Natural Language Processing Coursework

Jakub Cabala (jCabala)

Coursework Repository

February 2026

0 Introduction

This report is part of the Natural Language Processing coursework at Imperial College London. The corresponding repository containing all necessary artifacts can be found at: `https://github.com/jCabala/ICL_NLP_Coursework`

For the purposes of leaderboard identification, the shortname I want to use is **jCabala**.

This report and repository address each coursework exercise as follows:

- **Exercise 1: Critical Paper Review**
  Covered in Section 1.
- **Exercise 2: Exploratory Data Analysis**
  Addressed in Section 2 and in the repository notebook (EDA.ipynb).
- **Exercise 3: Describe your proposed approach**
  Presented in Section 3.
- **Exercise 4: Model Training**
  Implemented in the repository notebook BestModel.ipynb.
- **Exercise 5.1: Global Evaluation**
  Contained in the repository files dev.txt and test.txt.
- **Exercise 5.2: Local Evaluation**
  Written in Section 4.
- **Exercise 6: Coursework Submission**
  Hopefully, covered by the report and repository.

1 Critical Paper Review

This paper makes three main contributions. First, it introduces "Don't Patronize Me!", a new expert-annotated dataset for studying patronizing and condescending language (PCL) toward vulnerable communities, containing over 10,000 news paragraphs with both paragraph-level and span-level labels. Second, it proposes a clear taxonomy of PCL with seven categories grouped into three broader types (Saviour, Expert, Poet), enabling more detailed analysis than simple binary harmful-language tasks. Third, it provides baseline experiments using traditional and transformer-based models, showing that detecting PCL is possible but challenging and that some categories are easier for models to recognise than others.

The primary technical strength of the paper lies in how it turns a vague and subjective phenomenon into a clearly defined NLP task through careful labelling and expert annotation. Another strength is that the experimental study not only reports results but also analyses which types of PCL are harder for models to detect. The authors show that categories with clear linguistic cues are easier to identify, while those involving metaphor, presuppositions, or shallow solutions are more difficult because they require broader context and real-world understanding. Overall, the paper is significant for the NLP community because it introduces a new benchmark for studying subtle forms of harmful language that are not captured by existing datasets.

The paper has several limitations that affect the strength of its empirical claims. Inter-annotator agreement is only moderate ($\kappa \approx 0.41$) [1], indicating that PCL is subjective and difficult to label consistently, which may introduce noise into the dataset. The authors note that agreement improves when borderline cases are removed, but they miss the fact that this comes at a cost of removing more than half of the examples labelled as PCL. The data comes exclusively from English news articles, creating potential bias and making it unclear whether the findings generalise to other settings such as social media. Although the taxonomy is well motivated, it is not fully validated empirically: results show confusion between some categories (e.g., presupposition and authority voice), but the authors do not examine whether this could indicate overlapping categories or the need for a simpler taxonomy. Finally, the dataset is relatively small, which likely contributes to the weaker performance observed for larger architectures and limits the strength of conclusions that can be drawn from the experiments.

2 Exploratory Data Analysis

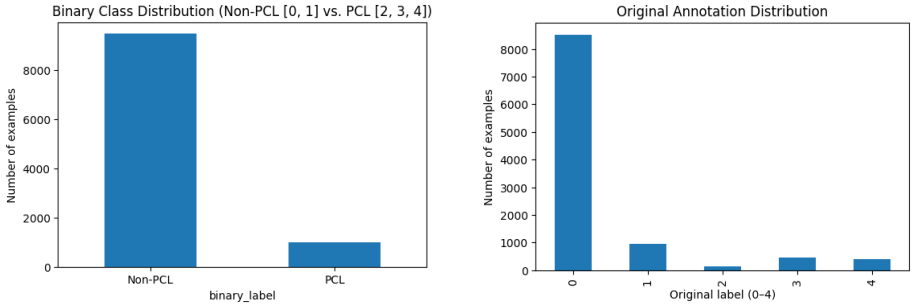
I tried all the techniques listed in the Appendix, including basic statistics, lexical analysis, syntactic and semantic exploration, and data cleaning checks. I ran all of these in a separate notebook (link: [FULL NOTEBOOK LINK HERE]).

Before starting the analysis, I cleaned the text to reduce noise and make the results more reliable. I removed stopwords, HTML tags, extra spaces. I also normalised punctuation, fixed spacing around punctuation marks, and replaced URLs and email addresses with placeholder tokens. For most of the analysis steps - including tokenisation, n-grams and POS tagging - I used the NLTK library.

From everything I explored, below I present two techniques that resulted in findings that seemed most useful for understanding the dataset and for deciding how to approach the PCL classification task.

2.1 Labels Distribution

2.1.1 Visual Evidence



(a) Binary class distribution (PCL vs Non-PCL)

(b) Original label distribution (0–4)

Figure 1: Distribution of labels in the dataset.

2.1.2 Analysis

I visualized a distribution of the dataset across the original annotation labels. The dataset is heavily skewed toward non-patronising language. Most examples fall into labels 0 and 1 (non-PCL), and within this group label 0 dominates by a large margin. Label 1 is already a minority, and the PCL classes (labels 2-4) are much smaller.

As a reminder, the 0-4 labels are derived from how two annotators judged each paragraph (no PCL / borderline / clear PCL). Label 0 means both agreed there is no PCL, label 4 means both agreed it clearly contains PCL, and the intermediate labels reflect different stages of partial agreement.

The main takeaway is that PCL is very rare, and most articles are clearly non-PCL - both annotators explicitly agreed on "no PCL".

2.1.3 Impact Statement

First, I will take extra care to handle the skewed class distribution during training by using techniques such as class weighting, resampling or try to somehow augment the dataset with new examples, so the model does not become biased toward predicting the majority non-PCL class.

Second, because clearly agreed PCL examples are rare, there are probably no obvious lexical patterns to learn. This means the model must most likely capture subtle contextual cues in how language is used. For this reason, I will prioritise contextual models (e.g. transformer-based models) rather than relying on simpler keyword-based approaches.

2.2 Embedding Visualisation

2.2.1 Visual Evidence

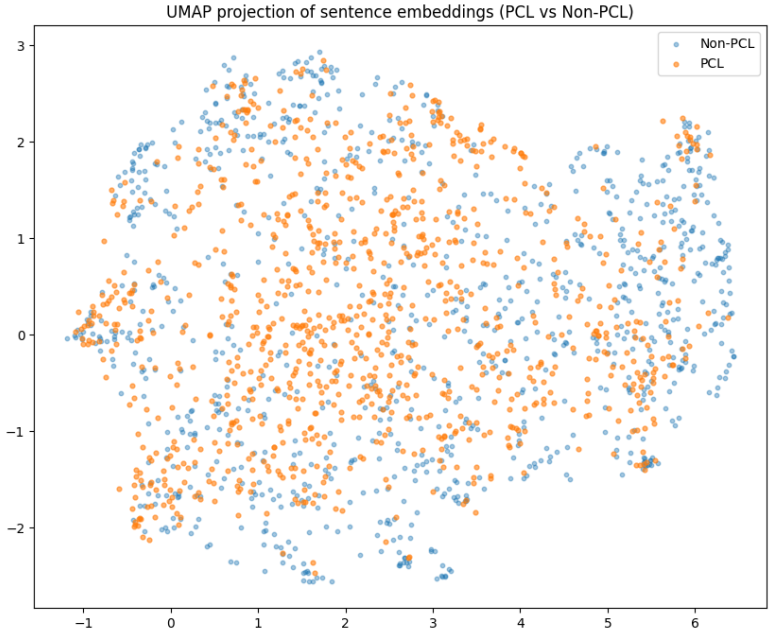


Figure 2: UMAP projection of sentence embeddings (Sentence-Transformers: all-MiniLM-L6-v2), coloured by class (PCL vs Non-PCL).

2.2.2 Analysis

I generated sentence embeddings using a pre-trained transformer model and projected them into two dimensions using UMAP. The plot does not show clearly separated groups. Instead, PCL and non-PCL examples overlap heavily.s. This suggests that the difference between PCL and non-PCL is not driven by simple vocabulary or topic differences but more by how the message is expressed i.e tone and context.

2.2.3 Impact Statement

Because the classes do not form clean clusters, the task does not reduce to a separation problem. This reinforces the earlier finding that PCL most likely depends on context rather than topic or vocabulary.

As a result, I am going to prioritise contextual models that can capture tone and framing and am unlikely to focus on testing separating models such as SVMs.

2.3 Other Findings

Other techniques I explored such as n-gram analysis and POS tagging also did not reveal clear syntactic patterns that consistently distinguish PCL from non-PCL examples. Common words and phrases appeared in both classes, and the overall POS distributions were very similar.

These observations strengthen the earlier findings from the class distribution and UMAP analysis: PCL is not characterised by specific keywords or simple structures. Instead, it depends more on subtle contextual cues, tone, and framing. This aligns with the observations reported in the original dataset paper.

3 Novel Approach

Based on the EDA, I focus on improving the baseline in two areas: data distribution and model choice.

For the data distribution, the baseline downsamples the non-PCL class. Instead, I try other approaches, namely upsampling the minority class and using a weighted loss. The main reason is that downsampling removes a large part of the dataset, while these methods allow the model to use more of the available data.

For the model, I continue using transformer-based architectures. I test models that are expected to perform better than the RoBERTa baseline, such as RoBERTa-large and DeBERTa-v3. Although larger models were tested in the original paper, they were trained on a downsampled dataset. Since my approach keeps more data, better performance might be expected.

I also add a preprocessing step identified during the EDA. This step standardises the text and reduces noise, which should make learning easier for the model.

3.1 Best Result

After testing different combinations, the best result was achieved by the original RoBERTa baseline trained with upsampling and for two epochs instead of one.

RoBERTa-large achieved a similar score but did not clearly outperform the baseline, so the simpler model was selected. This may suggest that the dataset is still too small for larger models to show their full advantage.

4 Local Evaluation

4.1 Error Analysis

4.1.1 Labels

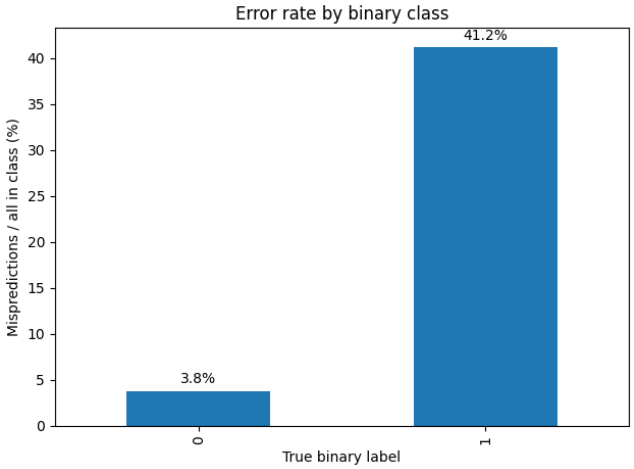


Figure 3: Error rate across binary labels.

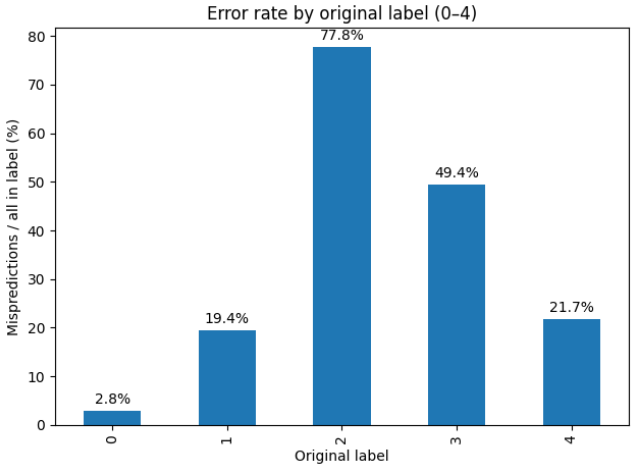


Figure 4: Error rate across original annotation labels (0-3).

Figure 3 clearly shows that the model more frequently fails to detect PCL when it is present (false negatives) than it incorrectly predicts PCL when it is absent (false positives).

Figure 4 shows the error rate across the original annotation labels (0–4). The highest proportion of mispredictions is observed for label 2, which corresponds to cases where both annotators judged the instance as borderline. These examples are inherently ambiguous and therefore particularly difficult to classify.

Label 3, where one annotator marked the instance as borderline and the other as PCL-exhibits the second-highest error rate. In contrast, label 1 (one annotator borderline, the other NO PCL) shows a substantially lower error rate (16.2%). This asymmetry suggests that when the model is uncertain, it tends to default to predicting the absence of PCL.

Overall, the model performs considerably better on the "tail" labels where annotators expressed high certainty. Nevertheless, even in these cases there are more mispredictions on the PCL side than on the non-PCL side (21.7% vs 2.8%). This pattern may indicate a general bias toward predicting the non-PCL class. A likely explanation is the class imbalance in the dataset: although positive instances were upsampled during training, the model may still have learned to favour negative predictions, potentially because repeated examples provide less signal.

4.1.2 Categories

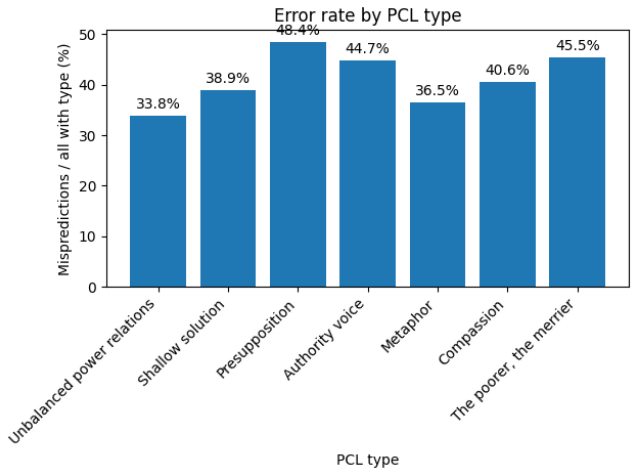


Figure 5: Error rate across PCL categories.

Figure 5 shows tahat the highest error rate is observed for the categories *"Presupposition"* and *"The poorer, the merrier"*. The latter includes texts that portray vulnerable groups as better or stronger because of their hardship. Such cases are often written in a positive and subtle way, which makes them harder for the model to recognise as patronizing.

For example, in one misclassified instance the text emphasises the talent and achievements of disabled people: "She hopes the book ... will show how

talented disabled people can be." At first glance, this reads as simple praise, which likely led the model to predict a non-PCL label.

Overall, we can draw a conclusion that the model struggles with patronizing language expressed through praise rather than clear negative or stereotypical statements. This is consistent with its comparatively better performance on the *Unbalanced power relations* category, where patronizing language is more explicit and therefore easier to detect.

For instance, correctly classified examples often contain direct references to providing assistance or protection. One such example states: "Parents of children who died must get compensation, free medicine must be provided to poor families across UP." Another example describes charitable support: "The former Chelsea star through his foundation gave out toys, bags and clothes to kids in need of a brighter holiday." In both cases, the language clearly frames vulnerable groups as recipients of help.

4.1.3 Keywords

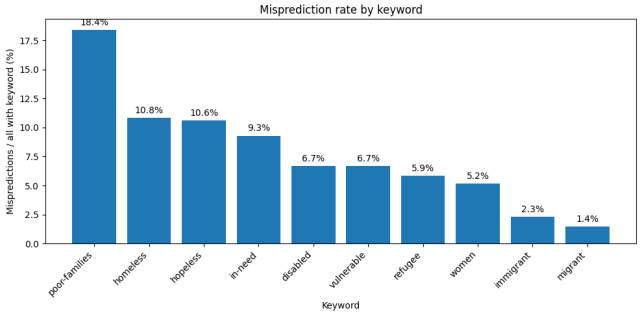


Figure 6: Error rate across keywords.

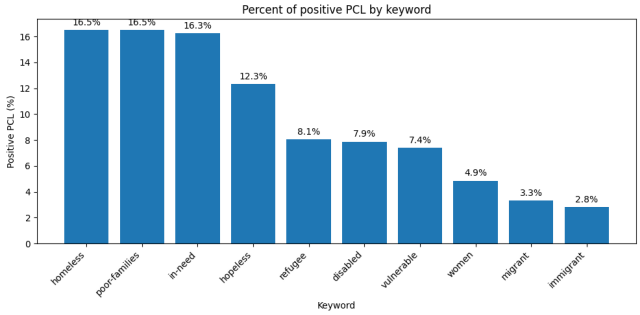


Figure 7: PCL rate accross keywords).

A surprising finding is that the model performs substantially worse on examples containing the keyword "poor-families" compared to other keywords (see

Figure 6).

Initially, I assumed this was due to class imbalance, since this keyword is associated with a relatively high proportion of PCL instances and the model generally tends to predict the non-PCL class. However, Figure 7 suggests that this explanation is incomplete. Other keywords, such as "in-need" and "homeless," have similar proportions of PCL examples, yet the model performs noticeably better on them.

This indicates that the issue might be related to some other interesting properties of text about poor families. A more detailed analysis would be needed to determine the exact cause; however, due to time constraints, I did not investigate it further. I just report it here as an interesting pattern in the model's behaviour.

4.2 Other Local Evaluation - Confusion Matrix Analysis

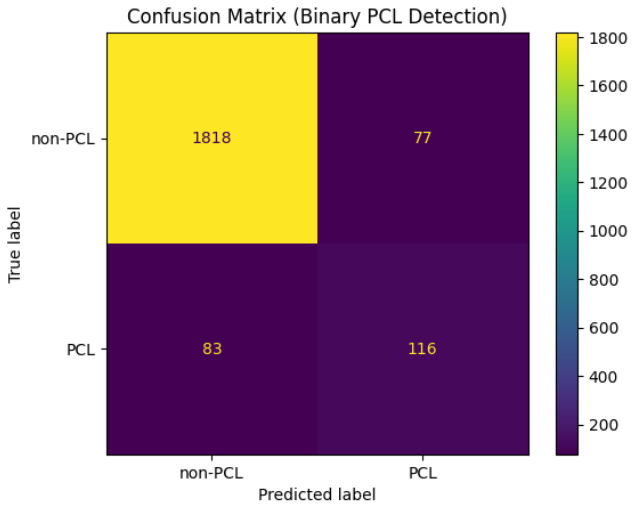


Figure 8: Confusion Matrix.

The confusion matrix shows that the model prefers precision over recall: it is careful not to label text as PCL when it is not, but as a result it misses many real PCL cases. This is in line with our previous findings from the error analysis section. A likely reason is that the dataset is imbalanced, with many more non-PCL examples, which pushes the model and the default decision threshold to favour the majority class.

References

[1] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.