KLASYFIKACJA WIADOMOŚCI E-MAIL TYPU SPAM/HAM

PODSTAWY SZTUCZNEJ INTELIGENCJI RAPORT Z PROJEKTU

MAJA BŁAŻEJEWICZ

1. Wyniki testowe i reningowe

Model został przetestowany na zbiorze danych podzielonym na zestawy treningowe i testowe. Kluczowe metryki, takie jak dokładność i funkcja straty, były monitorowane w trakcie treningu modelu i co 100 epok. Ostateczna dokładność modelu na zestawie testowym wyniosła **0.9539**. Poniżej przedstawiono wyniki dokładności w trakcie treningu:

Dokładność treningowa - osiągnęła wartość **0.9529**.
Dokładność testowa - utrzymała się na poziomie **0.9539**.

2. Uzasadnienie wyboru techniki/modelu

Wybrano model sieci neuronowej zaimplementowany w PyTorch ze względu na jego zdolność do skutecznego klasyfikowania tekstu, co zostało potwierdzone w wielu badaniach naukowych. Zalety zastosowanego podejścia to:

- Elastyczność: Sieci neuronowe mogą być dostosowane do różnych rodzajów danych i problemów.
- Skalowalność: Możliwość obsługi dużych zbiorów danych i skomplikowanych wzorców.
- Wysoka dokładność: Umożliwiają osiągnięcie wysokiej dokładności dzięki możliwości uczenia się z dużych ilości danych.

Ze względu na binarny charakter problemu klasyfikacyjnego, zbudowany został model regresji logistycznej - sieć z jedną warstwą ukrytą i sigmoidalną funkcją aktywacji.
Zastosowana funkcja straty to **BCELoss,** natomiast do optymalizacji wykorzystano algorytm **SGD (Stochastic Gradient Descent)** aby zapewnić skalowalność rozwiązania..

3. Strategia podziału danych

Dane zostały podzielone na zestawy treningowy i testowy w następujący sposób:
- **Zestaw treningowy**: 80% danych
- **Zestaw testowy**: 20% danych

Podział danych został dokonany za pomocą funkcji train_test_split z bibliotek sklearn z ustawionym random_state w celu zapewnienia powtarzalności wyników.

4. Opis danych wejściowych

Dane wykorzystane w projekcie pochodziły z pliku tekstowego `spam_ham.txt`, zawierającego wiadomości sklasyfikowane jako spam lub ham (nie-spam). Każda linia w pliku była sformatowana jako `Category <tab> Text`, gdzie `Category` oznaczała typ wiadomości (`spam` lub `ham`), a `Text` zawierał treść wiadomości.

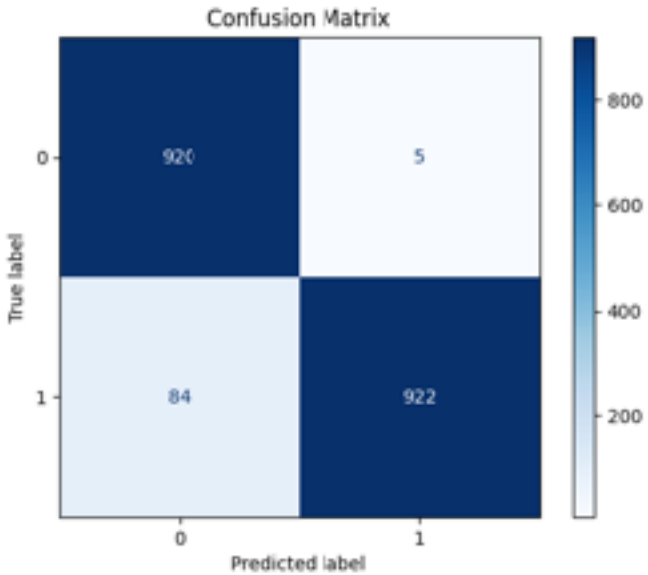
Oryginalne dane były mocno niezrównoważone (747 wiadomości spam i 4827 wiadomości niespamowych). W celu zrównoważenia zbioru danych, przeprowadzono augmentację danych tekstowych wykorzystując bibliotekę **NLTK w celu przeprowadzenia tokenizacji i stemmingu**.

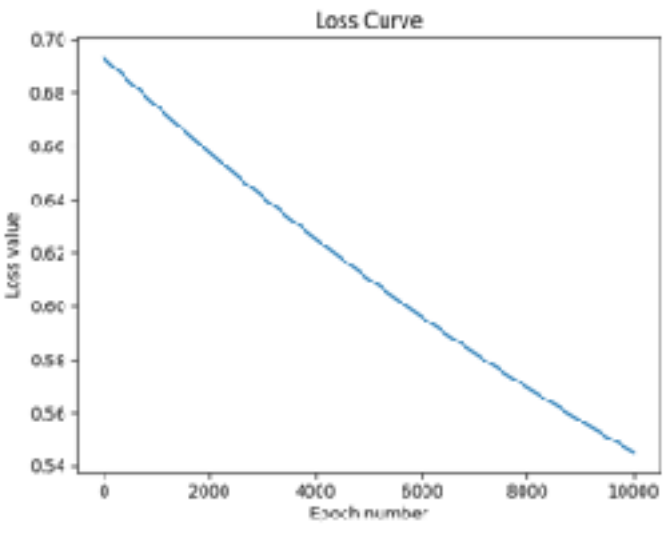
Ostatecznie jako dane wejściowe dla modelu wykorzystano zbiór 4827 wiadomości każdego z typów.

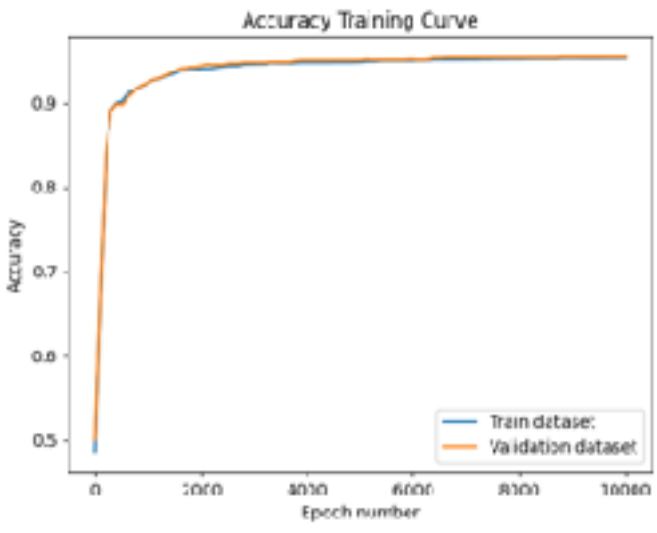
5. Analiza wyników i propozycja dalszych kroków

Osiągnięte wyniki są bardzo satysfakcjonujące, z wysoce zadowalającą dokładnością na zbiorze testowym po 10001 epok treningu. Analiza macierzy pomyłek wskazuje na wysoką skuteczność modelu w klasyfikowaniu zarówno wiadomości spam, jak i ham.

Poniżej znajdują się wykresy obrazujące krzywe strat i dokładności oraz confusion matrix dla zaugumentowanego zbioru testowego:







Potencjalne dalsze kroki obejmowały by działania umożliwiające lepszą generalizację modelu, sama dokładność jest moim satysfakcjonująca. W zależności od use caseu można by też dokonać optymalizacji pod kątem redukcji ilości FP bądź TN, jeśli byłaby taka konieczność.

Dodatkowo zastosowanie modeli transformatorowych z mechanizmem atencji mogłoby pozytywnie wpłynąć na dokładność modelu. Warto oczywiście też rozważyć optymalizację hiperparametrów, inne modele i rozszerzyć zbiór danych.