Классификация новостных статей РБК.ру с использованием нейронных сетей

Выполнила студентка
группы МО23-2м
Гордеева Наталия

Задачи исследования

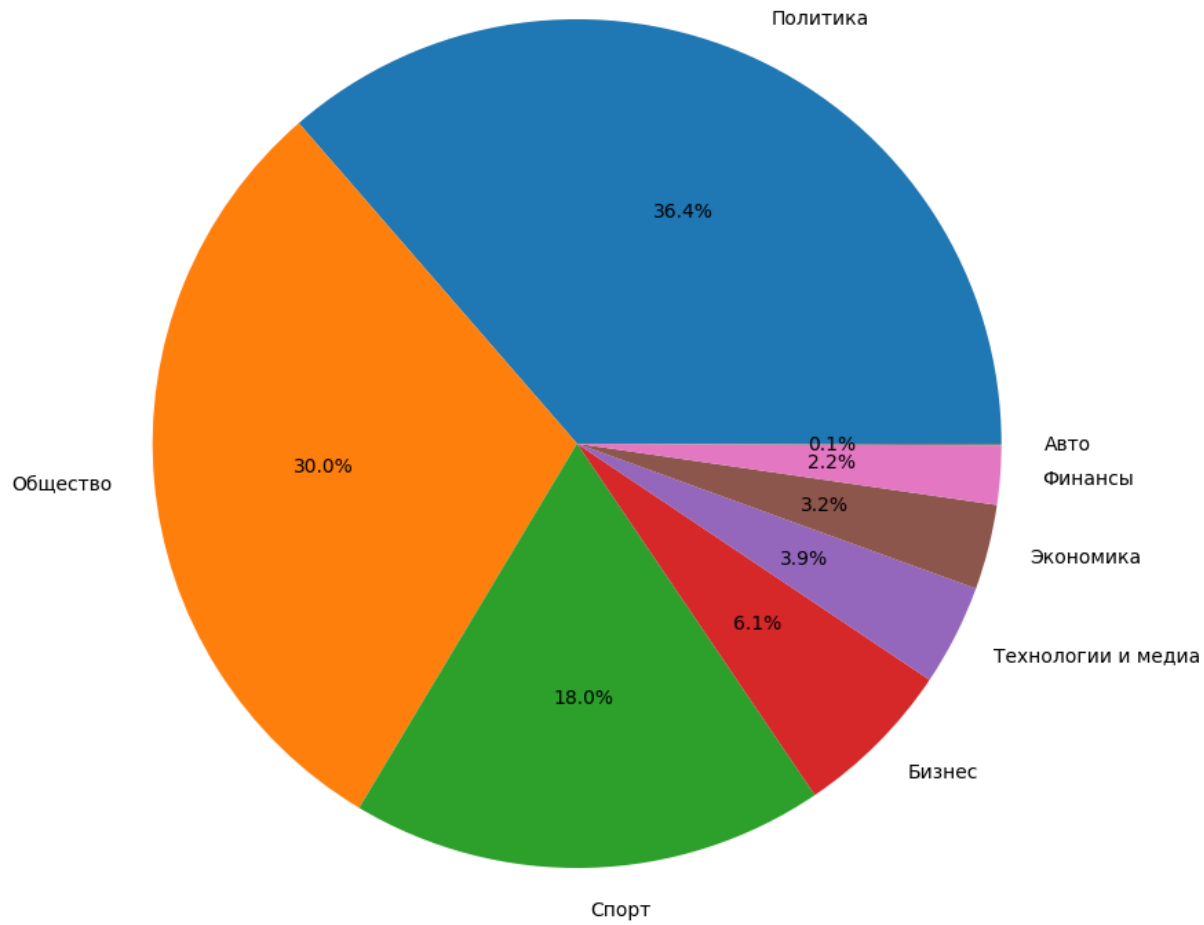
- Провести парсинг новостных статей с сайте РБК.ру;
- Сделать предобработку текстов, удалив стоп-слова и знаки пунктуации, проведя токенизацию;
- Построить архитектуры нейронных сетей с Embedding, BoW, выделить лучшую;
- Проверить работу сети на новостной сети.

rbc.ru

Политика 2171
Общество 1793
Спорт 1077
Бизнес 366
Технологии и медиа 230
Экономика 193
Финансы 134

5967 статей

За апрель-май 2024 г.



Предобработка текстов

[Служба безопасности Украины (СБУ) использовала для подрыва Крымского моста в октябре 2022 года самодельное взрывное устройство с мощностью, эквивалентной 10 тоннам тротила …]

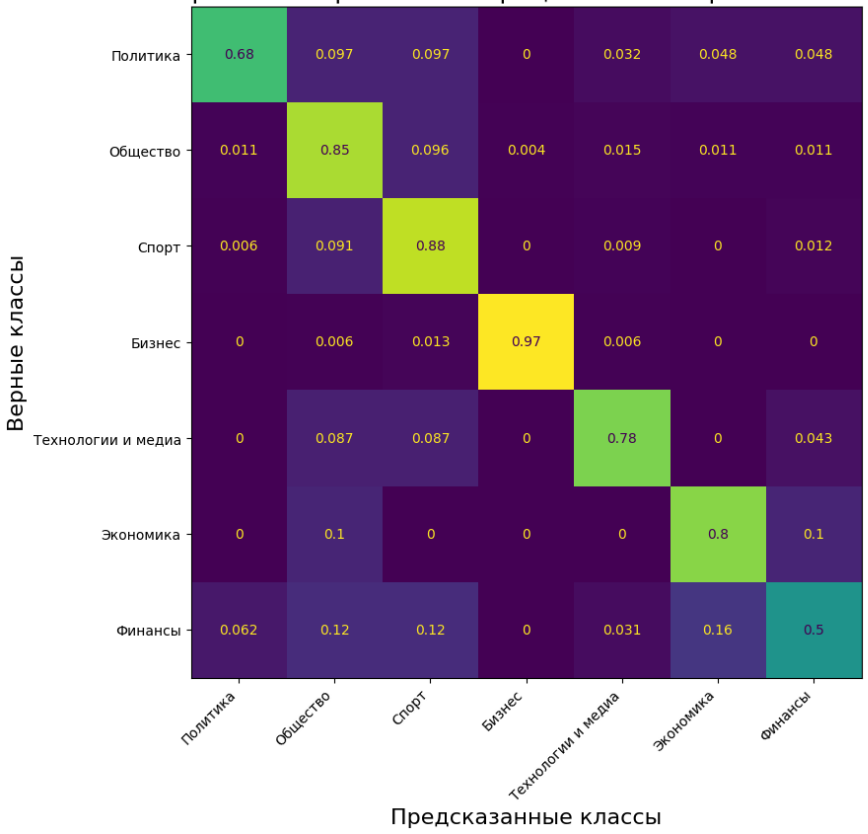
- Удаление стоп-слов и пунктуации;
- Построение словаря в виде пар слово – индекс;
- Векторное представление слов:

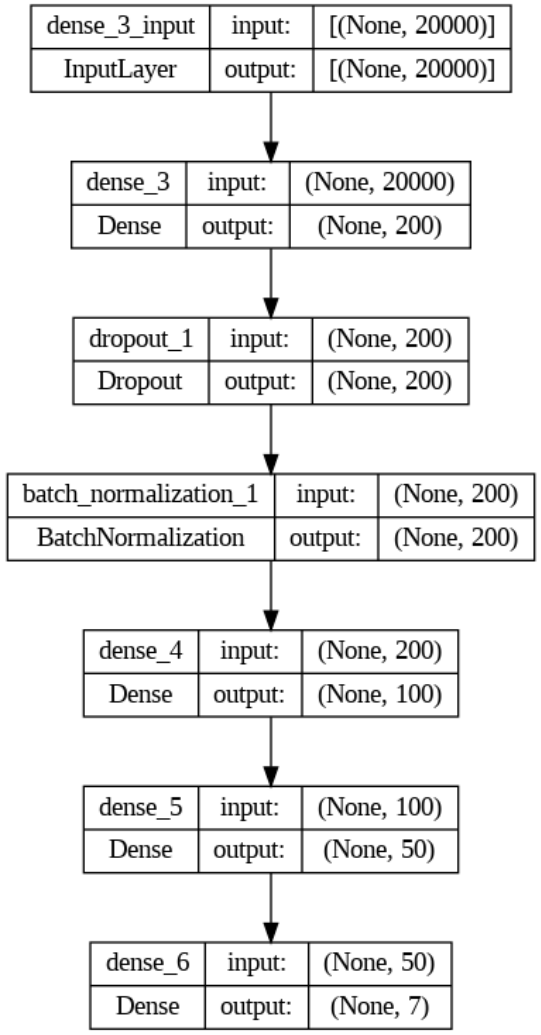
[3884, 1611, 11772, 4003, 4053, 3083, 351, 490, 264, 780, 2396, 20, 169, 3378, 534, 490, 2743, 9148 …]

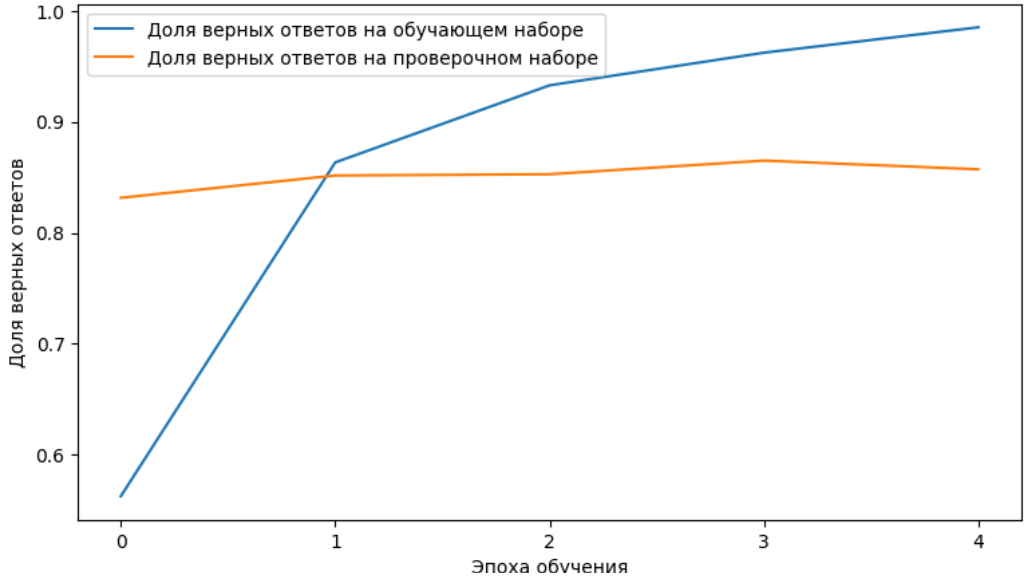
Simple BoW

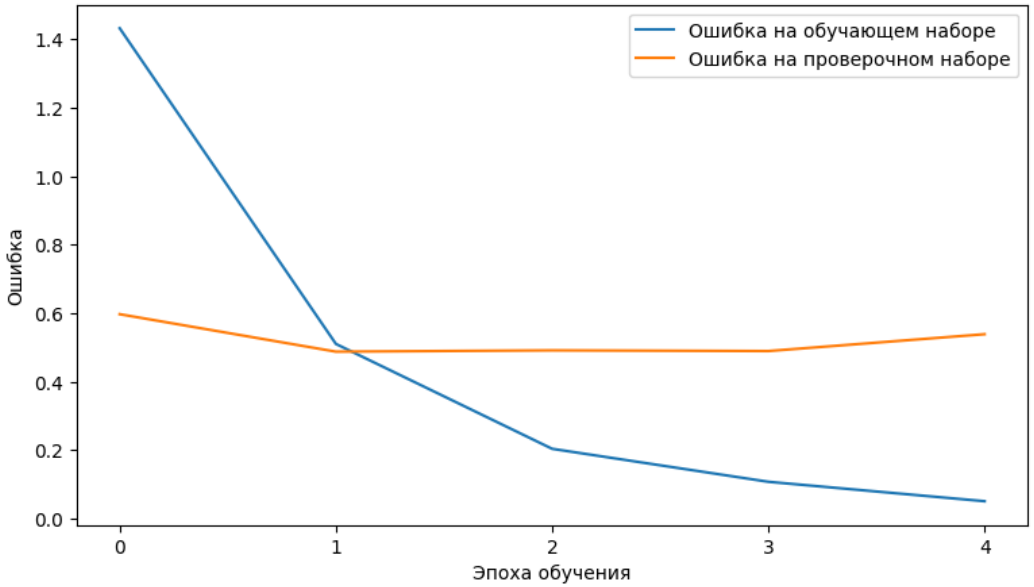
Лучшая архитектура нейронной сети.
Точность – **78%** на обучающей выборке

Нейросеть Simple BoW: матрица ошибок нормализованная









Выводы

- Чаще всего сеть путает классы «Финансы» и «Экономика», у которых есть большое количество схожих слов;
- Довольно низкая точность у категорий «Финансы» и «Политика», что может быть связано с введением весов классов в функцию компиляции;
- В целом, модель хуже всего относит статьи раздела «Финансы» к истинному классу, иногда относя их к «Политике», «Спорту», «Обществу» и «Экономике».