Sensacja: Wiemy co Nowak lubi komentować!!! [Zobacz zdjęcia]

Więcej danych znajdziesz na

Głos Danych

Wydanie specjalne
nr 1/2020

Środa, 29 stycznia 2020

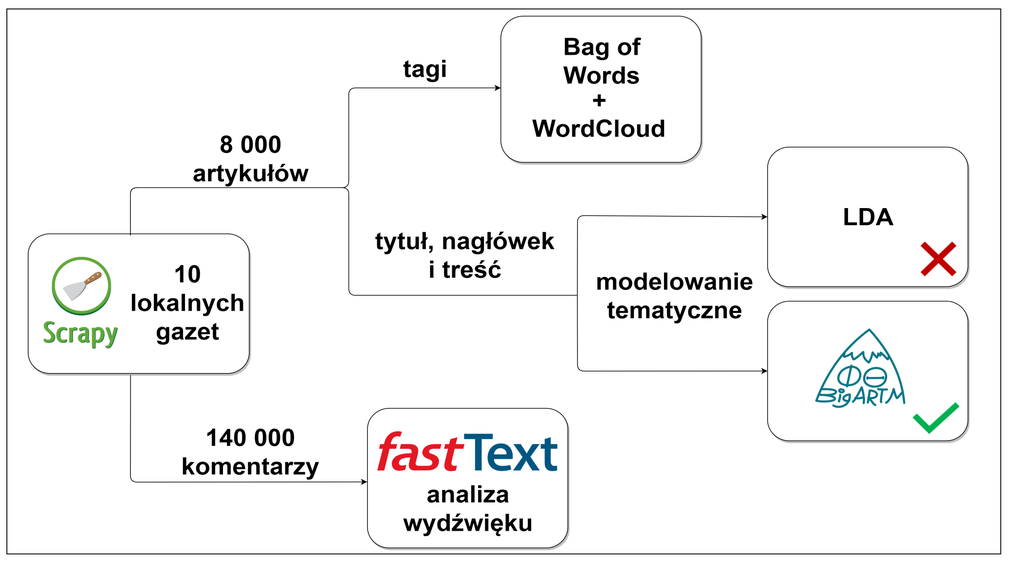
Redakcja: Joanna Komoszyńska, Martyna Laszczyńska, Michał Gołębiewicz
Kontakt: {248873, 229819, 229809}@student.pwr.edu.pl

Opis problemu

Celem projektu było zebranie artykułów i komentarzy z 10. lokalnych gazet z całej Polski. Na zebranych danych podjęto próbę analizy wydźwięku i modelowania tematycznego.

Rys. Chmura słów dla tagów z artykułów.

Przepływ danych

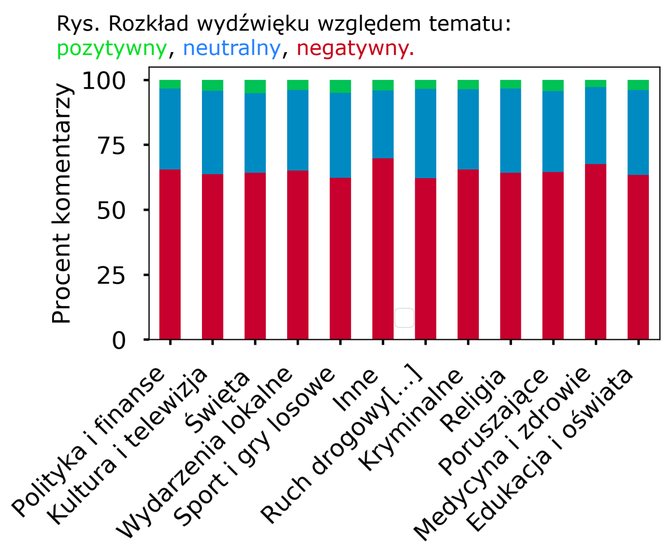


Rys. Schemat przetwarzania danych.

Rys. Przykład komentarza.

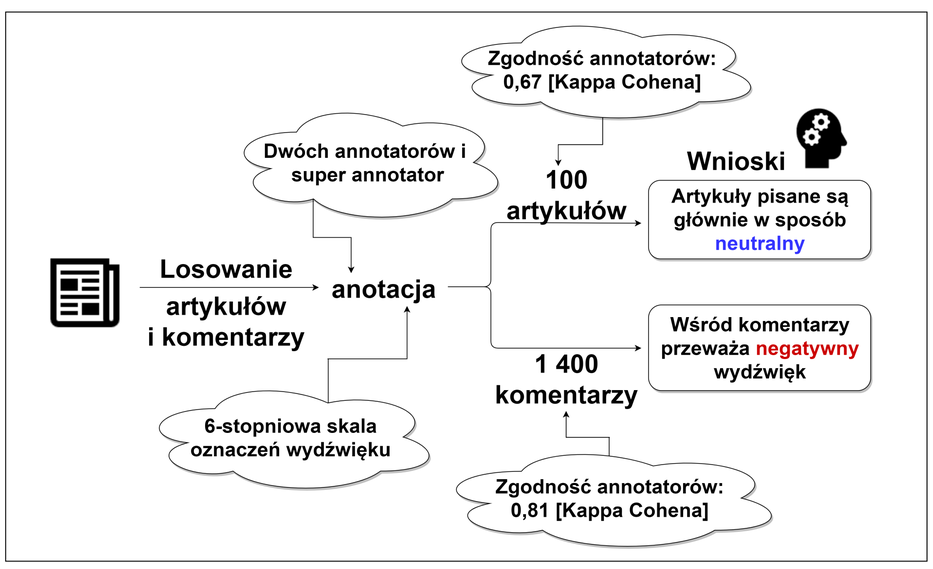
Analiza wydźwięku

Rys. Rozkład wydźwięku względem tematu: pozytywny, neutralny, negatywny.



Oznaczonych komentarzy użyto do trenowania modelu FastText. Przeprowadzono eksperymenty z różnymi parametrami oraz z dodatkowymi danymi, spoza zebranego zbioru (m. in. zbiór PolEmo). W finalnej wersji zastosowano uproszczoną klasyfikację, składającą się z wydźwięku pozytywnego (1), neutralnego (0) i negatywnego (-1).

Oznaczanie danych

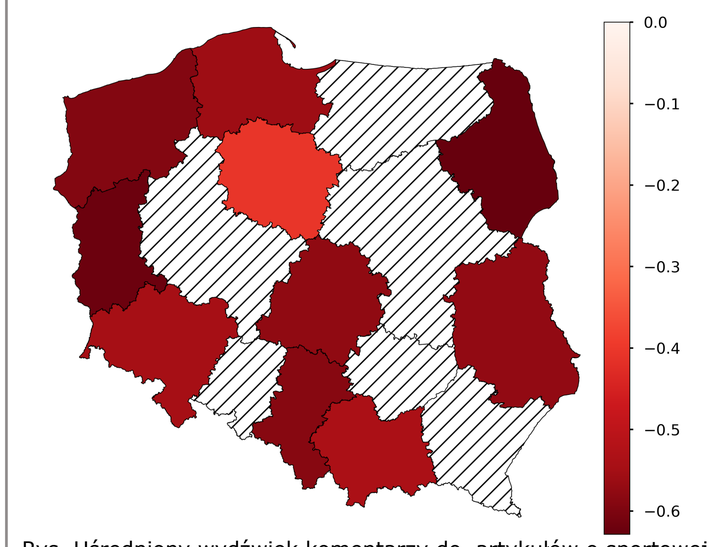


Rys. Proces oznaczania danych.

Dwóch annotatorów nadawało jedną z 6 klas: słabo i mocno negatywny / pozytywny, neutralny oraz ambiwalentny (wewnętrznie sprzeczny).

Mocno pozytywne: gratulacje, pochwały lub kilka mniej wyrazistych, ale pozytywnych zwrotów.

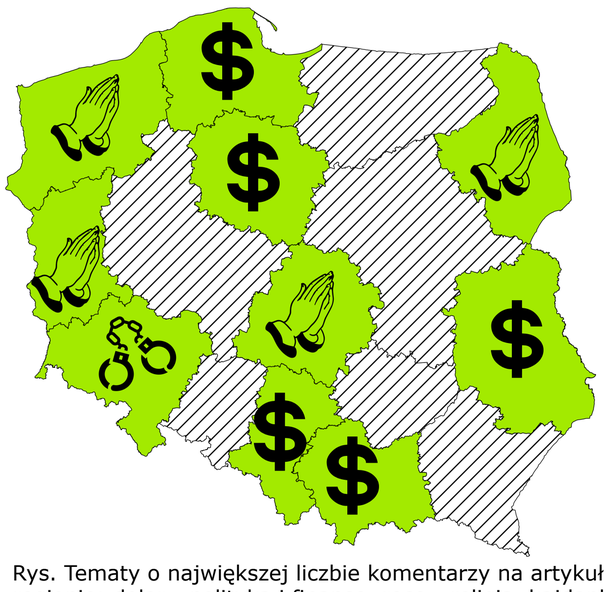
Mocno negatywne: bezpośrednie ataki na konkretną osobę, wzmożone użycie przekleństw lub życzenie śmierci.



Rys. Uśredniony wydźwięk komentarzy do artykułów o sportowej tematyce. Zakreskowanie oznacza brak gazety w danym regionie. Pozostałe tematy mają zbliżony rozkład wydźwięku.

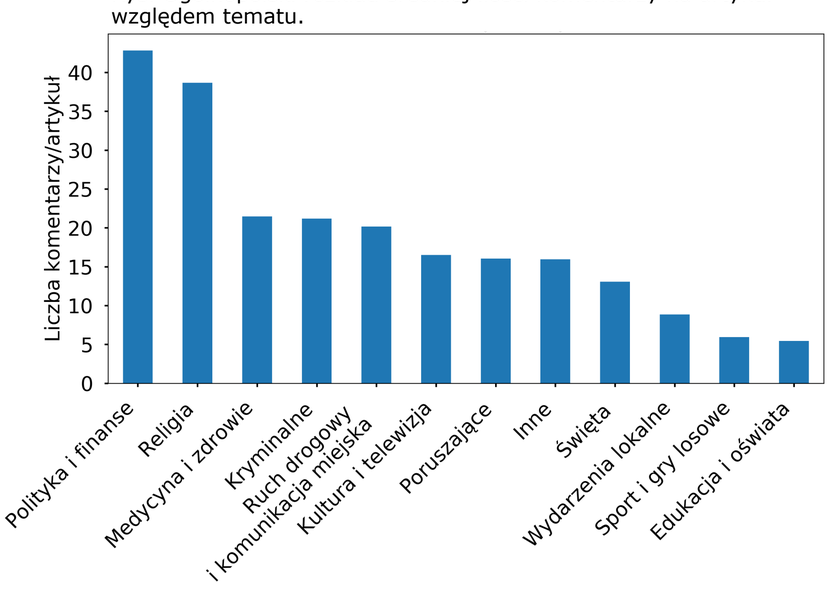
Modelowanie tematyczne

Do podziału artykułów wg tematu przebadano algorytmy BigARTM i LDA. Najpierw słowa w artykułach zamieniono na formy podstawowe za pomocą tagera MorphoDiTa. Następnie wygenerowano 50 tematów, które ręcznie zebrano w 12 grup. Lepsze wyniki otrzymano dla BigARTM. Na wylosowanej próbce 120 artykułów algorytm poprawnie wskazał temat dla 98 z nich.



Rys. Tematy o największej liczbie komentarzy na artykuł w regionie: dolar - polityka i finanse, ręce - religia, kajdanki - kryminalne. Obszary zakreskowane - brak gazety w regionie.

Rys. Ogólnopolski rozkład średniej ilości komentarzy na artykuł względem tematu.



Napotkane problemy

- mocno niezbalansowany wydźwięk komentarzy
- dobór kryteriów oznaczania komentarzy i artykułów
- występowanie wielu tematów w jednym artykule

Wnioski

- niezależnie od regionu Nowak komentuje najczęściej negatywnie
- największy ułamek pozytywnych komentarzy otrzymały święta
- polityka i religia przyciągają najwięcej komentarzy na artykuł
- dodatkowe zbiory danych nie poprawiły jakości klasyfikacji wydźwięku