Raport 4 - MSI

Sieci rekurencyjne

Tomasz Sewastynowicz

Uniwersytet im. Adama Mickiewicza w Poznaniu

10.12.2023 r.

1. Cel pracy

Raport skupia się na zastosowaniu sieci rekurencyjnych w celu identyfikowania sentymentu wiadomości tekstowych na platformie Twitter, obecnie noszącej nazwę X. Zbiór zawiera ponad 162 000 tweetów wraz z etykietą sentymentu, jakim danym twitt się wyróżnia. Sentyment tekstu posiada 3 zmienne: negatywny, neutralny i pozytywny.

W ramach pracy został wykorzystany zbiór danych "Twitter Sentiment Dataset", z platformy Kaggle *(https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset/data)*. Zbioru tego użyto do trenowania i testowania sieci rekurencyjnych i stworzenia modeli, który będą jak najpoprawniej klasyfikowały sentyment jakim dany twitt się charakteryzuje.

2. Zbiór danych

Zbiór ten zawiera łącznie 162 979 krótkich wiadomości tekstowych pochodzących z platformy Twitter i 3 zmienne sentymentu. Tabela 1. przedstawia 5 przykładowych twittów wraz z odpowiadającym im sentymentem.

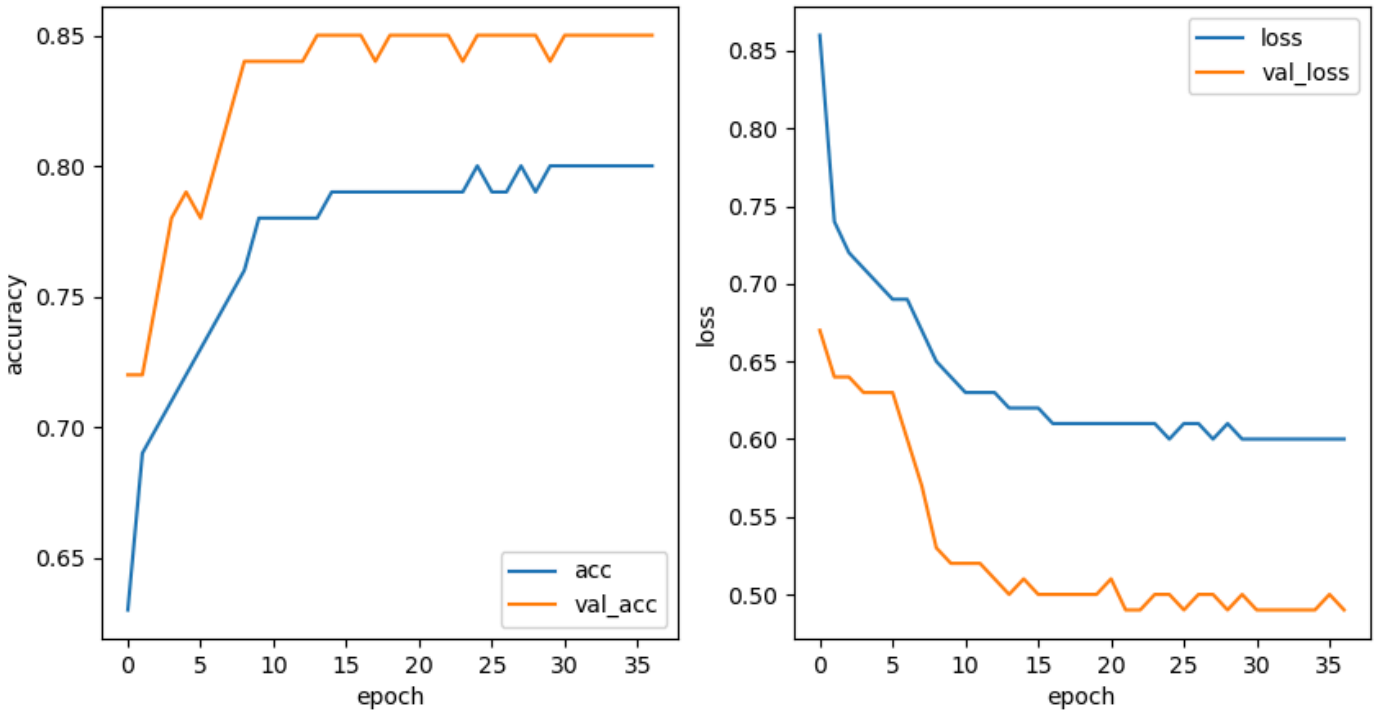
Tabela 1. Wiadomości tekstowe wraz z odpowiadającym im sentymentem.

Treść tweeta	Sentyment
talk all the nonsense and continue all the drama will vote for modi	neutralny
what did just say vote for modi welcome bjp told you rahul the main campaigner for modi think modi should just relax	pozytywny
how such people are being made amazedn fear that frustation him may not result vote against sir waste ministerdisgrace entire modi cabinet	negatywny
this the new india modi trying build with these leaders his party why have live with these deplorable characters	negatywny
before 2014 hindustan has seen the worst for hindus own maj hindu rashtra who thrashed the rascal faces these anti indian politiciansantinationals urban naxals wait watch after modis win pakistan mein bhi hindu hona garv baat hogi ✌	pozytywny

3. Modele sieci rekurencyjnych

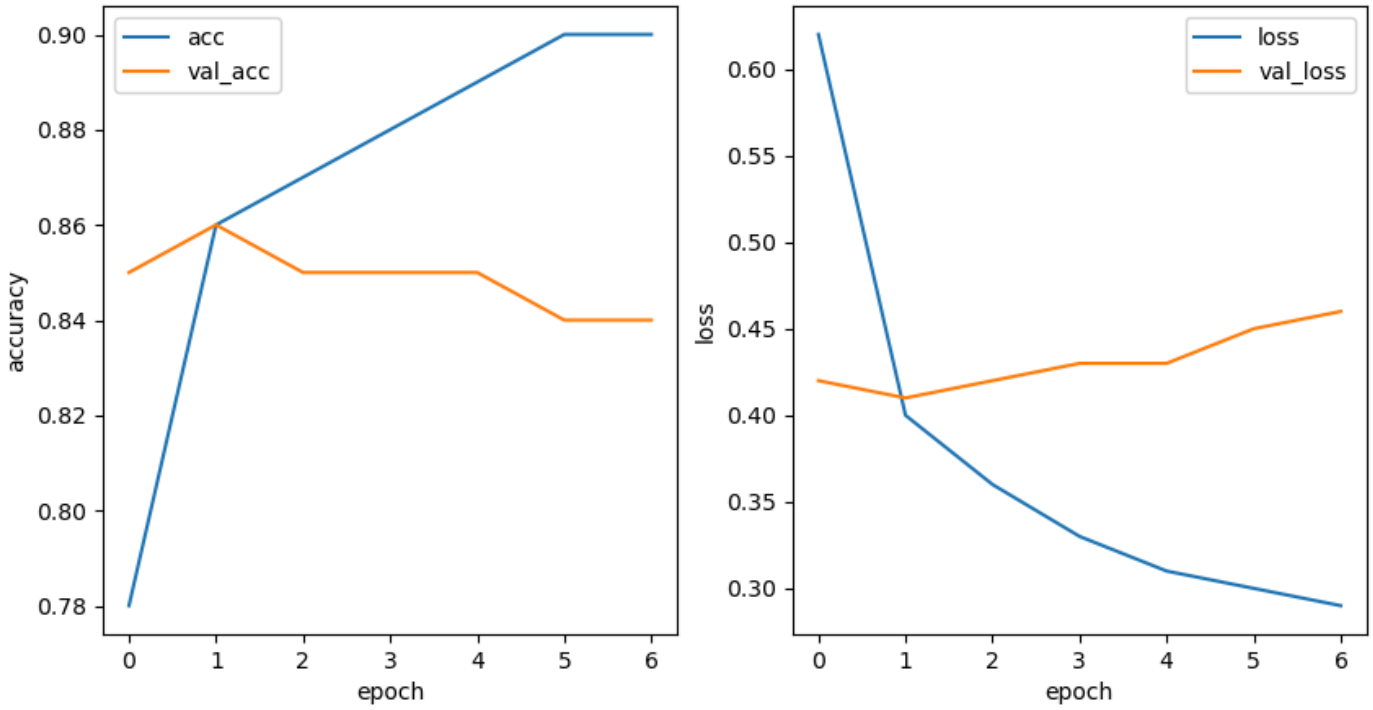
Łącznie zostały stworzone 4 modele, które osiągnęły accuracy dla zbioru walidacyjnego w okolicach 85%.

Na **rysunku 1.** został przedstawiony wynik pracy pierwszego i najprostszego modelu. Jest on najprostszym modelem nie posiadającym warstwy rekurencyjnej. Posiada on 4 warstwy. Mimo swojej prostoty zdołał osiągnąć wynik na poziomie 85% accuracy dla zbioru walidacyjnego.



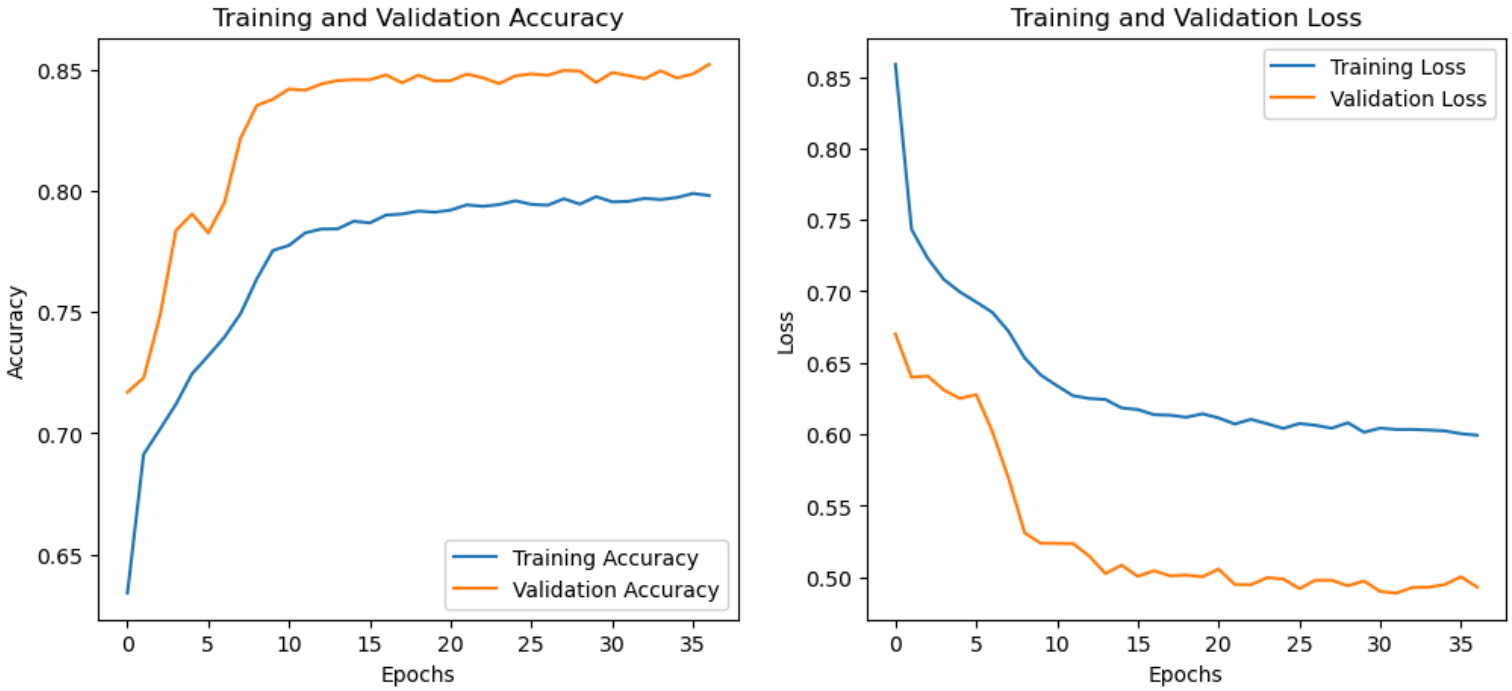
Rysunek 1. Accuracy i loss pierwszego modelu bez warstwy rekurencyjnej.

Następny model **(rysunek 2.)** posiada 7 warstw i również nie posiada warstwy rekurencyjnej. Osiągnął wyniki bardzo zbliżone do pierwszego modelu z accuracy dla zbioru walidacyjnego w okolicach 85%



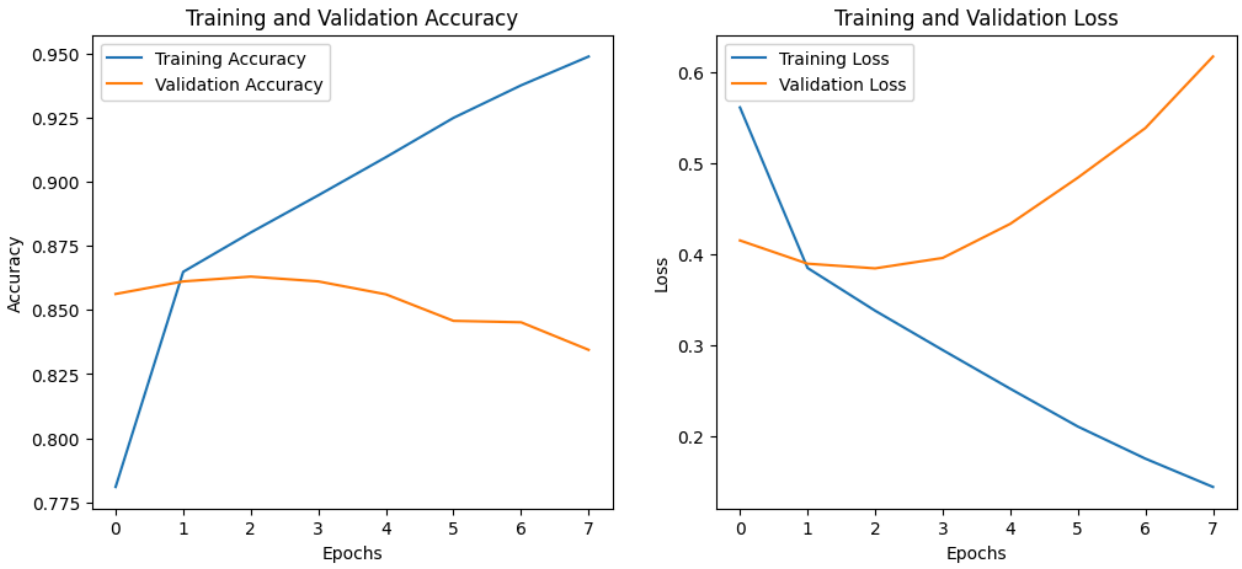
Rysunek 2. Accuracy i loss drugiego modelu

W 3 modelu widocznym na **rysunku 3.** zastosowano sposoby zapobiegania przeuczenia modelu wykorzystując dropout na poziomie 0.5, co osiągnęło swój cel i model mógł uczyć się przez większą ilość epoch przy niewystąpieniu przeuczeniu. Accuracy dla zbioru walidacyjnego wciąż osiągnęło wynik w okolicach 85%



Rysunek 3. Accuracy i loss trzeciego modelu - brak przeuczenia.

Ostatni model widoczny na **rysunku 4.** zawiera dodatkowo warstwę rekurencyjną LSTM. Mimo to osiągnął efekty przeciwne od zamierzonych. Nie zwiększył accuracy modelu, a pojawił się efekt, gdy model z każdą następną epochą “oduczał się” zmniejszając swoje accuracy dla zbioru walidacyjnego przy zwiększaniu loss zbioru walidacyjnego. Warstwy zapobiegające przeuczeniu z dropout lub regularyzator L2 nie zdołał wpłynąć na ten efekt.



Rysunek 4. Accuracy i loss czwartego modelu.

4. Podsumowanie

Różne sposoby modyfikowanie zbioru, zmiana liczby warstw, stosowanie metod zapobiegania przeuczeniu i warstwa rekurencyjna nie miała wpływu na accuracy modeli. Nie były to złe wyniki, dla każdego modelu accuracy dla zbioru walidacyjnego wynosiło około 85%, co jest wynikiem bardzo dobrym w klasyfikacji sentymentu krótkich testów pochodzących z platformy Twitter przy 3 różnych klasach: sentyment pozytywny, neutralny i pozytywny. Trudno określić czy jest możliwe zwiększenie accuracy dla tego zbioru danych, ani stwierdzić czemu

accuracy było bardzo podobne dla każdego modelu niezależnie od jego struktury. W przyszłości należałoby się przyjrzeć etapowi przygotowania zbioru przed stworzeniem modelu.