Zaawansowane techniki uczenia maszynowego - projekt

Marcin Zakrzewski

Adam Wawrzeńczyk

Cel projektu

- Zadanie polega na klasyfikacji recenzji leków w skali ocen 1-10 (minimalizacja Mean Absolute Error) oraz trzystopniowej skali low-medium-high (maksymalizacja accuracy)

	name	condition	opinion	rate	rate1
0	Zegerid	GERD	"Using it as a replacement for Nexium, since i...	10	high
1	Ethosuximide	Seizures	"This medicine is very good at controlling me ...	10	high
2	Tri-Sprintec	Birth Control	"I just started taking Tri Sprintec after my l...	9	high
3	Levaquin	Pneumonia	"This medicine made me feel absolutely horribl...	5	medium
4	Methylphenidate	ADHD	"I've been taking Concerta since 2003. Fo...	9	high

Preprocessing

```
"I've been taking Concerta since 2003. For me it's the only ADHD medication that works. It calms me
 down, I talk at a speed others can understand and my thoughts don't run off. The side effects I get
from it are .loss of appetite(growing up I was underweight for this reason) . Trouble sleeping(I end
ed up needing a medication for this but am currently not on one) .Dry mouth (drink lots of wate
r).... I have also been diagnosed with major depression and bipolar after years of taking Concerta
 (diagnosed in 2007 and 2013 respectively (both are genetic in my family)). Drugs effect each person
differently. If you decide to take Concerta I would suggest keeping track of you side effect and how
often they occur and talk to you M.D or N.P."
```

```
i have been taking concerta since 2003 for me it is the only adhd medication that works it calms me
 down i talk at a speed others can understand and my thoughts do not run off the side effects i get
 from it are loss of appetite(growing up i was underweight for this reason) trouble sleeping(i ended
up needing a medication for this but am currently not on one) dry mouth (drink lots of water) i have
also been diagnosed with major depression and bipolar after years of taking concerta (diagnosed in 2
007 and 2013 respectively (both are genetic in my family)) drugs effect each person differently if y
ou decide to take concerta i would suggest keeping track of you side effect and how often they occur
and talk to you md or np
```

- Pierwszym krokiem było przygotowanie kolumny recenzji:
  - Zmiana dużych liter na małe
  - Usunięcie znaków interpunkcyjnych
  - Usunięcie wielokrotnych spacji
  - Zamiana skrótów typu "won't" na pełne formy "will not"
- Próbowaliśmy też usunąć tzw. stopwords i dokonać lematyzacji, ale ponieważ tak przetworzone recenzje, jak się okazało, nie stanowiły lepszego zbioru uczącego dla naszych modeli, zdecydowaliśmy się poprzestać na czterech modyfikacjach wymienionych powyżej

Pozostałe kolumny?

- Name oraz condition - zmienne kategoryczne o tysiącach kategorii
- Dane niskiej jakości - np. condition "99 users found this comment helpful"
- Mediana liczby wpisów dla różnych kategorii – 8
- Dodanie ich do trenowanych modeli, nawet po oczyszczeniu, obniżyło ich skuteczność

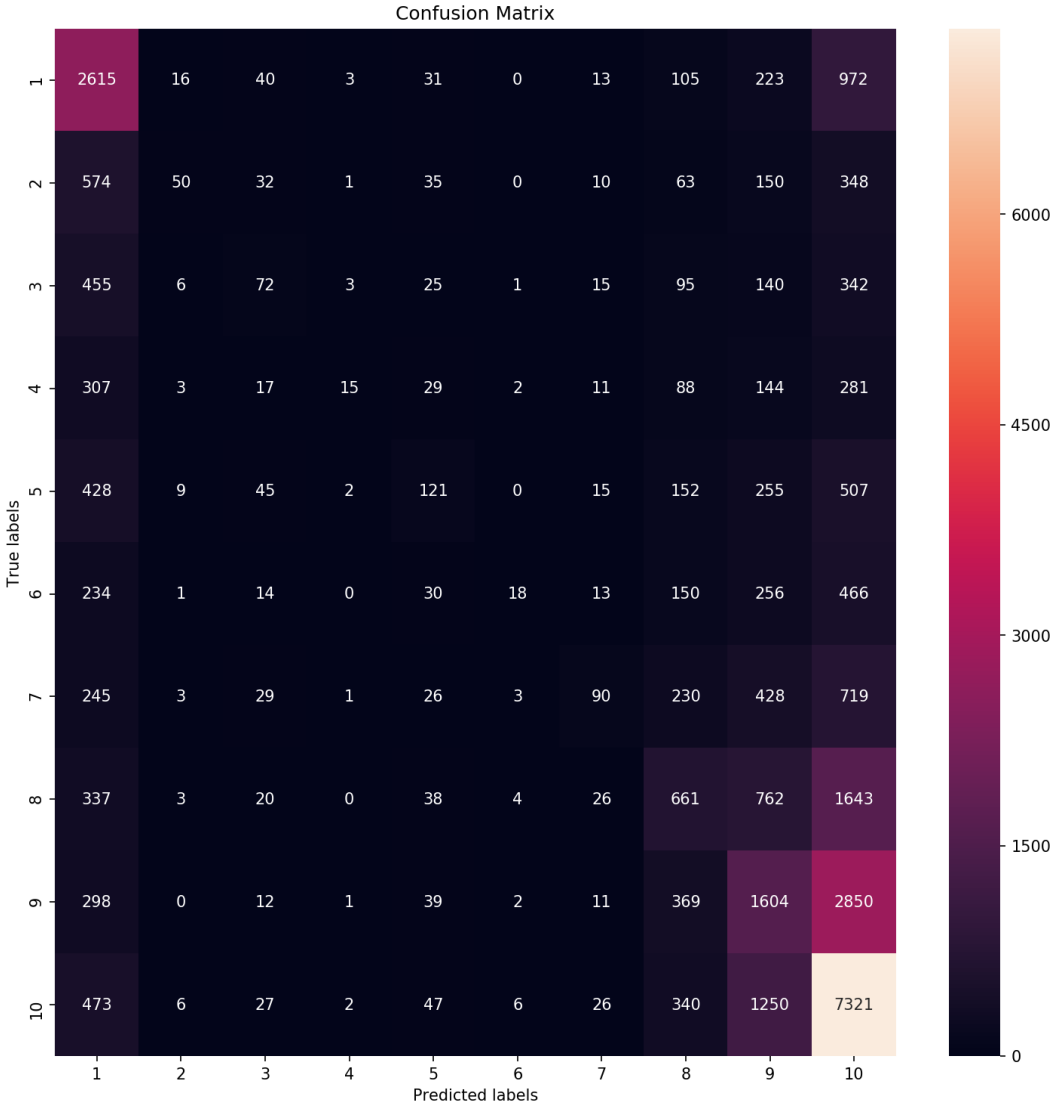
Naive Bayes

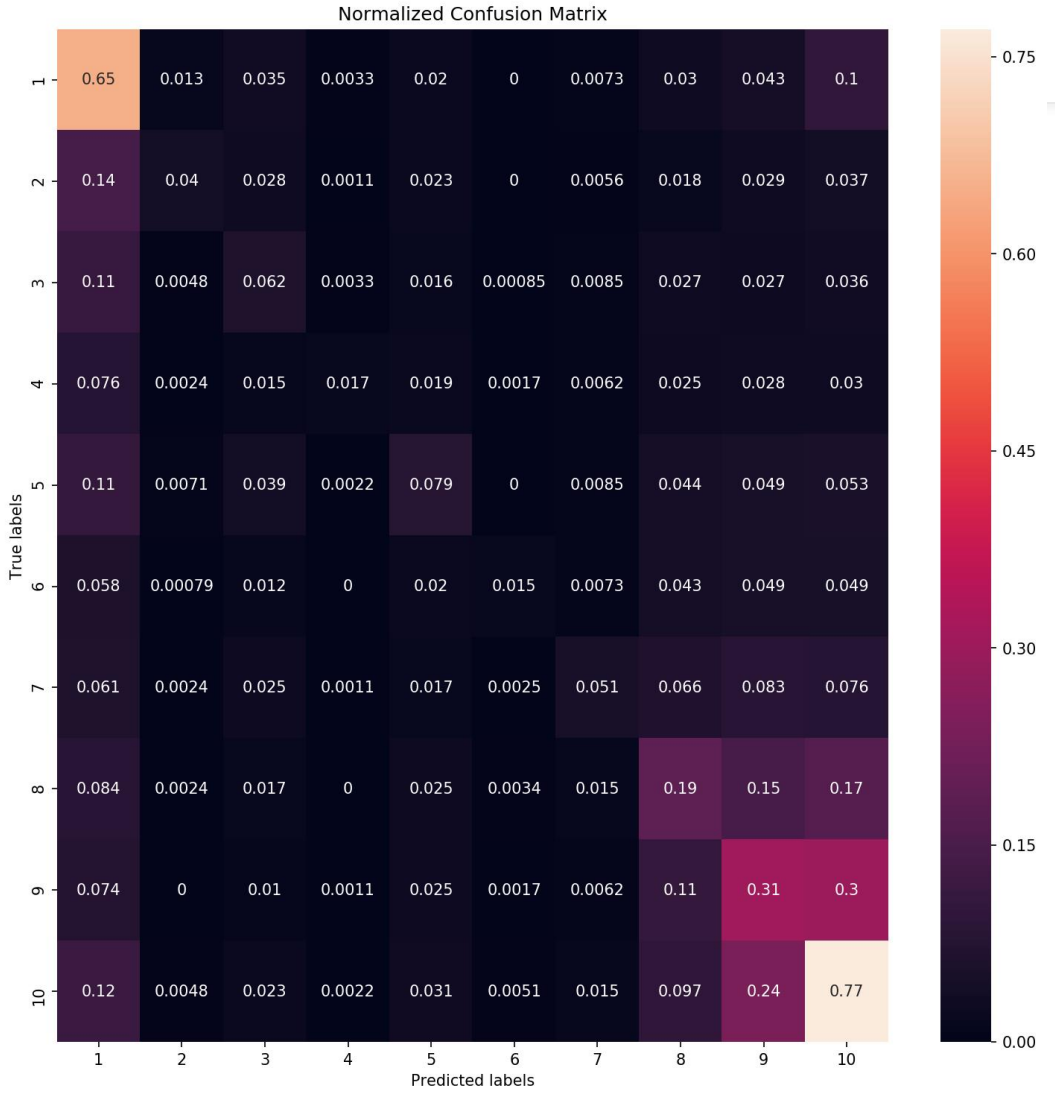
- Przetworzenie opinii metodą bag of words z minimalną częstością występowania słów
- Różne zakładane rozkłady:
  - Gaussowski
  - wielomianowy
  - Bernoulliego
- Różne progi częstości słów

Naive Bayes - wyniki

Rozkład	rate - MAE	rate - accuracy	rate1 - accuracy
Gaussa	1.797	15.2%	36.5%
wielomianowy	1.440	37.2%	66.0%
Bernoulliego	1.437	37.4%	66.2%
Bernoulliego (poprawione parametry)	1.393	42.9%	70.5%

Naive Bayes - macierz pomyłek





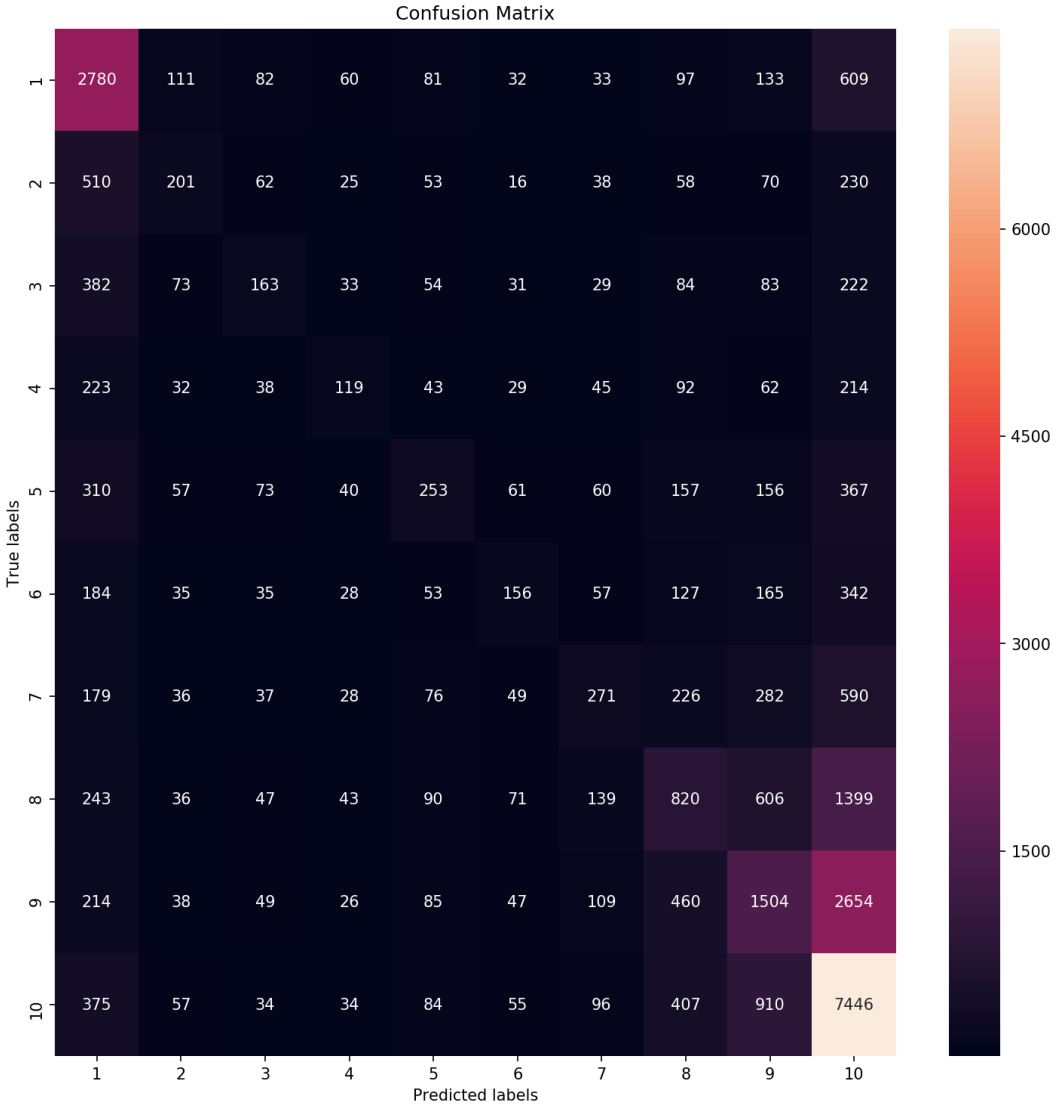
Regresja liniowa i logistyczna

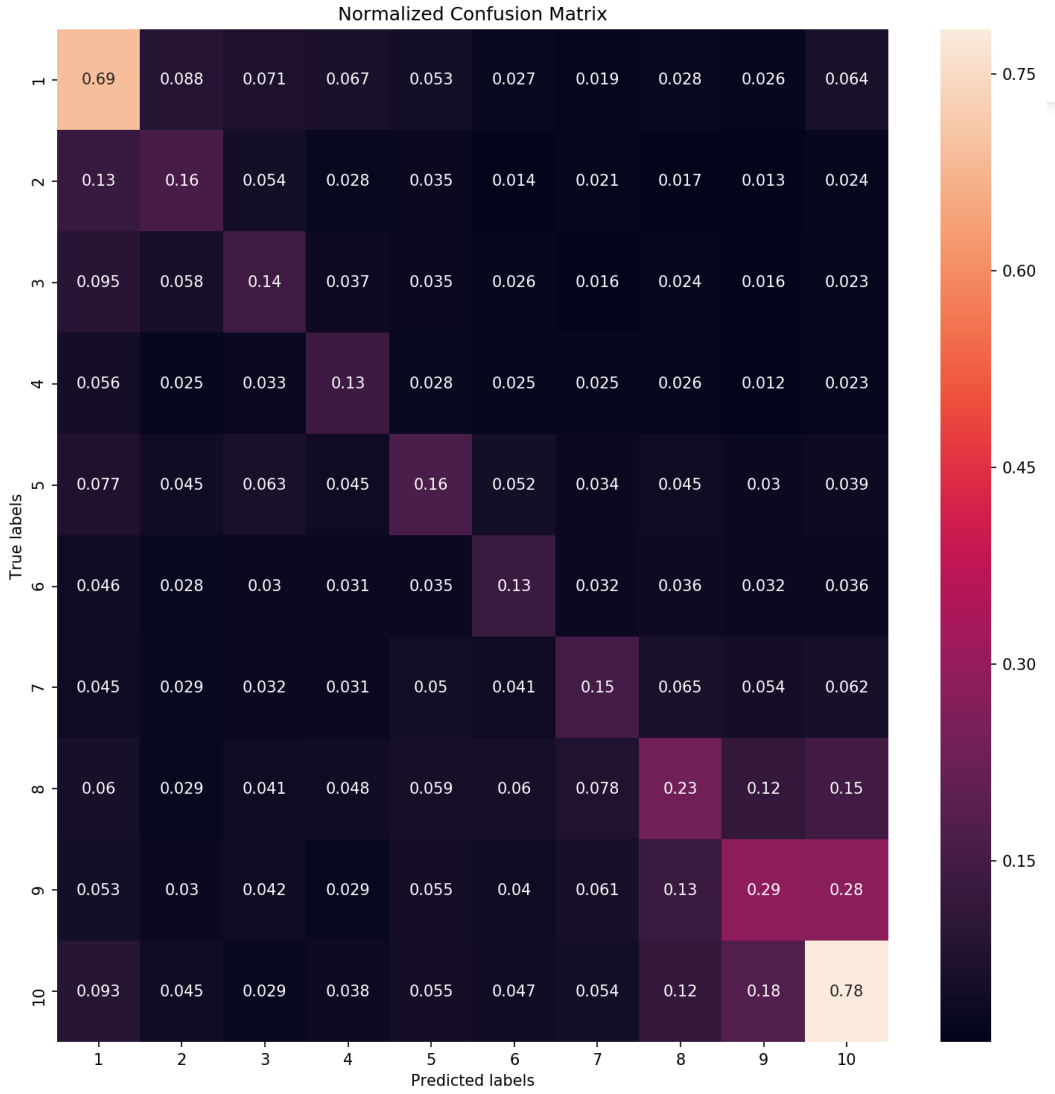
- Bag of words
- Regresja liniowa
  - Ridge
  - Lasso
- Regresja logistyczna
  - OVR – binarny problem dla każdej klasy
  - Wielomianowa

Regresja liniowa i logistyczna - wyniki

Algorytm	rate - MAE	rate - accuracy	rate1 - accuracy
Regresja liniowa (Ridge)	1.420	22.3%	54.7%
Regresja liniowa (Lasso)	1.657	6.5%	19.1%
Regresja logistyczna (multinomial)	1.248	47.8%	74.6%
Regresja logistyczna (OVR)	1.276	47.0%	74.0%

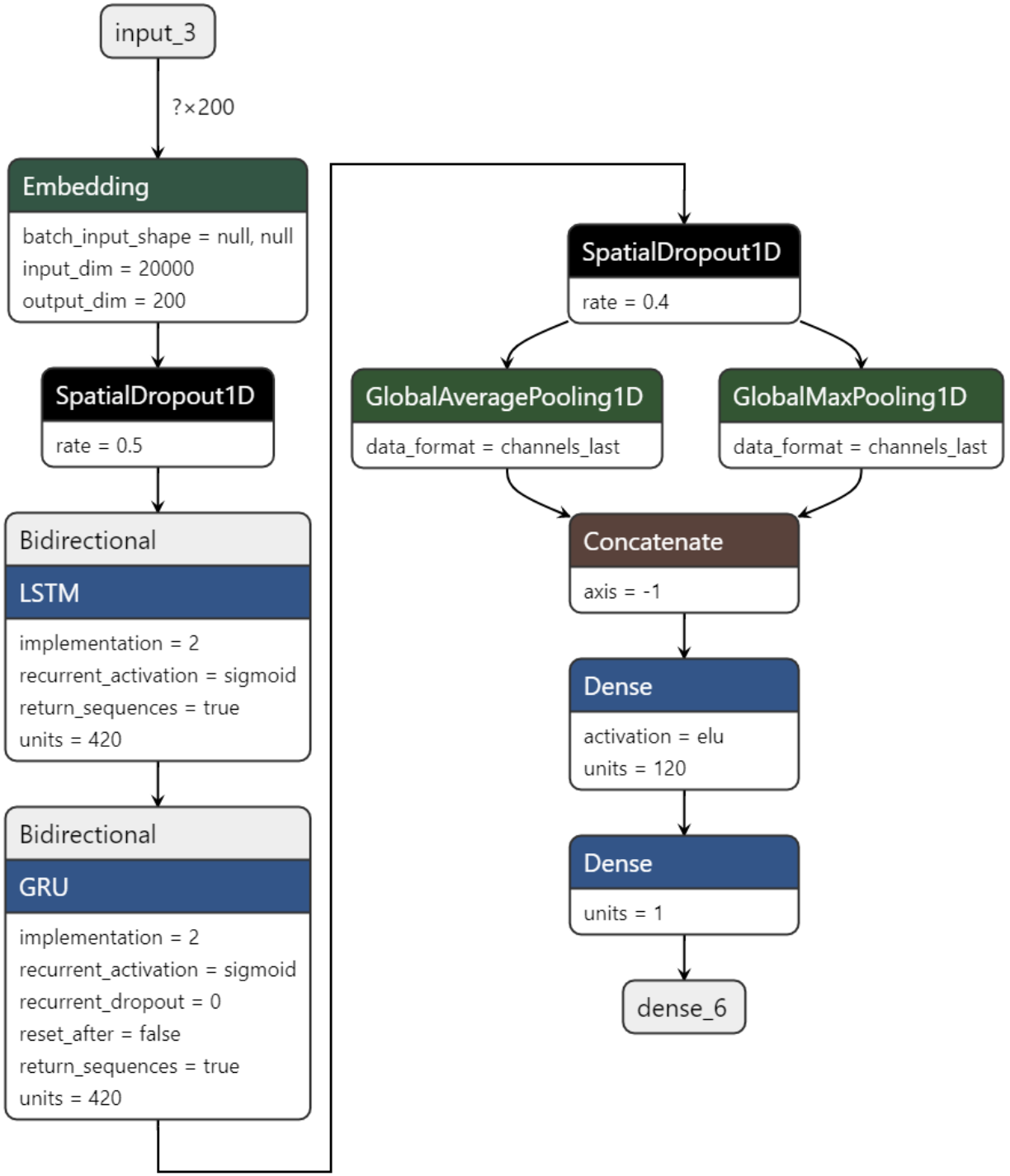
Regresja logistyczna - macierz pomyłek





Sieci głębokie - architektura

- Kolejnym rozwiązaniem, które zdecydowaliśmy się przetestować były sieci rekurencyjne LSTM i GRU.
- Za wzór dla architektury sieci posłużyła sieć zaproponowana w artykule https://www.kaggle.com/athoul01/predicting-yelp-ratings-from-review-text, przeznaczona do tego samego typu problemu.
- Po pewnych modyfikacjach mających dostosować sieć do problemu regresji, nie klasyfikacji, oraz długiej walce z problemem przeuczenia stworzyliśmy finalną wersję sieci, której schemat jest widoczny po prawo.



Sieci głębokie - model i uczenie

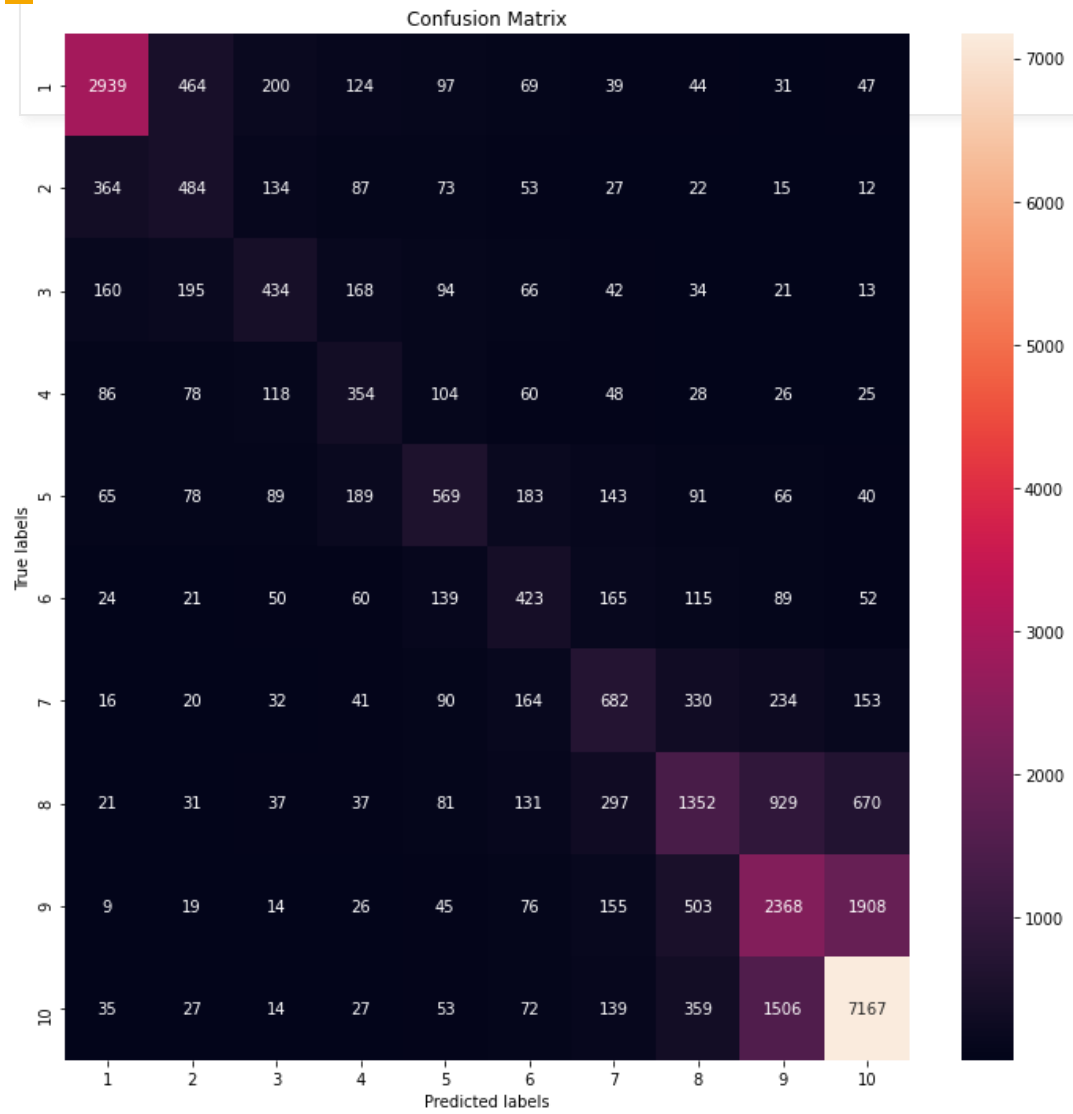
- Przedstawiona na poprzednim slajdzie sieć została zaimplementowana z wykorzystaniem bibliotek sklearn i keras.
- Ma 9,466,121 parametrów i była uczona metodą propagacji wstecznej z optymalizatorem *adam*.
- Proces uczenia był kontrolowany przy pomocy "Learning Rate Schedulera", który stopniowo zmniejszał współczynnik uczenia.
- Trenowanie docelowego modelu było możliwe dzięki platformie Google Colab i trwało w sumie około 13 godzin.

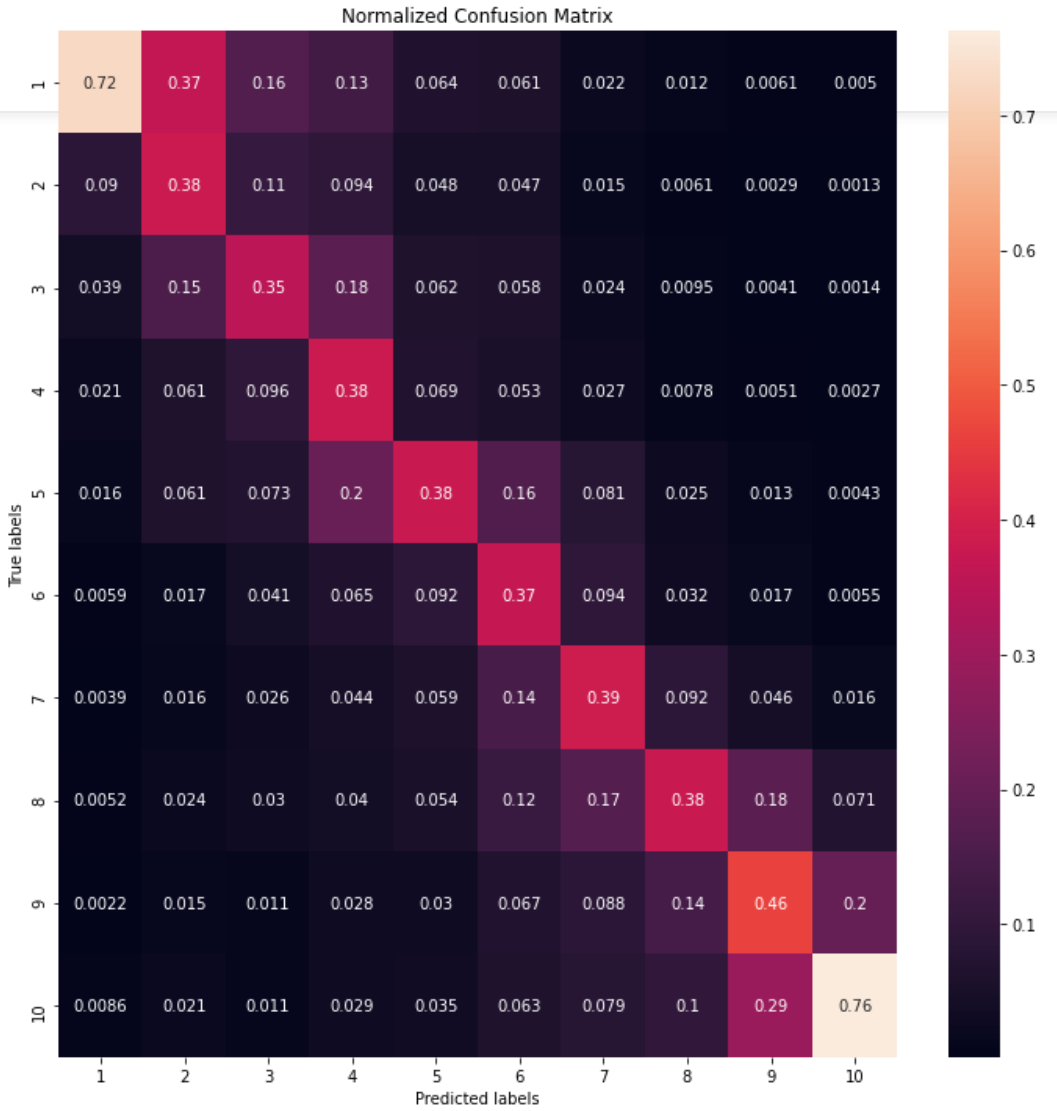
Sieci głębokie - wyniki

Wyniki zostały zebrane na liczącym 30 tysięcy rekordów zbiorze testowym

rate - MAE	rate - accuracy	rate1 - accuracy
0.818	55.9%	85.2%

Sieci głębokie - macierz pomyłek





Transfer learning – pre-training

- Wykorzystanie modelu wytrenowanego dla innego, bardziej ogólnego problemu
- Problem modelowania języka
  - Duże modele: BERT, GPT-2, XLNet – bardzo duża złożoność (setki milionów parametrów), duży koszt obliczeniowy
  - Wykorzystany prostszy model wytrenowany w ramach NAACL 2019

Transfer learning - adaptacja

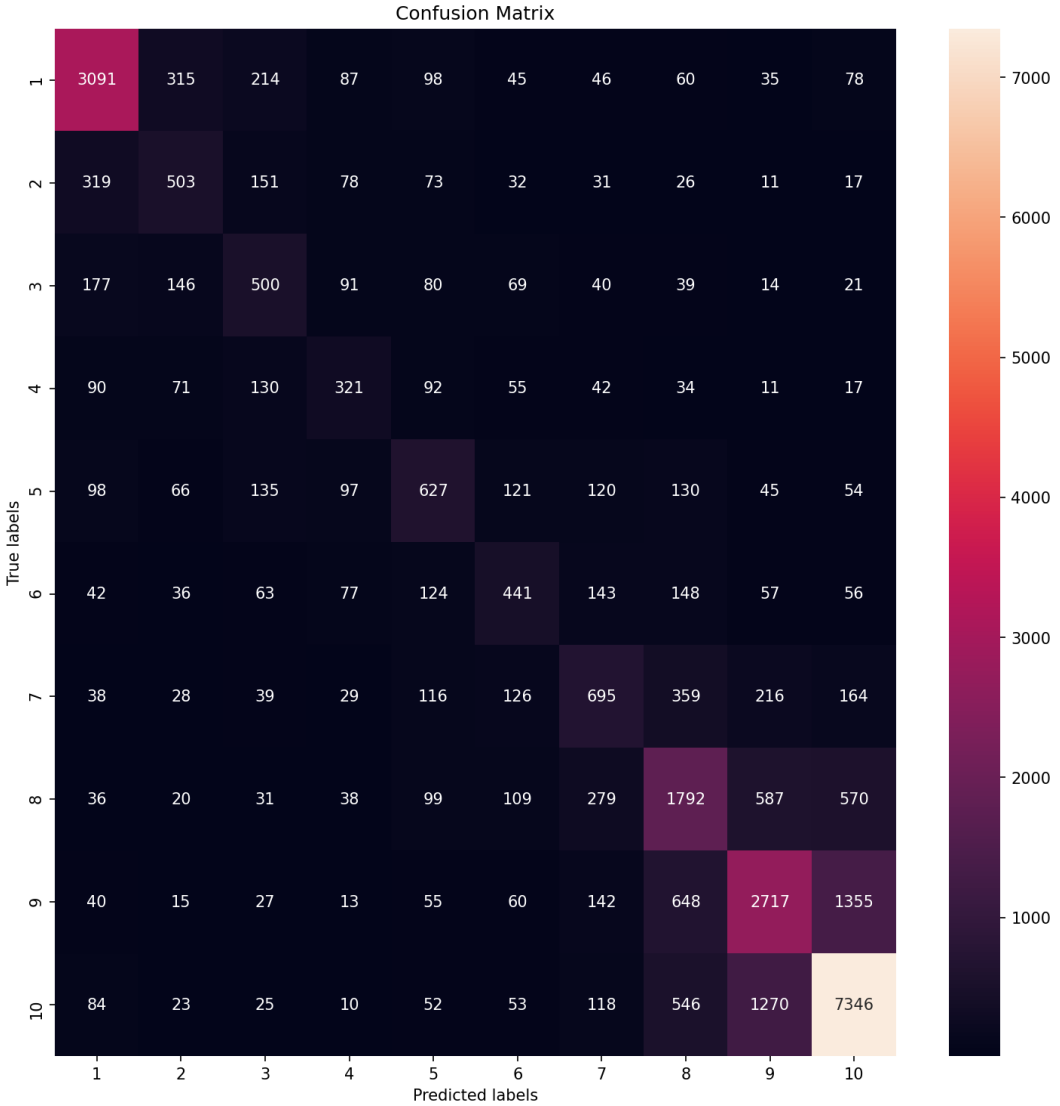
- Transformator
  - Odpowiada za dopasowanie modelu do nowego problemu
  - Wykorzystany transformator zaproponowany podczas NAACL 2019
- Warstwy gęste
  - Odpowiadają za klasyfikację

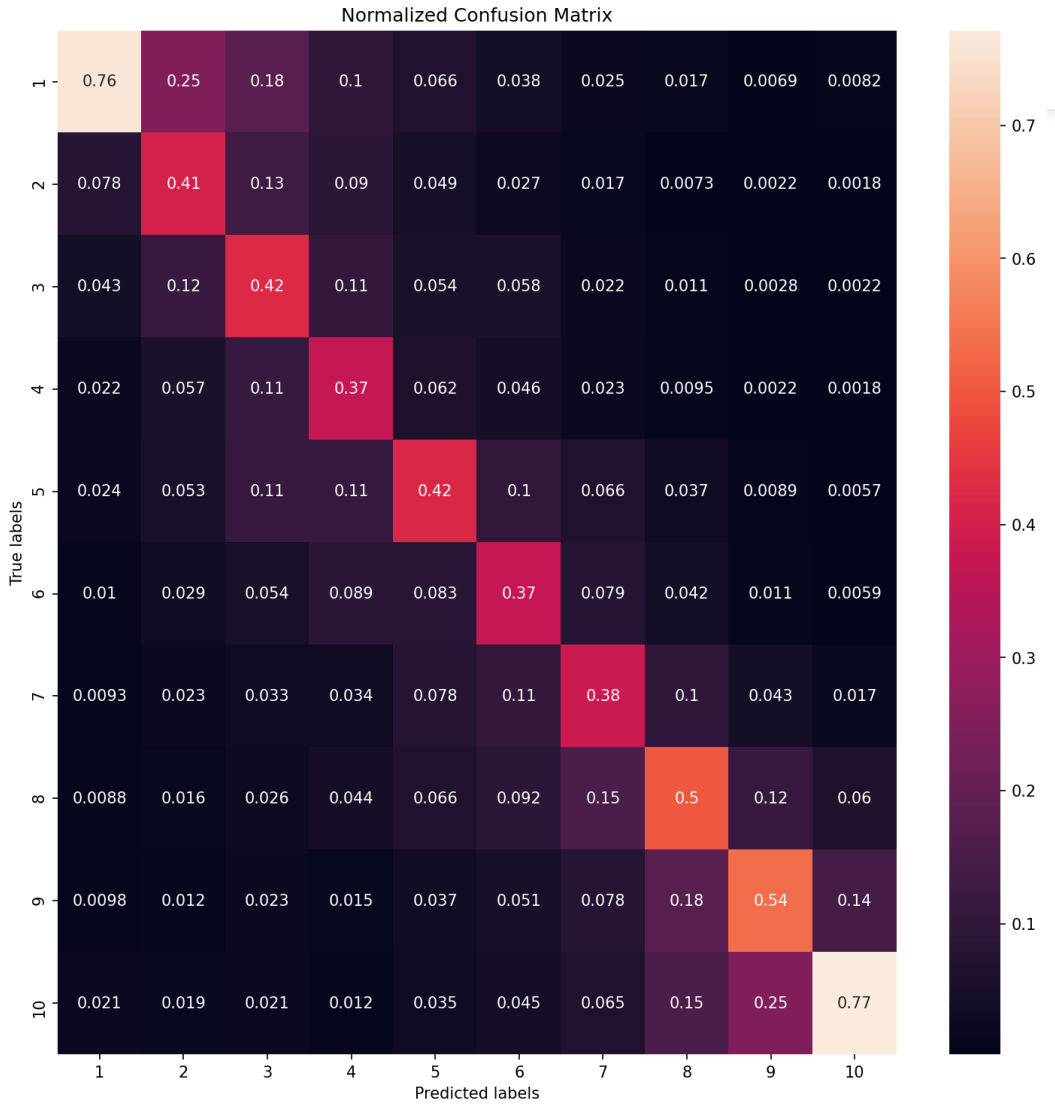
Transfer learning – wyniki

- Pretrenowany model - 50 milionów parametrów
- Adaptacja – 10/15 epok, ok. 50 minut / epoka (Colab)

Liczba epok	rate - MAE	rate - accuracy	rate1 - accuracy
10	0.874	55.4%	76.1%
15	0.816	60.1%	78.1%

Transfer learning - macierz pomyłek





Zaokrąglanie wyników

- Z ciekawości postanowiliśmy przetestować kilka metod zaokrąglania wyników regresji do liczb całkowitych.
  - Zaimplementowaliśmy metodę zaokrąglania opartą na liczności klas – im większa klasa tym większy przedział. Poskutkowało to wyższym MAE niż dla wartości niezaokrąglonych.
  - Zdecydowanie lepsze wyniki dało "naiwne" zaokrąglanie - średni błąd był niższy niż błąd dla wartości niezaokrąglonych.
- Po zaokrągleniu wartości zostały obcięte do zakresu $\{1,2,\ldots,10\}$, na wypadek gdyby regresor zwrócił coś większego lub mniejszego. Ten zabieg nie miał wpływu na MAE.

Porównanie najlepszych wyników

Algorytm	rate - MAE	rate - accuracy	rate1 - accuracy
Naive Bayes	1.393	42.9%	72.5%
Regresja logistyczna	1.248	47.8%	74.6%
Głębokie sieci neuronowe	0.818	55.9%	85.2%
Transfer learning	0.816	60.1%	78.1%

Wybrany algorytm

- Ostateczne rozwiązanie wykorzystuje algorytm oparty na **głębokich sieciach neuronowych**
  - Zastosowanie transfer learningu poprawiło MAE bardzo nieznacznie (duża poprawa accuracy nie ma znaczenia w kontekście zadania)
  - Macierz pomyłek dla tej metody jest bardziej skoncentrowana niż dla transfer learningu
  - Wybrany algorytm wykazuje zdecydowanie najlepszą poprawność predykcji zmiennej rate1