Imperial College London

NLP SemEval 2022 Task 4 Subtask 1

Department of Computing

Imperial College of Science, Technology and Medicine

Patronising and Condescending Language (PCL) Detection

Author:
Abhivir Singh

Supervisor:
Nuri Cingillioglu

March 4, 2026

Submitted in partial fulfillment of the requirements for the MEng Computing (AI and Machine Learning) of Imperial College London

Leaderboard Name: PCL-Hunter
Repository: `https://github.com/abhivir-42/PCL-Detection`

Contents

Chapter 1

Critical Paper Review

We review "Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities" by Pérez-Almendros, Espinosa-Anke, and Schockaert (COLING 2020) (1).

1.1 Primary Contributions (Q1)

1. **Novel dataset**: 10,637 paragraphs from news stories across 20 countries, annotated for patronising language directed at vulnerable communities.

2. **PCL taxonomy**: Seven PCL categories organised under three groups: *The Saviour* (unbalanced power, shallow solution), *The Expert* (presupposition, authority voice), and *The Poet* (compassion, metaphor, the poorer the merrier).

3. **Baseline experiments**: Binary detection (Task 1) and multi-label categorisation (Task 2) using SVM, BiLSTM, and transformers. RoBERTa achieves best F1 of 70.68 on Task 1.

1.2 Technical Strengths (Q2)

1. **Novel task framing**: Identifies an underexplored area in harmful language detection – language that is patronising rather than overtly hostile, which is "generally used unconsciously and with good intentions."

2. **Sound annotation scheme**: Two-step process with a 3-point scale per annotator, aggregated to 5 points. A third annotator resolves total disagreements. Inter-annotator agreement reported honestly: 41% Kappa overall, 61% excluding borderline cases.

3. **Informative error analysis**: Tables 5 and 6 provide concrete failure examples, identifying that categories requiring world knowledge (e.g., Metaphor, F1=43.4) are hardest to detect.

1.3 Key Weaknesses (Q3)

1. **Class imbalance unaddressed**: Only 9.4% of paragraphs are positive, yet no oversampling, class weighting, or stratified splitting is discussed. No confidence intervals across the 10 folds.

2. **No ablation or feature analysis**: Several models are benchmarked but there is no investigation into what linguistic features drive predictions. Claims about indicator words ("us," "they," "help") lack supporting feature importance analysis.

3. **Span annotations unused**: Span-level annotations are collected via BRAT but never evaluated – Task 2 is treated as paragraph-level classification only.

4. **Limited annotator bias discussion**: All three annotators share similar backgrounds. The keyword-based paragraph selection also misses paragraphs referencing vulnerable communities through indirect language.

Chapter 2

Exploratory Data Analysis

2.1 Technique 1: Class Distribution and Community Keyword Analysis

2.1.1 Visual Evidence

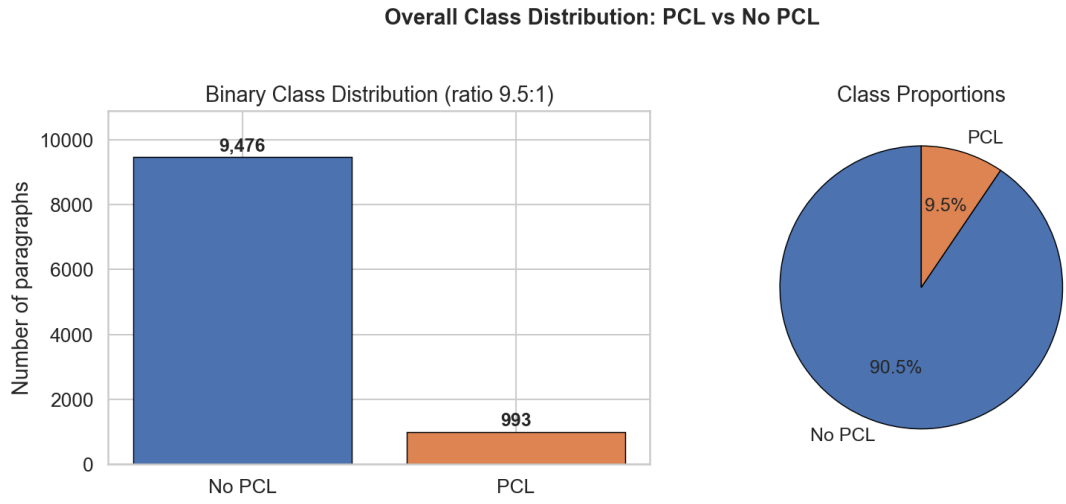

Figure 2.1: Binary class distribution: 9,476 No PCL (90.5%) vs 993 PCL (9.5%).

2.1.2 Analysis

The dataset has a 9.5:1 negative-to-positive imbalance (Figure 2.1). PCL rates vary 6× across community keywords (Figure 2.2): communities relating to economic vulnerability (*homeless*, *poor-families*, *in-need*) have rates around 16%, while *immigrant* and *migrant* are below 3.5%.

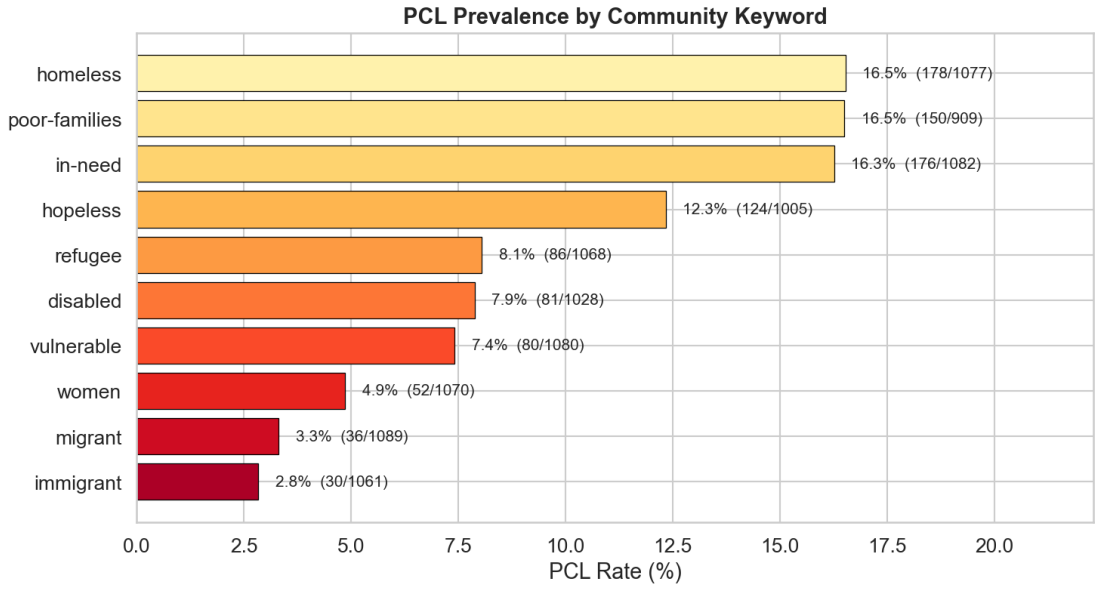

Figure 2.2: PCL rates by keyword: from 2.8% (immigrant) to 16.5% (homeless).

2.1.3 Impact on Approach

The 9.5:1 imbalance rules out standard cross-entropy. The baseline discards ∼85% of negatives via 2:1 downsampling; we instead use focal loss to train on all data. The 6× variation in keyword PCL rates motivates our community-aware input representation (Section 3.1.1).

2.2 Technique 2: Text Length Distribution

2.2.1 Visual Evidence

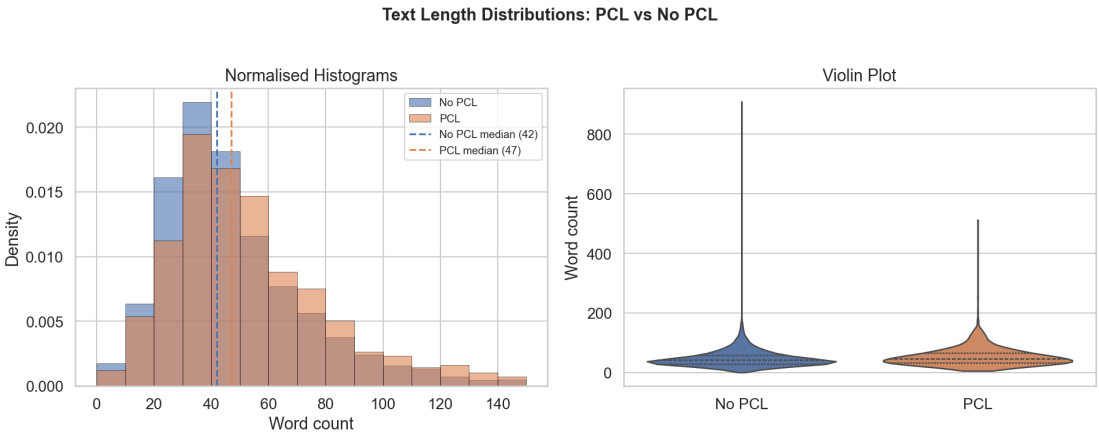

Figure 2.3: Text length distributions: PCL paragraphs are slightly longer (median 47 vs 42 words).

2.2.2 Analysis

PCL paragraphs have a higher median word count (47 vs 42) and a heavier right tail (95th percentile: 113 vs 100 words). This is intuitive: patronising language often involves elaborate justifications or excessive framing.

2.2.3 Impact on Approach

The 95th percentile is 102 words. Applying a sub-word expansion factor of $\sim$1.3$\times$ gives $\sim$133 tokens. We set `max_length=256` to capture 99%+ of paragraphs.

Chapter 3

Proposed Approach

3.1 Approach Description

We propose **Context-Enriched DeBERTa with Focal Loss**, with three components beyond the RoBERTa-base baseline.

3.1.1 Component 1: Community-Aware Input

We prepend a community context prefix to each paragraph:

`[Community: {keyword}] {paragraph text}`

This enables DeBERTa's attention mechanism to learn community-specific PCL patterns, motivated by the 6× variation in PCL rates across keywords (Section 2.1).

3.1.2 Component 2: Focal Loss

We replace the baseline's downsampling + cross-entropy with focal loss (3):

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{3.1}$$

where $\alpha = 0.75$ up-weights the positive class and $\gamma = 2.0$ down-weights easy examples. This trains on all 8,375 examples rather than discarding ∼6,000 negatives.

3.1.3 Component 3: Threshold Optimisation

After training, we sweep thresholds from 0.1 to 0.9 on the dev set, selecting the one that maximises F1 of the positive class.

3.1.4 Model Backbone

We use `microsoft/deberta-v3-base` (2), which offers disentangled attention and ELECTRA-style pre-training, providing advantages over `roberta-base` (4) on NLU benchmarks.

3.1.5 Training Configuration

Parameter	Value
Model	`microsoft/deberta-v3-base`
Max sequence length	256 tokens
Batch size	16 ($\times$2 gradient accumulation = 32 effective)
Learning rate	2×10^{-5} with linear warmup (10%) + decay
Epochs	5 (early stopping, patience = 2)
Optimiser	AdamW (weight decay = 0.01)
Focal loss α / γ	0.75 / 2.0

3.2 Rationale and Expected Outcome

Our approach addresses three limitations of the baseline: (1) undertrained model (1 epoch $\rightarrow$ 5 epochs with early stopping), (2) data wastage from downsampling ($\rightarrow$ focal loss on full dataset), and (3) suboptimal threshold ($\rightarrow$ systematic F1-maximising search). Based on SemEval 2022 results (5), where similar approaches achieved F1 $\sim$0.55–0.65, we expect to comfortably exceed the baseline.

Chapter 4

Evaluation and Error Analysis

4.1 Global Evaluation

Our model achieves F1 = 0.59 on the official dev set, a **22.8% relative improvement** over the baseline (F1 = 0.48). The optimal threshold is 0.38.

Table 4.1: Results on the official dev set.

Model	F1	Precision	Recall	Threshold
Baseline (RoBERTa-base)	0.480	–	–	0.50
Our model	**0.590**	0.520	0.680	0.38

Predictions are submitted as `dev.txt` (2,094 lines) and `test.txt` (3,832 lines). The model favours recall (0.68) over precision (0.52), which is preferable for PCL detection: missing patronising language is worse than over-flagging borderline cases.

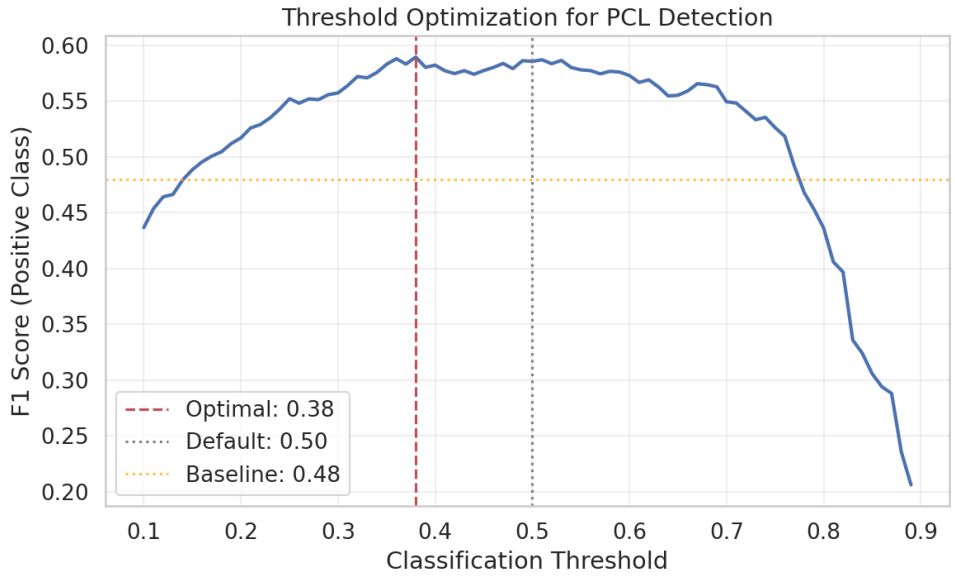

Figure 4.1: F1 vs threshold. Optimal = 0.38, yielding F1 = 0.59.

4.2 Error Analysis

The model produces 124 false positives and 64 false negatives.
False positives cluster around paragraphs that describe aid programmes without actually being patronising. The most confident FP ($p = 0.94$) discusses the Sermon on the Mount and charitable giving – the model triggers on religious/charitable framing co-occurring with a vulnerable community keyword.
False negatives involve subtle PCL requiring world knowledge. The most confident FN ($p = 0.04$) describes disability internship programmes with an original label of 3 (strong PCL). The model fails to recognise the implicit presupposition that disabled people need special "opportunities" framed as generosity.

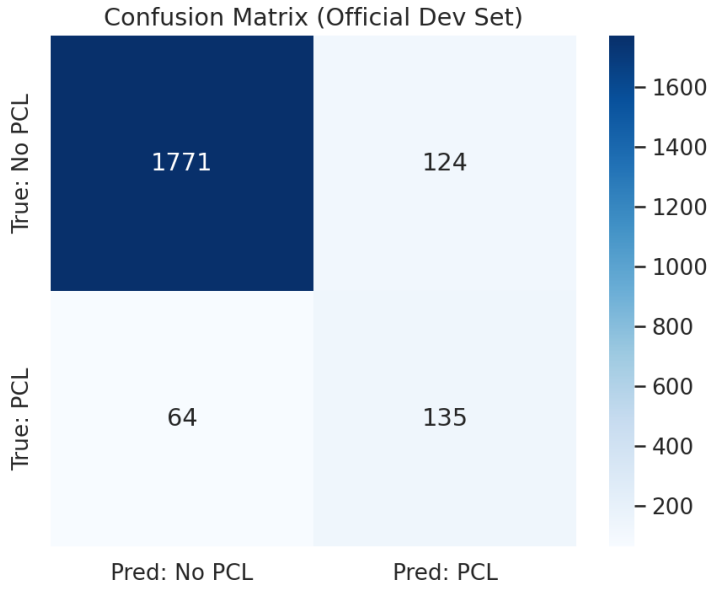

Figure 4.2: Confusion matrix at threshold 0.38.

4.3 Local Evaluation: Per-Keyword Performance

Performance correlates with PCL base rates: *in-need* (F1=0.71, 16.3% PCL rate) vs *immigrant* (F1=0.33, 2.8% rate). The model sees very few positive examples for *immigrant* during training. *Women* (F1=0.44) is notably weak despite a moderate PCL rate, suggesting PCL directed at women takes linguistically distinct forms (e.g., benevolent sexism) that differ from the charity/aid framing dominant in other categories.

Table 4.2: Per-keyword F1, precision, and recall.

Keyword	F1	Precision	Recall	Count
in-need	0.706	0.577	0.909	226
migrant	0.667	0.750	0.600	207
vulnerable	0.640	0.533	0.800	209
poor-families	0.602	0.556	0.658	190
homeless	0.563	0.476	0.690	212
hopeless	0.561	0.516	0.615	217
refugee	0.552	0.500	0.615	188
disabled	0.533	0.500	0.571	194
women	0.438	0.389	0.500	233
immigrant	0.333	0.400	0.286	218

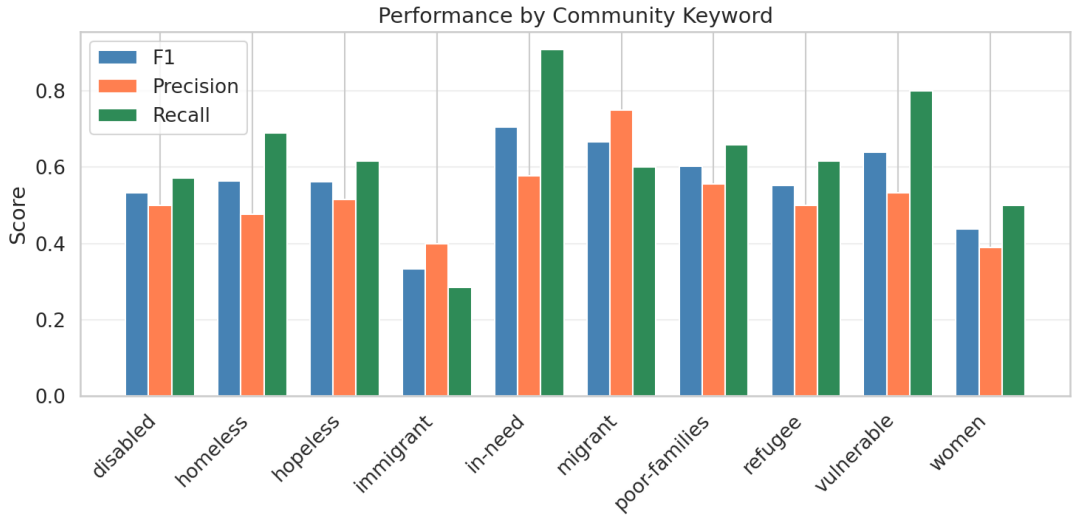

Figure 4.3: F1, precision, and recall by community keyword.

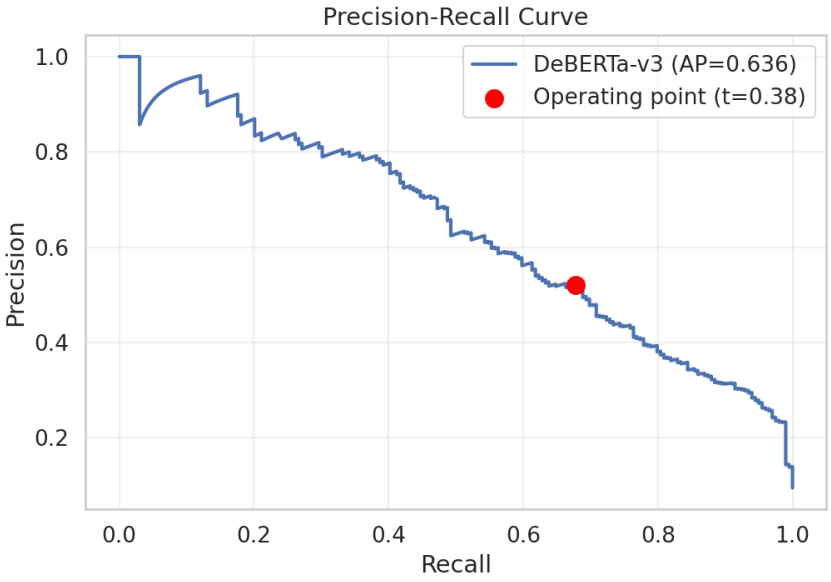

Figure 4.4: Precision–recall curve.

Bibliography

[1] Pérez-Almendros, C., Espinosa-Anke, L., & Schockaert, S. (2020). Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. *Proceedings of COLING 2020*, 5891–5902. pages 1

[2] He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Proceedings of ICLR 2023*. pages 6

[3] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *Proceedings of ICCV*, 2980–2988. pages 6

[4] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*. pages 6

[5] Pérez-Almendros, C., Espinosa-Anke, L., & Schockaert, S. (2022). SemEval-2022 Task 4: Patronizing and Condescending Language Detection. *Proceedings of SemEval-2022*, 298–307. pages 7