Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Metody Inteligencji Obliczeniowej

# Zastosowanie analizy SHAP w analizie sentymentu metodami NLP

Autorzy:
Maciej Leśniak
Paweł Dyjak
Aleksander Gaweł

czerwiec 2025

# 1 Wstęp

Celem projektu było zbudowanie modeli do analizy sentymentu wypowiedzi Donalda Trumpa na podstawie jego tweetów. Dane zostały pobrane z platformy `Kaggle`, ze zbioru *Trump Tweets*. Zdecydowaliśmy się na zastosowanie dwóch metod do wstępnego etykietowania sentymentu:

- **VADER** – leksykalny analizator sentymentu dostosowany do tekstów społecznościowych,
- **BERT (via Hugging Face)** – głęboki model transformerowy. (*cardiffnlp/twitter-roberta-base-sentiment*)

W dalszym etapie zbudowano modele klasyfikujące, przeprowadzono analizę wpływu poszczególnych tokenów na decyzję modelu z użyciem metody SHAP, a następnie omówiono uzyskane wyniki.

# 2 Dane i przygotowanie

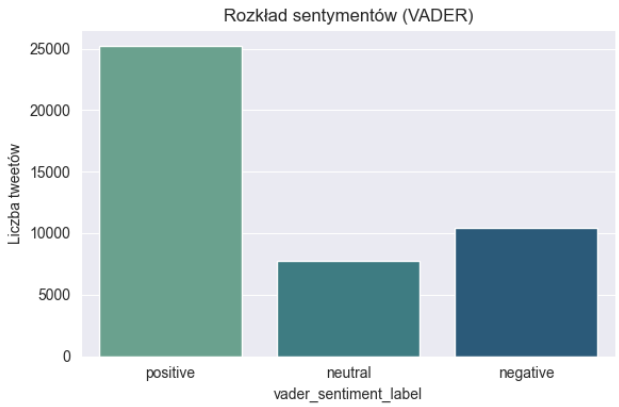
Zbiór danych zawiera ponad 40 tysięcy tweetów opublikowanych z konta Donalda Trumpa. Dane zostały przetworzone dla każdego z modeli osobno. Dla modelu **VADER**:

1. Zastąpienie linków do obrazków umieszczonych na Twitterze przez placeholder `<IMG>`
2. Usunięcie linków i adresów URL
3. Usunięcie znaków `@` i `#` z tekstu, zachowując słowa kluczowe
4. Usunięcie cudzysłowów pojedynczych i podwójnych
5. Redukcja wielokrotnych białych znaków do pojedynczej spacji.
6. Usunięcie wszystkich pozostałych znaków specjalnych i interpunkcji, pozostawiając jedynie litery i spacje.
7. Usunięcie tzw. *stopwords* (słów nieinformacyjnych) w języku angielskim.

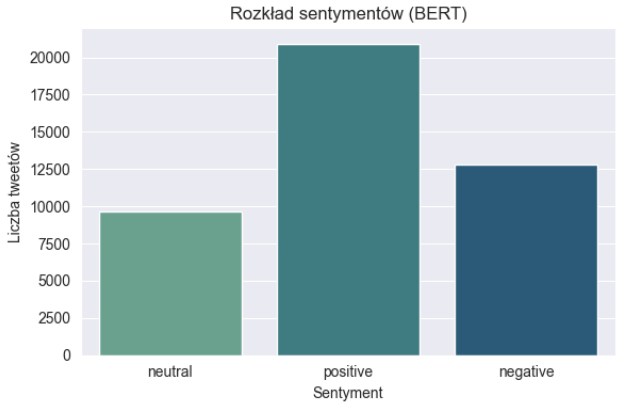
W związku z tym, że **BERT** lepiej rozumie kontekst wypowiedzi w jego przypadku zastosowaliśmy nieco mniej rygorystyczny preprocessing:

1. Zastąpienie pełnych linków adresem placeholder `http`
2. Zastąpienie linków do obrazków umieszczonych na Twitterze przez placeholder `<IMG>`
3. Usunięcie znaków `@` i `#` z tekstu, zachowując słowa kluczowe
4. Usunięcie cudzysłowów pojedynczych i podwójnych
5. Redukcja wielokrotnych białych znaków do pojedynczej spacji

Następnie użyto obu modeli do wstępnego etykietowania i uzyskano następujące wyniki:




Rysunek 1: Rozkład sentymentów tweetów według klasyfikatora Vader




Rysunek 2: Rozkład sentymentów tweetów według klasyfikatora BERT

## 2.1 VADER

Model VADER dokonuje oceny sentymentu bazując na słowniku słów z przypisanymi wagami emocjonalnymi. Przyjęto następujące progi:

- Compound score $> 0.05$ – sentyment pozytywny,
- Compound score $< -0.05$ – sentyment negatywny,
- W pozostałych przypadkach – neutralny.
- Sleepy Joe Biden refuses leave basement sanctuary tell Radical Left BOSSES heading wrong direction Tell get Seattle Liberal Governor JayInslee looking fool LAW ORDER JayInslee — **Negative**
- Domestic Terrorists taken Seattle run Radical Left Democrats course LAW ORDER — **Negative**
- Actually great book great highly respected historian Doug Wead — **Positive**

### 2.2 BERT

Model BERT zwraca rozkład prawdopodobieństw dla każdej z trzech klas sentymentu (`negative`, `neutral`, `positive`). Ostateczna etykieta tweeta została przypisana na podstawie klasy o największym prawdopodobieństwie, zgodnie z zasadą wyboru maksimum spośród podanych wartości.

- If Obama was smart, he would cancel the Muslim Brotherhoods WH visit later this month. He wont. — **Negative**
- And finally, Cruz strongly told thousands of caucusgoers (voters) that Trump was strongly in favor of ObamaCare and choice - a total lie! — **Negative**
- MAKE AMERICA GREAT AGAIN and then, KEEP AMERICA GREAT! — **Positive**

## 3 Wytrenowany Model

### 3.1 Dane

Do trenowania wykorzystaliśmy dane przetworzone taką samą metoda jak dane dla **VADERA**. Podzieliśmy dane na zbiór treningowy oraz testowy w stosunku 4:1.

### 3.2 Struktura własnego modelu klasyfikacyjnego

Dla każdego tweeta przypisano etykiety sentymentu (`positive`, `negative`, `neutral`) na podstawie wyników uzyskanych z gotowych modeli BERT oraz VADER. Następnie na bazie tych etykiet zbudowano i wytrenowano własny model klasyfikacyjny, którego celem było odtworzenie przypisanego sentymentu na podstawie tekstu tweetów.

Struktura modelu:

- **Warstwa wejściowa:** `TF-IDF Vectorizer`
  - Każdy tweet jest przekształcany w wektor cech o wymiarze maksymalnie 5000.
  - Każdy wymiar odpowiada wystąpieniu konkretnego tokena w tekście.
- **Warstwa wyjściowa:** `Logistic Regression (multinomial)`
  - Model liniowy z funkcją aktywacji softmax.
  - Przewiduje jedną z trzech klas sentymentu.
  - Parametry: solver `lbfgs`, maksymalnie 200 iteracji.

Łącznie pipeline można przedstawić w skrócie następująco:

$$Tweet \longrightarrow TF-IDF \longrightarrow LogisticRegression \longrightarrow Sentyment$$

Taki model umożliwia ocenę, które słowa lub zwroty mają największy wpływ na przypisanie danego sentymentu, co dodatkowo zbadano przy pomocy interpretacji SHAP.

## 3.3 Model trenowany na podstawie VADER

Model trenowany na podstawie etykiet wygenerowanych przez analizator VADER osiągnął wysoką skuteczność. Dokładność na zbiorze treningowym wyniosła **91.4%**, a na testowym **85.7%**, co świadczy o dobrej generalizacji i niewielkim przeuczeniu.

Najlepsze wyniki uzyskano dla klasy pozytywnej (F1 = 0.91 na teście), a najsłabsze – dla klasy neutralnej (F1 = 0.81). Model miał tendencję do nieco częstszego przypisywania klas pozytywnych, co wynika z przewagi przykładów pozytywnych w zbiorze treningowym. Niemniej jednak, nawet w klasach mniej licznych model zachował zadowalającą precyzję i czułość.

Macierze pomyłek dla tego modelu pokazują, że:

- Dla klasy pozytywnej model osiąga bardzo wysoką trafność – **4664** poprawnych predykcji w teście.
- Klasa neutralna jest stosunkowo dobrze rozpoznawana (**1096** trafień), ale model myli ją częściej z klasą pozytywną (**246**) niż negatywną (**161**).
- Najwięcej błędów dotyczy klasy negatywnej, gdzie aż **341** przypadków zakwalifikowano błędnie jako pozytywne.

**Test Accuracy:** 85.7%
**Train Accuracy:** 91.4%

## 3.4 Model trenowany na podstawie BERT

Model wytrenowany na etykietach pochodzących z BERT-a osiągnął niższą skuteczność niż model oparty na VADER. Dokładność na zbiorze treningowym wyniosła **85.1%**, a na testowym **79.1%**, co oznacza różnicę ponad 6 punktów procentowych — jest to sygnał umiarkowanego przeuczenia.

Pomimo że etykiety z BERT-a teoretycznie lepiej oddają kontekst i znaczenie wypowiedzi, ich reprodukcja za pomocą klasycznego modelu TF-IDF + regresja logistyczna okazała się ograniczona. Klasyfikator oparty na tokenach nie potrafi odtworzyć złożonych zależności kontekstowych, które BERT uchwyca. Problem szczególnie uwidacznia się przy klasie neutralnej, dla której f1-score na zbiorze treningowym wynosi **0.73**, a na zbiorze testowym nie przekracza **0.70**.

Analiza macierzy pomyłek wskazuje na wyraźne trudności w rozróżnianiu klasy neutralnej:

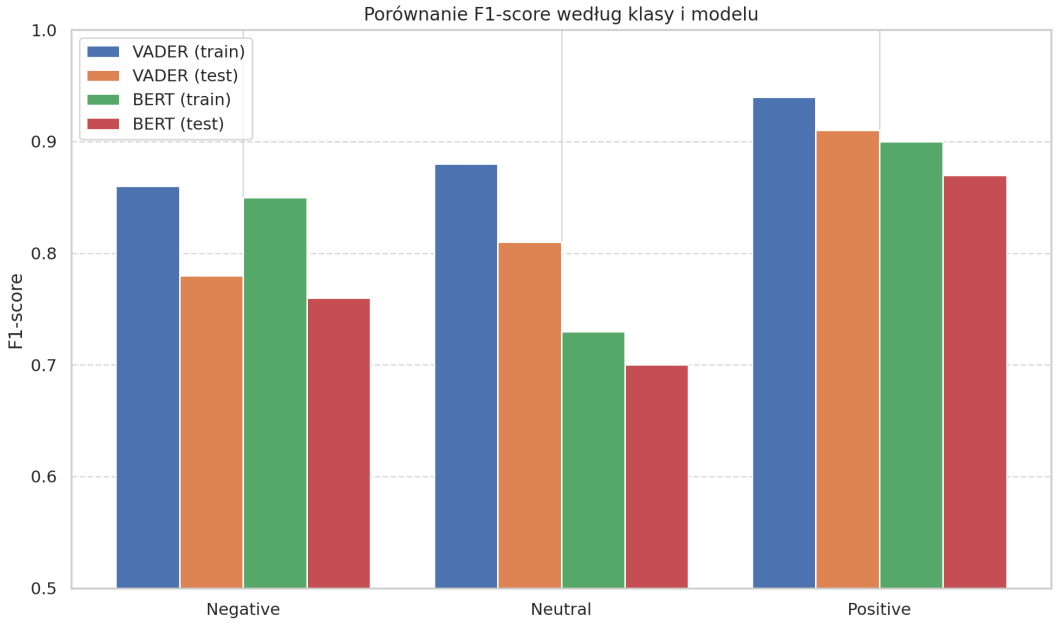
- Tylko **1141** tweetów neutralnych zostało poprawnie zaklasyfikowanych, podczas gdy **454** błędnie przypisano jako pozytywne, a **370** jako negatywne.
- Klasa pozytywna została rozpoznana najlepiej — **3656** poprawnych klasyfikacji, z relatywnie małą liczbą błędów.
- Klasa negatywna również była podatna na błędne klasyfikacje – pomylono ją **214** razy z pozytywną i **274** razy z neutralną.

Ogólnie model wykazuje przyzwoitą ogólną skuteczność, lecz trudność w klasyfikacji klasy „środkowej" (neutralnej) wskazuje na brak równowagi klas oraz brak rozumienia subtelności wypowiedzi. W tym podejściu szczególnie brakuje komponentu kontekstowego, którego wymagałyby etykiety BERT-a.

**Test Accuracy:** 79.1%
**Train Accuracy:** 85.1%

## 3.5 Porównanie F1-score według klas

Aby lepiej zrozumieć zachowanie obu modeli, porównano wartość metryki F1-score osobno dla każdej z trzech klas: **negative**, **neutral** i **positive**. Na wykresie przedstawiono wartości F1-score zarówno dla zbioru treningowego, jak i testowego, osobno dla modeli trenowanych na etykietach z VADER oraz BERT.




Rysunek 3: F1-score dla każdej klasy sentymentu: porównanie modeli VADER i BERT

Na podstawie wykresu można sformułować następujące wnioski:

- **Model trenowany na etykietach VADER osiąga lepsze wyniki w każdej klasie** na zbiorze testowym, mimo prostszej natury źródłowych etykiet.
- **Klasa neutralna jest najtrudniejsza do rozpoznania** dla obu modeli, a szczególnie dla modelu opartego na etykietach BERT-a (F1-score spada do około 0.70).
- **Widoczne są oznaki przeuczenia**, zwłaszcza w modelu BERT, gdzie różnica F1 między train a test osiąga nawet 0.1 punktu.
- **Klasa pozytywna jest najlepiej klasyfikowana**, co wynika z jej dominacji w zbiorze danych oraz obecności wyraźnych, nacechowanych emocjonalnie tokenów.

### 3.6 Porównanie skuteczności modeli

| **Model** | **Train Acc.** | **Test Acc.** | **Δ** |
|---|---|---|---|
| TF-IDF + LR (VADER) | 91.4% | 85.7% | 5.7 pp |
| TF-IDF + LR (BERT) | 85.1% | 79.1% | 6.0 pp |

Tabela 1: Porównanie dokładności modeli na zbiorze treningowym i testowym

Główne obserwacje:

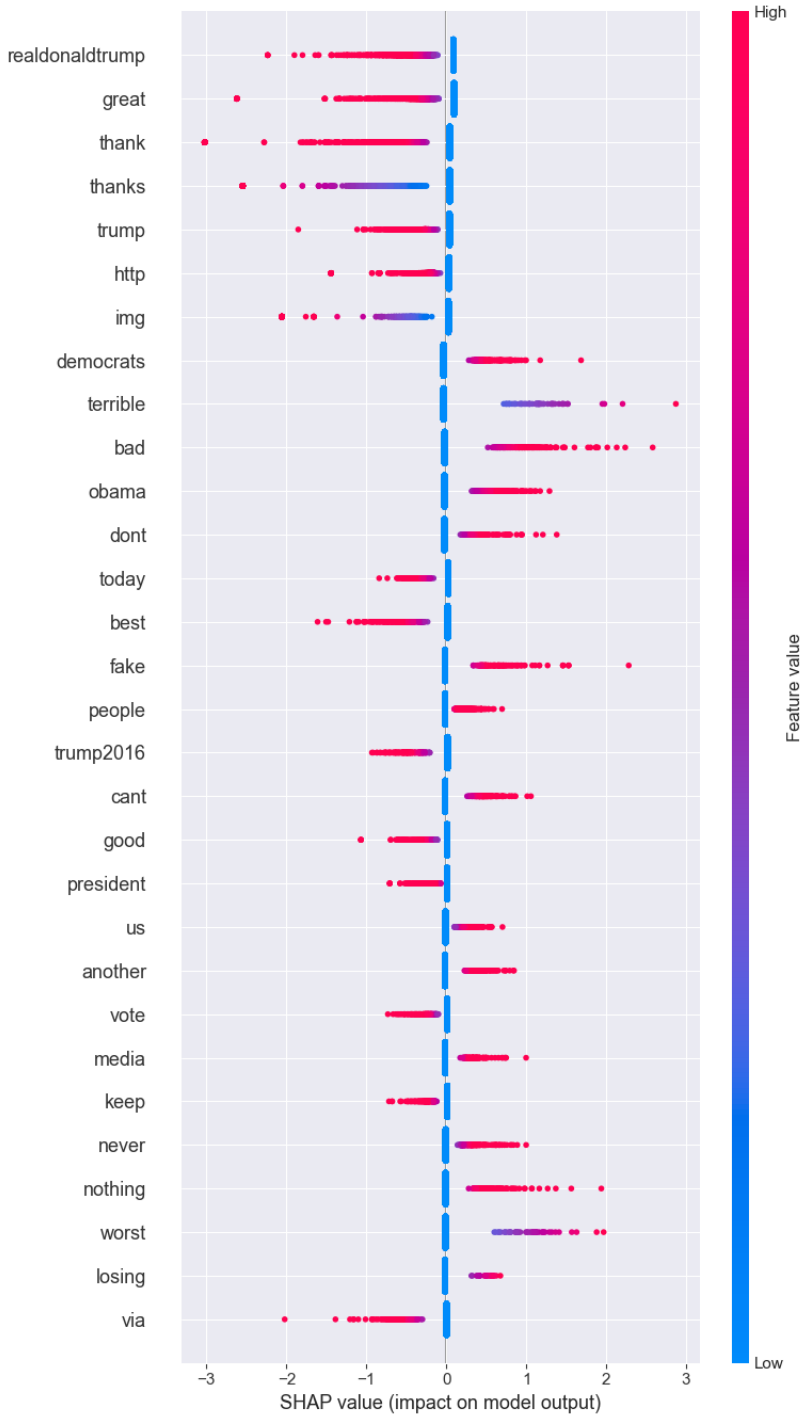
- Model uczony na etykietach VADER wykazuje mniejsze przeuczenie i lepszą skuteczność ogólną.
- Klasy neutralne są słabo rozpoznawane w obu przypadkach, szczególnie przez model trenowany na etykietach BERT-a.
- BERT generuje bogatsze, kontekstowe etykiety, ale klasyfikator bazujący na TF-IDF nie jest w stanie ich w pełni wykorzystać – brak mu rozumienia kontekstu.
- Pomimo przewagi BERT-a jako źródła etykiet, jego przewaga nie przekłada się jednoznacznie na lepsze rezultaty w klasyfikatorze liniowym.
- W przyszłości warto rozważyć fine-tuning samego BERT-a lub zastosowanie modeli typu ensemble.

## 4 Analiza SHAP

Dla obu modeli przeprowadzono analizę interpretowalności predykcji przy użyciu metody SHAP (SHapley Additive exPlanations). Wygenerowano wykresy typu beeswarm dla każdej klasy (negatywna, neutralna, pozytywna), pokazujące wpływ tokenów na wyjście modelu.

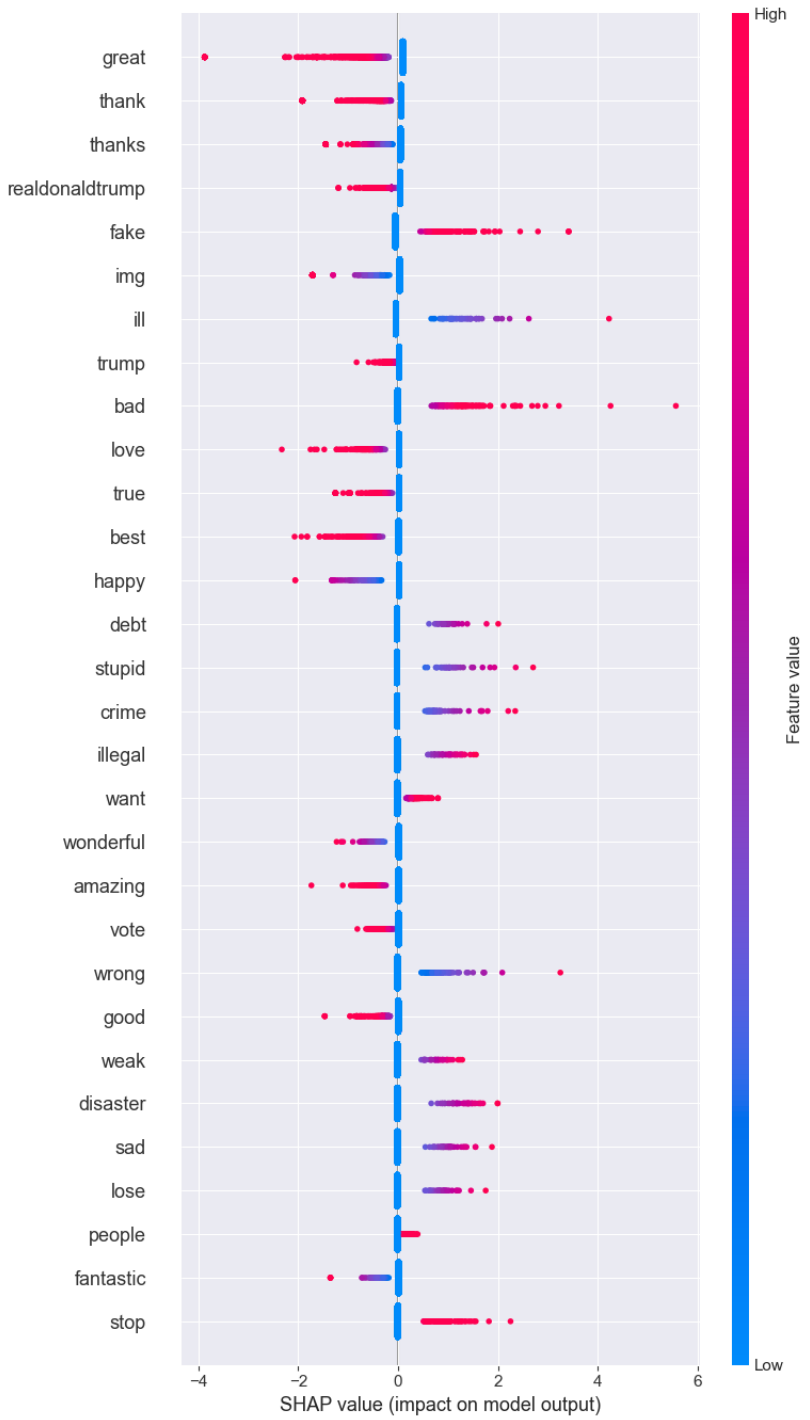
## 4.1 Klasa: Negatywna

### 4.1.1 BERT




Rysunek 4: SHAP – wpływ słów na klasyfikację jako pozytywną

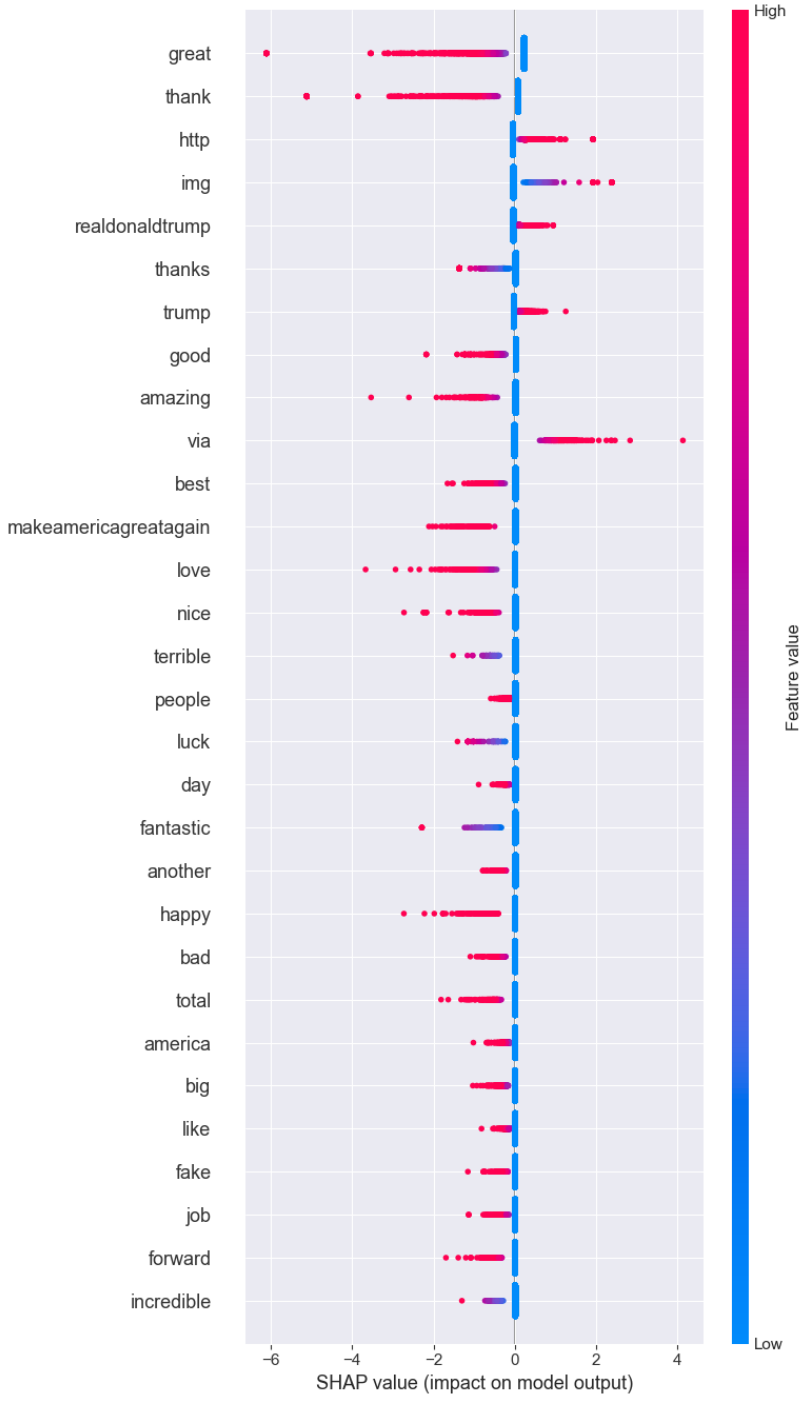
### 4.1.2 VADER




Rysunek 5: SHAP – wpływ słów na klasyfikację jako negatywną

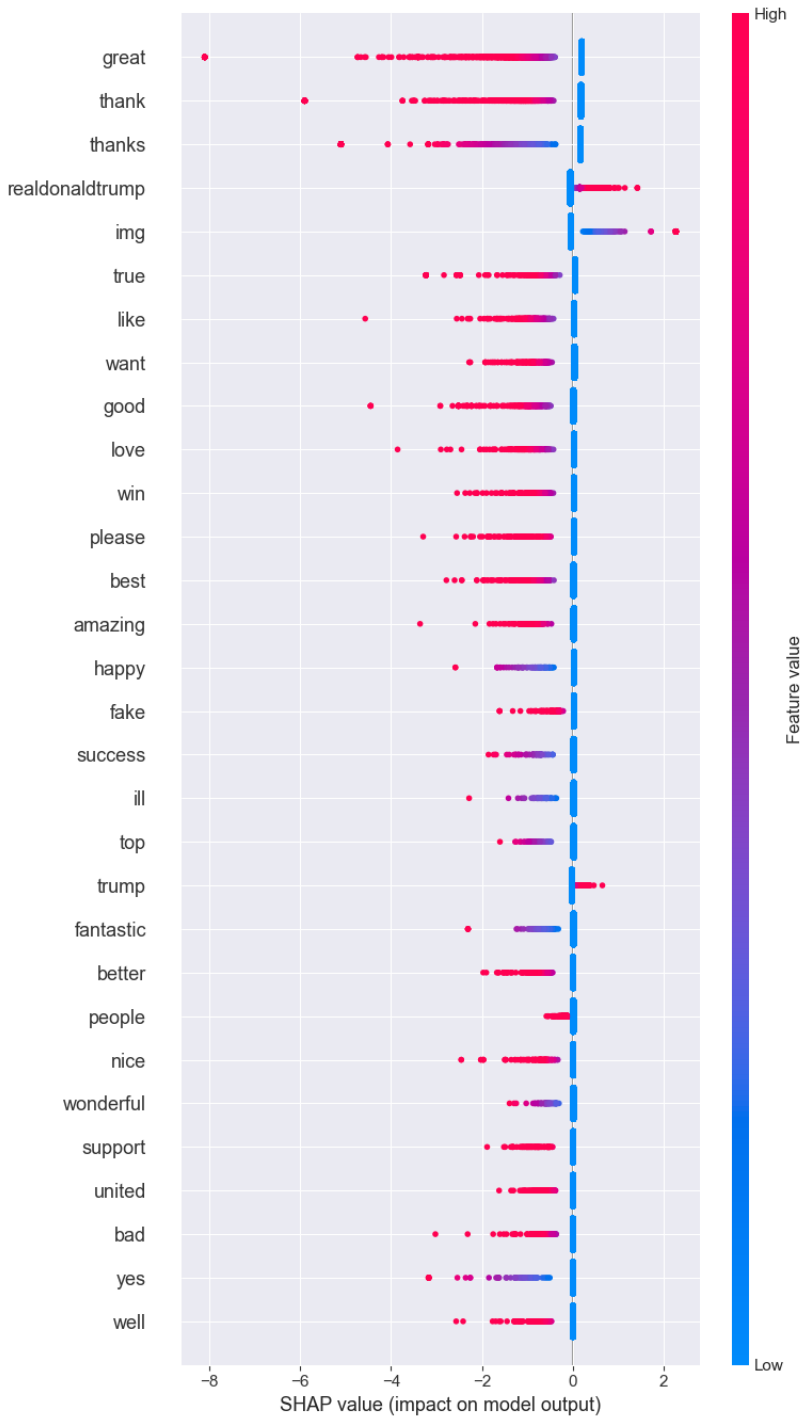
## 4.2 Klasa: Neutralna

### 4.2.1 BERT




Rysunek 6: SHAP – wpływ słów na klasyfikację jako neutralna

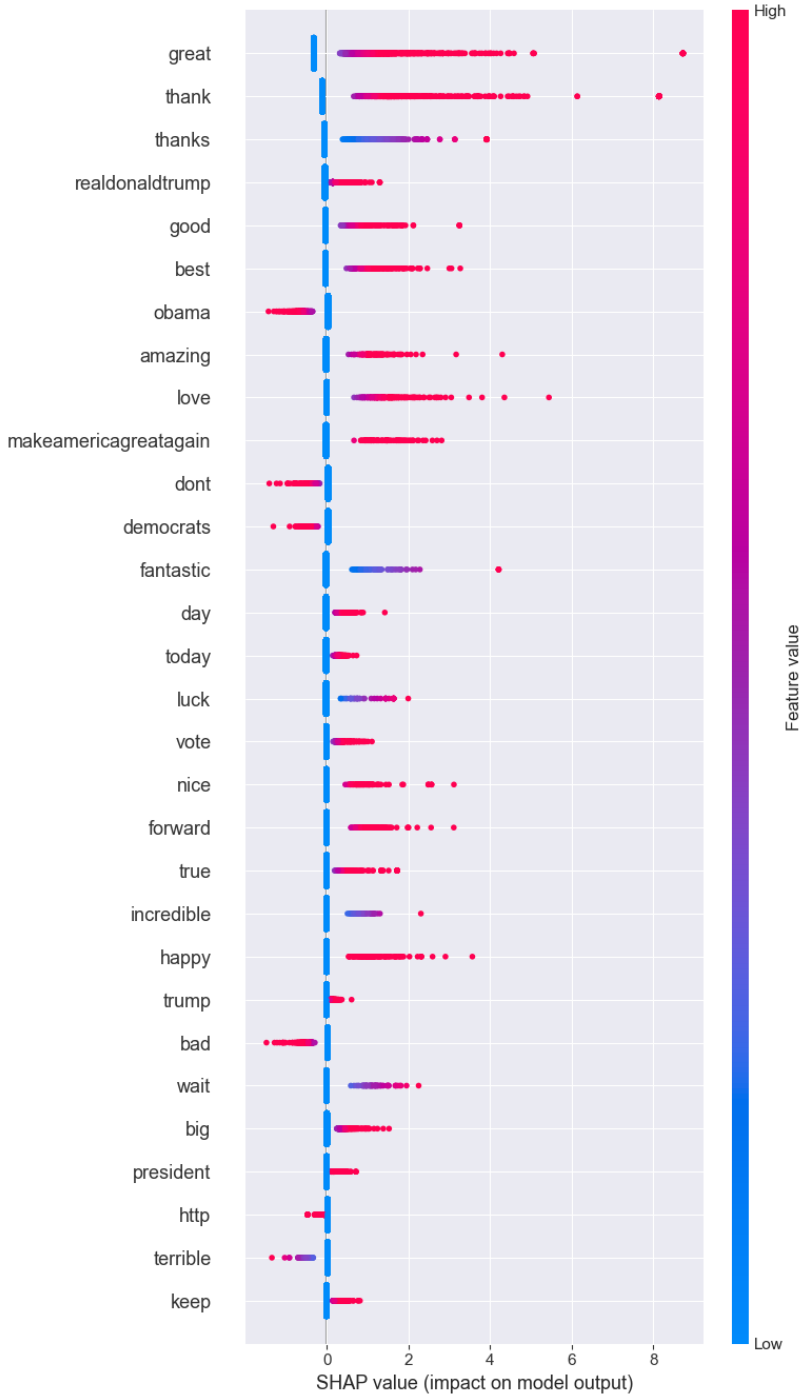
### 4.2.2 VADER




Rysunek 7: SHAP – wpływ słów na klasyfikację jako neutralna

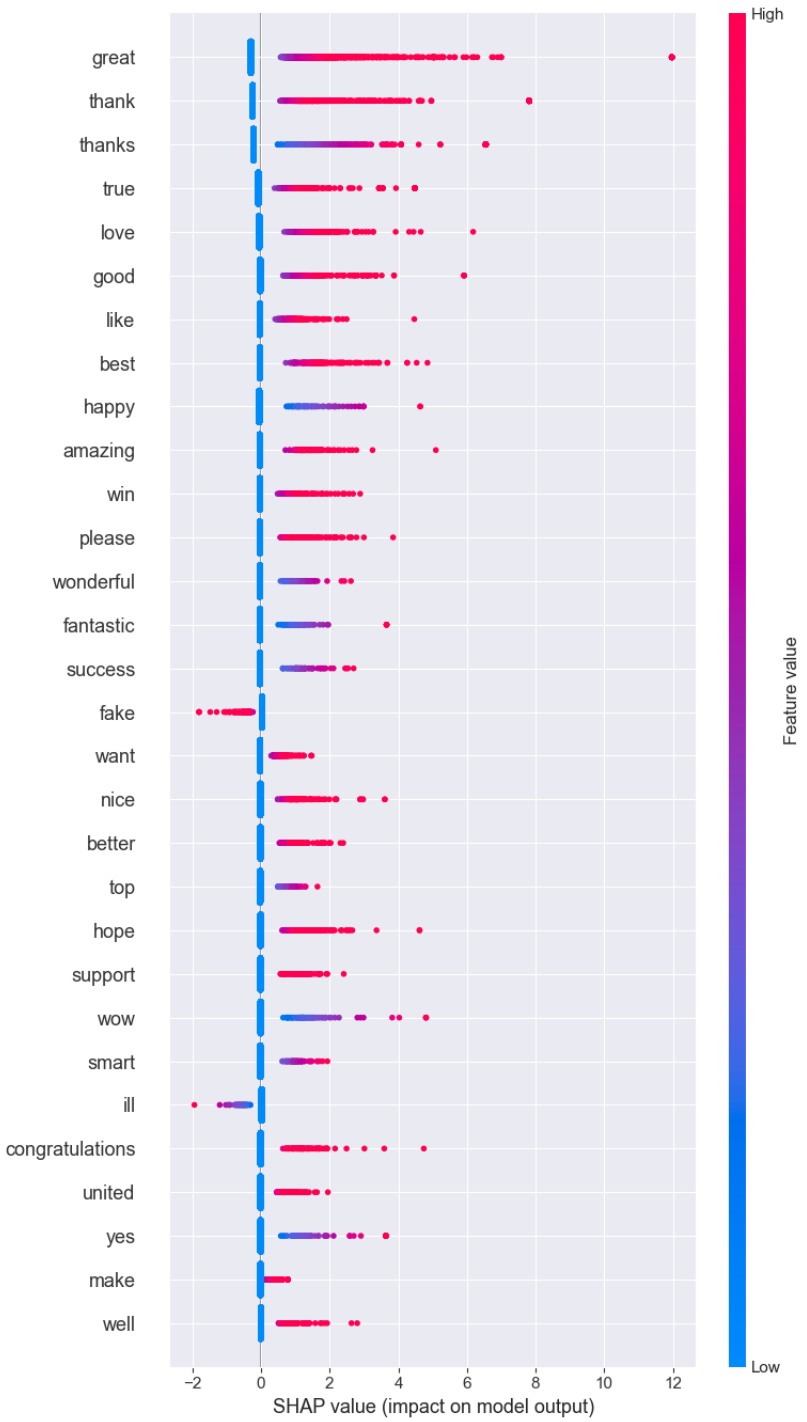
## 4.3 Klasa: Pozytywna

### 4.3.1 BERT




Rysunek 8: SHAP – wpływ słów na klasyfikację jako pozytywną

### 4.3.2 VADER




Rysunek 9: SHAP – wpływ słów na klasyfikację jako pozytywną

# 5 Wnioski

Zastosowanie analizy SHAP pozwoliło zrozumieć, które konkretnie słowa miały największy wpływ na ocenę klasyfikatora. Analizator ujawnił, że negatywne lub pozytywne nacechowanie wyrazów, zastosowane negacje oraz niektóre słowa kluczowe miały znaczący wpływ przy rozpoznaniu sentymentu. Przejrzystość decyzji modelu jest bardzo oczekiwana, dzięki temu taki system oceny i analizy jest praktyczny i łatwiejszy w ocenie.

Warto podkreślić, że model oparty na VADERze bazuje głównie na pojedynczych słowach o jednoznacznym wydźwięku emocjonalnym, takich jak *great* czy *amazing*. Z kolei model wykorzystujący BERTa potrafi uwzględniać pełny kontekst, w jakim użyte są dane wyrazy, co pozwala mu poprawnie interpretować bardziej złożone sformułowania i odniesienia polityczne, w których pojawiają się terminy takie jak *Obama* czy *Democrats*, oraz wnioskować z nich właściwy sens całego zdania.

# 6 Link do repozytorium

`https://github.com/maciekgangus/sentiment-analysis-shap`.