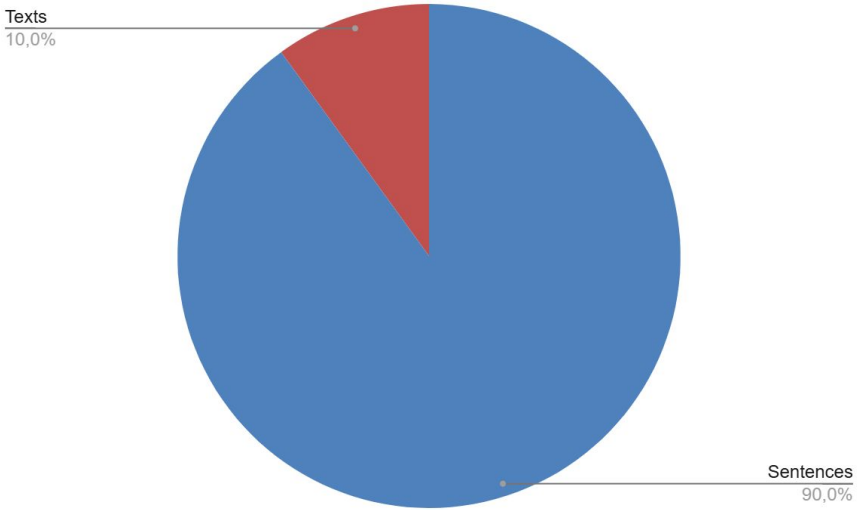
Ensemble as a Variance Reduction Method for Emotion and Sentiment Recognition

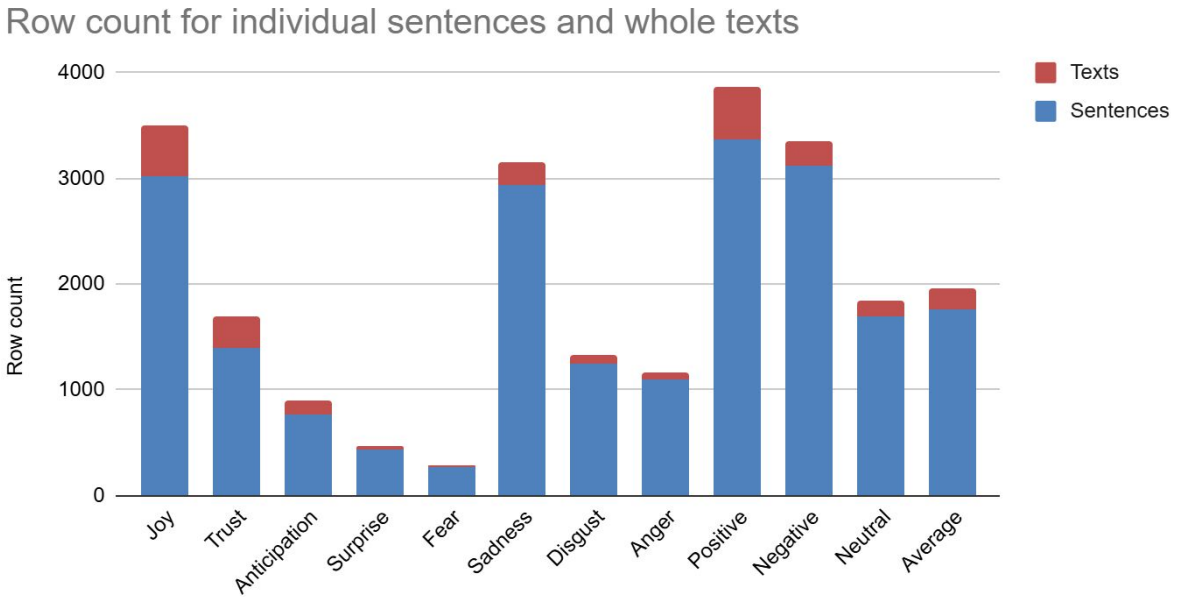
Tomasz Warzecha

The problem to be solved

- 11 classes to predict
  - 8 for emotions
  - 3 for sentiment
- 2 subtasks
  - individual sentences (task A), 6452 examples
  - whole review texts (task B), 717 examples
- At the first glance 7k dataset looks to be sufficient for training



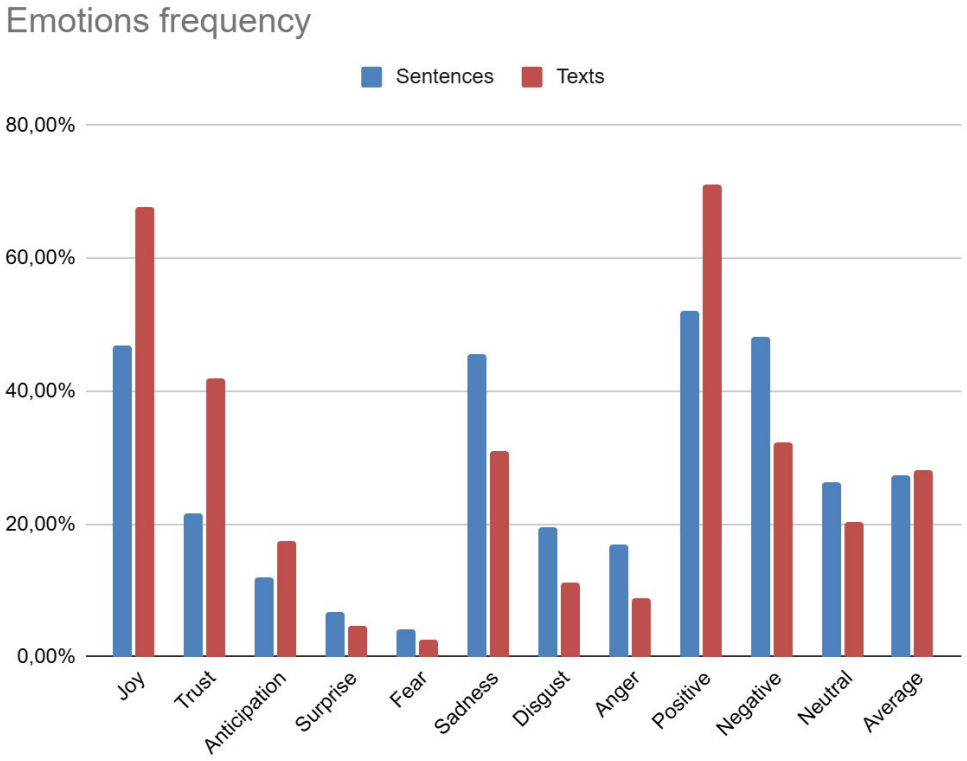
Is there enough data to train?



	Joy	Trust	Anticipation	Surprise	Fear	Sadness	Disgust	Anger	Positive	Negative	Neutral	Average
Sentences	3020	1391	771	433	265	2933	1256	1098	3363	3115	1699	1759
Texts	486	300	125	33	19	222	81	64	510	231	146	202

Everything is going to be alright!

- Differences in emotion distribution between whole texts and individual sentences
- No matter how bad it was, at the and everything is going to be alright!
- Will the training be alright?



What to expect from validation data?

Having 0.9/0.1 train/validation split, there is not enough examples to assess the quality for texts for certain class alone.

The big variance in the validation during training is expected, high level picture if training works.

	Joy	Trust	Anticipation	Surprise	Fear	Sadness	Disgust	Anger	Positive	Negative	Neutral
Sentences	433	265	111	27	17	190	63	49	456	200	137
Texts	42	21	9	7	5	44	24	23	44	43	13

Row count per emotion for validation dataset

What to expect from test data?

Test-A and Test-B datasets are similar

Variance in the final results can be expected, ex. for Fear having +-1 false positive or +-1 true positive may result in even 20pp differences in F1 scores for this class. As macro average is used, that may lead to 1pp final score differences.

	Joy	Trust	Anticipation	Surprise	Fear	Sadness	Disgust	Anger	Positive	Negative	Neutral
Sentences	592	273	151	85	52	575	246	215	659	610	333
Texts	113	70	29	8	4	52	19	15	119	54	34

Expected row count per emotion for test-B dataset

The approach

- Train as a multi-label problem
- Train sentences and whole texts together
- Explore various BERT-like models
- Ensemble to overcome the variability

Data preparation

Each dataset row contains textual data - either individual review sentence or review text as a whole

```
<sentence 1 of the first review>
<sentence 2 of the first review>
...
<sentence N of the first review>
<sentence 1 + sentence 2 + ... + sentence N of the first review>
<sentence 1 of the second review>
```

Add context to each row as **context is important!**

- ***It was unique***. *It was extraordinary and definitely worth seeing!*
- *The conference? Well... Let's say...* ***It was unique.***

Context gluing

Sentence without context:

- ***I'm so happy.***

Previous sentence as a context:

- *I love this product! [SEP]* ***I'm so happy.***

Whole review as a context:

- *I love this product! I'm so happy. 5 stars! [SEP]* ***I'm so happy.***
- ***I'm so happy.*** *[SEP] I love this product! I'm so happy. 5 stars!*
- *I love this product! [SEP]* ***I'm so happy.*** *[SEP] 5 stars!*

Base models selection

- **HerBERT**
- **Polish RoBERTa-v2**
- **XLM-Roberta**
- RemBERT
- mDeBERTa

- Large versions of models were used
- Models have up to 550M parameters and max input length of 512 tokens
- Limited experiments also with XLM-Roberta-XL

Training methods

Fine-tuning each base model:

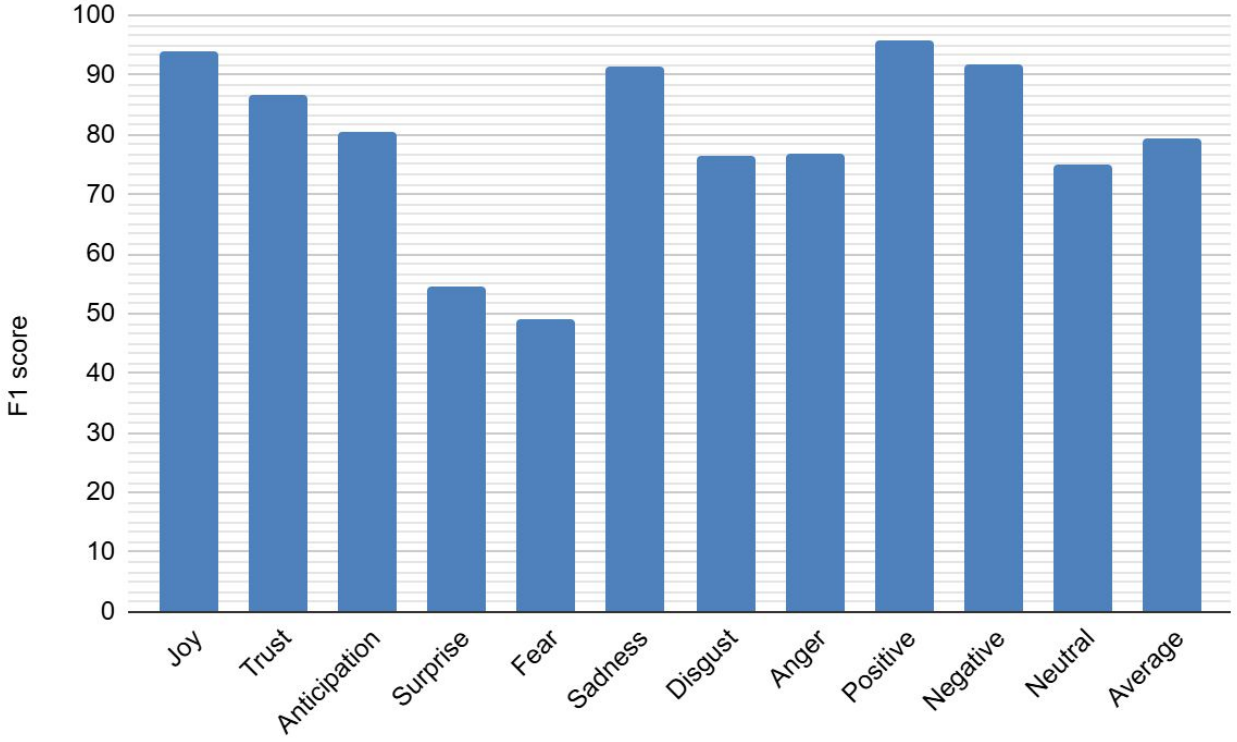
- fine-tuning for 4 different ways of providing context
- fine-tuning without any context
- additional fine-tuning for a specific task

Important training parameters:

- batch size: 32
- learning rate: 2e-5
- optimizer: AdamW (Hugging Face, with default parameters)
- loss: Binary Cross Entropy with weights inversely proportional to class frequency
- learning rate schedule: linear with warmup (0.05)

Training for 20 epochs, best checkpoints selected

Average individual model



Without Fear and Surprise, individual models on average achieve 85% F1

Fear and Surprise struggles from highest variation (stddev of 8.3 and 4.8 pp respectively)

Small ensembles

Table 4: F1 scores for individual sentences subtask, Test-A

	A	B	C	D	avg
XLM-RoBERTa$_{large}$	77.30	75.79	74.10	77.08	**76.07**
HerBERT$_{large}$	77.51	76.68	74.50	77.48	**76.54**
Polish RoBERTa-v2$_{large}$	77.69	78.38	75.80	75.21	**76.77**
3 models average	77.50	76.95	74.80	76.59	**76.46**
3 models ensemble	78.13	78.46	77.72	77.95	**78.07**

Table 5: F1 scores for whole review texts subtask, Test-A

	AT	B	CT	D	O	avg
XLM-RoBERTa$_{large}$	76.55	74.73	75.78	78.81	77.18	**76.61**
HerBERT$_{large}$	77.25	77.36	77.71	78.65	77.71	**77.74**
Polish RoBERTa-v2$_{large}$	78.05	74.04	76.32	78.20	76.32	**76.59**
3 models average	77.28	75.38	76.60	78.55	77.07	**76,98**
3 models ensemble	79.28	79.04	77.84	77.86	79.14	**78.63**

results

- Ensembles showed superior performance over individual models
- 2% better results on average.
- In 8 out of 9 cases three models ensemble gave better results
- In 7 out of 9 cases such three models ensemble achieved better results than the best of the individual models in the group.
- Stdev between small ensembles is 0,31 and 0,72 respectively which is 4 and 2 times less than between individual models

Final ensemble

The final solution consisted of two ensembles - 14 models targeting single sentence subtask and 16 models for whole review text subtask

	Sentences	Whole texts	Test-A score (avg)
Average of individual models	76.61	77.08	76.85
Ensemble of individual models	79.95	81.54	80.75

but only 78.82 Test-B

- **5% gain over individual models' average**, 2.7% over best performing ones
  - *more cautious estimate would be 3-4% and 2% over best performing ones
- Such ensemble achieved best result for Test-A, but **only 78.82 for Test-B**

Ensemble limits

- combining checkpoints from the same training
- ensembles of up to 180 individual results
- majority voting vs logits sum
- checkpoints meeting quality criteria (F1 threshold)
- results did not show significant improvement (although one of them was 79.31)

Summary

- Small models works well for well represented classes
- Even a small ensemble is an easy way to improve results
- To achieve better results best would be to gather a dataset that better represents rare classes

Q/A

Thank you!