oraz złożoności semantycznej tekstu. W pracy opracowano nowe, spersonalizowane podejścia, które uwzględniają preferencje użytkowników w celu precyzyjniejszej klasyfikacji obraźliwości tekstu, w odróżnieniu od tradycyjnych metod uogólnionych, które nie dostosowują się do indywidualnych oczekiwań odbiorców. Eksperymenty obejmowały szerokie testy porównawcze z wykorzystaniem różnych metod generowania reprezentacji wektorowych tekstów, takich jak CBOW, Skipgram, BERT, DeBERTa, MPNet, oraz XLM-RoBERTa. W badaniach zweryfikowano, że zaawansowane modele kontekstowe, takie jak DeBERTa, osiągnęły istotnie lepsze wyniki w zadaniu rozpoznawania obraźliwości w porównaniu do tradycyjnych metod opartych na modelach bezkontekstowych. Ponadto, przeanalizowano wpływ rozmiaru zbioru danych uczących na stabilność i tempo poprawy jakości predykcji, co podkreśliło dodatkową korzyść płynącą z wykorzystania modeli uwzględniających personalizację. W ramach pracy opracowano dwie, autorskie metody spersonalizowanego aktywnego uczenia wykorzystujące autorską miarę kontrowersyjności tekstu oraz stopień dyspersji anotacji. Obie metody pozwoliły na optymalizację procesu anotacji względem metody referencyjnej oraz innych metod spersonalizowanego aktywnego uczenia takich jak \textit{Ratio Distance} oraz \textit{Stranger Count}. Zastosowanie opracowanych metod skutkowało znacznym zwiększeniem efektywności modelu przy mniejszych zbiorach danych, co czyni te techniki szczególnie przydatnymi w sytuacjach, gdzie dostępne są ograniczone zasoby anotacji. Ponadto, praca badała zastosowanie modelu generatywnego ChatGPT-3.5, w kontekście zarówno uogólnionego, jak i spersonalizowanego rozpoznawania treści obraźliwych. Wyniki wskazują, że modele dedykowane, trenowane na specyficznych zbiorach danych, przewyższają skutecznością model generatywny w tym zadaniu, zwłaszcza w odniesieniu do personalizacji. Rozprawa przedstawia istotne wnioski dotyczące wpływu personalizacji na rozpoznawanie obraźliwych treści tekstowych oraz proponuje nowe techniki, które mogą znaleźć zastosowanie w systemach moderacji treści online.