Классификатор негативных комментариев

Цель проекта:
Создание эффективного классификатора негативных комментариев с акцентом на сарказм и иронию.

Милена Камская - обучение и оценка BERT, RoBERTa
Стефания Харская - обзор литературы, LLM, реализация ToXCL

В предыдущих сериях...

Предзащита

В предыдущих сериях...

1 **Обзор литературы**

Прочитали несколько статей, выявили «законодателей трендов» – BERT RoBERTA – и выбрали RoBERTa за её лучшие результаты.

2 **Выбор данных**

"Tweets with Sarcasm and Irony" за способность разграничивать сарказм и иронию, а также за достаточный объем данных для обучения моделей.

3 **Первые эксперименты**

BERT и RoBERTA for Sequence classification
F1 на таргетных классах (ирония и сарказм) около 0.8

Неосновные, но подходы

LLM

GPT-4-turbo и DeepSeek. Подавали им текстовый форматы на вход. По 20 текстов за раз. Однако не всё вышло хорошо:

- GPT-4-turbo начинал придумывать классы новые (“peace”, “war”, “sarcastic”)
- DeepSeek не все тексты брал в расчёт (вместо 20 выдавал 17 лейблов)
- В обоих случаях не смогли прийти к дообучению

ToXCL и Fuzzy rough nearest neighbour methods

- ToXCL: возникли проблемы с доступом к предобученным моделям
- ToXCL: формат фреймворка нам не совсем подошёл, так как распознаёт он скорее социальные предубеждения + много связано с генерацией текста
- FRNN: не получилось воспроизвести
- FRNN: результат по иронии (сарказм не смотрели отдельно) был 0.94 (F1). Решили всё же попробовать сделать что-то своё.

Неосновные, но подходы

LLM

Classification Report GPT EMOJI:

	precision	recall	f1-score	support
figurative	0.03	0.12	0.05	26
irony	0.68	0.49	0.57	139
regular	0.40	0.71	0.51	56
sarcasm	0.79	0.44	0.57	179
accuracy			0.47	400
macro avg	0.48	0.44	0.42	400
weighted avg	0.65	0.47	0.53	400

Classification Report DEEPSEEK EMOJI:

	precision	recall	f1-score	support
figurative	0.04	0.27	0.07	15
irony	0.66	0.49	0.56	136
regular	0.66	0.70	0.68	94
sarcasm	0.77	0.50	0.60	155
accuracy			0.53	400
macro avg	0.53	0.49	0.48	400
weighted avg	0.68	0.53	0.59	400

1. Наши первые результаты. RoBERTa:
   figurative **0.0**, irony **0.8**, regular **1.0**, sarcasm **0.8**
2. Фремворк ToXCL:
   Не разделяли иронию и сарказм. Общее: **78.19**
   Не наши данные. Другой формат вывода, но наш метод такую точность точно должен пробивать.
3. Fuzzy rough nearest neighbour methods:
   На данных с соревнования по иронии (сарказм не смотрели отдельно) был **0.94**. Плюс, это была RoBERTa с дополнением FRNN.

Наши бейзлайны (F1)

Важное о субъективности разметки наших данных

Класс figurative

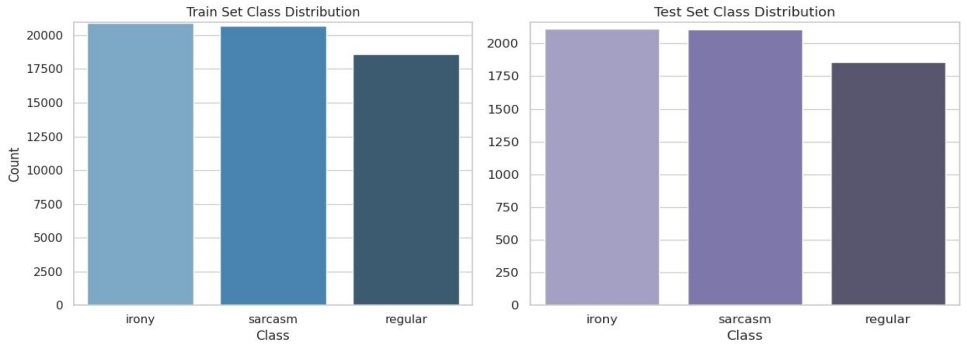
(1) My life today 😊👌 #sarcasm
(2) Women are so.....simple..... #females #complex #irony

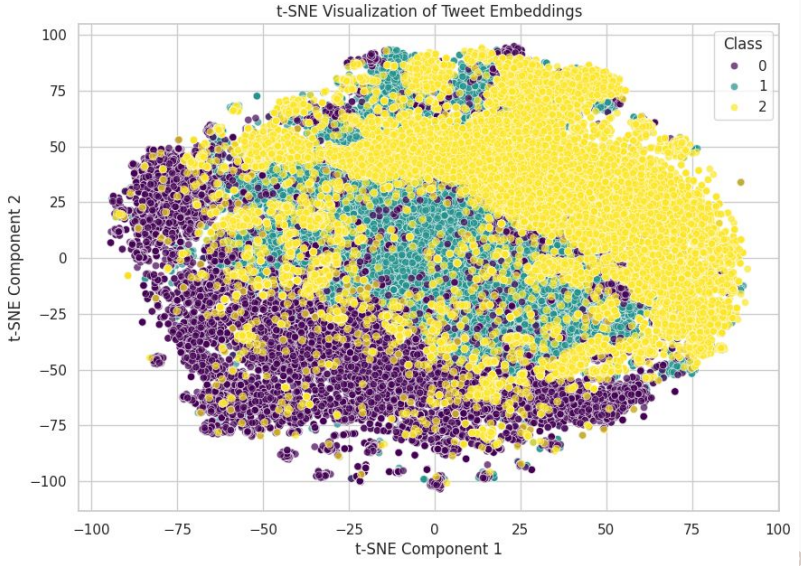
Создаётся ощущение, что разметка оказывается слишком субъективной или даже случайной. Т.к. это не наш целевой класс, мы от него отказались

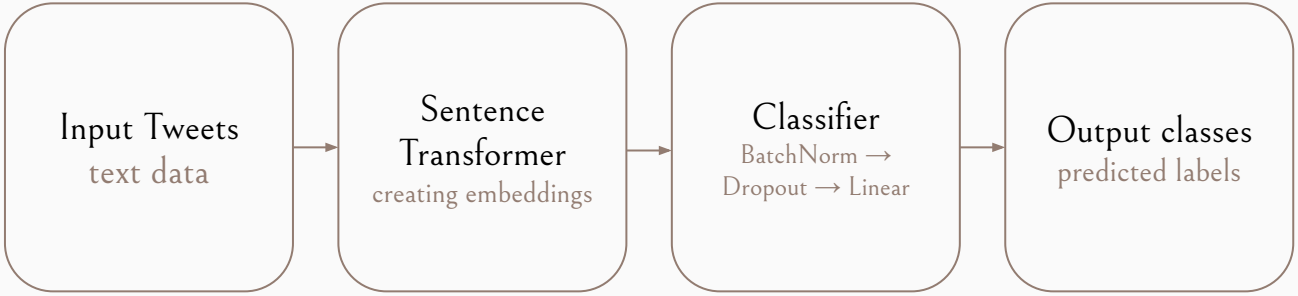
Данные

Tweets with Sarcasm and Irony

- Провели классификацию твитов с использованием модели Sentence Transformer, дообучив её на наших данных.
- Применили метод для устранения классового дисбаланса. Мы использовали взвешенную функцию потерь, чтобы увеличить значимость меньшего класса и уменьшить смещение модели в сторону преобладающего класса.







Мы также используем взвешенную функцию потерь, которая учитывает дисбаланс классов, позволяя модели уделять больше внимания меньшим классам и улучшая общую производительность классификации.

Sentence-RoBERTa

Мечта?

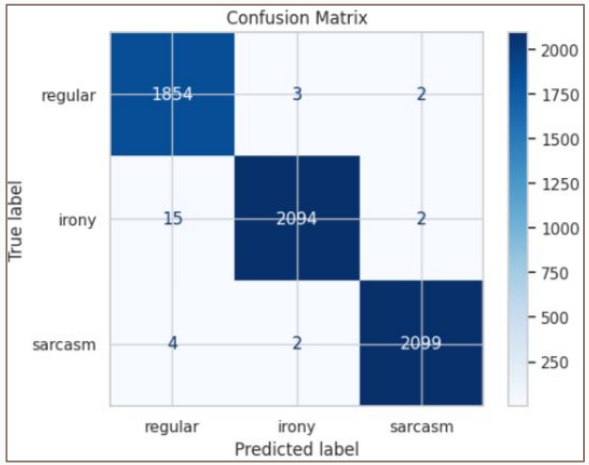
Classification Report:	precision	recall	f1-score	support
regular	1.00	1.00	1.00	1860
irony	1.00	1.00	1.00	2089
sarcasm	1.00	1.00	1.00	2068
accuracy			1.00	6017
macro avg	1.00	1.00	1.00	6017
weighted avg	1.00	1.00	1.00	6017

Результаты

- На тестовой выборке наша модель достигла практически идеальной точности, полноты и F1-меры для всех трех классов, что свидетельствует о её высокой эффективности в классификации твитов.
- Все же были некоторые ошибки, которые мы успешно проанализировали здесь. Общий тренд в анализе ошибок на тестовых данных показывает, что многие ошибки связаны с субъективностью разметки и интерпретацией текста.

Classification Report:

	precision	recall	f1-score	support
regular	0.99	1.00	0.99	1859
irony	1.00	0.99	0.99	2111
sarcasm	1.00	1.00	1.00	2105
accuracy			1.00	6075
macro avg	1.00	1.00	1.00	6075
weighted avg	1.00	1.00	1.00	6075



Спасибо за внимание!

Ждем ваших вопросов!

References

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. Cornell University.

Hoang, N., Do, X., Do, D., Vu, D., & Luu, A. (2024). ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 6460–6472). Association for Computational Linguistics.

Kaminska, O., et al. (2023). Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony. *Information Sciences, 625*, 521–535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. Cornell University.