Podstawy Sztucznej Inteligencji

Ćwiczenie nr 2

Klasyfikacja tekstu

Mateusz Praski

Informatyka Rok 3
AGH WIET

Spis treści

1 Wstęp

Celem ćwiczenia jest wytrenowanie pre-trenowanego modelu *HerBert base case* oraz *Naiwnego klasyfikatora Bayesowskiego* w celu klasyfikacji tekstu dla zbioru opisów wybranych filmów z gatunków komedia oraz thriller z serwisu *FilmWeb* oraz klauzul abuzywnych udostępnionych w konkursie organizowanym przez *UOKIK*.

2 Przygotowanie notatnika

Początkowe próby pracy na notatniku *Google Collab* zakończyły się niepowodzeniem - następowały ciągłe rozłączenia z narzędziem skutkujące utratą wyników pracy. W związku z tym postanowiłem przenieść obliczenia na maszynę lokalną o następującej specyfikacji:

- CPU: Intel i7 7700HQ 4x2.8GHz
- GPU: Nvidia GTX 1050 4GB
- RAM: 16 GB
- System: Linux Manjaro 5.10.79-1

3 Zbiór filmów z FilmWebu

Pierwszym zbiorem danych była baza filmów FilmWeb z gatunków komedia oraz thriller. Założono, że gatunki są zbiorami rozłącznymi. Rozmiar datasetu wyniósł 2556 tekstów, a rozkład klas $49.8 - 50.2[\%]$ (komedia - thriller).

Następnie zbiór został podzielony w stosunku $80 : 10 : 10$ na treningowy, walidacyjny oraz testowy. Wstępne przetwarzanie danych polegało na zamianie tekstowych nazw klas na numery (0 i 1) oraz dodaniu tytułu filmu do opisu, tworząc dane wejściowe.

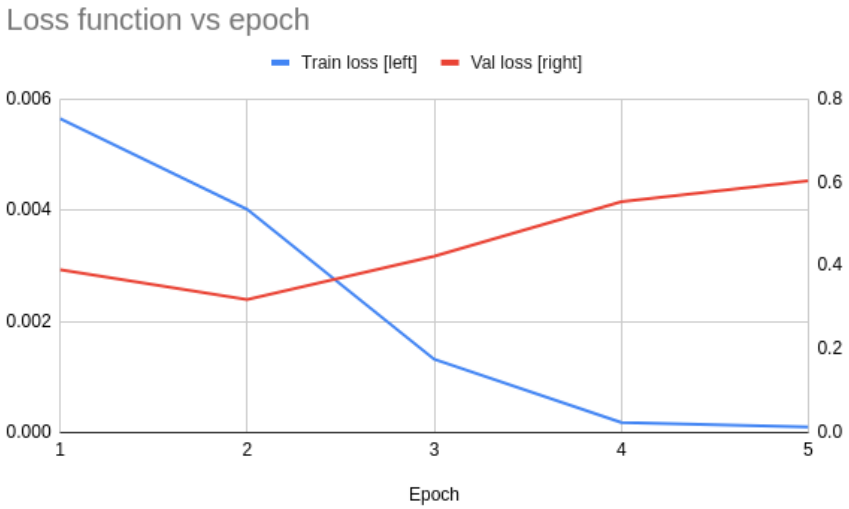
3.1 Dobór parametrów klasyfikatora

Pierwszą koncepcją w procesie doboru parametrów dla sieci neuronowej było wykorzystanie algorytmów przeszukiwania przestrzeni hiperparametrów modeli. Wybrana została biblioteka *Weights&Biases*, która okazała się nie działać z innymi wykorzystywanymi narzędziami. Z tego powodu postanowiłem skorzystać z prostej autorskiej implementacji algorytmu *Grid Search*.

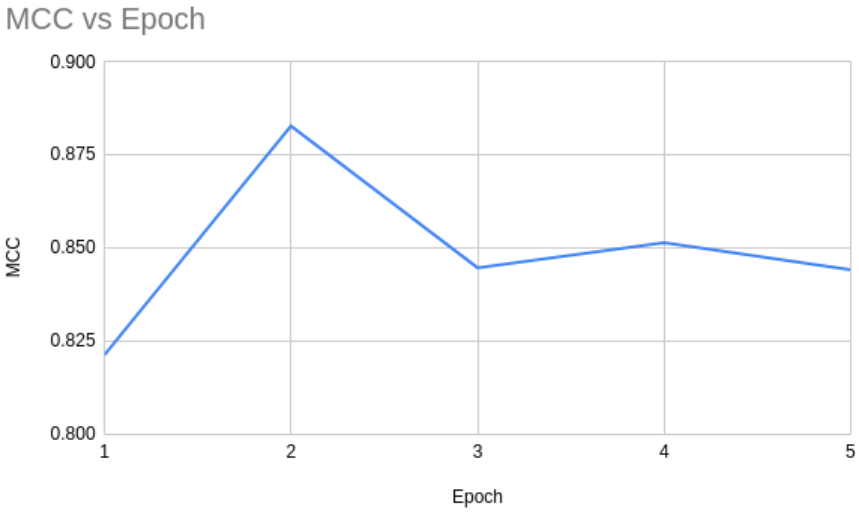
Jako pierwszy został wytrenowany oddzielnie model z następującymi parametrami:

- Liczba epok: 5
- Rozmiar batcha: 6
- Stała ucząca: $5 \cdot 10^{-4}$

Postępy procesu nauczania prezentują się następująco:



Rysunek 1: Zmiana funkcji straty od epok



Rysunek 2: Zmiana MCC od epok

Na podstawie przebiegu trenowania powyższego modelu możemy zauważyć, że od 3. epoki wzrasta wartość straty dla zbioru walidacyjnego, podczas gdy strata treningu kontynuuje spadek. Może to sygnalizować przetrenowanie modelu na podanym zbiorze treningowym, co skutkuje spadkiem generalizacji. Analogiczne wnioski możemy wysunąć patrząc na zmianę współczynnika korelacji Matthewa. W związku z tym podczas przeszukiwania hiperparametrów brane pod uwagę były epoki w zakresie od 1 do 2.

3.1.1 Przeszukiwanie przestrzeni parametrów

W następnym kroku przestrzeń została przeszukana z wykorzystaniem następującego algorytmu *Grid Search*:

for epoch ∈ epoch set **do**
  for batch size ∈ batch set **do**
    Wytrenuj sieć z parametrami (batch size, epoch)
    Dokonaj ewaluacji wyników i zapisz wyniki
  end for
end for

Jako parametry przeszukiwania zostały wybrane:

- batch set: [4, 8]
- epoch set: [1, 2]

3.2 Wynik treningu

W poniższej tabeli zostały zaprezentowane wyniki trenowania początkowego modelu oraz modeli z użyciem *Grid Search*:

Model (batch, epoch)	accuracy	precision	recall	F1-Score
(6, 5)	93.35%	94.40%	92.18%	93.28%
(4, 1)	92.18%	95.24%	89.55%	92.31%
(8, 1)	94.14%	91.27%	96.63%	93.88%
(4, 2)	92.58%	96.83%	89.05%	92.78%
(8, 2)	92.58%	96.03%	89.63%	92.72%

Tablica 1: Wyniki różnych metryk na przetestowanych modelach

Wykorzystując metryki *accuracy* oraz *F1-Score* jako decydujące możemy zauważyć, że model z parametrami (batch=8, epoch=1) sprawdził się najlepiej.

3.3 Naiwny klasyfikator Bayesowski

Dla tego problemu został również naiwny klasyfikator Bayesowski, z wykorzystaniem algorytmu *Grid Search* zaimplementowanego w bibliotece *SciKit*. Zbiór przeszukiwanych hiperparametrów składał się z następujących elementów:

- maksymalna częstotliwość dokumentów: [0.25, 0.5, 0.75]
- przedziały n-gramów: [(1, 1), (1, 2), (1, 3), (2, 3), (1, 4)]
- współczynnik alfa: [0.1, 0.01, 0.001]
- fit prior: [True, False]

Najlepszy klasyfikator otrzymaliśmy dla parametrów 0.5, (1, 2), 0.1, False. Jego wyniki prezentują się następująco:

accuracy	precision	recall	F1-Score
93.00%	93.00%	93.00%	93.00%

Tablica 2: Wyniki najlepszego naiwnego klasyfikatora Bayesowski

Klasyfikator dla tego problemu poradził sobie porównywalnie do rozwiązania opartego o sieci neuronowe. Warto również zauważyć, że liczył się on znacząco szybciej niż powyższe modele (kilka minut w stosunku do kilkudziesięciu).

4 Zbiór danych UOKIK

Drugim zbiorem danych był zbiór fragmentów regulaminów pochodzący z konkursu organizowanego przez *UOKIK*. Teksty były podzielone na dwie klasy: **KLAUZULA ABUZYWNA** i **BEZPIECZNE POSTANOWIENIE UMOWNE**. Dataset był podzielony na dwie części, treningową o rozmiarze 4284 elementów oraz testową zawierającą 3453 tekstów. Rozkład klas wynosił dla nich kolejno $54.58 - 45.42[\%]$, $67.56 - 32.44[\%]$ (KLAUZULA ABUZYWNA - BEZPIECZNE POSTANOWIENIE UMOWNE).

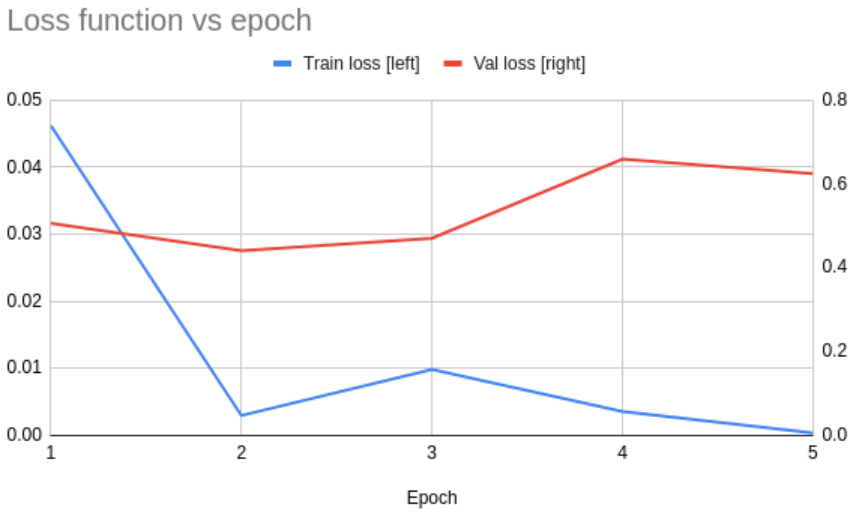
W celu przygotowania zbioru danych nazwy klas zostały zamienione na numery (0 i 1). Następnie ze zbioru treningowego został wydzielony zbiór walidacyjny w proporcjach $85 : 15$ (treningowy : walidacyjny).

4.1 Dobór parametrów klasyfikatora

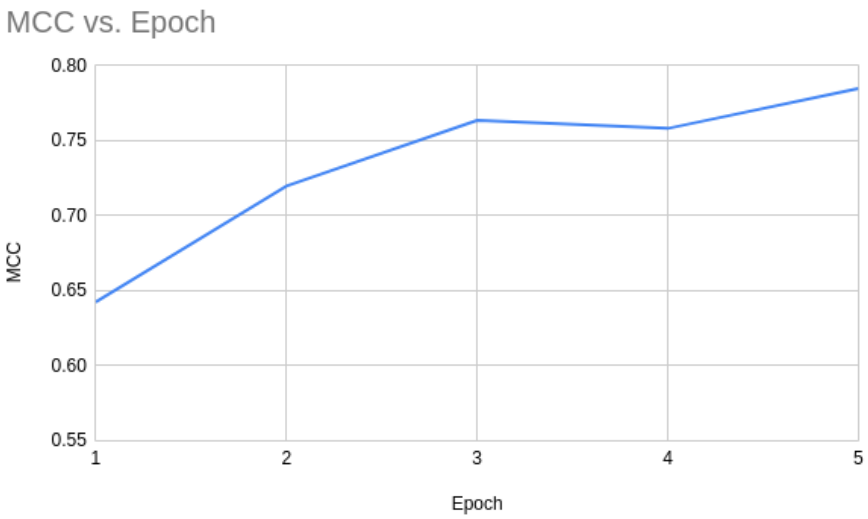
Na samym początku w celu wstępnej analizy hiperparametrów został wytrenowany oddzielnie model z domyślnymi parametrami:

- Liczba epok: 5
- Rozmiar batcha: 6
- Stała ucząca: $5 \cdot 10^{-4}$

Postępy procesu nauczania prezentują się następująco:



Rysunek 3: Zmiana funkcji straty od epok



Rysunek 4: Zmiana MCC od epok

4.1.1 Przeszukiwanie przestrzeni parametrów

W następnym kroku przestrzeń została przeszukana z wykorzystaniem następującego algorytmu *Grid Search*:

for epoch ∈ epoch set **do**
  for batch size ∈ batch set **do**
    for learning rate ∈ learning rate set **do**
      Wytrenuj sieć z parametrami (batch size, epoch, learning rate)
      Dokonaj ewaluacji wyników i zapisz wyniki
    end for
  end for
end for

Jako parametry przeszukiwania zostały wybrane:

- batch set: [4, 8]
- epoch set: [1, 2]
- learning rate set: $[10^{-4}, 10^{-5}]$

4.2 Wynik treningu

W poniższej tabeli zostały zaprezentowane wyniki trenowania początkowego modelu oraz modeli z użyciem *Grid Search*:

Model (batch, epoch, learning rate)	accuracy	precision	recall	F1-Score
$(6, 5, 5 \cdot 10^{-5})$	87.75%	87.14%	94.29%	90.58%
$(4, 1, 10^{-5})$	84.51%	83.15%	93.18%	87.88%
$(4, 1, 10^{-4})$	67.56%	100.0%	67.56%	80.64%
$(8, 1, 10^{-5})$	83.87%	82.51%	92.81%	87.36%
$(8, 1, 10^{-4})$	83.35%	77.28%	97.56%	86.25%
$(4, 2, 10^{-5})$	85.14%	81.96%	95.41%	88.17%
$(4, 2, 10^{-4})$	67.56%	100.0%	67.56%	80.64%
$(8, 2, 10^{-5})$	84.80%	84.44%	92.40%	88.24%
$(8, 2, 10^{-4})$	85.70%	84.01%	94.19%	88.81%

Tablica 3: Wyniki różnych metryk na przetestowanych modelach

W przypadku metryk *accuracy* i *F1-Score* model domyślny okazał się najlepszy. Jednakże możemy rozważyć również zastosowanie modelu (8, 1, 10^{-4}) ze względu na lepszą metrykę *recall*. Oznacza to, że ten model znajdowałby więcej klauzul abuzywnych kosztem większej liczby fałszywych ujemnych. W przypadku tego zagadnienia metryka ta może być preferowana, gdyż pozwala wykryć więcej przypadków klauzul abuzywnych.

4.3 Naiwny klasyfikator Bayesowski

Dla tego problemu został również naiwny klasyfikator Bayesowki, z wykorzystaniem GridSearcha zaimplementowanego w biblioteke *SciKit*. Zbiór przeszukiwanych hiperparametrów składał się z następujących elementów:

- maksymalna częstotliwość dokumentów: [0.25, 0.5, 0.75]
- przedziały n-gramów: [(1, 1), (1, 2), (1, 3), (2, 3), (1, 4)]
- współczynnik alfa: [0.1, 0.01, 0.001]
- fit prior: [True, False]

Najlepszy klasyfikator otrzymaliśmy dla parametrów 0.25, (1, 2), 0.1, False. Jego wyniki prezentują się następująco:

accuracy	precision	recall	F1-Score
84.00%	84.00%	84.00%	84.00%

Tablica 4: Wyniki najlepszego naiwnego klasyfikatora Bayesa

Klasyfikator okazał się gorszy w porównaniu do rozwiązań opartych o sieci neuronowe, jednakże jego wytrenowanie zajęło znacząco mniej czasu (kilka minut w stosunku do kilkudziesięciu). Może to być spowodowane faktem, że problem klasyfikacji klauzul prawnych jest znacznie trudniejszy niż klasyfikacja gatunków filmowych.