Detecting Patronising and Condescending Language Using DeBERTa

Lisa Faloughi
lf524@ic.ac.uk

Cesare Bergossi
cb624@ic.ac.uk

Oliver Shakespeare
ols24@ic.ac.uk

Abstract

Detecting patronising and condescending language (PCL) is a nuanced challenge in NLP, as it relies on subtle phrasing rather than overtly offensive content. We improve PCL detection using DeBERTa-v3-Small, balancing efficiency and performance. Our approach incorporates data augmentation, class balancing, and error analysis to mitigate dataset bias and enhance generalisation. Comparing against statistical baselines, we find that explicit patronisation is easier to detect, while subtle cases, such as positive framing or indirect condescension, remain challenging. Performance varies across social categories, reflecting biases in media representation. Despite achieving an F1-score of 0.52 on the dev set, distinguishing borderline cases remains difficult, highlighting the need for further refinements. All code and resources are available at https://gitlab.doc.ic.ac.uk/ols24/nlp_cw.

1 Introduction

Detecting patronising and condescending language (PCL) presents unique challenges in natural language processing. Unlike explicit hate speech or toxic comments, PCL is often subtle, relying on exaggerated pity, indirect framing, or well-intentioned but condescending phrasing. This nuance makes it difficult for both humans and machine learning models to consistently identify.

In this work, we build on existing approaches by leveraging transformer-based models to correctly classify PCL language, as shown in the original challenge paper (Perez Almendros and Schockaert, 2022), ultimately selecting DeBERTa-v3-Small for its balance of efficiency and performance. Through targeted data augmentation and class balancing, we improved the model's robustness while maintaining a fair representation of minority-class examples. Our methodology includes a comparative analysis with statistical baselines, error analysis to identify key failure cases, and an evaluation of performance across different categories of patronisation.

2 Data Analysis of Training Data

2.1 Analysis of Class Labels

A strong class imbalance is present in the dataset (Figure 1), with non-PCL instances greatly outnumbered PCL examples. This imbalance poses a challenge for model training, as classifiers tend to favor the majority class, making it harder to correctly identify patronising language.

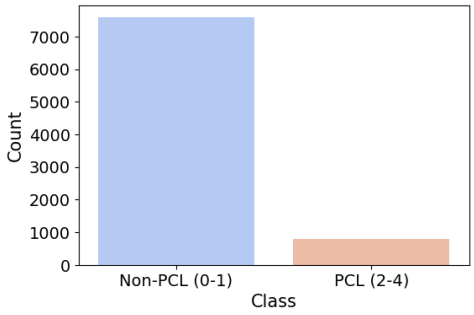

Figure 1: Binary Class Distribution: PCL vs. Non-PCL

To explore how PCL manifests across contexts, we examined its distribution within topic categories (Figure 2). While patronising language is consistently a minority across all groups, certain categories like *homeless* and *poor families* contain a higher proportion of PCL cases, suggesting it is more prevalent in topics where individuals are often framed as vulnerable or in need.

Additionally, we analysed text length distributions for both classes (Figure 3). Both follow a similar skewed distribution, with the majority of texts containing fewer than 100 words. However, PCL texts tend to be slightly longer on average, suggesting that patronising statements may

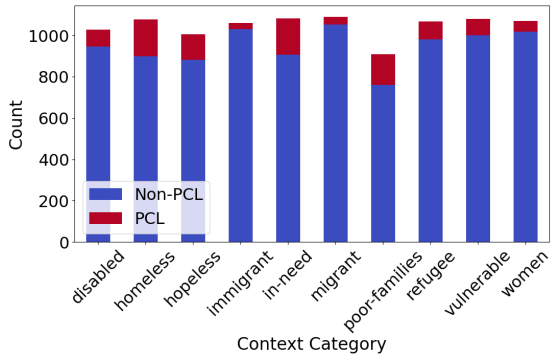

Figure 2: PCL vs. Non-PCL Count per Category

involve more elaboration, justification, or emotional framing, whereas neutral statements tend to be more direct.

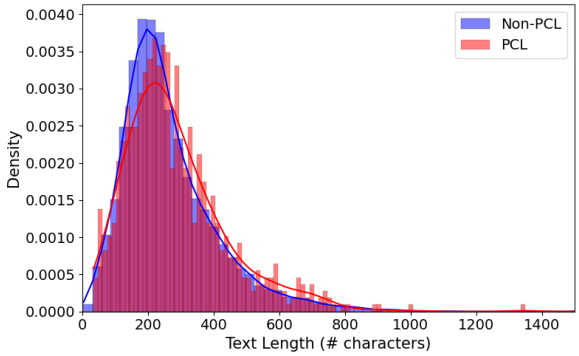

Figure 3: Histograms of Text Lengths

2.2 Qualitative Analysis of Patronising Language

Identifying patronising language is highly subjective, as condescension often depends on phrasing, context, and personal interpretation. While some sentences clearly frame individuals as helpless, others depend on tone and phrasing, leading to differing interpretations. For instance:

"Bruce has done a fantastic job in training youngsters, especially the youngsters that come from the poorest of the poor families, who would never have even had an opportunity to even buy a cricket bat."

While this highlights inequality, *poorest of the poor* adds an emotional weight that can make it seem overly dramatic rather than neutral.

Other cases are more ambiguous, making it unclear whether they should be classified as patronising. Take, for example:

"Today, homeless women are still searching for the same thing. A place to sleep and be safe."

The phrase *still searching* subtly suggests an endless struggle, which can come across as patronising. However, others might argue that it simply describes a persistent issue rather than belittling the individuals affected. Such subjectivity means even human annotators may disagree on classification, as personal perspective and cultural background shape interpretations.

2.2.1 Biases in the Dataset

Our analysis suggests societal biases in the dataset (Figure 2). PCL language appears more frequently in categories related to *poverty* and *homelessness*, reinforcing narratives of helplessness, while topics like *refugees* and *women* are often framed around resilience or empowerment.

These biases likely reflect real-world media portrayals, where certain issues are more commonly discussed using emotive or paternalistic language. Consequently, models trained on this data may over-classify patronisation in contexts where it frequently appears. This highlights why simple keyword-based models like BoW struggle with PCL detection. Effective classification requires a deeper contextual understanding, as meaning depends on framing and connotation rather than individual words.

3 Modelling

RoBERTa (Liu et al., 2019), an improved version of BERT, enhances masked language modeling with dynamic masking and larger training datasets. While effective, it struggles with subtle, context-dependent patronisation.

3.1 Preprocessing

We followed a RoBERTa-based preprocessing approach, first splitting the dataset into training and test (dev) sets, then further dividing the training set into 80% training and 20% validation while preserving the class distribution using stratification. This split was saved to ensure that future dataset improvements only affect the training set while keeping validation and dev sets unchanged.

Since we use Hugging Face's pretrained ALBERT and DeBERTa models, their tokenisers handle tokenisation, WordPiece splitting, and special token handling, eliminating the need for manual preprocessing. To preserve linguistic nuances, we avoided stemming, lemmatisation, and stopword filtering.

3.2 Further Improvements

3.2.1 Data Augmentation

We applied augmentation only to the training set, preserving validation and test integrity. Initially, we tested back-translation, synonym replacement, and contextual augmentation on 30% of the data. Based on effectiveness, we expanded back-translation and synonym replacement to 40%, introducing greater linguistic diversity while maintaining dataset size.

- **Back-translation**: Sentences were translated into and back from French, German, or Spanish, subtly altering phrasing while preserving meaning.
- **Synonym Replacement**: Key words were substituted with WordNet synonyms to introduce lexical variation without altering intent.

Each augmented sample retained its original label, reducing overfitting and improving generalisation.

3.2.2 Handling Class Imbalance

Due to the imbalance favoring neutral examples, we used class-weighted loss to improve recall for patronising language. We computed class weights via inverse frequency and applied a log transformation to smooth disparities. These weights were integrated into CrossEntropyLoss, making the model more sensitive to minority-class examples while maintaining classification balance.

3.3 Model Selection

3.3.1 ALBERT-v2

We initially fine-tuned ALBERT (`albert-base-v2`) due to its efficiency and reduced parameter count through parameter sharing (Lan et al., 2020). Its lightweight design allowed for faster training, but its limited capacity hindered its ability to capture nuanced patronising language. The model underperformed on our validation set, failing to generalise effectively.

3.3.2 DeBERTa-v3-Small

DeBERTa-v3-Small (`microsoft/deberta-v3-small`) was chosen for its advanced attention mechanism and improved contextual representation (He et al., 2021). Unlike ALBERT, it better captured implicit biases and subtle framing, leading to improved F1 scores. This made it the preferred model for detecting patronising language, balancing efficiency and performance effectively.

3.4 Hyperparameter Tuning and Training

We fine-tuned DeBERTa-v3-Small using `DeBERTaV2ForSequenceClassification`, optimising key hyperparameters:

- Dropout rates testing 0.2 and 0.4.
- Batch size of 32 for stable optimisation.
- Learning rates testing 1×10^{-6}, 5×10^{-6}, and 1×10^{-5},
- AdamW optimiser, weight decay of 0.01.

Training ran for 10 epochs with early stopping after two epochs of no validation loss improvement. A cosine learning rate scheduler with a 10% warm-up was applied.

The best-performing configuration used a **dropout of 0.2** and a **learning rate of 1×10^{-5}**.

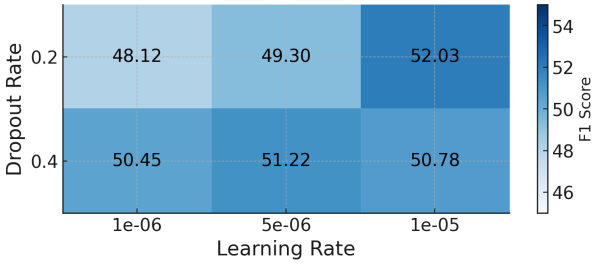

Figure 4: Hyperparameter tuning results for DeBERTa-v3-Small.

The loss curves in Figure 5 illustrate the model's learning progression across epochs. Initially, both training and validation losses decrease, indicating effective learning. However, after epoch 4, validation loss starts to rise while training loss continues to decrease, suggesting that the model is overfitting. This pattern indicates that while the model continues to fit the training data better, it generalises less effectively to unseen validation data.

3.5 Baseline Comparison

To assess our model's effectiveness, we compared DeBERTa-v3-small against two statistical baselines and the provided RoBERTa-base model. These baselines were chosen to represent traditional text classification methods that rely on word frequency rather than contextual meaning.

3.5.1 Statistical Baselines

- **Bag-of-Words (BoW) + Naïve Bayes:** Represents text as word frequency vectors, as-

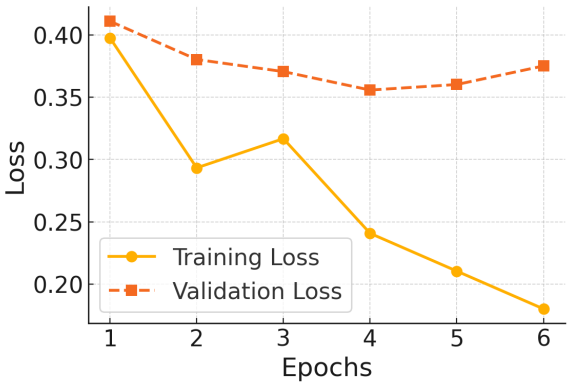

Figure 5: Training vs Validation Loss Across Epochs of deBERTa-v3-small

suming word independence. (Mikolov et al., 2013)

- **TF-IDF + Logistic Regression:** Improves upon BoW by weighting words based on document-wide significance, providing better feature representation. (Das et al., 2023)

These baselines were chosen because they are widely used in NLP and serve as a benchmark for evaluating contextual models. They offer a fair comparison by highlighting how much contextual embeddings improve performance over frequency-based representations.

3.5.2 Baseline Limitations

Both approaches struggled with implicit patronisation. For example, BoW misclassified the phrase: *"Many, many vulnerable people still lack access to basic needs."* Here, the repetition (*many, many*) subtly conveys condescension, but BoW, treating words independently, focused on token frequency and failed to recognise the patronising tone. TF-IDF assigned higher importance to terms like *vulnerable*, but without context, it also misclassified the sentence.

3.5.3 Final Evaluation on DeBERTa-v3-Small and Other Baselines

DeBERTa-v3-Small outperformed the three baselines by leveraging contextual embeddings and disentangled attention mechanisms, correctly identifying implicit patronisation. Table 1 shows the F1-score comparison, demonstrating the necessity of deep contextual modeling.

By analysing misclassified examples, we can better understand why these baseline models fall short and how a transformer-based approach like DeBERTa can better capture linguistic subtleties..

Model	F1-Score
RoBERTa Baseline	0.48
ALBERT	0.47
DeBERTa-small (Final)	0.52
BoW	0.22
TF-IDF	0.31

Table 1: Baseline Comparison on Dev-set

4 Results Analysis

While our DeBERTa-v3-small model improved over traditional baselines, its performance varies across different levels of patronisation, input lengths, and data categories.

4.1 Performance Across Patronisation Levels

A interesting question is whether the model performs better on highly patronising examples compared to subtler forms of condescension. The results (Figure 6) show a clear trend: as the level of patronisation increases, classification accuracy improves. Sentences where both annotators strongly agreed on high patronisation (*Label 4*) are shown as the hardest to classify correctly, with only 59.8% accuracy, whereas instances deemed neutral by both annotators (*Label 0*) are classified with 85.8% accuracy. Borderline cases, such as those where one annotator labeled a sentence as slightly patronising while the other did not (*Label 1* or *Label 3*), show a gradual decline in accuracy, indicating that subtle condescension remains challenging for the model.

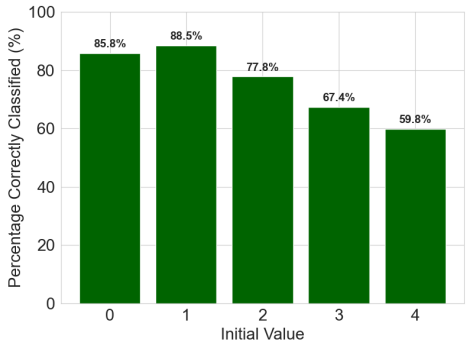

Figure 6: Model accuracy across different levels of patronisation. Accuracy decreases as patronisation becomes more ambiguous.

This difficulty is particularly evident in cases where positive framing or indirect language is used. For example:

"Kasun may not be a superhero or a super genius

but he is a great human being who has overcome challenges and hardships of life without even having the full functional capability of his body."

While this sentence discusses adversity, its emphasis on resilience rather than helplessness makes it harder for the model to classify as patronising. These results suggest that while the model effectively captures overtly patronising language, it struggles with more implicit forms where condescension is conveyed through framing rather than explicit wording.

4.2 Impact of Input Length on Performance

To assess whether longer inputs affect classification, we analysed F1-scores across different text lengths. While both PCL and non-PCL texts tend to be short, longer sentences show slightly higher misclassification rates. This is likely because longer inputs contain more context, which can introduce conflicting linguistic cues—a sentence may start neutrally but introduce subtle condescension later.

Additionally, the model uses 512-token truncation, meaning longer texts are cut off, potentially losing key information. Further investigation could explore segmenting long texts before classification to retain all relevant content.

4.3 Effect of Data Categories on Performance

Since the dataset contains predefined categories (e.g. homeless, poor families, refugees), we examined whether the model's performance varies across topics. PCL is more frequently detected in topics related to in-need and vulnerability, reflecting dataset biases. In contrast, categories like refugees and disbaled are often framed more neutrally, making classification harder.

This suggests the model captures societal patterns beyond pure linguistics. Emotionally loaded terms in discussions of poverty increase PCL detection, while resilience-focused refugee narratives make patronisation less obvious.

5 Discussion & Future Work

This work highlights the complexity of patronising language detection and the effectiveness of transformer-based models in addressing it. While DeBERTa-v3 improves classification, subtle forms of PCL remain challenging, requiring deeper contextual understanding.

5.1 Summary of Findings

DeBERTa significantly outperformed statistical baselines, achieving an F1-score of 0.52 on the validation set. Its contextual embeddings allowed for better recognition of explicit patronisation, where dramatic phrasing made classification more straightforward.

However, the model struggled with more ambiguous cases. Patronising language framed positively, such as narratives emphasising resilience, was often misclassified, while neutral statements about aid were sometimes flagged as PCL due to associations with vulnerability-related terms. Despite class weighting and augmentation, class imbalance still affected generalisation, particularly for subtler instances.

While limitations persist, these results highlight the potential of transformer-based models in capturing condescension beyond surface-level word patterns, laying the groundwork for further improvements.

5.2 Future Improvements

While our model effectively detects patronising language, there is room for improvement. One key direction is leveraging the dataset's predefined PCL categories, which capture different types of condescension. Rather than treating PCL as a single label, incorporating these categories could help the model distinguish between explicit pity, subtle framing, and exaggerated emphasis, improving generalisation and reducing misclassifications.

Another improvement with additional computational resources is using a larger model like DeBERTa-v3-large, which has demonstrated superior performance in capturing nuanced language. Its enhanced contextual embeddings and attention mechanisms could further refine detection, particularly for subtle cases where condescension is implied rather than explicit.

References

Mamata Das, Selvakumar K., and P. J. A. Alphonse. 2023. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.

2020. Albert: A lite bert for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Carla Perez Almendros and Steven Schockaert. 2022. Identifying condescending language: A tale of two distinct phenomena? In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 130–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.