Predykcja zainteresowania postami w social mediach z użyciem metod NLP

Metody Inteligencji Obliczeniowej 2023

Jakub Wójcik
Mateusz Cichostępski
Jakub Urbański

IS
WFiIS
AGH
21.06.2023 r.

1 Opis

Celem projektu jest wykorzystywanie sieci neuronowej typu NLP do budowy modelu predykcyjnego służącego do analizy zainteresowania (w formie retweetów) tweetów zebranych w zbiorze *tweetów Donalda Trumpa* stosując reprezentację występujących w tekście słów, a następnie - dla zbudowanego modelu - przeprowadzenie analizy SHAP.

1.1 Github

Repozytorium na githubie znajduje się pod linkiem:
`https://github.com/Matiixx/MIO-Project`.

2 Wstęp teoretyczny

W ostatnich latach przetwarzanie języka naturalnego, czyli NLP (ang. Natural Language Processing), zyskało na znaczeniu jako ważne narzędzie do analizy i predykcji różnych aspektów danych tesktowych, m.in. predykcji zainteresowania, zaangażowania oraz popularności postów i wiadomości umieszczanych w social mediach.

2.1 Model predykcyjny

Do skonstruowania modelu predykcyjnego najpierw w ramach preprocessingu obsługiwane są brakujące wartości i następuje tokenizacja przetworzonego tekstu na słowa (lub ich fragmenty), które stanowią podstawowe jednostki naszej analizy. Kolejnym krokiem było trenowanie modelu regresji (u nas XGBoost) przy użyciu tweetów zreprezentowanych wektorami TF-IDF. Model jest sprawdzany z użyciem metryki błędu średniokwadratowego oraz metryki błędu średniowymodulowego zarówno na danych treningowych, jak i testowych. Zapisywane do zewnętrznych plików są: wytrenowany model oraz używany do wytrenowania *vectorizer* TF-IDF.

2.2 Analiza SHAP

W celu uzyskania wglądu w znaczene i udział różnych słów w przewidywaniu liczby retweetów kolejnym krokiem po wytrenowaniu naszego mod-

elu była analiza SHAP. Jej wartości zapewniają ujednoliconą miarę priorytetów funkcji opartą na zasadach teorii gier. Określają ilościowo wpływ każdego słowa na przewidywany wynik i zapewniają zrozumiałe wyjaśnienie poszczególnych prognoz. Natępuje więc wczytanie wytrenowanego modelu oraz *vectorizer'a*, a także odczytanie zredukowanego zboru danych. W kolejnym kroku dane są przygotowane do analizy poprzez transformację danych testowych za pomocą wspomnianego już *vectorizer'a*, a następnie w celu uzyskania wartości SHAP zastosowanie ma eksplainer SHAP. Na koniec generowane są wizualizacje, takie jak wykres podsumowujący, wykres kaskadowy i wykres rozmieszczenia za pomocą wartości SHAP i nazw cech.

3 Dane wejściowe

W projekcie wykorzystano dane `Trump Tweets` z portalu *Kaggle*.

3.1 Opis danych

Zbiór danych zebrany jest w pliku: `trumptweets.csv`. Plik ten zawiera szereg informacji dla każdego tweet'a, w tym pola takie jak `content` (treść tweet-a), `favorites` (liczba ulubionych), `retweets` (liczba retweetów) oraz `date` (data utworzenia tweet-a).

3.2 Przygotowanie danych

Do dobrej analizy danych konieczne jest przeprowadzenie redukcji zbioru danych do konkretnych kolumn - `content` (treść tweet-a) oraz `retweets` (liczba retweetów) - istotnych dla naszej analizy zainteresowania postami.

3.3 Znaczenie danych i rozmiar zbioru

Kluczowe dla predykcji zachowania są dane związane z treścią tweet'ów, ponieważ zawierają one informacje wpływające na reakcje innych użytkowników. Redukcja zbioru danych do istotnych kolumn pozwoliła na skrócenie czasu przetwarzania i optymalizację analiz.

4 Technologie wykorzystane w projekcie

W całym projekcie zastosowane zostały technologie oparte o język `Python`, przy użyciu narzędzia *Jupyter Notebook*, które to umożliwia interaktywną analizę danych i tworzenie modeli.
Wykorzystane biblioteki:

- `pickle` - wykorzystana do zapisu i odczytu modelu predykcyjnego oraz narzędzi przetwarzających dane. Umożliwia serializację obiektów Pythona i przechowywanie ich w plikach,
- `shap` - użyta do analizy SHAP. Poprzez pomoc w zrozumieniu znaczenia cech wykorzystanych przez model pozwoliła na identyfikację kluczowych czynników wpływających na zainteresowanie postami w social mediach,
- `pandas` - posłużyła do wczytywania i przetwarzania danych w celu przygotowania ich do modelowania,
- `xgboost` - wykorzystany został algorytm *gradient boosting*, który ma swoje zastosowanie w problemach regresji i klasyfikacji,
- `sklearn` - wykorzystana została do podziału zbioru na dane treningowe i dane testowe, przekształcenia tekstu tweetów na wektory cech, oceny wydajności modelu czy przemieszania danych (w celu uniknięcia wpływu ich ułożenia na wynik).

5 Wyniki

```
$ python main.py create_model
Mean Squared Error: 3312260.1300347582
Mean Absolute Error: 1199.6926859184236
Mean Squared Error: 4972288.110227588
Mean Absolute Error: 1447.7751927596964
```

Figure 1: Ewaluacja modelu

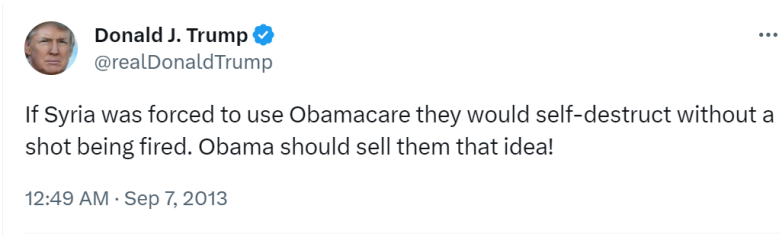

Figure 2: Losowo wybrany tweet dla którego przetestowany został model. Ilość retweetów wg naszych danych: **1045**.

```
$ python main.py predict_popularity
1076.1805
```

Figure 3: Przewidziana przez model liczba retweetów dla powyższego tweeta

Wykres beeswarm SHAP przedstawia rozkład wartości cech w zbiorze danych. Na osi Y znadują się pogrupowane cechy, a na osi X wartości SHAP. Im wyżej na osi Y znajduje się dana cecha, tym bardziej wpływa na predykcję modelu. Dla każdej grupy kolor punktu oznacza wartość danej cechy, a miejsce na osi X wartość do dodania do predykcji.

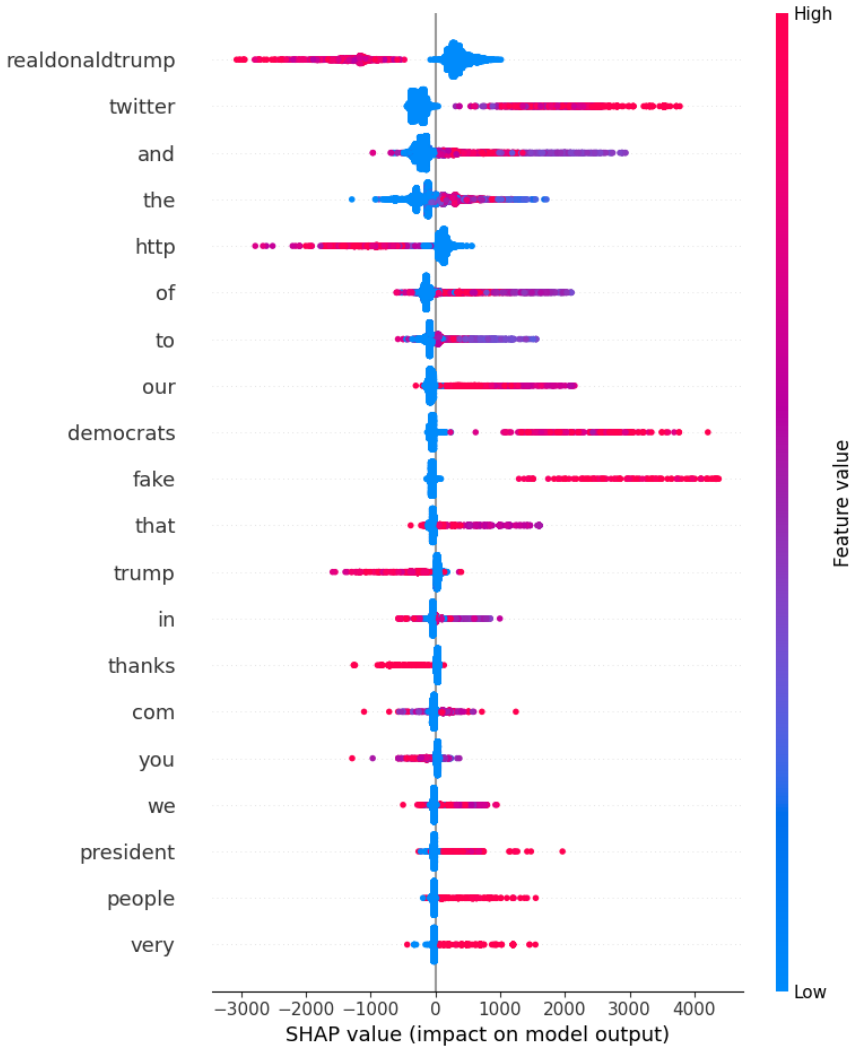

Figure 4: Wykres beeswarm SHAP przedstawiający wpływ słów na ilośc retweetów

Wykres waterfall SHAP przedstawia wpływ poszczególnych cech danego przykładu na predykcję modelu. Zaczynamy od wartości referencyjnej, do której od dołu dodajemy lub odejmujemy wartości dla każdej z cech (słów) do końcowej predykcji.

2586	twitter
2023	realdonaldtrump
2451	the
170	and
2542	trump
515	com
1198	http
1737	of
684	democrats
2498	to
1786	our

Table 1: Tabela wyjaśniająca, jakie 'Feature' odpowiada jakiemu słowu

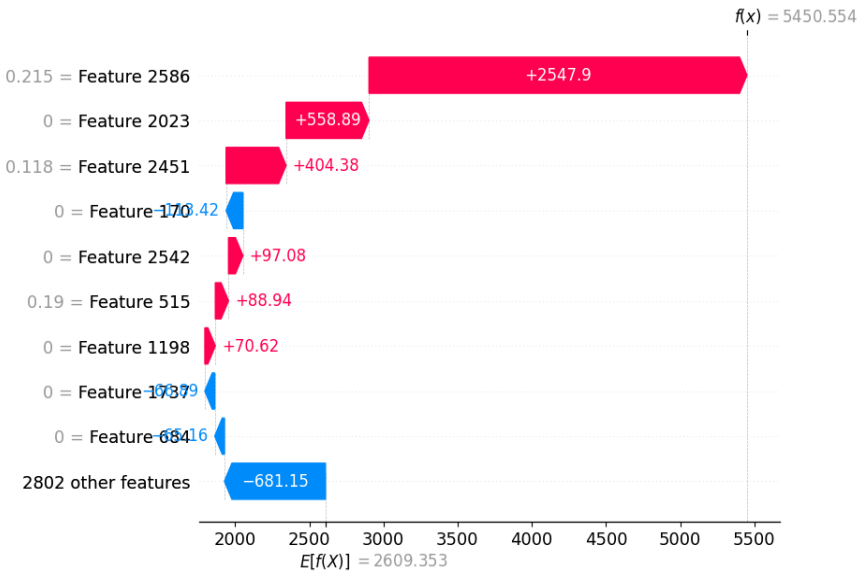

Figure 5: Wykres waterfall SHAP przedstawiający w jaki sposób słowa wpłynęły na przewidzenie liczby retweetów względem przeciętnego przewidywania

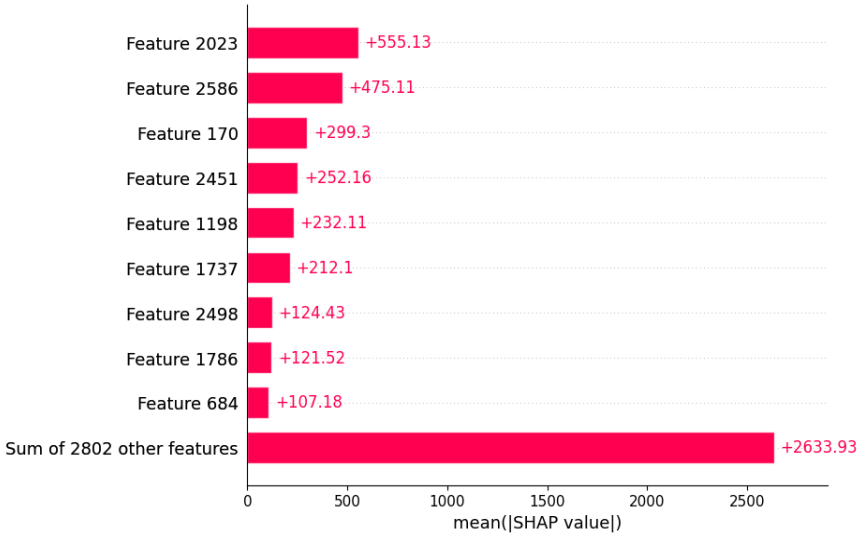

Figure 6: Wykres bar przedstawiający wpływ słów na predykcję modelu oraz ich średni wpływ na predykcję.

6 Podsumowanie

W ramach projektu osiągnęliśmy zamierzony cel, a zbudowany model predykcyjny okazał się całkiem skuteczny w przewidywaniu popularności tweetów na podstawie ich treści.
Dodatkowo, dzięki analizie SHAP, zidentyfikowane zostały istotne czynniki wpływające na zainteresowanie postami w social mediach.
Był to udany przykład wykorzystania technik NLP i sieci neuronowych w celu analizy zainteresowania tweetami.

6.1 Podział pracy

W projekcie współpracowaliśmy w miarę równomiernie nad każdym aspektem, chociaż najwięcej pracy w zaprojektowanie modelu i przeprowadzenie analizy włożył **Mateusz Cichostępski**, a w napisanie dokumentacji - **Jakub Wójcik** oraz **Jakub Urbański**.

7 Bibliografia

- towardsdatascience.com/introduction-to-shap-with-python
- towardsdatascience.com/using-shap-values-to-explain-how...
- link.springer.com/article/10.1007
- towardsdatascience.com/how-to-build-a-neural-network...
- arxiv.org/pdf/1807.10854.pdf