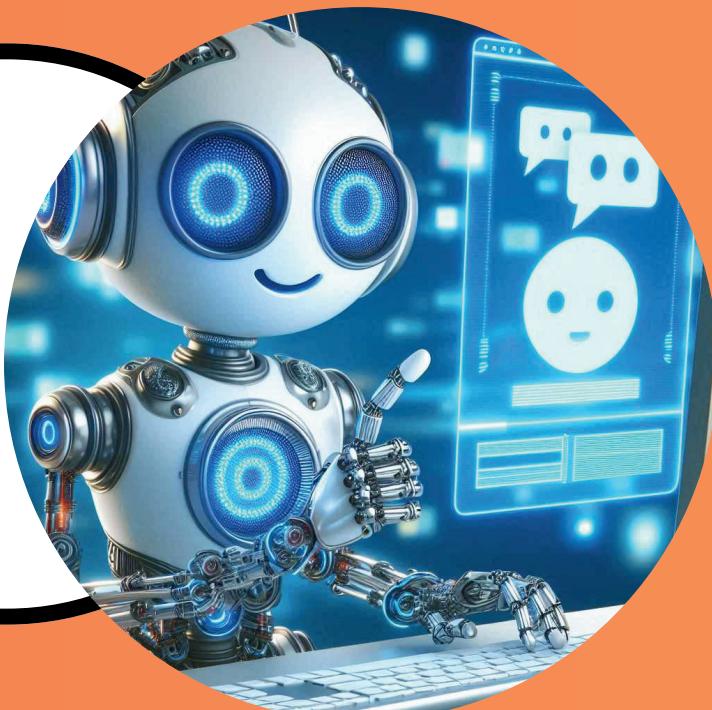




Czy po drugiej stronie jest człowiek?



1 Opis Problemu



Wraz z rozwojem technologii, boty stają się coraz bardziej zaawansowane i trudne do odróżnienia od ludzi. Ich zdolność do imitowania ludzkiego wydźwięku oraz charakterystyki postów i komentarzy stanowi poważne wyzwanie. Sztucznie generowana zawartość może wypiąć ludzką aktywność w social mediach.

Pytanie badawcze: Czy jesteśmy w stanie stwierdzić, czy dany post lub komentarz (publikacja) została napisana przez bota?

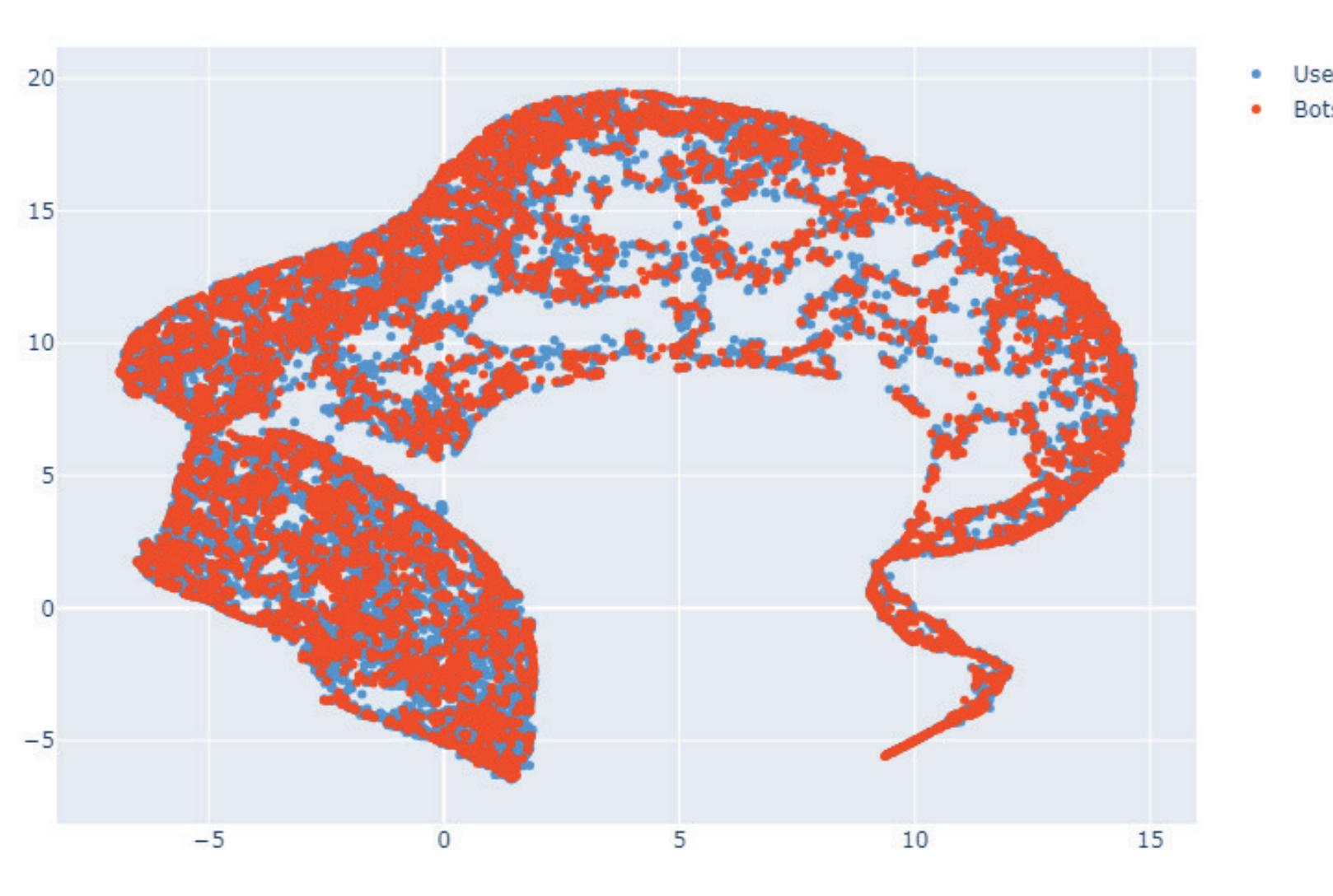
Cel: Analiza podobieństwa publikacji redditowych stworzonych przez boty jak i ludzi.

4 Czy są podobni?

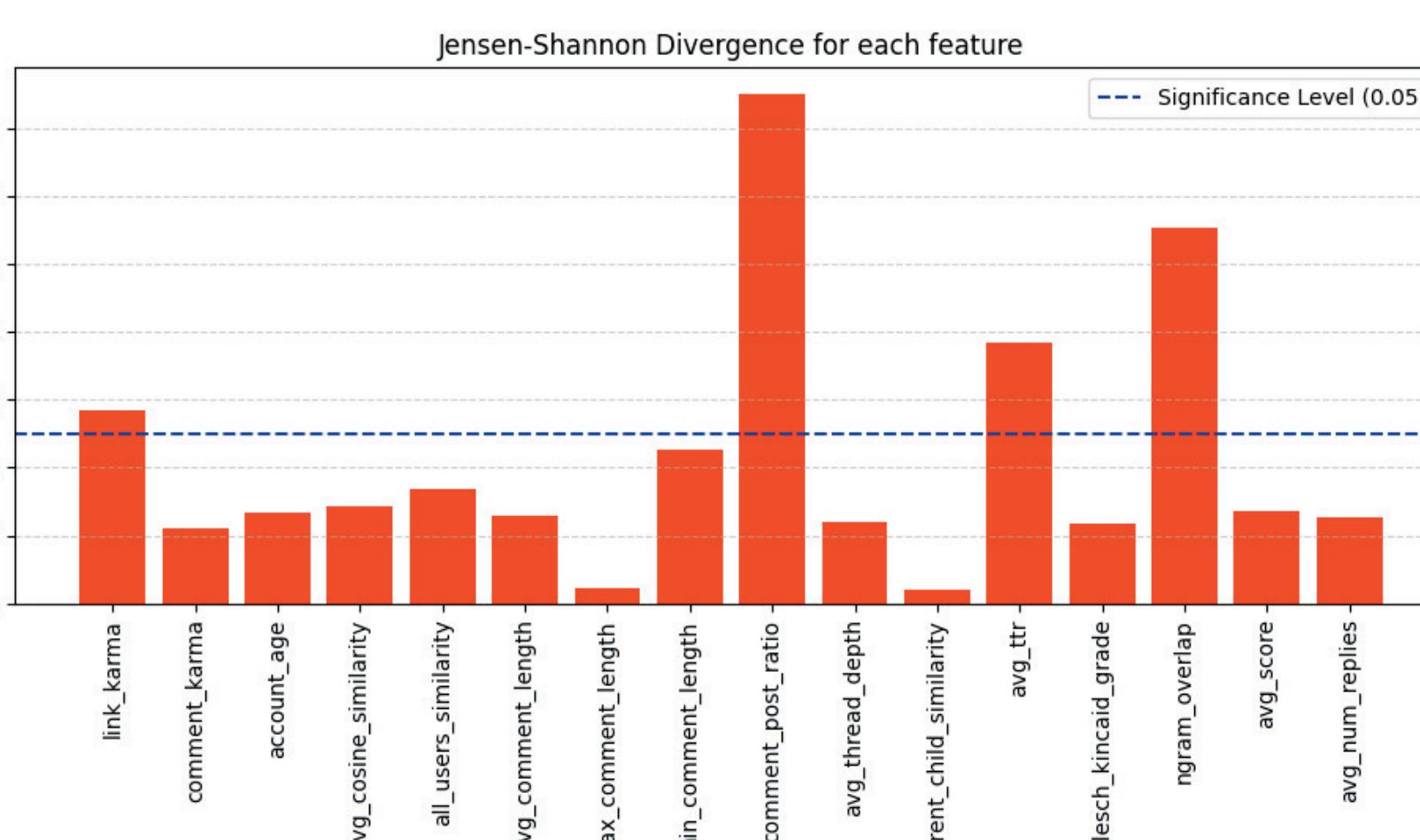


Analiza z użyciem klasyfikatorów *Random Forest*, *CatBoost* i *SVM* wykazała znaczne trudności w przypisywaniu etykiet osiągając najlepsze wyniki na poziomie metryki *F1* równej 0.58. Wyniki te spowodowane mogły być rozmyciem granicy decyzyjnej oraz występowaniem botów niejawnych w klasie użytkowników. Z tego powodu analiza została pogłębiona o wykorzystanie klasyfikatorów jednoklasowych takich jak: *OneClassSVM* i *LocalOutlierFactor*, które również osiągnęły jakość klasyfikacji na poziomie *F1* równej 0.6.

Kolejnym krokiem pogłębionej analizy była wizualizacja danych z pomocą nieliniowej metody redukcji wymiarowości **UMAP**. Obrazując rzutowanie przestrzeni wielowymiarowej na przestrzeń dwuwymiarową, zauważalna staje się niska separowalność klas (boty jawne i użytkownicy wraz z botami niejawnymi) dla rozkładu cech, co sugeruje ich podobieństwo.



Wykres 2: UMAP 2D dla danych Użytkowników i Botów.

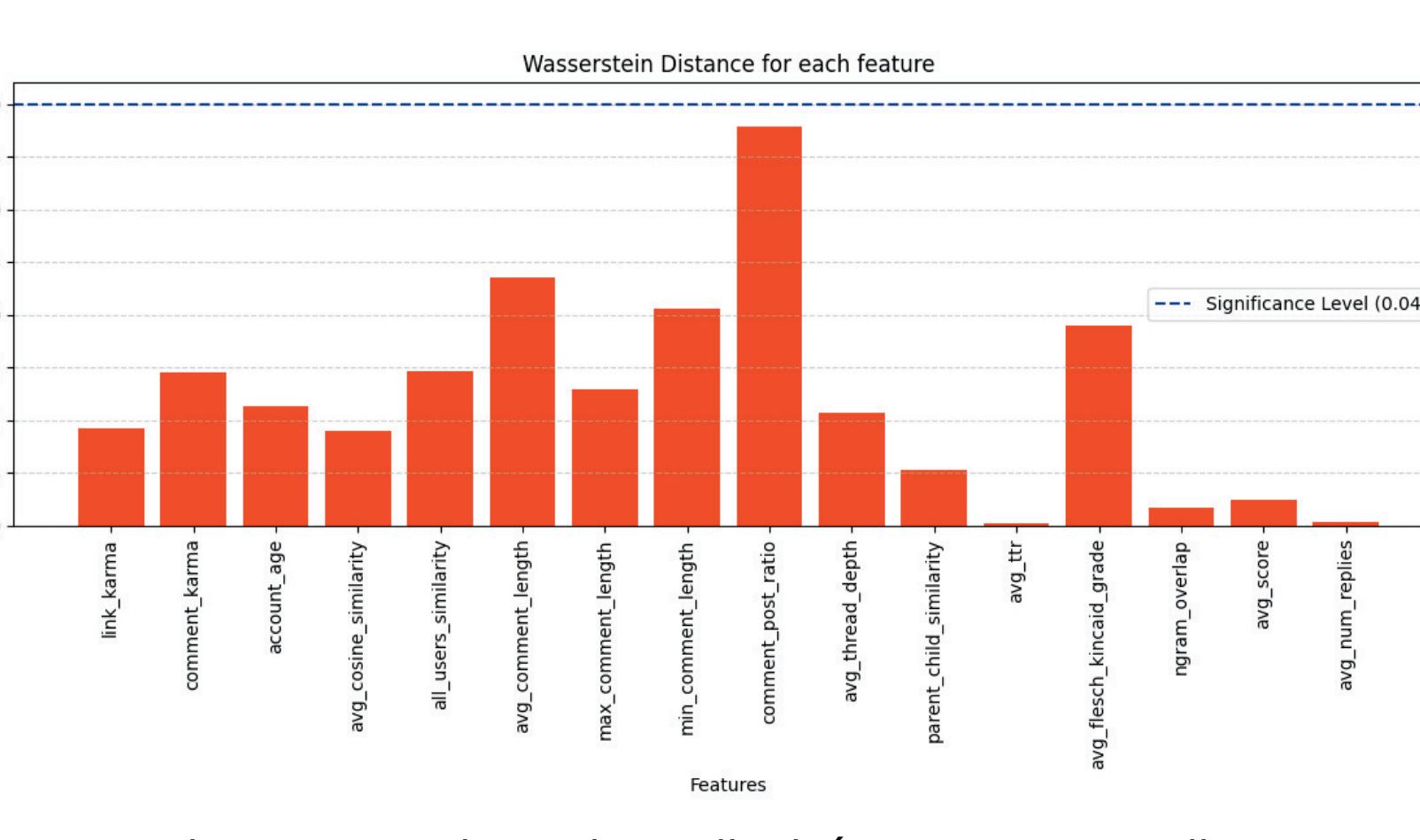


Wykres 3: Wyniki analizy dywergencji Jensen-Shannon dla wszystkich rozpatrywanych cech.

W celu zmierzenia różnicy w rozkładach cech między botami a użytkownikami użyliśmy dywergencji **Jensen-Shannon**, która mierzy podobieństwo między dwoma rozkładami prawdopodobieństwa.

W wynikach wyższe wartości oznaczają większe różnice między rozkładami dla danej cechy. Na wykresie kluczowe różnice zaobserwowano dla *min_comment_length* oraz *comment_post_ratio*, co sugeruje, że długość posta i liczba komentarzy najlepiej różnicują boty od ludzi. Ogólnie wyniki dla większości cech są bardzo małe (poniżej 0.05) co sugeruje, że rozkłady prawdopodobieństw są do siebie bardzo podobne.

Następnie aby udowodnić podobieństwo rozkładów wykorzystana została miara **odległości Wasserstein**, która mierzy minimalny koszt przekształcenia jednego rozkładu w inny. Odległości uzyskane dla wszystkich cech były poniżej progu 0.04, co sugeruje, że ich rozkłady są bardzo podobne w obu grupach.



Wykres 4: Wyniki analizy odległości Wasserstein dla wszystkich rozpatrywanych cech.

5 Niepokojąca przyszłość social mediów

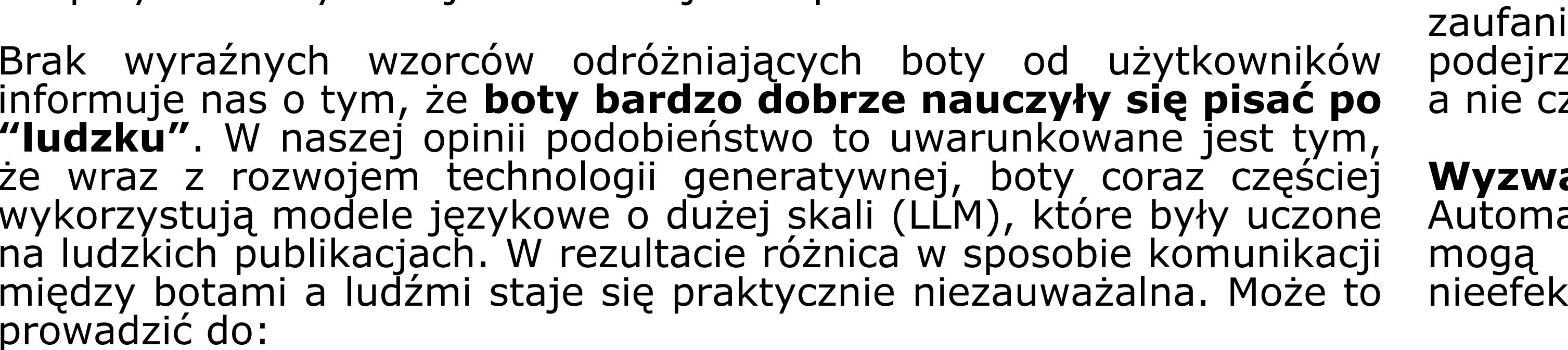
W przeprowadzonych badaniach nad wykrywaniem nieoznaczonych botów wśród zwykłych użytkowników za pomocą analizy różnych parametrów jak i podejść, doszliśmy do wniosku, że **boty są niemal nieroróżniczalne od ludzi**. Taka sytuacja stwarza poważny problem dla przyszłości cyfrowej komunikacji i bezpieczeństwa online.

Brak wyraźnych odróżniających boty od użytkowników informuje nas o tym, że **boty bardzo dobrze nauczyły się pisać po "ludzku"**. W naszej opinii podobieństwo to uwarunkowane jest tym, że wraz z rozwojem technologii generatywnej, boty coraz częściej wykorzystują modele językowe o dużej skali (LLM), które były uczone na ludzkich publikacjach. W rezultacie różnica w sposobie komunikacji między botami a ludźmi staje się praktycznie niezauważalna. Może to prowadzić do:

Manipulacji informacji: Boty mogą być wykorzystywane do masowego szerzenia dezinformacji w sposób trudny do wykrycia.

Obniżenia zaufania społecznego: Brak zaufania ludzi do interakcji online, podejrzewając, że rozmówca może być bot, a nie człowiek.

Wyzwań w moderacji treści: Automatyczne algorytmy i moderacja ręczna mogą okazać się niewystarczające i nieefektywne w identyfikowaniu botów.



Katedra
Sztucznej
Inteligencji



Politechnika Wrocławska

Projekt wykonany w ramach zajęć "Analiza mediów cyfrowych" na kierunku Sztuczna Inteligencja w roku 2024

Autorzy: Dawid Kopeć, Dawid Krutul,
Maciej Wizerkaniuk

2 Skąd pochodzą dane?

Jako źródło danych posłużyły wypowiedzi użytkowników z pięciu najpopularniejszych subredditów na platformie Reddit. Dane te zostały zebrane za pomocą autorskiego skryptu, który umożliwia bezpośredni dostęp do danych przy użyciu biblioteki *PRAW* oraz interfejsu programistycznego platformy (*API*).



W celu przeprowadzenia analizy zebrano obszerny zbiór danych z pięciu najpopularniejszych subredditów na platformie Reddit: *r/funny*, *r/AskReddit*, *r/gaming*, *r/worldnews* oraz *r/todayilearned*. Proces zbierania danych, obejmujący zarówno posty, komentarze, jak i informacje o użytkownikach, trwał nieprzerwanie przez 6 dni, co pozwoliło na stworzenie kompleksowego zbioru danych. Zawierającego oznakowane przez Reddit'a **boty jawne** oraz użytkowników "ludzkich" wraz z niezidentyfikowanymi **botami niejawnymi**.

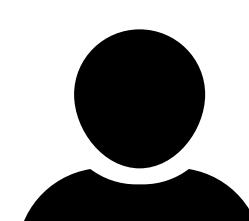
W ten sposób udało nam się zebrać następującą liczbę danych:



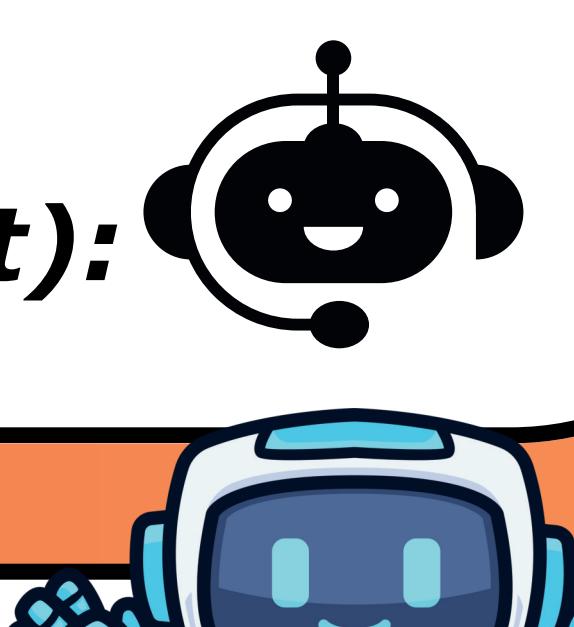
komentarze: 800 000



posty: 4 500

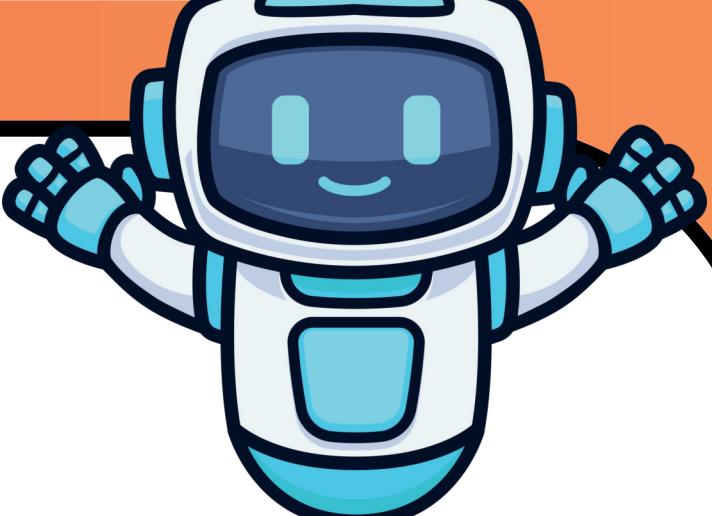


**użytkownicy (w tym boty niejawne):
370 000**



**Boty jawne (oznaczone przez Reddit):
10 000**

3 Jak opisać boty?



Na podstawie przeglądu literatury oraz analizy cech użytkowników platformy Reddit stworzyliśmy 16 kluczowych wskaźników identyfikujących boty. Wskaźniki bazujące na analizie tekstu oparte zostały na porównaniu reprezentacji tekstu uzyskanej modelem *TfidfVectorizer*. Wskaźniki zostały podzielone na 4 grupy (Po więcej szczegółów zapraszamy na nasze repozytorium):

Aktywność: Boty wyróżniają się nietypowymi wartościami oceny aktywności (*link_karma*, *comment_karma*), młodym wiekiem konta (*account_age*) oraz specyficznym stosunkiem komentarzy do postów (*comment_post_ratio*).

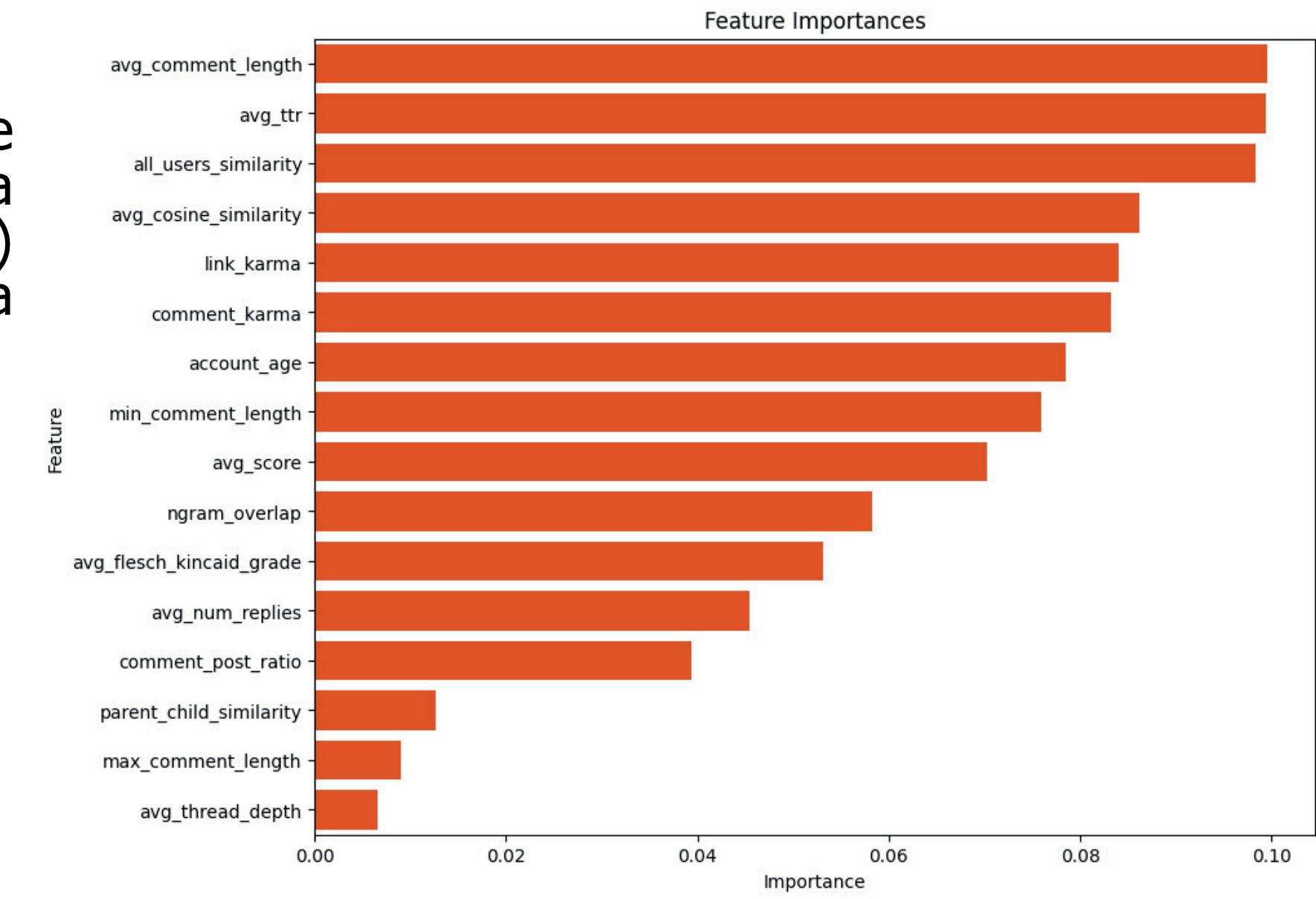
Styl językowy i interakcje: Generują powtarzalne treści (*avg_cosine_similarity*, *all_users_similarity*, *parent_child_similarity*), które wyliczone zostały za pomocą miary podobieństwa cosinusowego, charakteryzuje się skrajnymi długościami komentarzy (*avg_comment_length*, *min_comment_length*, *max_comment_length*), oraz określona głębokością wątków w, których biorą udział (*avg_thread_depth*).

Złożoność językowa: Tworzą teksty o niskiej różnorodności (*avg_ttr*), która obliczona została za pomocą metryki type-toke-ratio, specyficznej złożoności językowej i czytelności tekstu (*avg_flesch_kincaid_grade*), obliczonej za pomocą metryki flesch-kincaid grade i wysokiej powtarzalności wzorców (*ngram_overlap*), obliczonej za pomocą techniki ngram overlap.

Zaangażowanie: Otrzymują nietypowe oceny punktów na platformie (*avg_score*) oraz mniej odpowiedzi na komentarze (*avg_num_replies*).



Dowiedz się więcej na repozytorium projektu (GitHub)



Wykres 1: Znaczenie cech wyznaczone modelem RandomForest.