Emotion and Sentiment Recognition in Polish Texts Using Large Language Models

A Winning Approach to PolEval 2024

Krzysztof Wróbel

Enelpol

>_ SpeakLeash

/'spix.lɛʂ/ a.k.a Spichlerz

BIELIK

Data

- Polish consumer reviews from 4 domains:
  - Hotels
  - Medicine
  - Products
  - School
- Dataset split:
  - Training: 776 reviews (6,393 sentences)
  - Two test sets: 167 reviews each (~1,250 sentences each)

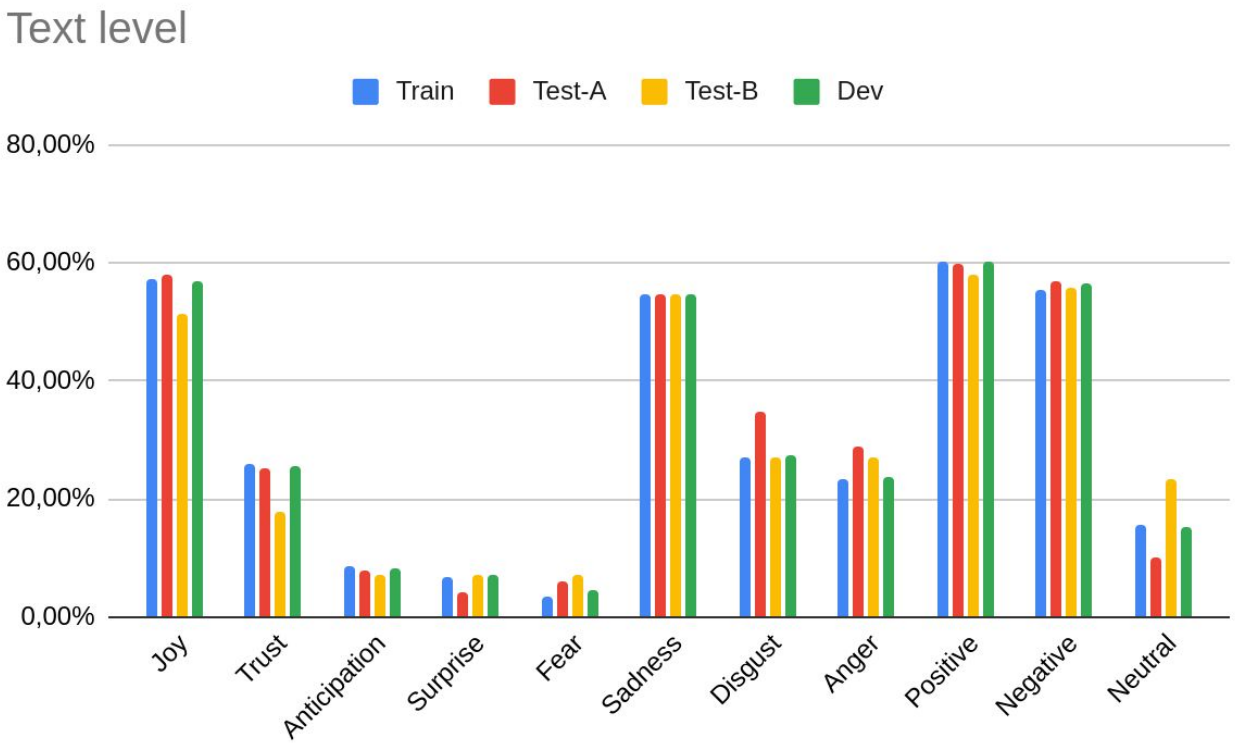
Annotation Schema

- Emotions (Plutchik's wheel):
  - Joy,
  - Trust,
  - Anticipation,
  - Surprise
  - Fear,
  - Sadness,
  - Disgust,
  - Anger
- Sentiment polarity:
  - Positive,
  - Negative,
  - Neutral
- Labels applied at both sentence and review (text) levels

Data Distribution Highlights

- Most Common Labels
  - Joy: ~48% (sentences), ~57% (texts)
  - Sadness: ~43% (sentences), ~54% (texts)
  - Positive: ~53% (sentences), ~60% (texts)
- Rare Labels
  - Fear: ~4% (both levels)
  - Surprise: ~6% (both levels)

Random vs Stratified split class distribution



	Test-A	Test-B
Anticipation	**13**	**12**
Surprise	**7**	**12**
Fear	**10**	**12**
Neutral	**17**	**39**

Length of examples

Average token length: ~412; max: 6,043

- 85.44% (663) of samples contain fewer than 512 tokens, making them suitable for processing with standard BERT-style models
- 93.69% (727) of samples are under 1,024 tokens
- 97.42% (756) of samples are under 2,048 tokens
- 99.36% (771) of samples are under 4,096 tokens

Evaluation

$$Final\ score = \frac{F1_{macro}\ sentences + F1_{macro}\ texts}{2}$$

Technical Approach

- Model Selection:
  - Bielik-11B as base model
  - Handles longer sequences effectively
  - Supports Polish language processing
- Training Techniques:
  - Low-Rank Adaptation (LoRA)
  - Supervised Fine-Tuning (SFT)
  - Custom multilabel classification head

Model Architecture

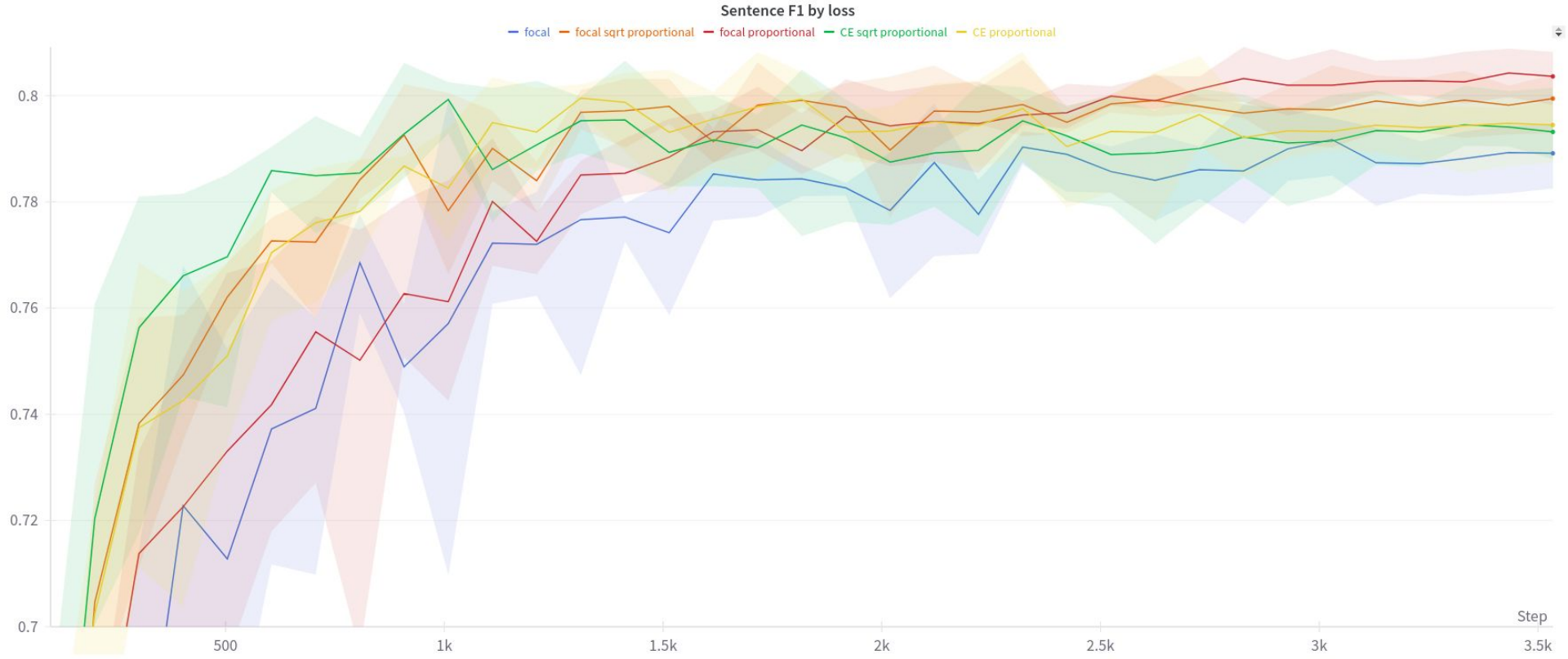
- Custom classification head with 11 binary classifiers
- Processes both sentence and text-level inputs
- Structured prompt format:

```
Lewy kontekst: {left_context}
Tekst do oceny: {text_to_evaluate}
Prawy kontekst: {right_context}
Oznacz tekst do oceny względem emocji i sentymentu: radość, zaufanie,
oczekiwanie, zaskoczenie, strach, smutek, obrzydzenie, gniew, pozytywny,
negatywny, neutralny.
```

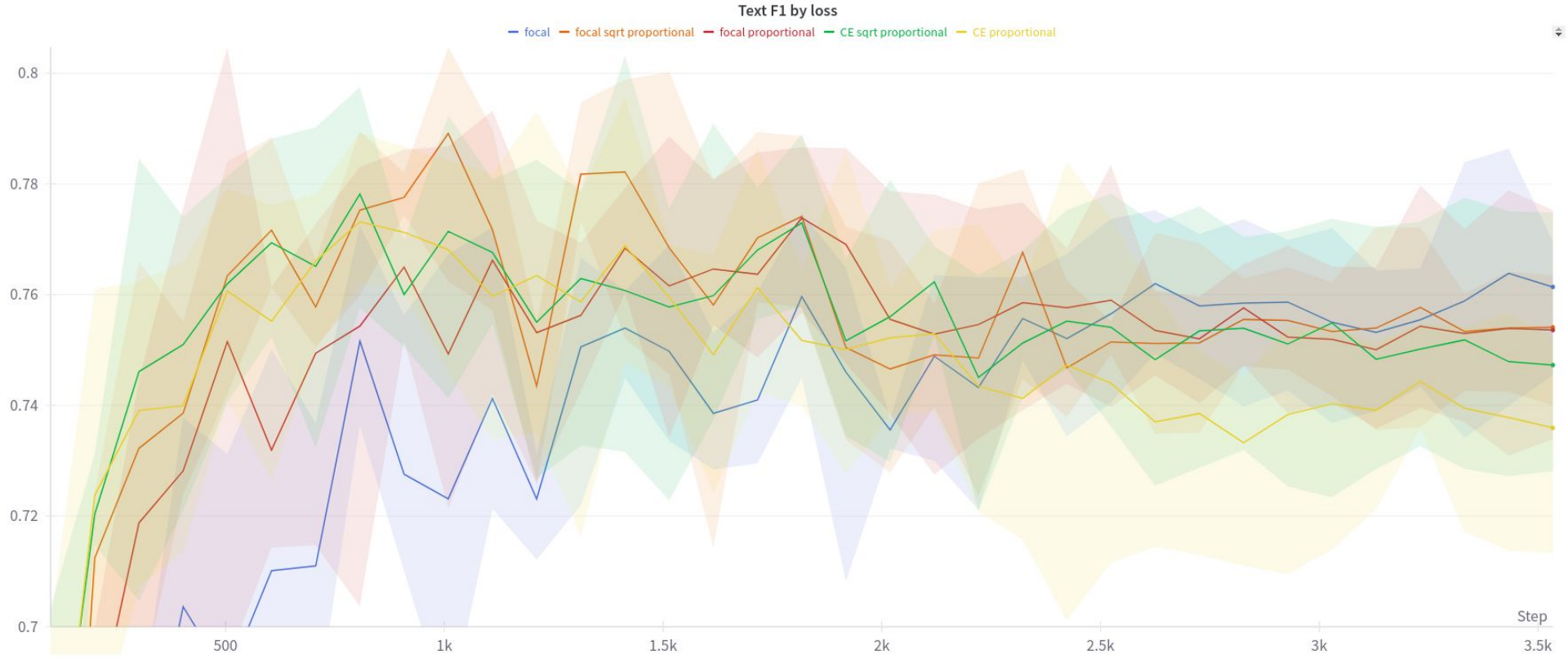
Experimental Setup

- learning rate: 1e-4 to 2e-4
- batch size: 16, 32
- epochs: 3, 5, 8, 10
- LoRA rank: 16, 64
- context window: 0, 1, 2, 3 sentences
- loss: binary cross-entropy, focal loss
- class weights: uniform, proportional to inverse of class frequency, proportional to inverse of square root of class frequency
- weight for text loss: 1.0, 5.0, 10.0
- train on sentence level, train on text level or both
- using additional dataset: XED (Öhman et al. 2020)
- 5 weight initialization seeds
- max sequence length: 1024, 2048, 4096

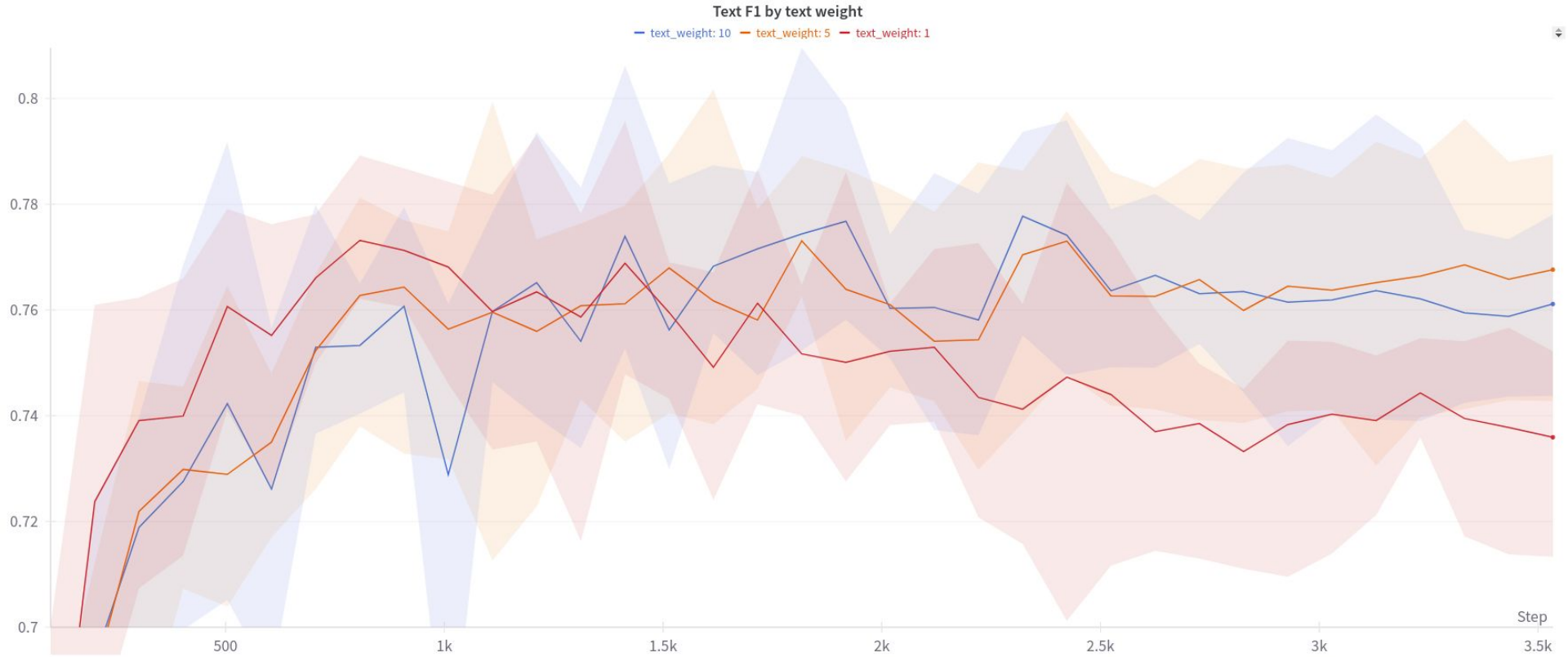
Sentence F1 by loss



Text F1 by loss



Text F1 by text weight



Results

Model	Dev			Test-A			Test-B		
	Sent.	Text	Avg	Sent.	Text	Avg	Sent.	Text	Avg
1 Single unified	81.19	78.81	80.00	80.78	77.86	79.32	80.72	74.27	77.49
2 Sentence specialist	82.31	76.78	79.54	80.68	76.81	78.74	81.06	74.40	77.73
3 Text specialist	79.24	80.17	79.71	78.36	**79.40**	78.88	77.82	**78.48**	78.15
5 Per-label ensemble	83.99	**85.67**	**84.83**	80.84	78.19	79.51	80.80	73.92	77.36
7 Majority voting	82.84	81.06	81.95	**81.62**	77.77	79.70	81.51	74.95	78.23
9 Hybrid ensemble	**84.21**	84.75	84.48	80.71	78.32	79.51	**81.72**	76.07	78.90
10 Test-A optimized	82.84	80.17	81.51	**81.62**	**79.40**	**80.51**	81.51	**78.48**	**79.99**
11 Full data retrain	–	–	–	80.83	77.82	79.32	81.50	76.12	78.81

Competition Results

Rank	Submitter	Affiliation	Entries	test-A scores			test-B scores		
				Sent.	Text	Final	Sent.	Text	Final
1	Krzysztof Wróbel	Enelpol, UJ, AGH	10	81.62	79.40	80.51	81.51	78.48	79.99
2	T	-	196	78.87	81.54	80.20	79.34	79.28	79.31
3	Cezary Kęsik	UW	25	74.94	76.42	75.68	76.66	79.33	77.99
4	Jakub Pokrywka	-	15	78.65	75.93	77.29	79.43	75.77	77.60
5	Paweł Lewkowicz	-	10	74.29	77.73	76.01	77.27	77.20	77.23
6	ka	-	32	75.94	77.47	76.70	76.11	77.76	76.94
7	Cezary Kęsik	University of Warsaw	5	73.62	79.12	76.37	75.94	70.43	73.19
8	Jakub Kosterna	-	4	50.47	28.71	39.59	52.19	28.71	40.45
9	Paweł Cyrta	Metamedia Technologies	5	33.04	32.74	32.89	31.86	34.28	33.07

Finetuned Bielik-1.5B-Instruct: Final 75.16%

Key Findings

- Ensemble approaches showed best performance
- Text specialist models excelled at text-level predictions
- Context window size significantly impacts results
- Model initialization seed affects performance stability
- Balanced handling of rare emotions is crucial

Future Directions

- Test smaller model architectures
- Explore 22-output prediction approach
- Improve handling of rare emotion categories
- Develop better validation set creation methods

Thank you

Source code and models are available at:

https://github.com/enelpol/poleval2024-task2

Contact:

contact@enelpol.com

https://www.linkedin.com/in/wrobelkrzysztof/

Supported by PLGrid ACK Cyfronet AGH (Grant PLG/2024/017168)

Enelpol

>_ SpeakLeash

/'spix.lɛʂ/ a.k.a Spichlerz

BIELIK