Emotion Classification at the essay-level

Final results

Женя Егорова, Настя Чижикова

https://github.com/jeka-e/WASSA2022_EMO

Задача

Соревнование **WASSA-2022**, Track 2

Дано: текст длиной от 300 до 800 символов - реакция на новостную статью, содержащую в себе информацию о вреде человеку или группе людей

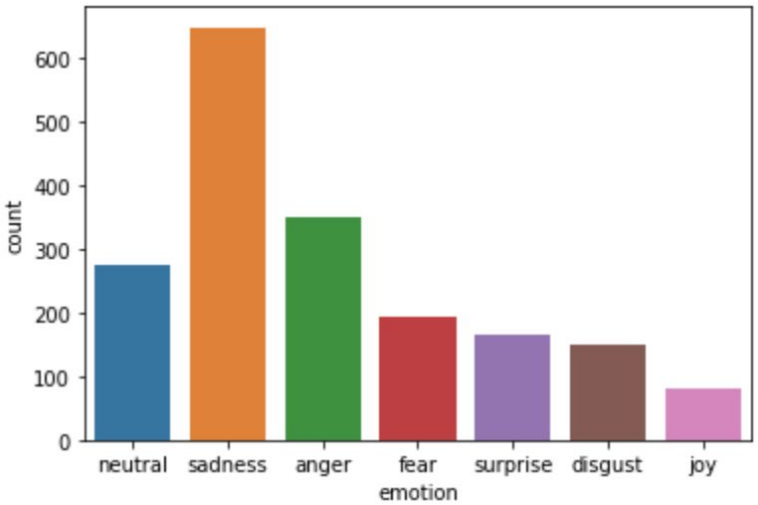
Задача: мультиклассовая классификация эмоции, выраженной в тексте (***sadness, anger, fear, joy, surprise, disgust***), или ее отсутствие (***neutral***)

Дополнительно: некоторая метаинформация об авторе текста (*гендер, уровень образования, раса, возраст, доход*), текст статьи, которой посвящена реакция

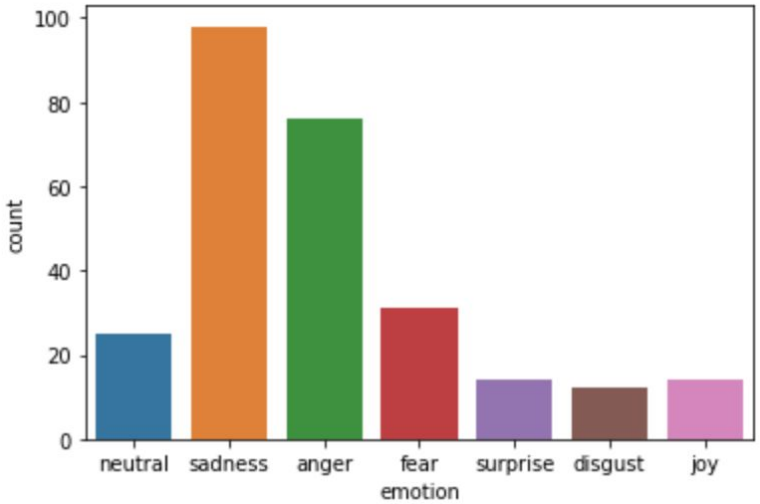
	message_id	response_id	article_id	essay	gender	education	race	age	income	emotion
0	R_1hGrPtWM4SumG0U_1	R_1hGrPtWM4SumG0U	67	really diheartening read immigrants article dr...	1	4	1	33	50000	sadness
1	R_1hGrPtWM4SumG0U_2	R_1hGrPtWM4SumG0U	86	phone lines suicide prevention line surged ele...	1	4	1	33	50000	sadness
2	R_1hGrPtWM4SumG0U_3	R_1hGrPtWM4SumG0U	206	matter heritage able serve country thai herita...	1	4	1	33	50000	neutral

О данных

Train-dev-test: 1860-270-526 texts



Распределение классов в трейне



Распределение классов в dev

Бейзлайн

Мешок слов (TF-IDF) + SVM

Почему такой?

- Используется в нескольких статьях с такой же задачей
- Прост в исполнении
- При этом неплохо работает
- Классика

Задача - обычная классификация, **метрики стандартные** для классификации

	precision	recall	f1-score	support
anger	0.49	0.38	0.43	76
disgust	0.09	0.08	0.09	12
fear	0.68	0.55	0.61	31
joy	0.20	0.07	0.11	14
neutral	0.25	0.24	0.24	25
sadness	0.57	0.79	0.66	98
surprise	0.20	0.14	0.17	14
accuracy			0.49	270
macro avg	0.35	0.32	0.33	270
weighted avg	0.47	0.49	0.47	270

https://github.com/jeka-e/WASSA2022_EMO/blob/main/Notebooks/Baseline.ipynb

ML-подходы

- SVM, Logistic Regression, Random Forest, XGBClassifier на предобученных эмбеддингах FastText и BERT, эмбеддинг предложения как среднее эмбеддингов по словам;
- 3 лучшие модели:

SVM_BERT

	precision	recall	f1-score	support
anger	0.59	0.63	0.61	76
disgust	0.50	0.17	0.25	12
fear	0.68	0.55	0.61	31
joy	0.17	0.07	0.10	14
neutral	0.38	0.40	0.39	25
sadness	0.61	0.74	0.67	98
surprise	0.50	0.29	0.36	14
accuracy			0.57	270
macro avg	0.49	0.41	0.43	270
weighted avg	0.56	0.57	0.56	270

LogReg_BERT

	precision	recall	f1-score	support
anger	0.58	0.59	0.59	76
disgust	0.31	0.33	0.32	12
fear	0.62	0.52	0.56	31
joy	0.33	0.21	0.26	14
neutral	0.40	0.40	0.40	25
sadness	0.67	0.74	0.71	98
surprise	0.45	0.36	0.40	14
accuracy			0.58	270
macro avg	0.48	0.45	0.46	270
weighted avg	0.57	0.58	0.57	270

XGB_BERT

	precision	recall	f1-score	support
anger	0.59	0.59	0.59	76
disgust	0.23	0.25	0.24	12
fear	0.67	0.52	0.58	31
joy	0.60	0.21	0.32	14
neutral	0.36	0.36	0.36	25
sadness	0.65	0.79	0.71	98
surprise	0.44	0.29	0.35	14
accuracy			0.58	270
macro avg	0.51	0.43	0.45	270
weighted avg	0.58	0.58	0.57	270

Простые подходы

Двухслойный перцептрон на предобученных эмбеддингах

FastText:

	precision	recall	f1-score	support
0	0.41	0.36	0.38	25
1	0.42	0.95	0.58	98
2	0.54	0.20	0.29	76
3	0.00	0.00	0.00	31
4	0.00	0.00	0.00	14
5	0.00	0.00	0.00	12
6	0.00	0.00	0.00	14
accuracy			0.43	270
macro avg	0.20	0.22	0.18	270
weighted avg	0.34	0.43	0.33	270

BERT:

	precision	recall	f1-score	support
0	0.41	0.28	0.33	25
1	0.66	0.78	0.71	98
2	0.57	0.62	0.59	76
3	0.70	0.45	0.55	31
4	0.36	0.36	0.36	14
5	0.33	0.42	0.37	12
6	0.33	0.14	0.20	14
accuracy			0.58	270
macro avg	0.48	0.43	0.45	270
weighted avg	0.57	0.58	0.56	270

LSTM-классификатор

Однослойный BiLSTM:

	precision	recall	f1-score	support
0	0.12	0.28	0.17	25
1	0.60	0.60	0.60	98
2	0.35	0.53	0.42	76
3	0.00	0.00	0.00	31
4	0.00	0.00	0.00	14
5	0.00	0.00	0.00	12
6	0.00	0.00	0.00	14
accuracy			0.39	270
macro avg	0.15	0.20	0.17	270
weighted avg	0.33	0.39	0.35	270

+attention-механизм:

	precision	recall	f1-score	support
0	0.08	0.04	0.05	25
1	0.54	0.64	0.59	98
2	0.45	0.37	0.41	76
3	0.32	0.35	0.34	31
4	0.12	0.21	0.15	14
5	0.00	0.00	0.00	12
6	0.00	0.00	0.00	14
accuracy			0.39	270
macro avg	0.22	0.23	0.22	270
weighted avg	0.37	0.39	0.38	270

ML-подходы

- disgust и anger часто путаются между собой и ни с чем другим;
- fear сам по себе неплохо отделяется или же путается с sadness;

Но в целом все достаточно рандомно.

Классификаторы с объединением некоторых классов:

2 classes	**macro F1**	**Acc**
anger vs all	0.71	0.79
sadness vs all	0.76	0.8
fear vs all	0.76	0.93
sadness+fear vs all	0.79	0.8

2 classes	**macro F1**	**Acc**
anger+disgust vs all	0.76	0.8
anger+disgust+joy+ + surprise vs sadness+fear+neutral	0.79	0.8
anger+disgust+joy+ +neutral vs sadness+fear+surprise	0.77	0.77

3 classes	**macro F1**	**Acc**
anger+disgust vs fear+sadness vs joy+surprise+neutral	0.66	0.69
fear vs sadness+neutral vs anger+disgust+joy+ +surprise	0.69	0.73

Выводы из первых попыток

- Опробованные алгоритмы не справляются с сильным дисбалансом классов в данных
- Данных в целом не очень много, модели склонны к переобучению, обучение эмбеддингов с нуля - плохая идея
- Эксперименты с объединением некоторых классов неудачные
- Нужно пробовать сложные модели и экспериментировать с аугментацией данных

Расширение датасета

EmoEvent

https://huggingface.co/datasets/fmplaza/EmoEvent

- 5112 текстов в английском сабсете
- Датасет твитов, говорящих о разных событиях
- Те же классы эмоций, что у нас
- Замусорен хештегами и другими артефактами твитов
- Средняя длина текста ~127 символов

```
tweet: 'The #NotreDameCathedralFire is indeed sad and people call all offered donations humane acts, but please if you have money to donate, donate to humans and help bring food to their tables and affordable education first. What more humane than that? #HumanityFirst'

emotion: 'sadness'
```

Go Emotions

https://huggingface.co/datasets/go_emotions

- Около 50 000 текстов
- Для мультиклассовой и мультилейбловой классификации
- Комментарии с Реддита
- 28 классов эмоций, все наши представлены
- Очень короткие тексты, средняя длина около 50 символов
- В итоге отказались от этого датасета

```
comment: ‘Man I love reddit’

emotion: ‘love’
```

BERT-based модели

- дообучение различных BERT-based моделей с huggingface на нашу задачу: BERT-base, RoBERTa, их дистиллированные версии, а также модели, дообученные на других датасетах на сантимент анализ или классификацию эмоций;
- эксперименты с используемыми для классификации выходами берта: CLS-токен с последнего слоя, конкатенация или агрегирование выходов с n последних слоев;
- эксперименты с гиперпараметрами, предобработкой текстов;

качество на текстах без аугментации: **f1 0.45-0.5**

качество с аугментацией: **f1 0.5-0.53**

Лучшая модель - bert-base-uncased и сконкатенированные выходы с последних слоев

T5 модель

HuggingFace:

T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. Every task – including translation, question answering, and classification – is cast as feeding the model text as input and training it to generate some target text.

	precision	recall	f1-score	support
0	0.80	0.51	0.62	76
1	0.21	0.42	0.28	12
2	0.65	0.71	0.68	31
3	0.70	0.50	0.58	14
4	0.43	0.48	0.45	25
5	0.73	0.79	0.76	98
6	0.50	0.71	0.59	14
accuracy			0.64	270
macro avg	0.57	0.59	0.57	270
weighted avg	0.68	0.64	0.64	270

Результаты T5 для многоклассовой классификации

Обучение бинарных классификаторов

Идея с бинарными классификаторами не получилась

Бинарная классификация на T5

	precision	recall	f1-score	support
joy	0.75	0.21	0.33	14
neutral	0.96	1.00	0.98	256
accuracy			0.96	270
macro avg	0.85	0.61	0.66	270
weighted avg	0.95	0.96	0.94	270

Финальный ансамбль

3 многоклассовых модели (лучшие из имеющихся) делают предсказания: bert-base-uncased, distilroberta, T5

Финальный ответ: голосование по большинству

Результат на dev - лучше, чем у всех первоначальных моделей:

accuracy			0.67	270
macro avg	0.57	0.57	0.57	270
weighted avg	0.68	0.67	0.67	270

Выводы и дальнейшие улучшения

- Мы заняли 7ое место из 14;
- У победителя этого года большой отрыв по качеству - f1 **0.698**,
  2-11 места имеют скор в пределах **0.53-0.58**, наш скор на тесте - **0.544**;
- Максимальный скор при переборе параметров и дообучении готовых моделей колеблется в районе **0.5-0.6**, для получения большего надо менять подход;
- Сложные модели работают лучше, но не идеально, аугментация данных и комбинирование моделей помогают, но не значительно.

Распределение задач

Настя: простые нейросетевые подходы (перцептроны, ЛСТМ, ЛСТМ+attention), расширение датасета, работа с T5

Женя: ML-подходы, BERT-based модели, эксперименты с комбинациями классов

Литература

1. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
2. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi and Puneet Agrawal, https://aclanthology.org/S19-2005.pdf
3. Sboev A., Naumov A., Rybka R., 2021. Data-Driven Model for Emotion Detection in Russian Texts, https://www.sciencedirect.com/science/article/pii/S1877050921013247
4. Polignano, Marco and Basile, Pierpaolo and de Gemmis, Marco and Semeraro, Giovanni, 2019. A comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention - https://dl.acm.org/doi/pdf/10.1145/3314183.3324983