Metody Inteligencji obliczeniowej

Zastosowanie analizy SHAP w analizie sentymentu metodami NLP

Michał Orlewski, Kamil Kaproń

Informatyka Stosowana

Wydział Fizyki i Informatyki Stosowanej

Akademia Górniczo-Hutnicza w Krakowie

1. Opis Projektu

Celem projektu było przeanalizowanie tweetów Donald'a Trump'a, zebranych na przestrzeni lat oraz sprawdzenie sentymentu jaki niosą ze sobą wpisy, a następnie wykonanie analizy SHAP użytego modelu. Wszystkie potrzebne dane znajdowały się na stronie:

https://www.kaggle.com/datasets/austinreese/trump-tweets

Kod projektu dostępny jest w repozytorium: https://github.com/m-orlewski/nlp-sentiment-analysis

2. Realizacja Projektu

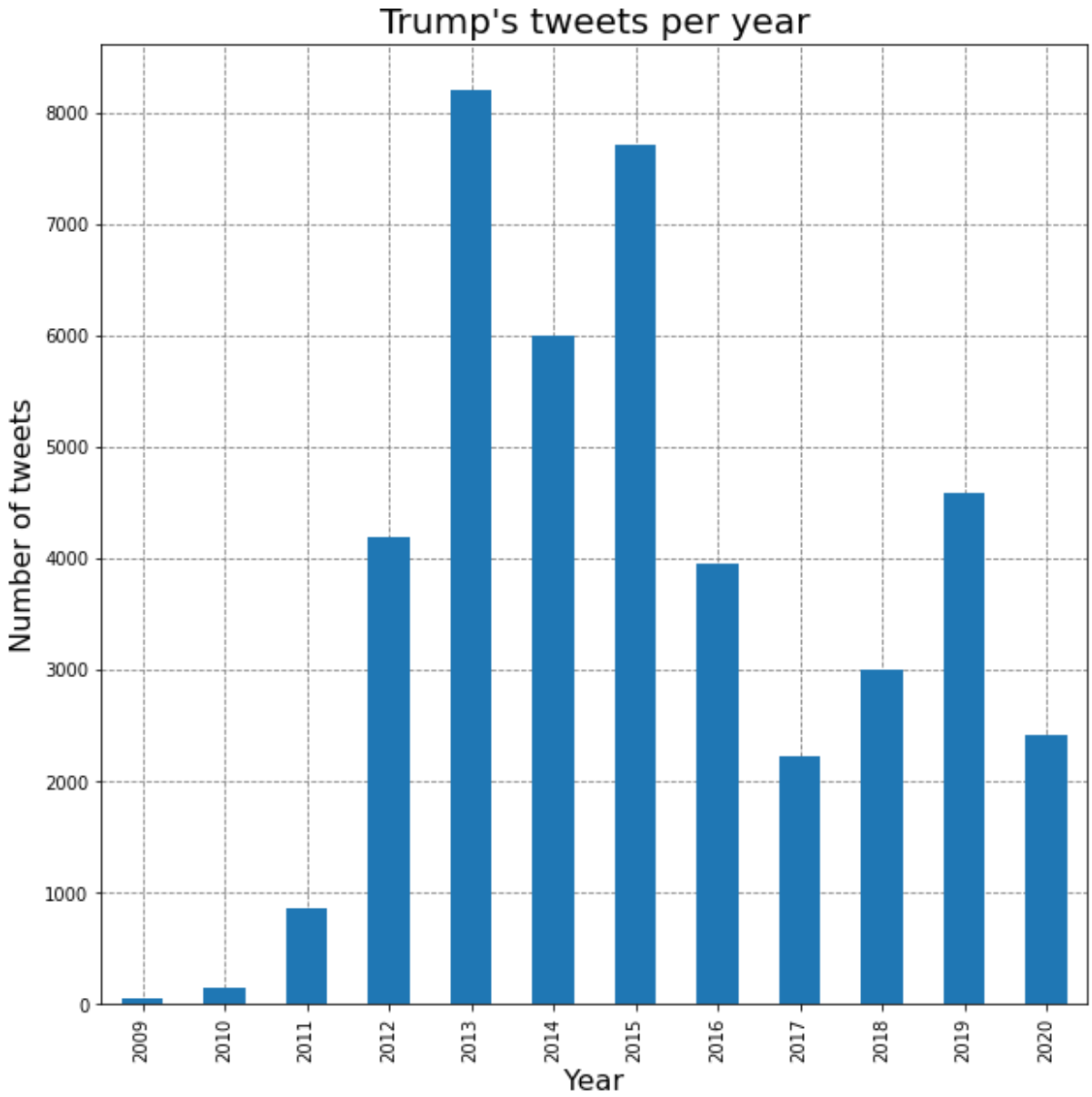
Projekt w całości został wykonany w języku Python przy użyciu Jupyter Notebook z wykorzystaniem następujących bibliotek:

- *NumPy* - Biblioteka NumPy, w Python, została stworzona, aby umożliwić szybkie i sprawne operacje na macierzach.
- *Pandas* - Pandas, jest jednym z najbardziej rozbudowanych pakietów, do analizy danych, w Pythonie
- *Matplotlib* - biblioteka do tworzenia wykresów dla języka programowania Python i jego rozszerzenia numerycznego NumPy.
- *PyTorch-Transformers* – biblioteka najnowocześniejszych, wstępnie przeszkolonych modeli przetwarzania języka naturalnego NLP.
- *SHAP* – biblioteka Python, która używa wartości Shapley do wyjaśnienia danych wyjściowych dowolnego modelu uczenia maszynowego.

Pierwszym krokiem było pobranie danych niezbędnych do wykonania analizy. Po pobraniu i wprowadzeniu do programu tweety zostały przygotowane do pracy. Odczytane informacje wyświetlono na wykresach oraz tabelach. Odpowiednie przygotowanie danych było kluczowe ze względu na ilość tweetów, która wynosiła 43352.

	id	content	date
0	1698308935	Be sure to tune in and watch Donald Trump on L...	2009-05-04 13:54:25
1	1701461182	Donald Trump will be appearing on The View tom...	2009-05-04 20:00:10
2	1737479987	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 08:38:08

Tabela 1: Pierwsza seria danych przed zmianami.



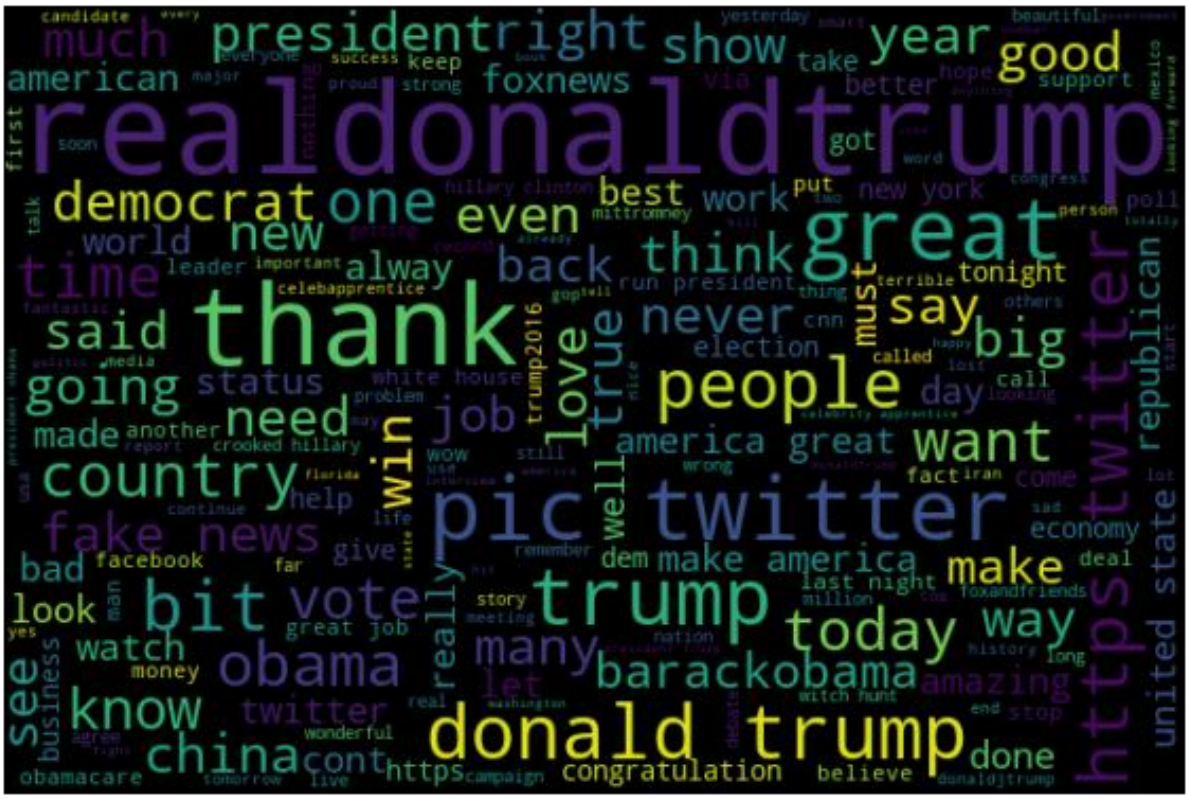
Wykres 1: Liczba tweetów Trump'a na przestrzeni lat.

W ramach analizy należy odpowiednio przygotować treść tweetów. Dane zawierały nie koniecznie interesujące od strony analizy sentymentu słowa. Były to między innymi odnośniki do innych stron, pseudonimy użytkowników czy słowa, które nie wpływają na emocjonalność wpisów. Przy użyciu biblioteki nltk dokonano tokenizacji tweetów. Za ich pomocą wykluczono zbędne nie niosące sentymentu słowa oraz wykluczone zostały wyrazy, zbyt krótkie, aby miały znaczenie (a, an, the).

	id	content	date	text_token	text_string
0	1698308935	Be sure to tune in and watch Donald Trump on L...	2009-05-04 13:54:25	[sure, tune, watch, donald, trump, late, night...	sure tune watch donald trump late night david ...

Tabela 2: Pierwszy wpis po odpowiednim przygotowaniu.

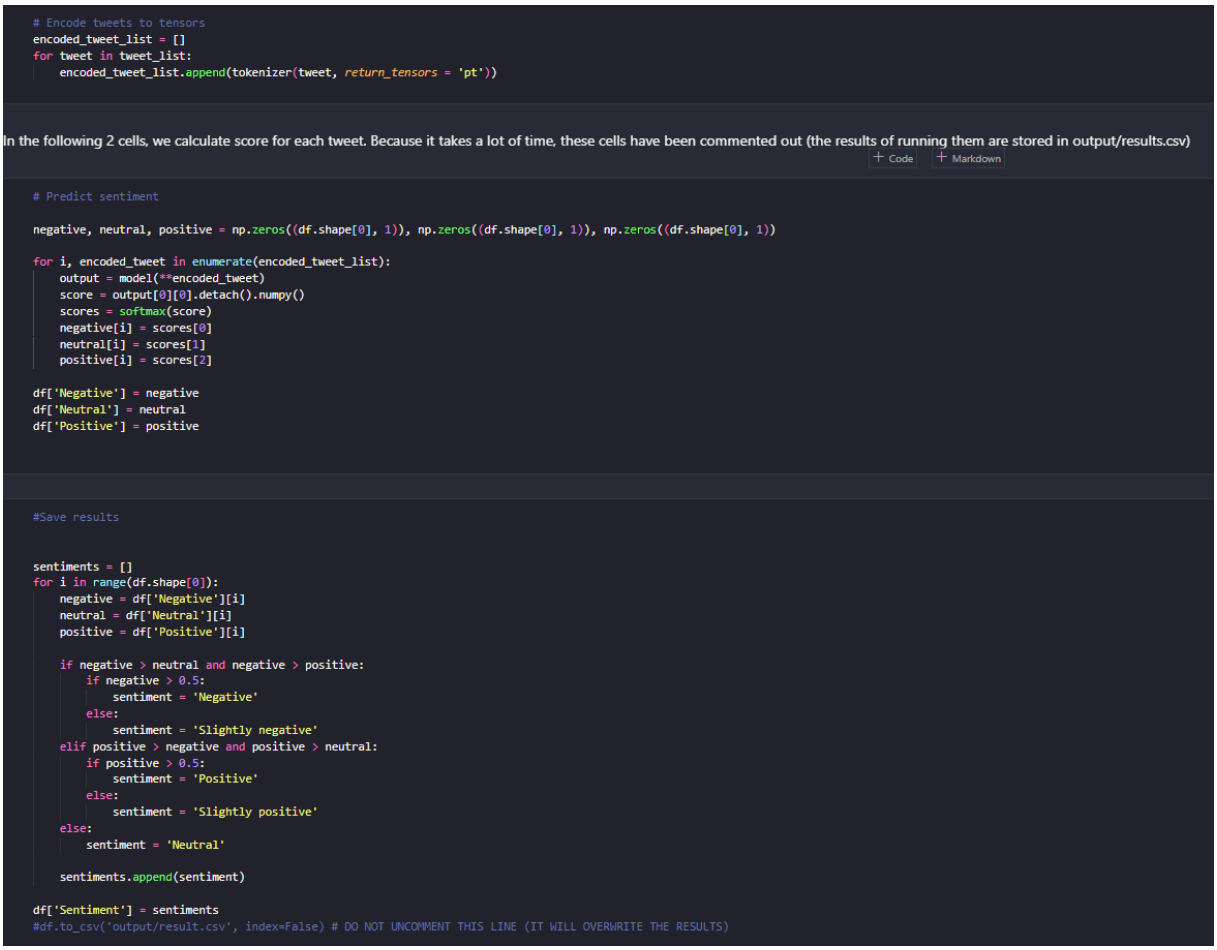
W ramach przedstawiania danych zespół przygotował również WordCloud (chmurę słów). Przedstawia ona słowa, których rozmiar na obrazie zależy od częstotliwości ich występowania w zbiorze danych.



Rysunek 1: Word Cloud.

Po przygotowaniu danych można było wykonać odpowiednią analizę. W tym celu zespół skorzystał z gotowego modelu, przygotowanego specjalnie pod analizę tweetów. RoBERTa to wydany przez Google w 2018 roku model, specjalnie wytrenowany na ~58 milionach tweetów i zoptymalizowany pod analizę sentymentu w języku angielskim. Model dostępny jest do pobrania: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment.

Przygotowane dane zostały przepuszczone przez model, a rezultaty zapisane do pliku.



```
# Encode tweets to tensors
encoded_tweet_list = []
for tweet in tweet_list:
    encoded_tweet_list.append(tokenizer(tweet, return_tensors = 'pt'))
```

In the following 2 cells, we calculate score for each tweet. Because it takes a lot of time, these cells have been commented out (the results of running them are stored in output/results.csv)

+ Code + Markdown

```
# Predict sentiment

negative, neutral, positive = np.zeros((df.shape[0], 1)), np.zeros((df.shape[0], 1)), np.zeros((df.shape[0], 1))

for i, encoded_tweet in enumerate(encoded_tweet_list):
    output = model(**encoded_tweet)
    score = output[0][0].detach().numpy()
    scores = softmax(score)
    negative[i] = scores[0]
    neutral[i] = scores[1]
    positive[i] = scores[2]

df['Negative'] = negative
df['Neutral'] = neutral
df['Positive'] = positive
```

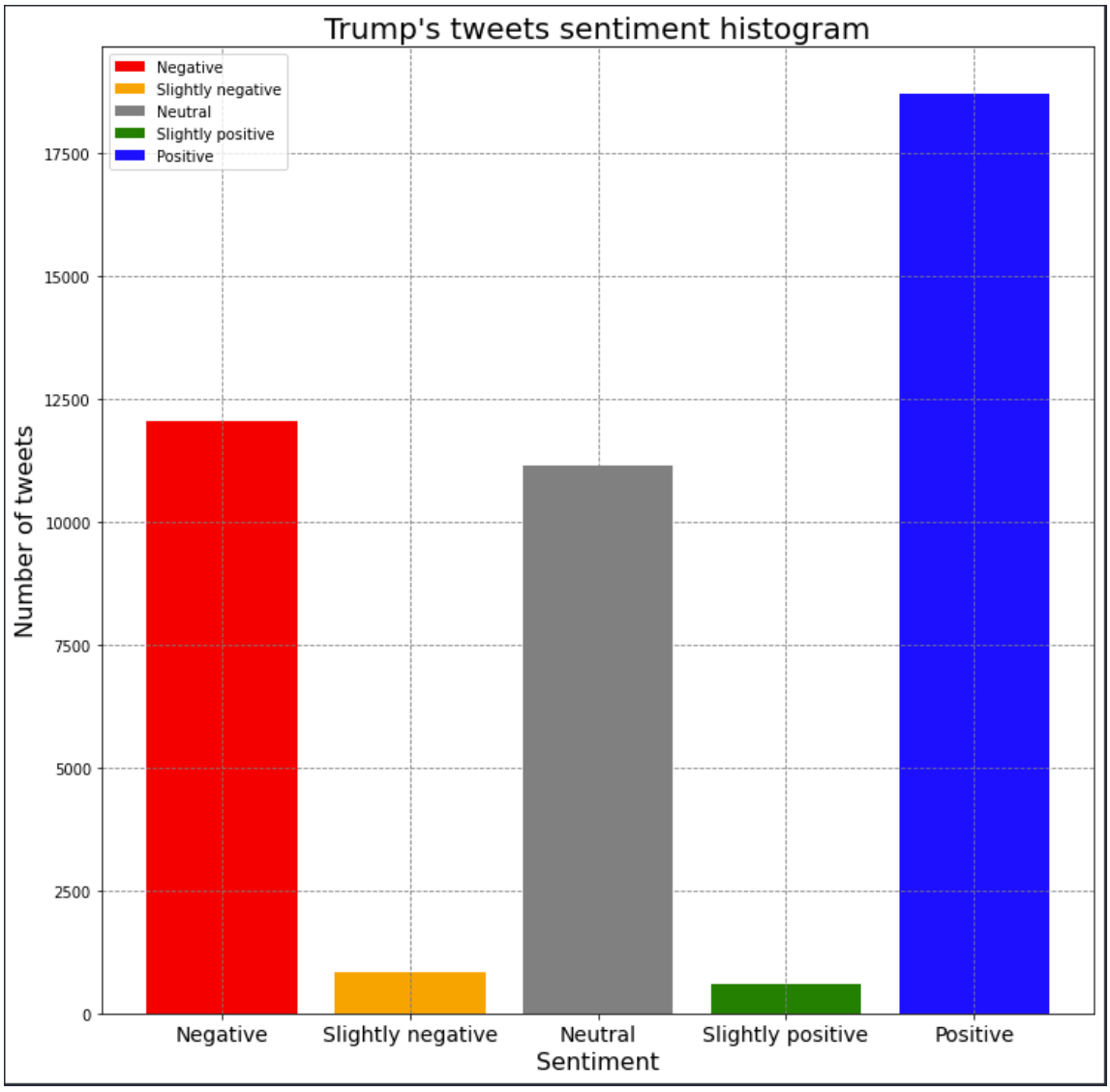
```
#Save results


sentiments = []
for i in range(df.shape[0]):
    negative = df['Negative'][i]
    neutral = df['Neutral'][i]
    positive = df['Positive'][i]

    if negative > neutral and negative > positive:
        if negative > 0.5:
            sentiment = 'Negative'
        else:
            sentiment = 'Slightly negative'
    elif positive > negative and positive > neutral:
        if positive > 0.5:
            sentiment = 'Positive'
        else:
            sentiment = 'Slightly positive'
    else:
        sentiment = 'Neutral'

    sentiments.append(sentiment)

df['Sentiment'] = sentiments
#df.to_csv('output/result.csv', index=False) # DO NOT UNCOMMENT THIS LINE (IT WILL OVERWRITE THE RESULTS)
```

Rysunek 2: Fragment kodu wyliczającego sentyment tweetów

Otrzymane wyniki reprezentowały przynależność każdego tweeta do każdej z trzech kategorii sentymentu: negative, neutral i positive (wartości z zakresu 0-1). W celu uzyskania dokładniejszych wyników, dodano 2 dodatkowe kategorie – slightly negative oraz slightly positive, a następnie na podstawie otrzymanych wyników przydzielono jedną z pięciu kategorii do każdego tweeta.

Positive	18728
Negative	12047
Neutral	11146
Slightly negative	841
Slightly positive	590

Tabela 3: Ilość tweetów o każdym sentymencie



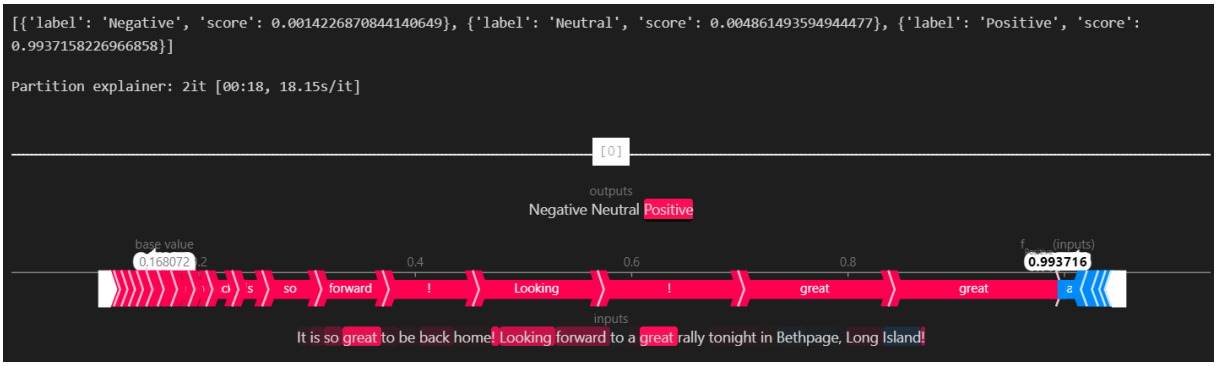
Wykres 2: Histogram tweetów z podziałem na sentymenty

W ramach dalszej części projektu zespół wykonał analizę SHAP. Pozwala ona na zobaczenie wpływu funkcji (lub w tym przypadku zmiennej) na ostateczny wynik poprzez porównanie względnego wpływu danych wejściowych ze średnią.

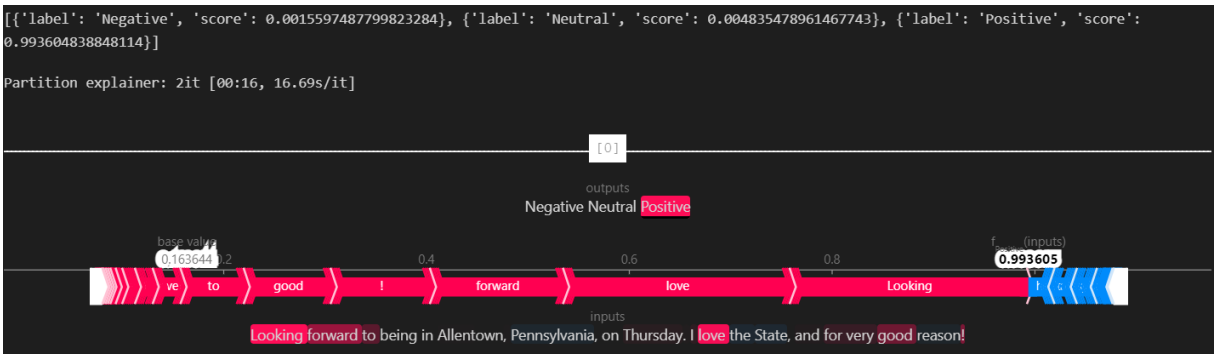
```
def score_and_visualize(text):
  prediction = pipe([text])
  prediction[0][0]['label'] = 'Negative'
  prediction[0][1]['label'] = 'Neutral'
  prediction[0][2]['label'] = 'Positive'
  print(prediction[0])

  explainer = shap.Explainer(pipe, output_names=['Negative', 'Neutral', 'Positive'])
  shap_values = explainer([text])

  shap.plots.text(shap_values)
```

Rysunek 3: Fragment kodu dokonującego analizy SHAP

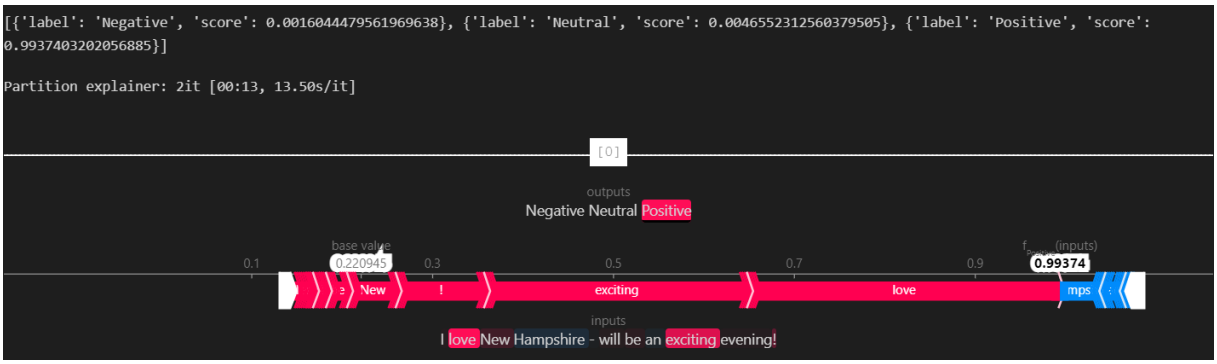
Poniżej pokazano wyniki analizy SHAP dla poszczególnie wybranych grup tweetów.



Rysunek 4: Najbardziej pozytywny tweet.



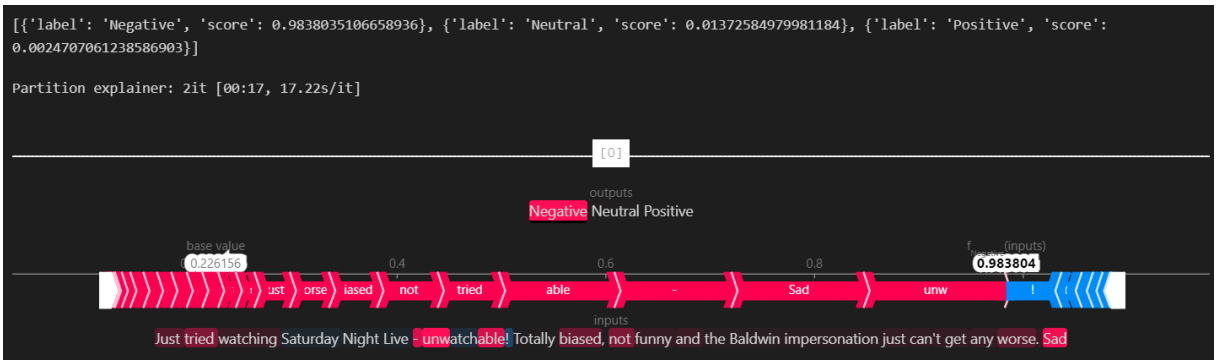
Rysunek 5: 2 Najbardziej pozytywny tweet.



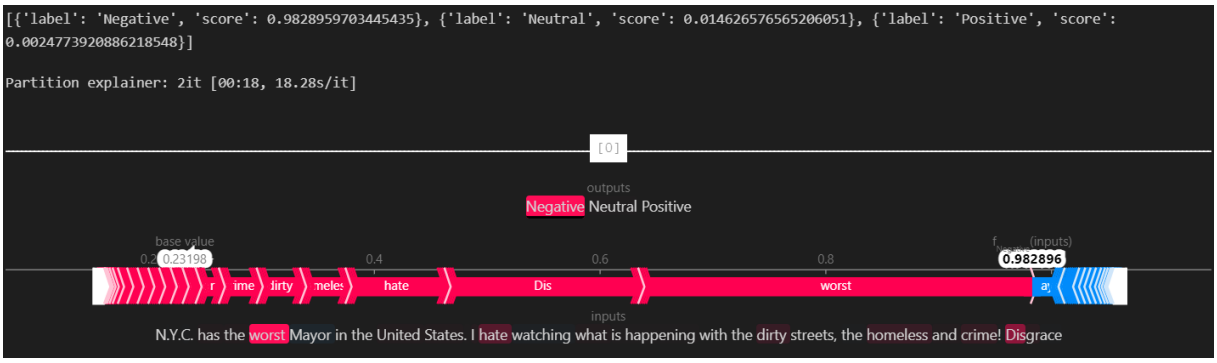
Rysunek 6: 3 Najbardziej pozytywny tweet.

Analiza wykazała, iż najbardziej wpływowymi słowami były love oraz great, więc analiza jest poprawna. Warto zwrócić uwagę, że znaki interpunkcyjne również wpływają na wydźwięk wpisów.

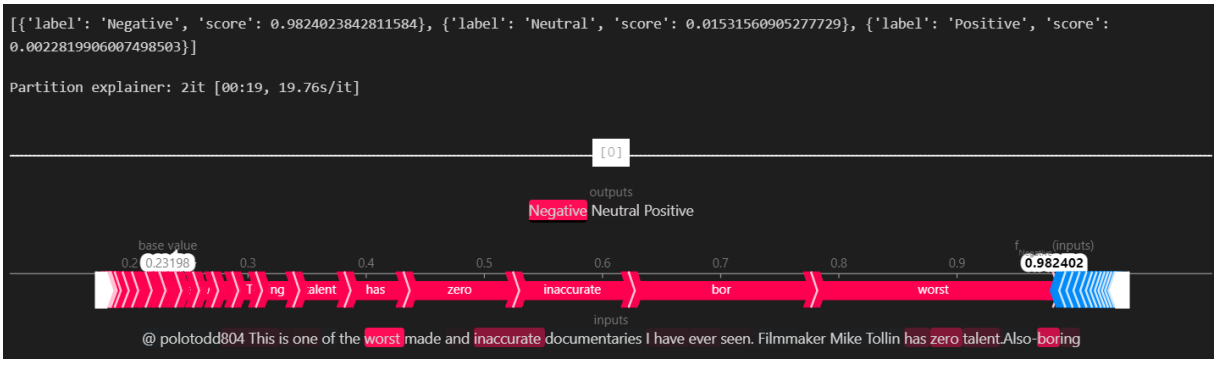
Poniżej przedstawiono 3 najbardziej negatywne wpisy.



Rysunek 7: Najbardziej negatywny tweet.



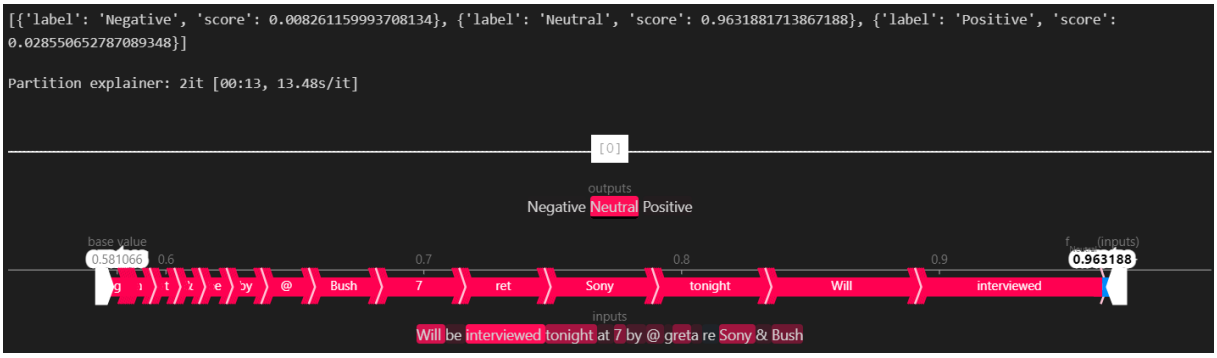
Rysunek 8: 2 najbardziej negatywny tweet.



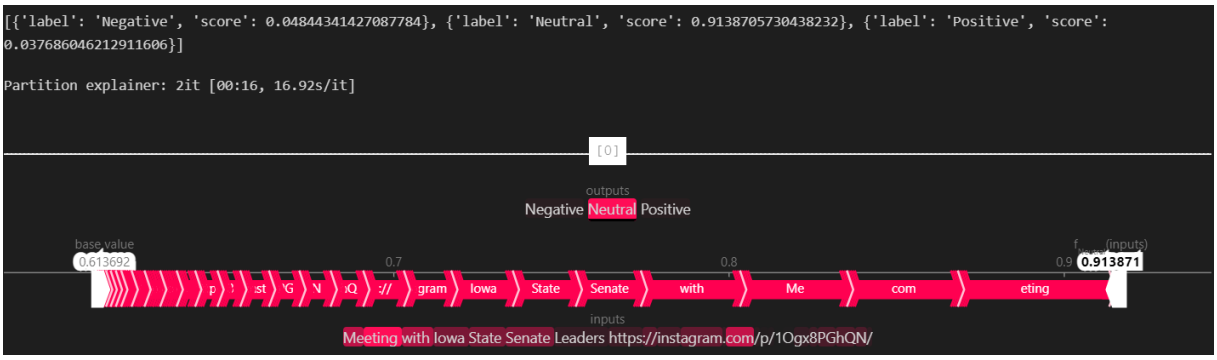
Rysunek 9: 3 najbardziej negatywny tweet.

Jak widać z otrzymanych danych w największym stopniu przyczyniły się do tego takie słowa jak worst, sad czy boring. Co potwierdza poprawność wykonanej analizy. Warto zwrócić uwagę, że analiza nie została poprzez analizę samych słów ale również ich składu. Słowo unwatchable zostało podzielone i wyciągnięte z niego un oraz able.

Poniżej przedstawiono 3 najbardziej neutralne wpisy



Rysunek 10: Najbardziej neutralny tweet.

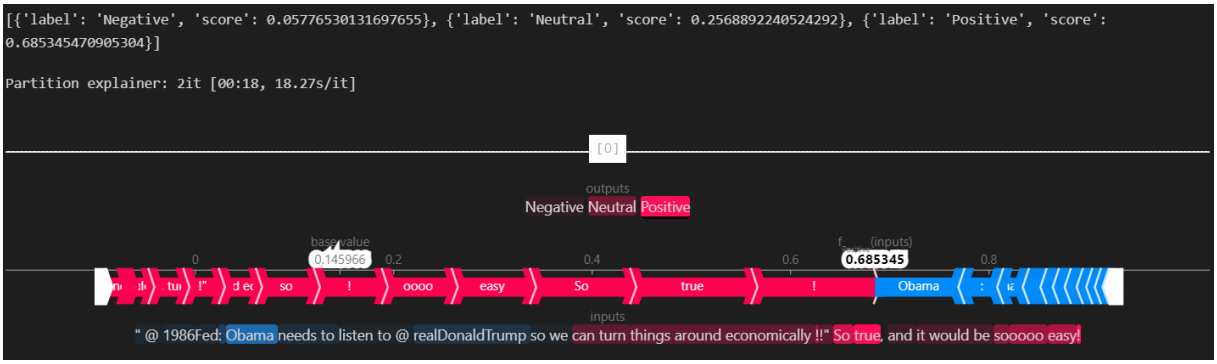


Rysunek 11: 2 najbardziej neutralny tweet.

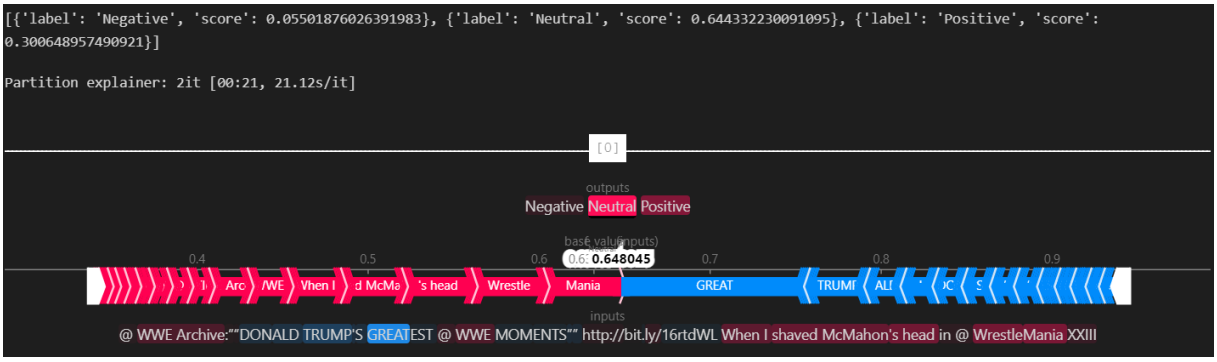
Rysunek 12: 3 najbardziej neutralny tweet

Z otrzymanych danych rzuca się w oczy, że najbardziej neutralne wpisy to te, które są po prostu informacjami o wydarzeniu. Słowa które mają największy wpływ to 18th, Meeting, will, czy interviewed. W tego typu zdaniach interesują nas najbardziej miejsce, kiedy oraz co się będzie działo lub co się działo.

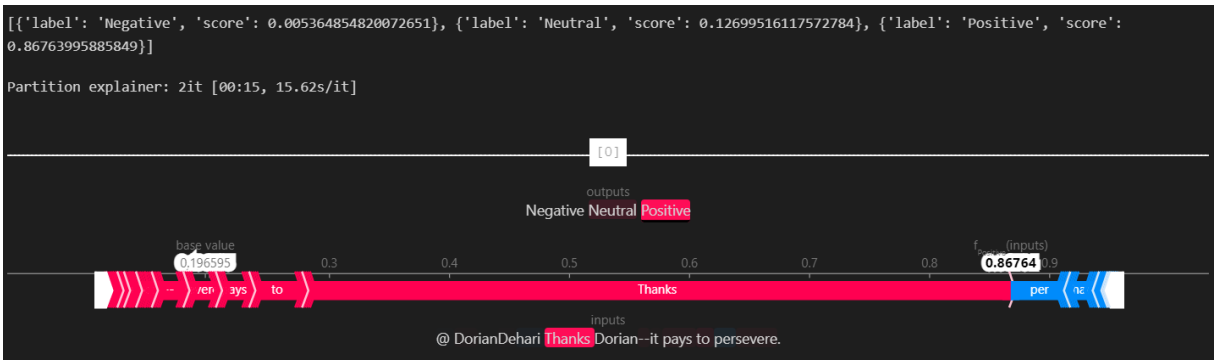
Dalej przedstawione są losowo wybrane tweety spośród całej bazy.



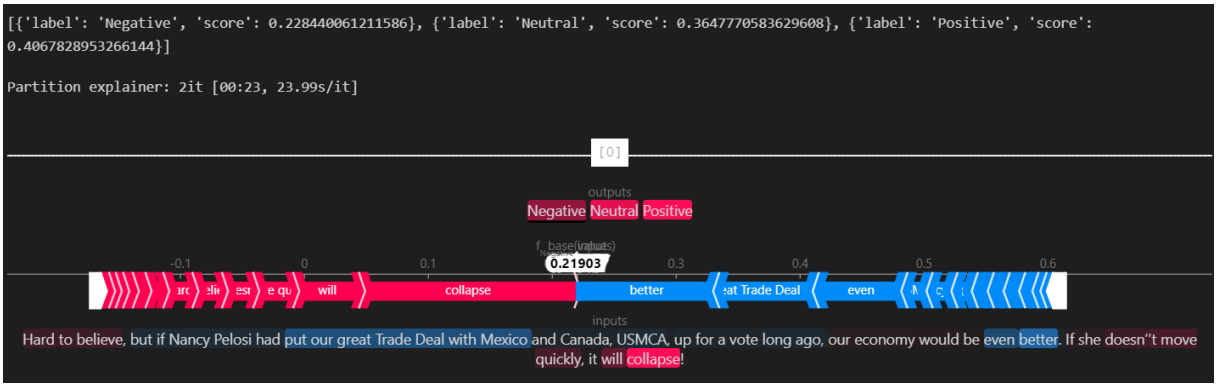
Rysunek 13: losowo wybrany tweet.



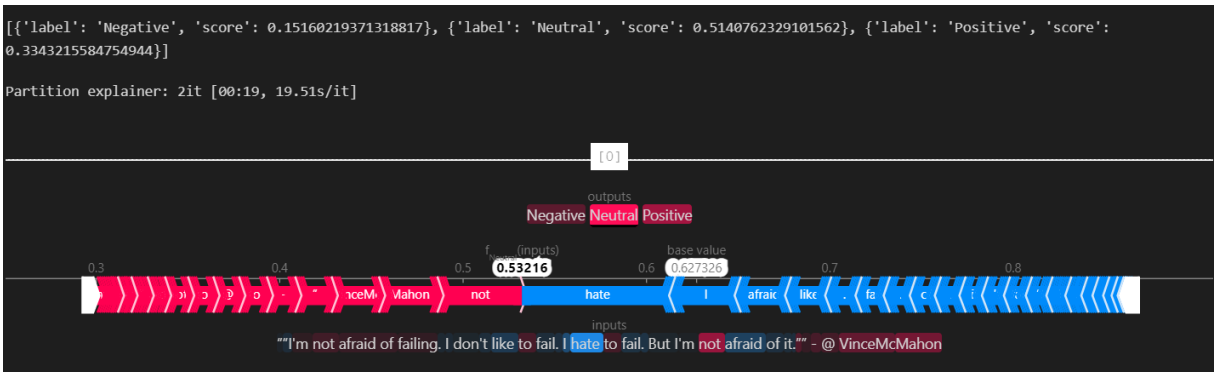
Rysunek 14: losowo wybrany tweet.



Rysunek 15: losowo wybrany tweet.

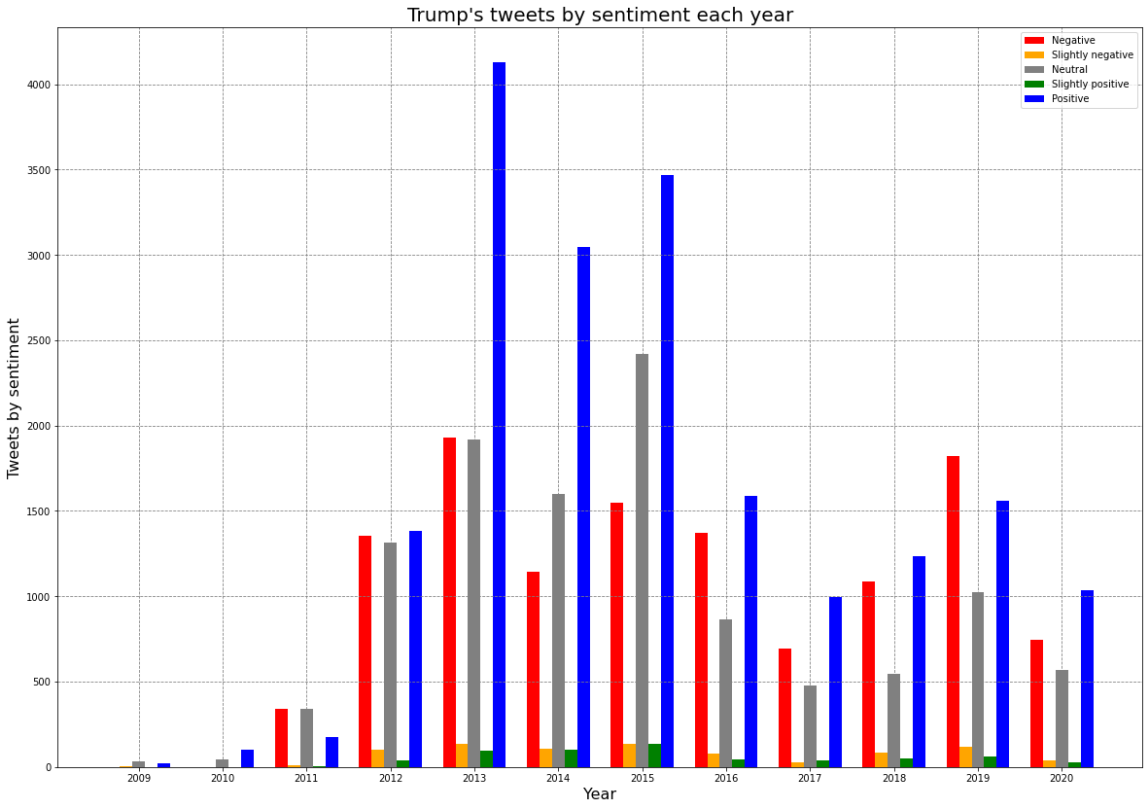


Rysunek 16: losowo wybrany tweet.



Rysunek 17: losowo wybrany tweet.

Na koniec dokonano analizy otrzymanych wyników – na wykresie przedstawiono ilość tweetów w każdej kategorii, w zależności od roku w jakim zostały opublikowane. Możemy zauważyć zwiększoną aktywność w latach 2013-2016, co może być związane z kandydaturą Donalda Trumpa na prezydenta Stanów Zjednoczonych w roku 2016. Z dużej ilości pozytywnych tweetów w tym okresie możemy wywnioskować, że Twitter był istotną platformą do zyskiwania przez Donalda Trumpa poparcia w trakcie kampanii wyborczej.



Wykres 3: Histogram tweetów o każdym sentymencie z podziałem na daty publikacji