Реализована модель для определения тональности сообщения. При выполнении была проведена предобработка текста, а именно: замена смайлов их текстовым описанием, замена повторяющихся больше двух раз букв, замена ссылок, ников пользователей и тд. В качестве архитектуры нейронной сети для обучения были выбраны CNN (сверточная нейронная сеть), LSTM (разновидность рекуррентной нейронной сети ). В качестве простых статистических моделей, выбраны логистическая регрессия и метод опорных векторов. Для метрики качества выбраны f1, Precision и Recall. Таблица результатов представлена ниже

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metrics | CNN(word2vec) | LSTM(word2vec) | Log regress(Bag of words) | Log regress(TF-IDF) | SVM(TF-IDF) |
| F1 | 0.84795 | **0.85059** | 0.82304 | 0.82397 | 0.82922 |
| Precision | 0.84872 | **0.85060** | 0.82344 | 0.82411 | 0.82927 |
| Recall | 0.84797 | **0.85058** | 0.82302 | 0.82395 | 0.82920 |

Как мы видим незначительное преимущества зафиксировано у LSTM модели. Она обгоняет CNN модель всего на долю процентов. И сам результат в 85 % является достаточно хорошим для классификации твиттов.

Ссылка на GIthab: <https://github.com/vraptor74/SentimentTwit/tree/master>

Ссылка на первичные данные: <http://study.mokoron.com/>

Ссылка на работу участника соревнований: <https://arxiv.org/abs/1704.06125>

Ссылка на word2vec модель: <https://yadi.sk/d/Xohf3dTuVTakDA>

КЭ - 120

Хоряков Никита

20.07.2020