Natural Language Processing Coursework: Identifying Patronising and Condescending Language

Aslihan Gulseren
ag724@ic.ac.uk

Asia Belfiore
ab6124@ic.ac.uk

Ginevra Cepparulo
gc1424@ic.ac.uk

Abstract

This project was aimed at designing a **Language Model** (ML) for the prediction of *Patronizing and Condescending Language* (PCL). We implemented and compared baseline models recreated from the original paper *Don't Patronize Me!* (Perez Almendros et al., 2020) alongside a custom BERT-based model, which outperformed both the baseline models and the best model in the paper.
Full code implementation of all the described models, processing pipeline and results can be found on GitLab at **ab6124/natural-language-processing**.

1 Introduction

Patronizing and Condescending Language (PCL) is a type of hate speech that targets vulnerable communities, without explicitly including hateful words. Due to the complexity and subtleness of human expressions, PCL detection represents a challenging sentiment analysis task(Perez-Almendros and Schockaert, 2022). The *SemEval-2022* **Don't Patronize Me!** Task (Perez Almendros et al., 2020) showed, however, that powerful language models like *transformer*-based models can identify PCL with a high level of accuracy. This paper compares the performance, strengths and faults of three separate models, Support Vector Machine (**SVM**), Bidirectional LSTM (**BiLSTM**) and a BERT-based transformer model (**RoBERTa**), on the original SemEval-2022 data, consisting of over 10 thousand paragraphs associated with vulnerable social groups.
Our final RoBERTa-base model achieved an F1 score of **0.57** on the official dev test, surpassing the score of 0.48 of the SemEval task baseline model.

2 Data Analysis

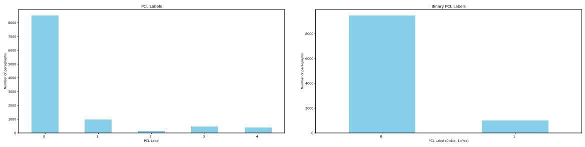
2.1 Context and Task Analysis

The Original *The Don't Patronize Me!* Task 1 Dataset comprises of 7,638 paragraphs about 11 vulnerable communities (*Disabled, Homeless, Immigrant, Migrant, Poor, families, Women, Hopeless, Vulnerable, In need, Refugee*), extrapolated from a vast collection of news stories published from 2010 to 2018 (Perez Almendros et al., 2020). Each article has been assigned a 'PCL score' by two separate annotators, as containing PCL (2), unsure or borderline case (1), or *not* containing PCL (0). Each paragraph's PCL label is calculated on a 5-point scale, as the sum of the scores of the individual annotators, with a minimum score of 0 (no PCL (0) detected by either annotator), to 4 (PCL (2) detected by both annotators).

2.2 Analysis of Class Labels

Following the class labels split of the original paper (Perez Almendros et al., 2020), we assigned **binary** PCL labels to each paragraph in the dataset, with a threshold of 2. This meant that every paragraph that had a label of 0 to 1 was given a label of 0 ('*no PCL*'), and any paragraph that had a label of 2 to 4 was assigned a label of 1 ('*positive PCL*'). We measured the models' performances with respect to the positive class in order to infer their ability to detect the presence of PCL. Figure 1 shows the distribution of positive and negative PCL paragraphs in the training dataset before (1a) and after (1b) binary label conversion.

The distribution of binary labels is evidently *unbalanced* and not uniform, with many more negative samples (without PCL) compared to positive ones (with PCL) in the dataset. We explored how this irregularity affects the distribution of other data fields, like *Country* and *Vulnerable Group*, in Figure 7 (Appendix) based on the original an-



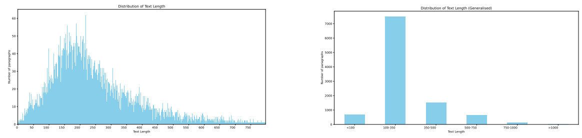
(a) Distribution of original label annotations.

(b) Distribution of binary labels.

Figure 1: Comparison of distribution of PCL labels.

notation of **PCL strength**. Overall, the majority of samples in the dataset were considered as not containing PCL (with an overpowering presence of class 0 paragraphs) for most fields.

We found no particular correlation between PCL strength and country (Figure 7a), as all countries show a somewhat equal distribution of mosttly negative PCL samples, with a decreasing number of higher PCL samples, and very few paragraphs with an ambiguous level of PCL. On the other hand, we found a correlation between PCL strength and the vulnerable group which the data sample refers to, as shown in Figure 7b. The samples pertaining to the groups *Homeless*, *Migrant*, *Immigrant* and *Refugee* show a higher proportion of PCL strengths above 2 when compared to other groups. Paragraphs referring to *Disabled* individuals, on the other hand, were found to have the highest incidence of negative PCL and the lowest number of samples with PCL level of 3 and above.



(a) Distribution of original paragraph lengths.

(b) Binned distribution of paragraph lengths.

Figure 2: Distribution of text lengths as paragraph word count.

We then analysed the relationship between PCL strength and text length, calculated as paragraph word count, as shown in Figure 2. The distribution of paragraph lengths is skewed towards the right and high peaked (figure 2a), meaning that most paragraphs in the dataset have lengths in the 'lower range'. To further investigate this, we generalised the text lengths into bins (figure 2b) and found that the majority of paragraphs had a word count range of 100-350 words.

Text length was found to be directly correlated to PCL strength (Figure 3), as on average longer paragraphs were found to have more samples that are categorized into higher PCL strengths.

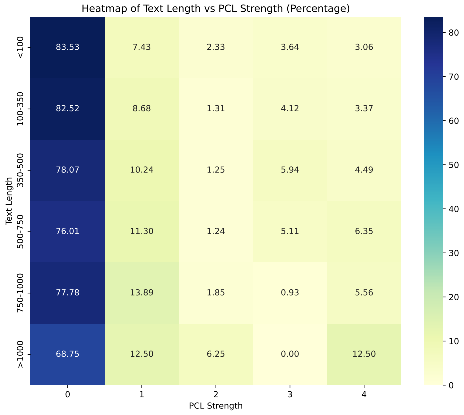


Figure 3: Correlation between text length and PCL Strength

2.3 Qualitative Assessment

Despite a defined taxonomy that highlights key PCL traits, PCL remains both interpretable and context-dependent. As further shown by the disagreements between the annotators in the original paper (Perez Almendros et al., 2020) on almost 1500 paragraphs. These discrepancies stem from PCL's ability to embed underlying patronizing attitudes within ostensibly positive or empathetic wording. This phenomenon is exacerbated by the fact that subjective life experiences play a crucial role in PCL perception, especially due to the diverse cultural, social and personal backgrounds. It may be easier for people belonging to vulnerable groups that are particular targets of PCL to detect PCL, compared to someone who does not belong to any targeted group (Nadal et al., 2019).

Building on these observations, we hypothesize that borderline cases (PCL labels 1 to 3) will represent the most challenging samples to classify, supported by the fact that more than half of the disagreements between the annotators (867 out of 1457) included borderline articles. The presence of certain linguistic patterns can help identify PCL, and have driven most of our data augmentation strategies, detailed below in Section 3.2. Among these, it is usual to find the distinctions between '*Us*' and '*Them*' to refer to vulnerable individuals in condescending texts, in sentences like '*You* can make a difference in *their* lives' (Perez Almendros et al., 2020). Similarly, it is common to find flowery adjectives and embellishments to paint scenes that raise feelings of pity in patronizing sentences, as shown in the example of (Perez Almendros et al., 2020): ‘Poor children might find more obstacles in their *race* to a *worthy* future’.

3 Modelling

3.1 Final Model

Our approach leverages a **RoBERTa**-based Transformer architecture. By training more, and removing the next-sentence prediction objective, RoBERTa expands upon BERT and captures richer contextual clues, spotting subtle patronization. After subword tokenization, RoBERTa processes embeddings through self-attention layers, followed by a classification head (feed-forward layers + sigmoid) to detect PCL. We decided to focus on fine-tuning RoBERTa because of its strong baseline performance in the SemEval Task, and how its self-attention mechanism helps capturing contextual word interactions.

	F1	Accuracy	Precision	Recall
(Test)	0.571	0.897	0.472	0.724

Table 1: Comparison of performance metrics of our final RoBERTa model on the Official Dev Set

3.2 Model Improvements

3.2.1 Preprocessing

The *community* keyword, such as “women” or “refugee”, is appended to the beginning of each training example. This choice was motivated by the correlation between community label and PCL strength uncovered in our data analysis. We hypothesized that explicitly including the community keywords in the sample could help the model to better distinguish between PCL classes.

3.2.2 Data Sampling and Augmentation

Because the dataset is highly imbalanced, where only a minority of paragraphs contain PCL, we used **downsampling** to achieve a 1:3 ratio of positive (PCL) to negative (non-PCL) examples. We trained and tested our model on different ratios, and found that 1:3 resulted in highest F1 score. Thus, all positive examples are combined with a subset of the negative examples of thrice that size. Limiting the number of negative instances prevented the model from overfitting to the majority class (0).

Due to the limited presence of positive PCL data, we applied different *Easy Data Augmentation* (EDA) techniques to increase diversity/ **Synonym Replacement** via WordNet was used to replace 10% of tokens with randomly sampled synonyms. **Backtranslation** using MarianMT was implemented to translate a random subset of 20% of paragraphs from English to German and back to English. This technique can paraphrase sentence structure and introduce additional lexical variety. Table 2 shows that the baseline model with downsampling, weighted loss and addition of community feature increased the F1 score on Test Set from 0.48 to 0.55. Implementing synonym replacement on top of this model improved the F1 further to 0.55. Finally, implementing backtranslation increased F1 score significantly to 0.571.

Model	F1 Score	Accuracy
Baseline (Downsampling + Weighted Loss + Community)	0.554	0.878
Synonym Replacement	0.555	0.887
Backtranslation	**0.571**	**0.897**

Table 2: Performance metrics on Official Dev Set on different EDA techniques.

3.2.3 Weighted Loss Function

To mitigate the found class imbalance, a *weighted cross-entropy* loss is used in order to give higher importance to the minority class (positive PCL). This means that misclassifying a PCL paragraph (i.e. false negavtives) led to a larger penalty than misclassifying a non-PCL paragraph (i.e. false positives). Class weights are computed based on the distribution of labels in the balanced set and are passed during the training routine.

3.3 Hyperparameter Tuning

We experimented with different hyperparameters combinations and evaluated the model performance for each combination, as shown in Table 4. The final selected hyperparemeters were: learning rate $= 2e^{-5}$, number of epochs $= 3$, batch size $= 16$. We did not implement Early Stopping due to the low number of epochs (the model was not trained for longer due to the high computational cost). Furthermore, the implemented RoBERTa Model is **cased**, meaning that it can distinguish between words with different capitalization. This is useful for the task at hand, as capitalization might be relevantly linked to higher PCL levels (e.g., shouting in all caps, proper nouns indicative of condescending language).

As hyperparameter tuning strategy we performed manual hyperparameter tuning, changing the learning rate, number of traning epochs and batch size. Table 3 shows our search space and 4 shows the performance of each tuning step, it seems that increasing decreasing batch size and increasing the number of training epochs favoured higher F1.

Hyperparameter	Value
Learning Rate	[1e-5, 2e-5, 3e-5]
Number of epochs	[2, 3, 4]
Batch size	[8, 16, 32]

Table 3: Hyperparameter tuning

LR	EP	BS	Dev F1	Dev Acc
2e-5	3	16	0.57	0.90
3e-5	3	16	0.53	0.88
1e-5	3	16	0.55	0.89
2e-5	5	16	0.57	0.90
2e-5	3	32	0.52	0.86
2e-5	3	8	0.57	0.90

Table 4: Hyperparameter tuning of Improved RoBERTa performance metrics on Dev Set

4 Models Comparison

All models (baseline and final) used a dataset split of 80% Training and 20% Validation. We refer to the 'Dev' metrics to indicate the model performance on the internal Validation set, and the 'Test' metrics to indicate the model performance on the official labelled Dev Set.

4.1 Baseline Models

Both baseline models were chosen as the best performing ones right after the RoBERTa model in the Task Paper (Perez Almendros et al., 2020). We followed the same model hyperparameters and architectures that were described in the paper in order to reproduce the original performances. The baseline models were trained on the non-augented original data.

Support Vector Machine (SVM): We used a TF-IDF weighted Bag-of-Words representation of the paragraphs as input to a SVM implemented with scikit-learn. The hyperparameters that were selected are C=10, gamma= 'scale' and kernel= 'rbf'. As further improvement on the baseline SVM we employed our Data Sampling and Augmentation method. The SVM achieves a Test F1 score of 0.32 as shown in Table 5.

Bidirectional LSTM (BiLSTM): We used the 300 dimensional **Word2Vec** skip-gram model trained on the Google News corpus (Mikolov et al., 2013) to get the embeddings of the words in each paragraph of the dataset.
The BiLSTM was created using the *keras* package. It comprises two bi-directional LSTM layers with 20 units and dropout rate of 0.25% each. We tested different versions of the model, based on the output labels. For binary labels (PCL / No PCL), we trained two models, one with size 1 output trained on binary labels and one with size 2 output trained on 2-dimensional one-hot encoded labels. The BiLSTM performance is shown in Table 7, with the model achieving a maximum F1 score of 0.43 on the internal Dev Set with binary 2-dimensional outputs.

4.2 Performances

Our final RoBERTa Model showed great improvements in F1 score compared to the other three baseline models, as shown in Table 5. SVM's reliance on TF-IDF features limits its ability to model semantic relationships. BiLSTM benefits from pre-trained word2vec embeddings. However, RoBERTa's large-scale pretraining captures nuanced contextual information, leading to stronger generalization on the test set.

Model	Dev F1	Test F1
SemEval Baseline	*	0.49
SVM	0.94	0.32
BiLSTM	0.43	0.37
Improved RoBERTa	**0.58**	**0.571**

Table 5: Comparison of performance metrics of different implemented models.

5 Analysis

5.1 Incorrect predictions made by improved model on Task 1

To gather further insights into the model performance and the dataset, Table 8 shows examples of mispredictions by our improved RoBERTa on the Official Test Set. There are 5 false positives and 3 false negatives. Based on the examples, it seems the model struggles to understand the nuance of being condescending or patronizing in sensitive contexts, which leads to false positives. For example, the first sample uses a lot of "they" vs

”us” and does include some terms which may remind of PCL; ”person who always needs help”. However, going more in depth, one can notice that the sample doesn’t exhibit a condescending or patronizing tone. On the other hand, from the false negatives shown, we can stem that they require greater knowledge to be fully interpreted. For the second sample paragraph; knowing that celebrities hold much more power and money than immigrants could better highlight the unbalanced power dynamics between them which is why the sample was labelled as PCL.

5.2 Different Levels of PCL

To what extent is the model better at predicting examples with a higher level of patronising content? Figure 4, shows how accuracy varies by original PCL strength (orig_label: 0–4).The model performs best on the extremes of the spectrum—level 0 (no PCL) and level 4 (very strong PCL). Hence, similar to human annotators, the model detects paragraphs with obviously no patronizing content or paragraphs with very strong condescending language more easily. However, borderline or moderate PCL is often ambiguous.

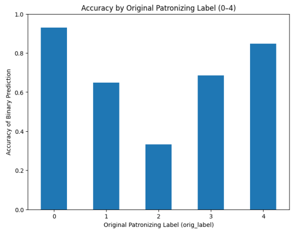


Figure 4: Prediction by PCL Strength

5.2.1 Input Sequence Length

How does the length of the input sequence impact the model performance? If there is any difference, speculate why.

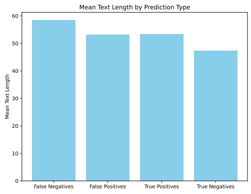


Figure 5: Prediction of Improved RoBERTa by Mean Text Length

As shown in 5, the mean text length for False Positives and Negatives is higher than the mean text length for True Positives and Negatives. The reason could be that with a greater amount of words the model struggles to pay attention to the terms related to PCL. Or perhaps, longer texts often convey concepts in a more subtle way with figuritive language or referrng to world knowledge. In other words, length of text is related to other confounders that make the text harder to interpret by the RoBERTa.

5.2.2 Data Categories

To what extent does model performance depend on the data categories?

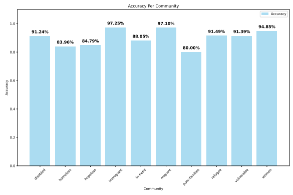


Figure 6: Accuracy of Improved RoBERTa by Vulnearable Group

The model performance does significantly depend on the data category (i.e. vulnerable group) the PCL text refers to. There is a significant difference in accuracy in detecting PCL between the groups *Immigrant, Migrant, Women* which achieve an accuracy above 94% and the groups *Poor-families, Homeless and Hopeless* which achieve an accuracy below 85%, as shown in Figure 6.

6 Conclusions

Our model increased the F1 score on the official Dev Set to 0.57. We got these results by downsampling, adding community label, a weighted loss function, synonym replacement and back-translation. We found that longer paragraphs and borderline PCL examples posed the most difficulty for the classifier, while extremely high or zero PCL texts were easier to predict. We also observed that performance varies by data category. A potential future experiment is to integrate the PCL level of each sentence into the weighted loss function. Class imbalance could be further addressed by assigning greater weights to level 4 PCL sentences.

References

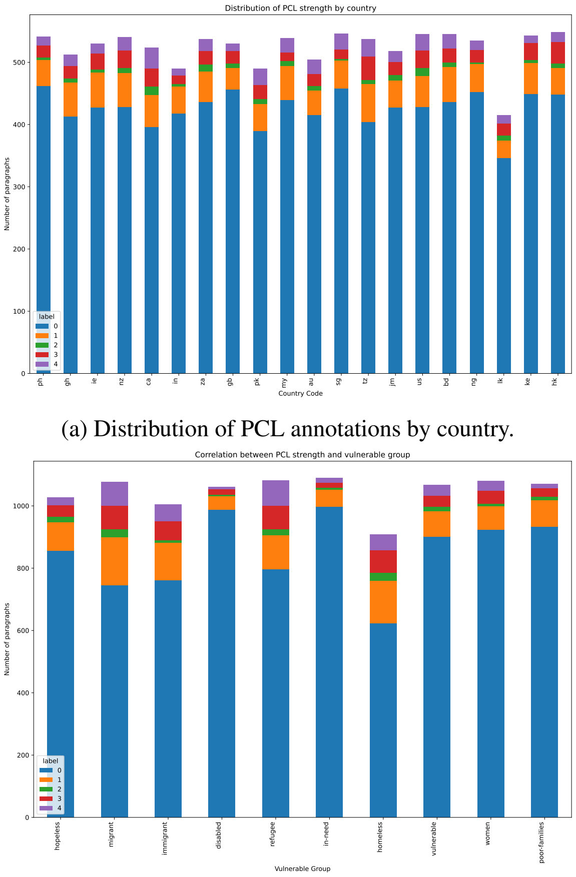
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Kevin L. Nadal, Terri Erazo, and David King. 2019. Elevated sensitivity to racial microaggressions in african american college students: Further validation of the racial microaggressions scale. Journal of Counseling Development, 97(4), 272–283.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carla Perez-Almendros and Steven Schockaert. 2022. Identifying condescending language: A tale of two distinct phenomena? Unpublished.

A Appendices



(a) Distribution of PCL annotations by country.

(b) Distribution of PCL annotations by Vulnerable Group.

Figure 7: Comparison of distribution of original PCL annotated labels by data field.

Test Metrics	SVM	BiLSTM
F1	0.32	**0.42**
Precision	**0.38**	0.33
Recall	0.28	**0.37**
Accuracy	**0.89**	**0.89**

Table 6: Performance metrics of Baseline Models compared.

	Dev F1	Test F1
Binary Labels	0.37	0.37
One-Hot Labels	**0.43**	0.37

Table 7: Performance metrics of BiLSMT on different class labels encoding (Binary and one-hot encoded).

B Supplemental Material

Full code implementation of all the described models, data analysis and processing pipeline and results can be found on GitLab at **ab6124/natural-language-processing**.

Pred	Paragraph	Label
1	**in-need**: That one person who always needs help - it 's nice to be there for people and having someone there for you when you need them. However, there are a few people who are constantly in need of help and have reached a stage where they can't get even the simplest of things done independently. We're not sure what to call them though, since they always manage to get their work done while you sit there wondering how, even after a day of hard work your checklist remains untouched. We wonder! Tell them to get an assistant, if they can't afford that, there are always interns.	0
0	**immigrant**: Many celebrities wore blue ribbons to support the American Civil Liberties Union, which is seeking to shed light on the plight of young immigrants facing the potential of being deported.	1
1	**migrant**: Clans of various surnames and functions were formed by Chinese immigrants, brought into then Malaya by the British to open up tin mines and jungles for rubber planting. Their mission then was to help fellow countrymen who were fleeing famine and civil wars to find jobs and shelter and send hard-earned money to their impoverished families in China.	0
1	**poor-families**: The living conditions of many poor families who collect waste for EcoPost have improved greatly because of the higher income they now receive. It also helps keep the streets a little tidier; there's so much waste plastic strewn across Kenya that Rutto jokes plastic bags have become the national flower	0
0	women: Maida noted that political issues should not twist women away from matters of development, the important thing is peace, and women must wake up, since they need economic revolution through peaceful means.	1
1	**in-need**: You also get to meet a lot of people enthuses Kanthi. Kanthi, has not only grasped the opportunity to meet a lot of people through dancing, but has also used the chance to reach out to those in need of help. The proceeds from 'Step by step' will be in aid of the Society for Uplift and Rehabilitation of Leprosy Patients (SUROL).	0
0	**immigrant**: Nearly 15,000 West African teenagers leave their homes every year to play football in Europe. Few make good on their dreams. Some are lured by corrupt agents, smuggled across the searing Sahara and discarded in the streets of Europe, resigned to selling fake designer bags as undocumented immigrants. Others are nabbed by academies and feeder teams affiliated with European clubs and often dumped like bad stocks.	1
1	**women**: Women in Sri Lanka have proven their skills and capabilities in competing equally with their main counterparts in various fields. Among them there are some women who are using their skills and capabilities for the betterment of the society instead of their own personal gain. This week the Empowering Women's corner brings you the story of such a woman who has become a well-known character in the society due to her social services. She is a model character who has proven herself by withstanding countless criticisms and challenges to continue her journey beyond the role of a traditional woman.	0

Table 8: Examples of incorrect predictions made by improved RoBERTa model on Test Data