|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Classification Accuracy (micro)** | **Precision** | **Recall** | **F1** |
| Random Forest Classifier – 100 e. | 0.56 | 0.56 | 0.72 | 0.55 |
| Multinomial Naïve Bayes | 0.64 | 0.64 | 0.66 | 0.65 |
| LinearSVC | 0.65 | 0.65 | 0.65 | 0.64 |
| Voting Classifier (the 3 above) | 0.66 | 0.66 | 0.66 | 0.66 |
| BiLSTM: (300d: LSTM 200) + Feedforward: (20: Dense 20 – 10) - Dense 32 | 0.766 | 0.77 | 0.77 | 0.76 |
| CNN: (300: Conv1D – 26 filters, kernel size 10, Leaky ReLU rectifier) + Feedforward: (20: Dense 20 – 10) - Dense 32 | 0.771 | 0.77 | 0.77 | 0.77 |
| Feedforward: Dense 32 units – Dense 32 units | 0.775 | 0.78 | 0.78 | 0.78 |
| (300: Embedding 300 - CNN 26-10 – LSTM 200) + (20: Feedforward 20-10) -> Dense 200-100  Emb parameters: Padding 100, dict 10000, с обучением (Trainable = True) |  |  |  |  |
| FastText | **0.85** | **0.85** | **0.85** | **0.84** |

В реализациях и примерах, которые я видел, LSTM используется в сочетании с эмбеддинг слоем, что я и применил.

Он и занял большую часть времени: обучение очень долгое, и откалибровать для оптимальных параметров сложно. В конечном счёте я решил получить хоть какой-то результат с малым объёмом словаря для относительно быстрого обучения (CNN здесь даже ускорил процесс), и всё равно 5 минут на эпоху.

Weighed-unweighted, комбо 310-320 и прочее пересчитать лучше будет уже на новых данных, на мой взгляд

Для Bag of words: те топовые слова (наиболее информативные), ибо иначе размерность вектора большая и он слишком разреженный.

По идее тут macro почему-то, но классы в итоговой выборке сбалансированы, так что норм

Топ 100 (BoW):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| ComplementNB | 0.71 | 0.71 | 0.75 | 0.70 |
| LinearSVC | 0.71 | 0.71 | 0.80 | 0.73 |
| Random Forest Classifier | 0.61 | 0.61 | 0.74 | 0.62 |
| Voting Classifier (gauss – linsvc – rfc 100) | 0.70 | 0.70 | 0.76 | 0.71 |

Топ 300 (BoW):

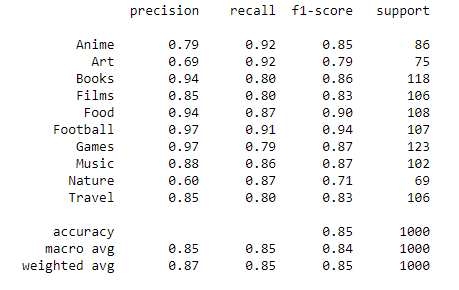
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| ComplementNB | 0.59 | 0.59 | 0.66 | 0.59 |
| LinearSVC | 0.61 | 0.61 | 0.78 | 0.65 |
| Random Forest Classifier | 0.54 | 0.54 | 0.71 | 0.56 |
| Voting Classifier  (gauss – linsvc – rfc 100) | 0.61 | 0.61 | 0.75 | 0.63 |

На топовом алгоритме тестили только лучшую репрезентацию потому что в итоге занимает очень много времени.

Fasttext

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Fasttext (calibrated; обучение после калибровки 34 минуты, калибровка на val.txt 20000 секунд) | 0.85 | 0.85 | 0.85 | 0.84 |

Fasttext classification report:



confusion matrix:

