NLP Coursework

Jack Hau
MSc AI
jhh23@ic.ac.uk

Irfaan Kaderbhai
MSc AI
ik620@ic.ac.uk

Kyoya Higashino
MSc AI
kh123@ic.ac.uk

1 Introduction

The task is to develop a binary classification model to predict Patronising and Condescending Language (PCL) in texts, defined as language that ”shows a superior attitude towards others or portrays them in a compassionately patronizing manner” [1]. This presents a notable challenge for natural language processing (NLP) models due to the subtlety and nuanced nature of PCL.

This challenge leverages a specially annotated dataset from the study ”Don’t Patronize Me! Patronizing and Condescending Language Towards Vulnerable Communities” [1]. Specifically, this dataset focuses on PCL use toward vulnerable communities, including Disabled, Homeless, Hopeless, Immigrant, In-need, Migrant, Poor families, Refugee, Vulnerable, and Women.

To build the dataset, extracts from various news media outlets were used, spanning 20 different countries to ensure a diverse and comprehensive set of examples for robust model training and evaluation. Here, PCL severity has been scored from 0 (no PCL) to 4 (clearly contains PCL). For binary classification, 0-1 is considered ”no-PCL”, and 2-4 is considered ”PCL”.

The code and data for this research are available on our GitLab repository: https://gitlab.doc.ic.ac.uk/jhh23/nlp_cw.

2 Data Analysis of Training Data

2.1 Analysis of Class Labels

In the analysis of class labels, there are pronounced class imbalances across multiple categories but limited connection to sentence length. On a macro level, the dataset comprises a significantly smaller subset of PCL texts (approximately 1000 instances) in comparison to no-PCL texts (approximately 9000 instances). This makes results during training heavily skewed toward the majority class.

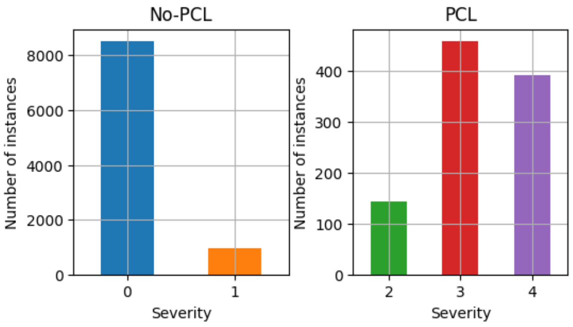

Figure 1: Frequency of PCL Severity

Class imbalances also exist in the PCL and no-PCL subsets themselves, with imbalances favouring more extreme scores as seen in Figure 1. Fortunately, this simplifies binary classification as the majority of instances are clearly containing PCL or no-PCL, removing additional ambiguity that even a human evaluator would find difficult.

The distribution of PCL and non-PCL texts also varies across different categories (keyword and country). While all categories contain mostly no-PCL texts, the ratio between PCL and no-PCL instances greatly differ. When considering texts by their keyword, the PCL distribution is highly varied, with immigrant-related texts having a ratio of 35 and homeless-related texts having a ratio of only 5. This trend suggests there are large biases in media coverage towards using PCL when reporting on certain vulnerable groups. Conversely, this variance is reduced when considering geography, with the largest and smallest ratios observed being 16 and 6, respectively. This underscores that although there are regional variations in PCL use, these differences are comparatively moderate.

On the other hand, sentence length appears to have minimal correlation with class labels, ex-

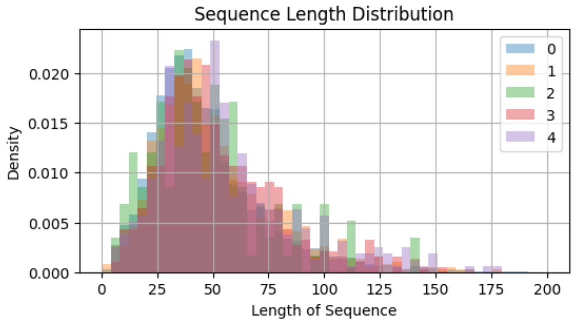

Figure 2: Sequence length distribution according to PCL severity

hibiting similar distributions across categories, as detailed in Figure 2. Although the median sentence length for PCL texts is marginally longer (47 words compared to 42), this minor difference is not considered significant enough to warrant further exploration in this analysis.

2.2 Qualitative Assessment of Dataset

This task of identifying PCL is very challenging, primarily due to the dataset's imbalances as well as the subtle and subjective nature of PCL itself.

As there are significantly fewer positive instances (texts labelled as PCL), the dataset's imbalance can lead to skewed model training, where the model becomes overly proficient at identifying the majority class (non-PCL texts) at the expense of accurately detecting the less represented PCL texts. Consequently, the model might struggle to generalise well on new, unseen data, especially in identifying positive instances.

Another major challenge is the subtlety and often unconscious bias inherent in PCL. Unlike explicitly offensive language, ”the use of PCL in the media is commonly unconscious, subtler and more subjective than the types of discourse.” [1] This subtlety is compounded by the fact that a statement's intent does not always align with its impact.For example, ”Monash doc stands up for women” appears to commend feminism but has arguably patronising undertones. As it could be interpreted as undermining women while casting the doctor in a saviour-like role, the text was scored as ”borderline-case” by expert annotators [1]. While the intent of a message is still a useful feature, its exploitative potential is greatly reduced, forcing models to analyse a variety of abstract features instead of solely language use.

Closely related to subtlety, this task is further complicated by PCL's inherent subjectivity. As highlighted in ”Don't Patronize Me!”, even the expert annotators do not always agree on what constitutes PCL, disagreeing in 13.6% of cases. For example, ”Pharrell Williams thinks women can save the world” was annotated as both ”borderline” and as ”clearly containing PCL”. This subjectivity is further influenced by cultural and societal differences, possibly exacerbating the biases introduced during the data annotation process. Consequently, this variability creates an inconsistent benchmark, impacting the accuracy and reliability of PCL-detecting machine-learning models.

3 Modelling

3.1 Model Description

Instead of RoBERTa, Microsoft's pre-trained Decoding-enhanced BERT with disentangled attention (DeBERTa) [2] model was chosen for its superior disentangled attention mechanism. In disentangled attention, attention weights are computed using content and position vectors, enhancing the model's ability to capture segment dependencies and understand contextual relationships more effectively. Furthermore, DeBERTa-cased was selected for its higher performance potential, able to handle proper nouns and capitalisation, improve on case-sensitive tasks, and retain information about sentence boundaries.

3.2 Training and Exploration Set-up

Before exploring improvements and hyperparameters, several model features require establishment. Firstly, model weights are frozen while a binary classification head is added, containing one fully connected layer with 768 in-features and 2 out-features. Secondly, the tokeniser is customised to use padding and truncation, standardising sequence length while limiting sparsity. Thirdly, an AdamW optimiser is chosen for its popularity and adaptive properties. Lastly, an internal validation set is obtained via a 80/20 split on the original training set, used for technique evaluation and early stopping. As early stopping is employed, the max number of epochs is fixed at 10.

Using this set-up, potential model improvements were investigated first, followed by a hyperparameter tuning process for the learning rate, scheduler, and weight decay. This is because optimal hyperparameters may be unique to applied technique.

4 Investigated Model Improvements

Investigated improvements encompass a variety of techniques, including pre-processing, sampling, and augmentation. To ensure a fair assessment, proposed techniques were tested in parallel, being applied to multiple DeBERTa-base models individually. The results of model variations are presented in Table 1.

Applied Technique	F1-Score	Accuracy
None (DeBERTa-base)	0.519	0.916
Lowercasing	0.526	0.908
Contraction Expansion	0.542	0.905
Correction of Spelling	0.522	0.903
Stopwords Removal	0	0.900
Punctuation Removal	0.508	0.911
Upsampling	0.551	0.900
Downsampling	0.533	0.968
Synonym Replacement	0.538	0.920
Context Replacement	**0.560**	**0.916**

Table 1: Results of all tested models on the unofficial validation set, varying by applied technique.

Among all trials, ”context replacement” is shown to be the most effective at improving performance, followed closely by data sampling methods. Conversely, ”stopwords removal” resulted in a 0 F1-score, performing poorly amongst other ineffective pre-processing techniques.

4.1 Data Pre-Processing

Lower-casing, contraction expansion, spelling correction, removal of stopwords, and removal of punctuation were explored as pre-processing techniques. Overall, these techniques performed relatively poorly, producing lower F1-scores on the internal validation set than the sampling and augmentation techniques. This is because removing punctuation, stopwords, and capital letters can obfuscate the semantic meaning of the text, hindering the model's ability to detect PCL.

4.2 Data Sampling

As the dataset is largely imbalanced, negative and positive instances respectively underwent random downsampling and upsampling. To avoid significant information loss or overfitting from aggressive sampling, a 2-to-1 ratio of negative-to-positive instances was enforced instead of an even distribution. As seen in Table 1, while both sampling approaches improve model performance, downsampling is less effective, likely due to its inherent removal of information. By contrast, upsampling is able to improve label distribution without decreasing training size or diversity.

4.3 Data Augmentation

To augment the data, two unigram replacement techniques were utilised: synonym and contextual. For both, all positive samples were augmented thrice before being added to the training set, effectively quadrupling the number of PCL instances. While synonym replacement exposes the model to similar themes, context replacement focuses on training the model on similar contexts, thereby improving the performance of the DeBERTa model by improving its understanding of linguistic nuance. For this reason, synonym replacement was only moderately effective while context replacement was the the most effective.

5 Hyperparameter Tuning

In lieu of an expensive grid search, a systematic manual search approach was used to tune hyperparameters. These results are shown in Table 2.

LR	WD	Scheduler	F1	Accuracy
10^{-5}	0.01	Linear	56.13	99.57
10^{-5}	0.01	Constant	55.97	91.64
10^{-5}	0.0	Linear	54.92	92.06
10^{-5}	0.0	Constant	54.39	90.69
10^{-5}	0.1	Linear	53.98	90.33

Table 2: Best 5 hyperparameters based on internal validation f1-score. (LR-learning rate, WD-weight decay)

First, the learning rate was explored. Larger learning rates (such as 10^{-3}) were initially tested, but were found to produce F1 scores of 0%, demonstrating that higher learning rates result in overshoots during gradient descent and fail to facilitate model convergence. Although lower rates did not result in 0% F1 scores, their performance was also sub-optimal, likely due to the model getting trapped in local minima and unable to discover more optimal solutions. Through tuning, a balanced learning rate of 10^{-5} was found to produce optimal F1 scores. Following this discovery, the weight decay and scheduler were explored.

Ultimately, it was empirically found that a weight decay of 0.01 produced the best results due to the addition of moderate regularisation and without over-constraining the model. Similarly, empirical results from testing linear, constant, and exponential schedulers showed that the linear scheduler supported superior performance.

6 Model Results

Using context replacement and optimal hyperparameters, the final model was trained on the full training set and evaluated on the official development set, achieving an F1-score of 0.550 and an accuracy of 0.914.

Unlike its significant improvement of the validation F1 score, the final model's performance only slightly increased when tested on the official development set, suggesting that it has slightly overfit to the validation data. This likely stems from the generated contextual replacements being too similar, reducing the dataset's overall diversity while predisposing the model to overfitting. In the future, this can be avoided by generating more diverse examples, tuning augmentation parameters to ensure more diverse context replacements.

6.1 Comparison against Baselines

For comparative analysis, the unmodified DeBERTa-cased model serves as a natural baseline, facilitating the direct assessment of the applied improvements. As an additional baseline, a unigram bag-of-words (BoW) model integrated with a Naive Bayes classifier is chosen to provide better context for comparing advanced NLP techniques against simple, statistically-based models. Along with the final model, these results on the official dev-set are summarised in Table 3.

Model	F1-Score	Accuracy
DeBERTa (custom)	0.550	0.914
DeBERTa (base)	0.540	0.918
BoW (cased)	0.07	0.91

Table 3: Performance of the custom DeBERTa model and two baseline on the official development set.

As the DeBERTa-base model already provides an easy and intuitive comparison to the final model, the BoW model is analysed instead.

6.2 Bag-of-Words and Naive-Bayes

In a unigram BoW model, text is represented by the frequency of individual words, disregarding their order. Consequently, while unigram models offer high simplicity and interpretability, they fail to capture context or linguistic nuance. Vectorised texts are also highly-dimensional and sparse, being the length of the training vocabulary while only populating several elements. However, resultant vectors can easily used as input for models such as a Naive-Bayes classifier. By calculating the label probabilities for each word in the sequence, overall sequence probabilities for multiple labels can be calculated.

6.3 Misclassified Example

To highlight the weaknesses of the BoW model, a false positive from the official development set can be analysed:

”A homeless couple is seen by the roadside along Jalan Tuanku Abdul Rahman”

Below in Table 4 are the calculated probabilities for each word and label. Note that because probabilities are incredibly small, their log probabilities are taken instead.

Word	**Log Probability**	
	No-PCL	**PCL**
along	-8.358	-9.270
by	-5.354	-5.926
couple	-9.141	-9.963
homeless	-6.232	-6.058
is	-4.678	-5.021
roadside	-11.366	-11.062
seen	-8.462	-8.664
the	-2.915	-3.393
Total	**-56.506**	**-59.357**

Table 4: Log probabilities of each word in the misclassified example sequence.

As shown in Table 4, this approach has resulted in a false positive, primarily attributed to the words ”along”, ”by”, and ”couple”. While significantly influencing the text's PCL classification, these words lack substantial meaning as words that obviously do not inherently contain PCL. This underscores the limitations of the BoW model as a purely statistical approach, providing a positive label simply because it has commonly observed some individual words in prior PCL instances.

7 Analysis

7.1 Detecting Highly Patronising Content

As seen in Table 5, the misclassification rates for each of the original PCL labels can be used to highlight model performance at varying levels of PCL. As expected, misclassification rates are low for non-PCL instances, following the majority class seen during training. While error rates are expectedly high for PCL instances, model performance is unexpected for varying levels of positive instances. Previously, it was assumed that stronger PCL samples would be easier to detect, containing straightforward language and less linguistic nuance. Instead, the opposite is observed, where increasing PCL levels result in higher misclassification rates. This is potentially due to an invalid assumption that strong levels of PCL are very straightforward, instead still being very nuanced and contextual.

PCL Level	**0**	**1**	**2**	**3**	**4**
Misclassification (%)	12	9	83	87	92

Table 5: Misclassification by PCL level

7.2 Impact of Input Sequence Length

As seen in Figure 3, the model's f1 scores are fairly similar for sequence lengths in the range 0-125 but have a slight decreasing pattern. There is then a slight increase for words in the range 125-150 before a very large increase for words in the range 150-175. The f1 score for 150-175 is 1, indicating perfect performance. However, it is worth noting that there is a far smaller amount of sequences of this length, making anomalous results more likely. Our model may be better at predicting PCL for longer sequence lengths due to DeBERTa's disentangled attention mechanism that is able to capture long range dependencies. This is particularly useful when trying to understand the nuance and complexity of PCL.

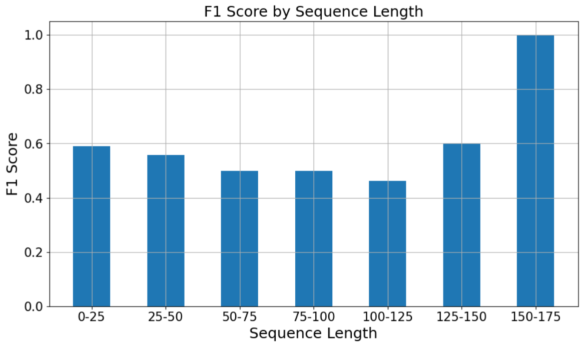

Figure 3: F1 score by sequence length

7.3 Impact of Keyword Categories

As shown in Figure 4, the false positive and false negative rates vary greatly between keyword categories but are closely related to the PCL distribution for each category (See Appendix A Figure 5). In categories where there are more PCL instances, such as 'poor-families', 'homeless', 'hopeless', there is a high false negative rate. This is due to the model's poor PCL detection, increasing false negative rates simply because there are more true positive instances to misclassify. Vice versa, these categories also have a low false positive rates, having proportionally less true negatives to misclassify. Similarly, the impact of PCL distribution also extends to categories with low PCL distribution, as seen in "immigrant" and "migrant". In these cases, the opposite is observed, having high false positive rates and low false negative rates.

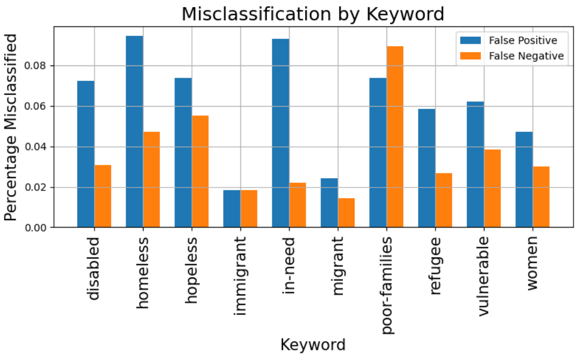

Figure 4: Misclassification by keyword category

8 Conclusion

In conclusion, the final custom DeBERTa model successfully outperformed the RoBERTa-base baseline model, achieving an F1-score of 0.54 on the official dev-set (improving by 0.06). While several techniques were tested, contextual replacement with upsampling proved to be the most effective. Furthermore, whilst model performance is not greatly affected by sequence length, it is highly reliant on the keyword and the level of PCL.

For future work, additional methods can be implemented to increase the model's overall performance. The first approach is to combine multiple data preprocessing and augmentation techniques and rigorously experiment on the collective effects. Lastly, ensemble methods can be employed to combine the predictions of multiple models to improve performance and generalisation.

References

[1] Carla Pérez, Almendros Luis, and Espinosa-Anke Steven Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. pages 5891–5902.

[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020.

A Appendices

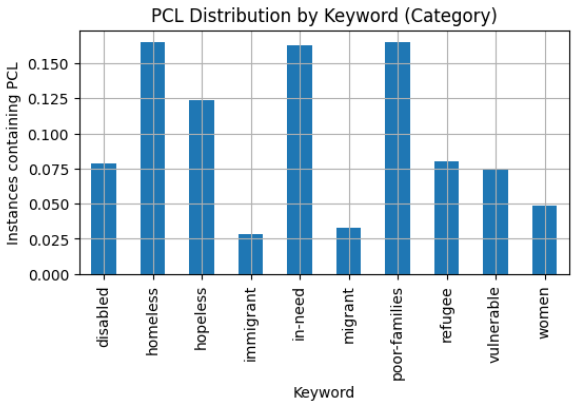

Figure 5: PCL distribution by category. Percentages are given in terms of the number of total instances in each category.