Kto to napisał?

Kilka postów na Twitterze wystarczy do identyfikacji twoich upodobań politycznych

01 Opis problemu

Aktualnie w Polsce dzieje się wiele istotnych dla obywateli zmian. Przeprowadzane są reformy, a wypowiedzi oraz czyny polityków powodują **skrajne reakcje** wśród społeczeństwa. Przekłada się to na różnorodność i mnogość wpisów w mediach społecznościowych, szczególnie na Twitterze.
W ramach pracy postawiliśmy następującą hipotezę:
Treść postów danego użytkownika ujawnia jego poglądy polityczne.

02 Metodyka

Zbiór postów pozyskaliśmy z platformy *Twitter* przy wykorzaniu oficjalnego API oraz własnego scrapera. Do klasyfikacji preferencji politycznych użytkownika użyliśmy głębokiej sieci neuronowej, a sam model poddaliśmy dokładnej analizie.

03 Dane

Proces zbierania danych był kilkuetapowy oraz zawierał opracowaną heurystykę oceny preferencji politycznych użytkownika. Każdy post musiał dotyczyć co najmniej jednego z popularnych ostatnio tematów:

- **Aborcja**
- **PolskiŁad**
- **LexTVN**
- **TSUE**

Zebrane tweety ograniczyliśmy do wpisów o długości co najmniej 5 słów.

Statystyki finalnego zbioru:

- > **190 tys.** tweetów
- 5 głównych partii politycznych
- 91 kont polityków
- > **10 tys.** aktywnych użytkowników
- średnio 18 wpisów na osobę

Zebranie tweetów z profili polityków wybranych partii

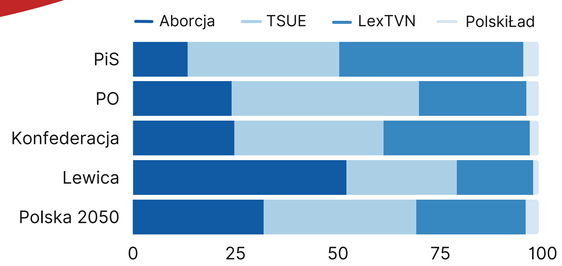
Zdobycie zbioru użytkowników lubiących wpisy polityków

Wybór aktywnych użytkowników > *10*

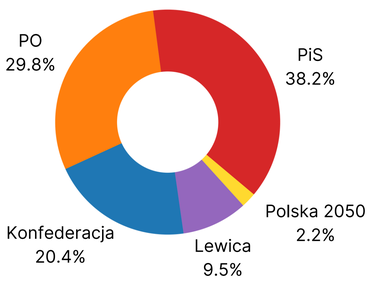
Przypisanie preferencji politycznych według największej liczby polubień postów polityków danego stronnictwa

Zebranie tweetów użytkowników

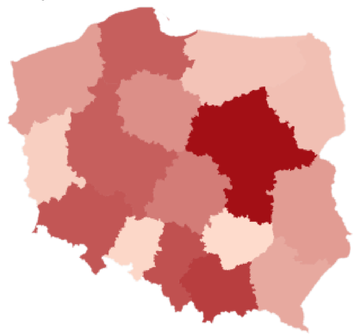
Rozkład tematów w zbiorze



Rozkład profili politycznych

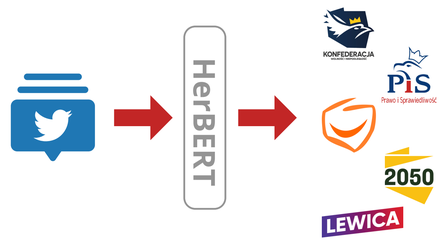


Aktywność Polaków na Twitterze



04 Model

Zastosowaliśmy udostępniony przez firmę *Allegro* kontekstowy model języka polskiego - **HerBERT**

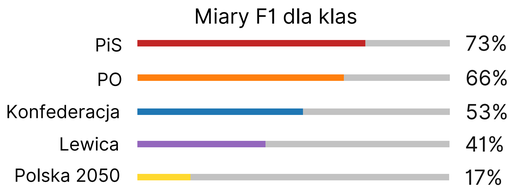


Przygotowanie danych:

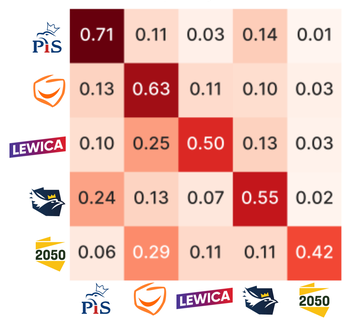
- usunięcie z tekstu # @ 🌐
- podział na 3 zbiory - treningowy, walidacyjny i testowy
- próbkowanie ważone z racji dużego **niezbalansowania klas**

05 Wyniki

Najlepszy uzyskany model osiągnął wynik na poziomie ok. **0.65** globalnej **miary F1**. Metrykę obliczyliśmy także dla każdej z klasy z osobna.



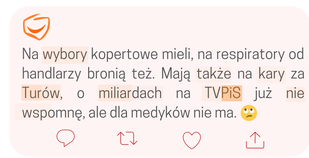
Macierz pomyłek modelu

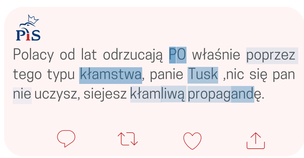


- Model najczęściej mylił się w przypadku postów zwolenników Lewicy i Polski 2050.
- Macierz pomyłek do pewnego stopnia oddaje różnice pomiedzy partiami obserwowane w polskim społeczeństwie.
- Najłatwiej jest rozpoznać zwolenników dwóch największych partii.
- Polska 2050 jest często mylona z innymi ugrupowaniami.

06 Wyjaśnialność

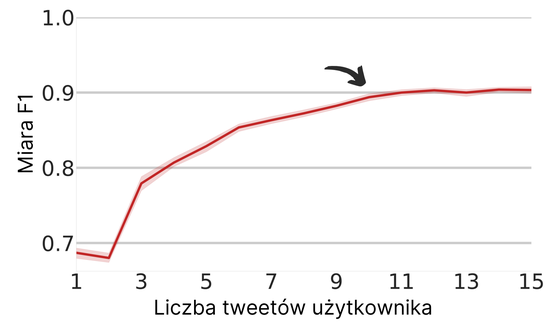
Do analizy działania modelu użyliśmy metody *Layer Integrated Gradients*, która pozwala na obliczenie wpływu wektorów osadzeń tekstu na ostateczny wynik. Intensywność koloru pokazuje istotność słowa przy predykcji poprawnej partii.





07 Analizy

Ilu tweetów potrzeba aby poznać twoje preferencje polityczne?



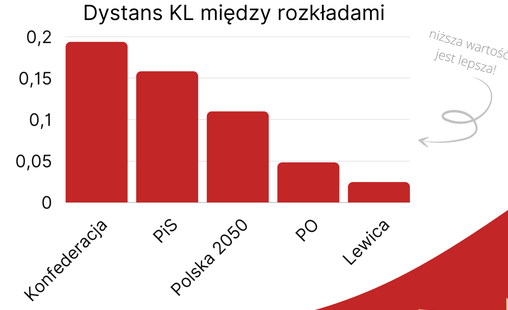
Do określenia poglądów użytkownika wystarczy jedynie **10 tweetów**!

Czy politycy danej partii piszą tak samo jak ich zwolennicy?

Obliczyliśmy odległość *Kullbacka-Leiblera* między rozkładem predykcji modelu dla wpisów oficjalnych przedstawicieli partii oraz tweetów jej zwolenników.

- Styl wypowiedzi posłów Konfederacji najbardziej odbiega od stylu zwolenników tej partii.
- Najbardziej zgodni z treściami społeczności są politycy Lewicy i PO.

Dystans KL między rozkładami



Opiekun projektu
mgr inż. Krzysztof Rajda

Autorzy
Joanna Baran *242505@student.pwr.edu.pl*
Michał Kajstura *242491@student.pwr.edu.pl*
Maciej Ziółkowski *242475@student.pwr.edu.pl*

Politechnika Wrocławska