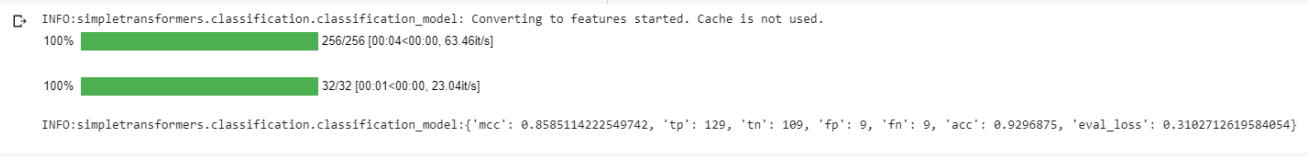
Przetwarzanie języka naturalnego (PJN)

Michał Maksoń
296622

1. Wyniki treningu modeli na dostarczonym zbiorze danych.

Uruchomienie modelu Roberta na przygotowanej konfiguracji przy num_train_epochs = 5.



Jak możemy zaobserwować dokładność osiągnęliśmy na poziomie niewiele niższym niż 0.93, a wartość utraty poprawnej ewaluacji w okolicach 0.31.

```
'acc': 0.9296875, 'eval_loss': 0.3102712619584054}
```

Osiągnięte wyniki dla klasyfikacji Bayesa dla wybranych najlepszych parametrów w przygotowej konfiguracji:

max_df=0.5, ngram_range=(1, 1), alpha=0.01

```
Applying best classifier on test data:
              precision    recall  f1-score   support

           0       0.92      0.91      0.91       118
           1       0.92      0.93      0.93       138

    accuracy                           0.92       256
   macro avg       0.92      0.92      0.92       256
weighted avg       0.92      0.92      0.92       256

[Parallel(n_jobs=3)]: Done  36 out of  36 | elapsed:    6.8s finished
```

2. Wyniki eksperymentów związanych z modyfikacją hiper-parametrów.

Dla modelu Roberta:

Num_train_epochs	Eval_batch_size	Accuracy score:	Eval loss:
5	**8**	**0.9297**	**0.3102**
7	**8**	**0.9179**	**0.4019**
10	**8**	**0.9297**	**0.5988**
5	**15**	**0.9307**	**0.3180**
7	**15**	**0.9453**	**0.3184**
10	**15**	**0.9297**	**0.4604**
2	**3**	**0.9063**	**0.2529**
5	**3**	**0.9258**	**0.3482**

3. Opis zbioru danych wraz z linkiem pozwalającym na jego pobranie.

Jako zbiór danych zdecydowałem się wybrać utwory Zbigniewa Wodeckiego oraz zespołu „Dżem”.
Jako pojedynczą encję danych traktowany jest wers utworu.
Treści utworów zostały pobrane z publicznie udostępnionych źródeł w Internecie.
Jako delimiter w pliku .csv zdecydowałem się użyć „;”, ze względu na często występujący w utworach znak przecinka „,”.
Pojedynczemu wersowi przyporządkowana jest wartość „dżem” lub „wodecki”.

Zbiór danych do pobrania z:

Github: https://github.com/Mervolt/MSI/blob/master/NLP/NLP/data.csv

Googledrive: https://drive.google.com/file/d/1X2ZofwZkihdSEc1xr-9-fdEUOPfJJujm/view?usp=sharing

Przykładowy fragment:

Daję Wam ogień, podajcie sobie ręce;dżem

Gdzie teraz białą farbą maluje wieczny czas;dżem

Polish Barbados i Galapagos;wodecki

Gubiąc gdzieś po drodze prawdę swą i sens;dżem

Choć tak łatwo zejść na psy!;dżem

Mój koncert skrzypcowy mistrzowsko się udał;wodecki

Dobranoc ci przez klamkę - niestety! zamknięta!;wodecki

4. Wyniki treningu modelu dla własnego zbioru danych.

Num_train_epochs	Eval_batch_size	Accuracy score:	Eval loss:
5	**10**	**0.7523**	**0.7685**
10	**10**	**0.7862**	**0.9689**
10	**5**	**0.8302**	**1.0740**

Uważam, że osiągniete wyniki są akceptowalne.
Najwidoczniej między utworami występują pewne podobieństwa co nie pozwoliło na osiągnięcie wyniku rzędu ~90% przy takich parametrach.

5. Przykłady błędnych klasyfikacji dla ustawienia: num_train_epochs = 10 i eval_batch_size = 5

Ironii czuję smak
wodecki
Nocą, jak biała dama, wstaje mgła
dżem
Nie rozgrzeszy rozłożone Pismo Święte
wodecki
I modlitwa zaraz potem

wodecki
Powiedział Pan, powiedział Pan:
wodecki
To naiwność, to błąd
dżem
Będę miał własny dom
wodecki
byłaś całkiem naga
wodecki
ale brakuje słów
wodecki
Nie spiesz się w ten słoneczny dzień, Ty wiesz:
wodecki
kiedy tak blisko obok siebie
wodecki
Choć figowy liść.
dżem
rzuci cień.
dżem
- Ja, Imogeno, niestety nie mogę...
dżem
Bo przecież serce kwili, jak młody złodziej
dżem
Żyj z całych sił
wodecki
Wspomnienia paru miejsc.
wodecki
Jak staw pełen łez
dżem
ojca matkę i rodziny smak,
wodecki
Ten ktoś jest, z tobą jest
dżem
W pasjansie zamiast damy król
dżem
Jak lód na Twojej twarzy.
wodecki
Miłości
dżem
Ale pogubiłem się,choć tak blisko był mój cel
wodecki
To przecież nic
dżem
Co kto robi jego rzecz.
wodecki
Trudno mi się zgodzić z tym
dżem