NLP - raport

Dawid Macek

1 Wyniki treningu modeli na dostarczonym zbiorze danych.

1.1 Neuronalny

Z domyślnymi parametrami. Osiągnięta dokładność to 90%.

```
INFO:simpletransformers.classification.classification_model:{'mcc': 0.8035303721977938, 'tp': 101, 'tn': 130, 'fp': 16, 'fn': 9, 'acc': 0.90234375, 'eval_loss': 0.37534565618261695}
```

1.2 Bayesowski

Osiągnięta dokładność to 89%.

	precision	recall	f1-score	support
0	0.90	0.90	0.90	146
1	0.87	0.87	0.87	110
accuracy			0.89	256
macro avg	0.89	0.89	0.89	256
weighted avg	0.89	0.89	0.89	256

2 Wyniki eksperymentów związanych z modyfikacją hiper-parametrów

Dla mniejszych batch sizów uruchamiało się za długo.

	Epochs: 1	Epochs: 5	Epochs: 10	Epochs: 15
Batch size: 10	0.878906	0.882812	0.898438	0.894531
Batch size: 20	0.878906	0.914062	0.890625	0.886719

	Epochs: 1	Epochs: 5	Epochs: 10	Epochs: 15
Batch size: 20	0.925781	0.929688	0.921875	0.914062
Batch size: 40	0.890625	0.937500	0.925781	0.937500

3 Opis zbioru danych wraz z linkiem pozwalającym na jego pobranie

Wersy z różnych polskich piosenek z następujących gatunków:

- disco polo
- hip-hop

Link: https://github.com/Palkovsky/msi/blob/master/05-nlp/dataset.csv W tym samym repo jest również skrypt użyty do stworzenia datasetu. Teksty pobierane były z tekstowo.pl.

Powyższy zbiór danych jest dość trudny, ponieważ disco-polo i polski hip-hop niczym się nie różnią.

	labels	text
3827	0	Środkowy palec wbijam w hejterów
392	1	Dlaczego, gdy szeptam do ucha, ty tego nie sły...
3884	0	I śmiał prosto w twarz, Ty otyły zjebie,
1383	1	Ilona, Ilona to radość nieskończona
308	1	Lecz jedno zawsze mam w pamięci.
...	...	...
2228	0	To dobry układ, jestem z tego rad, że ten drin...
3000	0	Ciemnoo już, noc nadchodzii, głuucha
1311	1	Nie realna,nie dotykalna jak wiatr,
1497	1	Zabiorę Cię do USA na ranczo heja hej
3565	0	Oni ścięci jak automatyczna sekretarka

4 Wyniki działania modelu dla własnych danych

Trenujemy model neuronalny z domyślnymi parametrami(liczba epok=5, wielkość batcha=8). Osiągnięta dokładność to 87%.

({'acc': 0.8759305210918115,
  'eval_loss': 0.6694126452511067,
  'fn': 25,
  'fp': 25,
  'mcc': 0.7517372233995367,
  'tn': 181,
  'tp': 172},