Analiza polityczna polskich mediów społecznosciowych

https://github.com/SMAPWr/project-r-polska_political_analysis

Marek Pokropiński Filip Strzałka

Cel projektu

Celem projektu było zbudowanie modelu do analizy nacechowania politycznego tekstu oraz analiza polskich mediów społecznościowych pod kontem nacechowania politycznego.

Pozyskanie danych

Dane do uczenia modelu zostały pozyskane z twittera. Dane zostały oznaczone automatyczne na podstawie hashtagów oraz wydźwięku.

Do analiz pobrano teksty:

- z twittera - zawierające polityczne hashtagi,
- z reddita - komentarze z subreddita r/Polska z flarą "Polityka"
- z wykopu - oznaczone #Polska

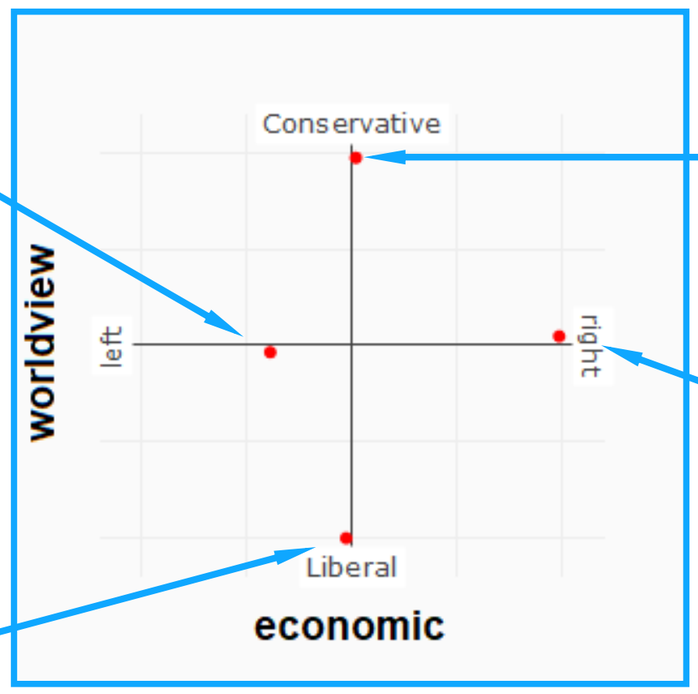
Modele

Podstawą budowy modeli był model HerBERT o architekturze *transformer*. Zbudowaliśmy dwa modele: do predykcji sentymentu oraz do predykcji nastawienia politycznego.

Empiryczna ewaluacja modelu

#Kraj I Dobra wiadomość! Coraz wyższe wynagrodzenia Polaków! #wynagrodzenia #wzrost #TVRepublika https://t.co/w0IfAvVWDe

Czerwony alarm dla Polek! Zakaz aborcji znów w Sejmie #czarnyprotest #ratujmykobiety Podpisz i podaj dalej: #StrajkKobiet #piekłokobiet #zakazaborcji #aborcja #AborcjaBezGranic



@lis_tomasz #LGBTtoldeologia i to bardzo zła ideologia. Dla zła nie ma miejsca. Mówienie o LGBT czy o ideologii nie jest oceną zwyklych normalnych osób o innej orientacji. Jeżeli ktoś mówi, że gej to LGBT to go poniża, utożsamia z nachalną, bez tolerancji wobec innych ideologią.

P. doktorze @DominikKucinsk Czy to prawda że VAT na nawozy wzrósł z 8% aż do 23% za rządów @pisorgpl ? □ #PiS #podatki #socjalizm #podwyżki #bezrobocie #kryzys #rolnicy

Wydźwięk (ang. *sentiment)*

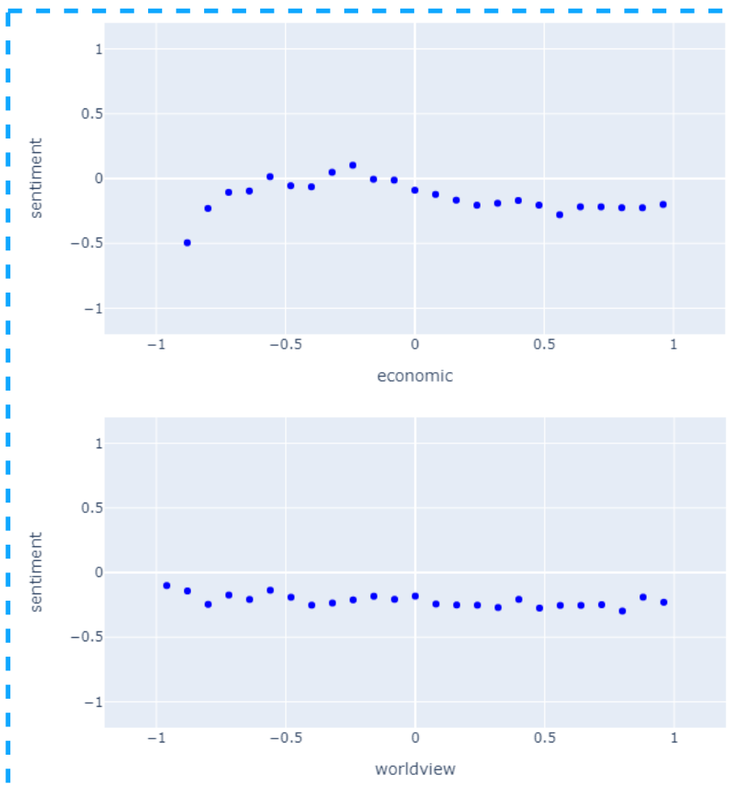
Korzystając z danych CLARIN-PL wyuczyliśmy model regresyjny wydźwięku potrzebny do automatycznego oznaczania danych. Dokonaliśmy też prostej analizy statystycznej sentymentu zebranych zbiorów postów.

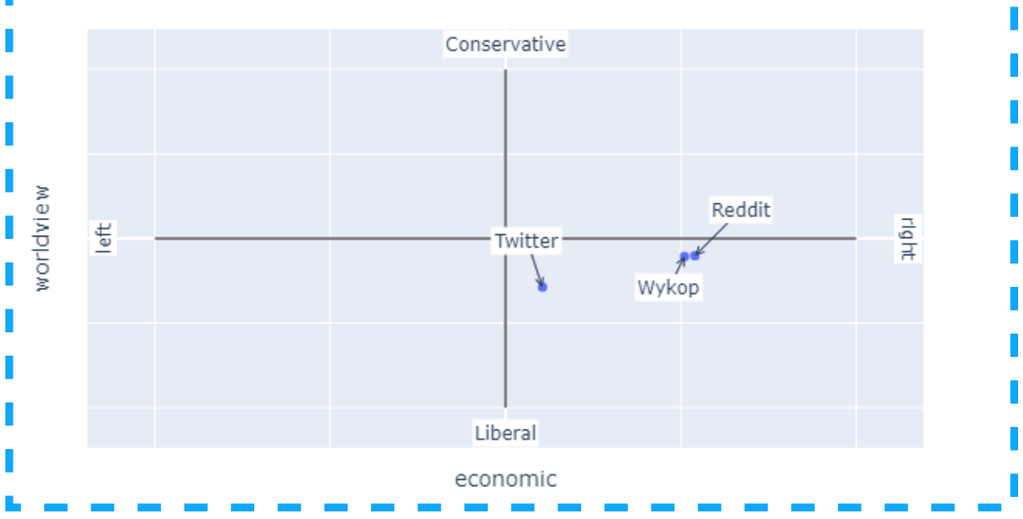
Z uwagi na brak podobnych modeli dla języka polskiego, upubliczniliśmy go na Githubie oraz w formie repozytorium PyPI.

https://github.com/philvec/sentimentPL

sentimentpl 0.0.6

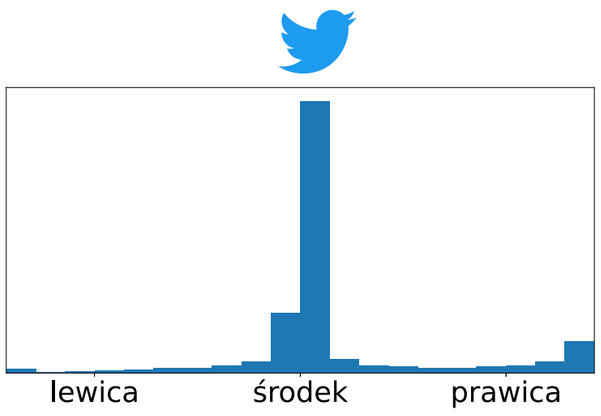
```
pip install sentimentpl
```

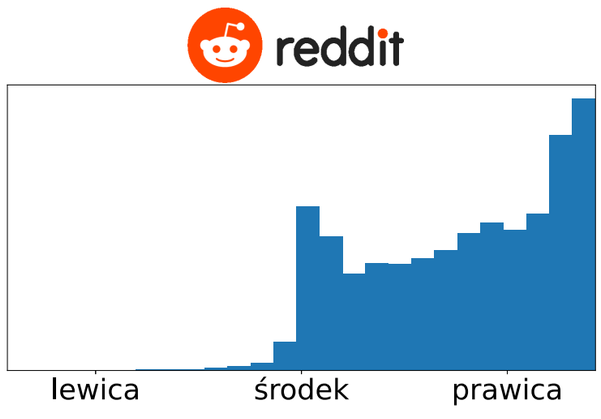


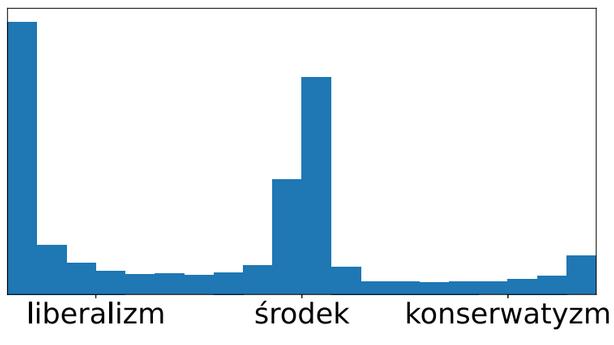


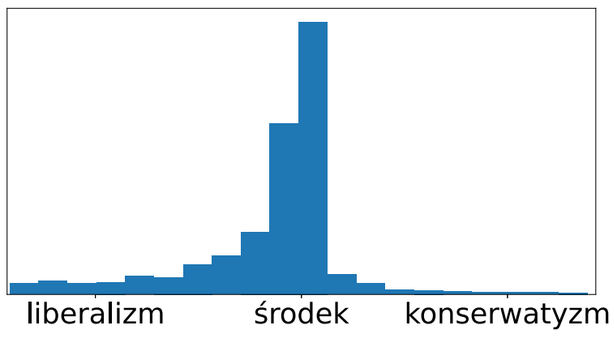
Porównanie polskich mediów

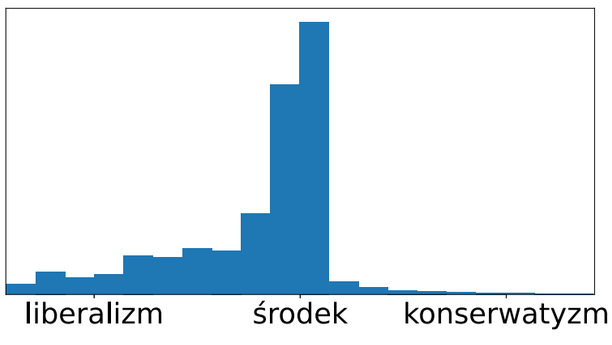
W celu analizy polskich mediów społecznościowych przeanalizowaliśmy teksty pozyskane z portali za pomocą zbudowanych modeli. W celu uzyskania pojedynczego punktu dla nacechowania politycznego uśredniliśmy predykcje dla każdego medium.











Rozkład nacechowania politycznego

Po przeprowadzeniu analizy okazało się, że reddit i wykop mają podobne rozkłady, jednakże twitter wygląda zupełnie inaczej niż pozostałe.