NLP Coursework-Patronizing and Condescending Language Detection

Joshan Dooki
CID:02182106
jd22email@domain

Sam Kelso
CID:01506756
smk22@ic.ac.uk

Pavel
CID: 01205961
pb1121@ic.ac.uk

1 Introduction

The aim of this study is to develop a binary classification model to predict the presence of patronizing or condescending language (PCL). For this task the ”Don’t Patronize Me!” dataset (5) has been used. This consists of paragraphs that refer to susceptible communities which have been published in media across 20 English-speaking countries. These paragraphs have undergone manual annotation to indicate if the text includes any patronizing, condescending, or belittling language, specifically when referring to communities identified as vulnerable to unfair treatment in the media. Two annotators categorised the data based on the strength of the PCL used.

PCL is frequently unconscious and unintentional, however, it causes harm by reinforcing stereotypes and promoting exclusion. By utilizing the capabilities of NLP models to recognize and pinpoint instances of PCL towards others, corrective measures can be taken to promote responsible communication.

2 Analysis of the Training Data

EDA: The dataset contains 10,469 paragraphs extracted from news stories, which have been annotated to indicate the presence of PCL. Each paragraph is labeled between 0-4, where 0 indicates no sign of PCL, to 4 which indicates strong signs of PCL.

PCL Label	Frequency	% Frequency
0	8529	81.47
1	947	9.05
2	458	4.37
3	391	3.73
4	144	1.38

Table 1: Class frequency.

For this binary classification task labels 0 and 1 are grouped as non-patronizing (9476) and labels 2, 3, and 4 are grouped as patronizing (993). Figure 1 shows that the training data for this study is significantly imbalanced.

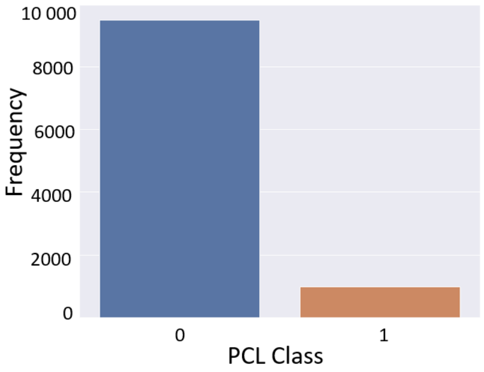


Figure 1: Frequency per class.

A correlation analysis was conducted to identify the dataset’s key features which can be attributed to PCL. The features analysed were keywords, country, and input length of sentence (using feature engineering techniques).

The input length feature contains several outliers (appendix, figure 6). These outliers were removed to ensure incorrect conclusions were not made on the underlying distribution.

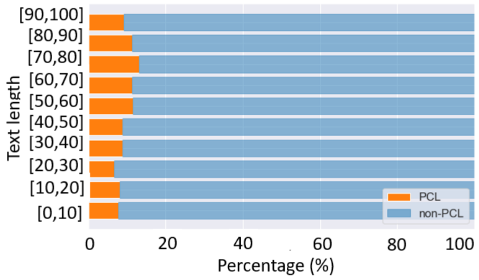


Figure 2: Text length correlation comparison.

The Spearman correlation coefficient between input length and the presence of PCL was 0.053. This metric as well as figure 2 above show that no clear correlation exists between input length and the presence of PCL. The same conclusions were drawn for the keyword and country features where evidence can be found in the appendix (figure 7). While figure 2 may show that input sentence lengths between 50-90 contain more observations of PCL, due to the large imbalance between the binary classes, no firm conclusions can be deduced.

Qualitative Analysis: The creators of the dataset emphasize the difficulty of the annotation task, due to the subtlety and subjectivity of PCL language (5). The labelling method used, attempts to mitigate the effect of the tasks subjectivity by using 2 annotators to annotate text as 0 (no PCL), 1 (borderline PCL) and 2 (clearly contains PCL). In the case where a paragraph is labelled 0 by one annotator and 2 by another, an additional annotator then judges the text sample which assigns a final label of 1, 2 or 3 depending on their choice. Therefore, within the dataset the confident cases occur at original label values 0 and 4 whereas all the cases annotated as 1, 2 or 3 contain higher levels of subjectivity. Therefore, referring to table 1, we note that only 14.5%, of the already minority positive class labels, are clear-cut annotations.

We can see the subjectivity in annotation by looking at some example texts. The blurred boundary between categories is most apparent between texts with the original labels of 1 and 2 because these are the most borderline cases. Take the example *"Currently, what's left to the Palestinian people is nothing more than the tiny territory of Gaza Strip, which is a de facto prison of Palestinian refugees, as well as the scattered and isolated villages and cities of the West Bank."* given an original label of 1. It could be said that this text contains PCL language, the use of the adjective "tiny" and metaphorical language which refers to the Gaza strip as a "prison" could be classed as condescending language. It is clear why this can be considered a borderline case. Conversely, the text *"Jenny Neal, regional director of the Grandmothers Advocacy Network for Saskatchewan and Manitoba, displays her orange scarf at the sculpture Prairie Wind at River Landing, Thursday, November 24, 2016, to promote 16 days of orange to raise awareness about violence against women. Greg Pender / The StarPhoenix"* was given an original label of 2. This piece of text could be interpreted as simply reporting a factual event which has taken place. From these two examples, it is clear that the borderline cases of PCL annotation are very difficult to categorise. When the text classification task is this difficult and subjective for human readers, it poses difficulty in accurately training a language model to detect PCL in unseen text. This displays the difficulty in finding a solution to the PCL classification problem.

3 Modelling

3.1 Model Comparison

	RoBERTa	DistilBERT	XLNet	BERT
Precision	0.556±0.04	0.582±0.02	0.511±0.018	0.620±0.02
Recall	0.548±0.02	0.446 ±0.02	0.517±0.01	0.427 ±0.01
F1 (Val.)	0.552±0.02	0.505±0.01	0.514±0.03	0.506 ±0.02

Table 2: Model results.

For this study, 4 state of the art language models were applied to the dataset. Each model was trialed on the up-sampled data augmented input set, as discussed in the next section. The results (using internal val. set) showed that these tuned models outperform the F1 score baseline of 0.48 to varying degrees. The RoBERTa-base model outperformed the other models when the upsampled dataset was used. Theoretically, this is expected due to the RoBERTa model's larger size and longer training time, which allows it to capture more complex language patterns and nuances. RoBERTa uses an enhanced version of BERT's training methodology by removing the next sentence prediction (NSP) task from pretraining and introducing dynamic masking so that the masked token changes during the training epochs (4). Based on these results, the RoBERTa model was chosen for this task, and all hyperparameter tuning and further improvements were made using this model. After further improvements, data augmentation strategies and hyperparameter tuning, the RoBERTa-base cased model achieved a Dev. set F1 score of 0.575 which outperformed the baseline F1 score of 0.48. N.B. In this report, Dev. set refers to the external dev. set provided, and Val. set refers to the internal dev. set used for hyperparameter tuning.

3.2 Model improvement methodologies

Several experiments were conducted in order to improve the class balance and model Val. F1 score (this was done before hyperparameter tuning). The following experiments were trialed: N.B. All model improvements and tuning was conducted using the internal Val. set.

SMOTE (Synthetic Minority Over-sampling Technique) - This method automatically creates synthetic samples for the minority class by interpolating between existing samples. This technique was the least effective which may be due to the

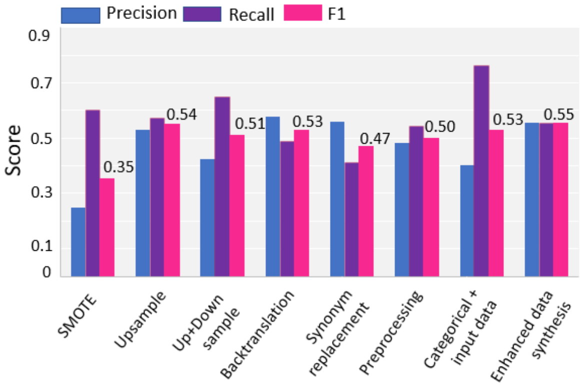


Figure 3: Model improvement strategies (Val. set F1).

algorithm creating instances that were too similar leading to overfitting and poor generalization.

Random upsampling - The minority class was upsampled, and although this method did improve the F1 score, there was significant over-fitting after the first epoch due to limited information gain.

Random up + down sampling - F1 score improved, however, precision decreased since the model was trained on fewer data samples.

Back-translations - Multiple back-translations were implemented on the minority class. This process generates new samples similar to the original but with some variants due to the translation process. Translations that achieved similarity scores less than 50% were removed:

Synonym replacements and random swapping - These simple augmentations performed poorly. This may be due to the complexity of the dataset and lack of diversity in the minority class.

Removing stopwords and punctuation - These preprocessing strategies marginally improved the F1 score. However, the model performed better with these features in the dataset. This may be because these features are important to the model as they provide important information about the structure and meaning of sentences.

Incorporating categorical data - The categorical data provided additional context and feature information that helped the model make more accurate predictions.

From conducting several experiments, the best result achieved was obtained from combining each of these individual techniques. The **enhanced data synthesis** consisted of 2 back translations (French + German), 2 upsamples of the minority class and the inclusion of the keyword categorical feature. This resulted in a more balanced training dataset of 6,462 and 5,859 datapoints in classes 0 and 1 respectively. This combination of techniques achieved a balanced precision and recall score, resulting in the highest F1 val. score of 0.552.

3.3 Model Setup & Hyper-parameter Tuning

Model setup: To fine-tune the pre-trained cased RoBERTa model for PCL detection, we added a classifier block consisting of two linear layers with two logit values as outputs. During training, cross-entropy loss was used which takes the logit values for positive and negative labels along with the binary ground truth label to output the loss value for that iteration. We attempted to address the data imbalance in the dataset by experimenting with a weighted binary cross-entropy loss that up-weighted positive samples, but it did not work as well as up-sampling positive samples within the training set and using unweighted cross-entropy loss.

Hyperparameter tuning: The tuned hyperparameters include the learning rate, number of hidden layers, attention heads, and dropout probabilities in the classifier block. The approach taken was to find an initial learning rate which showed good signs of learning. This was done by evaluating several models trained with different learning rates on a held-out val. set. The results of this experiment are shown in table 8 within the appendix. This showed that a learning rate close to 0.00001 would provide the best results, motivating our further tuning around this value as described later in this section.

Using this initial learning rate value, we plotted the training scores across two epochs both, with and without a learning rate scheduler. A linear learning rate scheduler was used as this gave the resulting plot shown in figure 4.

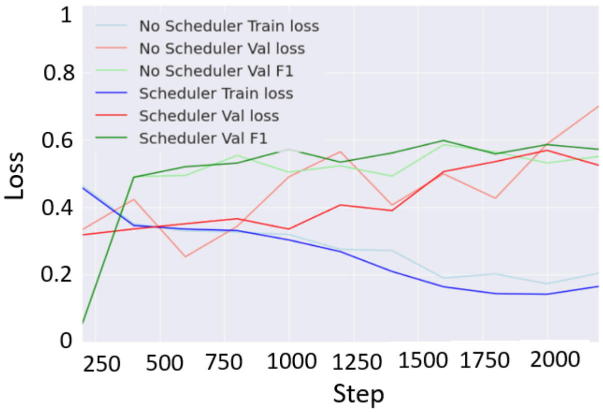


Figure 4: Loss and F1 score across 2 epochs with and without learning rate scheduler.

Including a learning rate scheduler improved the stability of the training loss and resulted in higher val. F1 scores (Figure 4). Therefore, we applied a linear learning rate scheduler to promote better model convergence. We observed a decrease in validation F1 score midway through the second epoch in figure 4. Therefore, early stopping was implemented based on the validation F1 score, with a patience value of 4, to prevent the validation F1 score from decreasing after its optimal value. In practice it was important that we monitored both the validation loss and F1 score; however, we performed early stopping based on the validation F1 score because the cross entropy loss implemented does not directly correspond to the F1 score we are optimising.

The final stage involved fine-tuning the remaining hyperparameters of the RoBERTa model using Optuna. Example results are shown in table 3, with the 23rd trial showing the best results. The trials showed that adjusting the number of hidden layers and attention heads away from the default value of 12 reduced the resulting performance of the fine-tuned model. We also separately tuned the dropout value in the classifier block, with a value of 0.5 producing the best results.

No.	Att. Head	Hid. Layer	LR	F1
0	16	10	2.05e-5	0.495
9	12	10	3.95e-5	0.540
23	12	12	2.23e-5	0.597

Table 3: Example trial results from further hyperparameter tuning.

The final hyperparameters were as follows: Hidden layers: 12, Attention heads: 12, learning rate: 2.23e-5, and dropout: 0.5. This configuration resulted in a final val. and dev. set F1 score of 0.597 and 0.575 respectively.

3.4 Model performance vs Baseline models

We compared our best RoBERTa model to the BoW (Bag of Words) and TF-IDF (Term Frequency-Inverse Document Frequency) models. We used SVM, XGBoost, and MLP Classifier. From table 4 we observed that our RoBERTa model significantly outperforms these baseline models (the best model being SVM in both cases).

A RoBERTa transformer model is expected to perform better than BoW models for sequence classification tasks since it is able to understand the contextual and semantic meaning of words within a sentence. In contrast, BoW models treat each word separately and ignore the connections between words and their context (1).

Model	F1 (BoW)	F1 (TF-IDF)
RoBERTa	0.575	0.575
SVM	0.252	0.252
XGBoost	0.157	0.065
MLP Classifier	0.241	0.237

Table 4: Dev set F1 score comparison between RoBERTa and models using BoW and TF-IDF.

Table 5 shows the top 4 features/words that influence the model's decisions. For example, if we input a sentence which contains the word 'underprivileged', the model will classify this sentence as class 0 (no PCL).

No.	Words	SVM Weights	Label
1	hurdle	1.111	0
2	dreamers	0.940	0
3	underprivileged	0.905	0
4	shiver	0.843	1

Table 5: Sharing the top 4 features of our SVM BoW model.

The sentence: "For jobless, hopeless Zimbabweans there is nothing much to cheer" is patronising. Describing a group of people using words such as "jobless" and "hopeless" may be seen as objectifying or dehumanizing because it reduces individuals to a single characteristic or situation. This can create a sense of separation between the speaker and the group and may be classified as PCL. Furthermore, the phrase "there is nothing much to cheer" can be interpreted as dismissive or minimizing the difficulties jobless and hopeless Zimbabweans face.

However, our SVM BoW model labels the sentence as class 0. This is because BoW models consider each word independently and do not consider the relationships and context between the words.

4 Analysis

4.1 Analysis 1- To what extent is the model better at predicting examples with a higher level of patronising content?

The accuracy scores in figure 5 show that the model's accuracy increases for higher levels of patronizing content. The model performance is the lowest for label class 2. This is due to the subjectivity of the binary split between the 4 classes. The model performs best on class 0, due to 81.5% of the original data consisting of this class label. Although the model does show an increase in accuracy for higher levels of patronizing text, this accuracy could be significantly improved by having more higher-level (non-synthetic) PCL content for the model to train on.

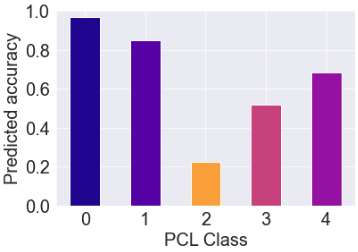


Figure 5: Accuracy vs PCL level.

4.2 Analysis 2- How does the input sequence length impact the model performance? If there is any difference, speculate why.

The dev. set length distribution was divided into 5 bins based on its quartile ranges and maximum length (see appendix, figure 10).

Length	Acc.	Prec.	Recall	F1 (dev.)
[0 – 30]	0.944	0.516	0.533	0.525
[30 – 41]	0.916	0.576	0.492	0.530
[41 – 58]	0.952	0.750	0.600	0.667
[58 – 100]	0.877	0.593	0.286	0.386
[> 100]	0.885	0.632	0.343	0.444

Table 6: Influence of sequence length on model performance.

The results in table 6 show that the model performs better at sentence lengths within the interquartile range of the length distribution (between length 30-58). The model may not perform as well on short sentences (<30) due to it not containing sufficient information. Similarly, the model's performance deteriorates for longer sentences (>58). This may be due to the model struggling to process all information and due to noise in these longer sentences. The training data consisted of 25% sentences of length > 58, which indicates that the model did have sufficient training data on long sentences. This may suggest that while RoBERTa is a state-of-the-art language model, it is not suitable for very long text sequences. The longformer model for sequence classification has been proven to outperform RoBERTa for longer sequences (2), which may be better suited for longer input sequences.

4.3 Analysis 3- To what extent does model performance depend on the data categories?

Table 7 shows that the performance of each keyword varies in contributing to the overall F1 score. This is due to data imbalance at the keyword level. For example, the "women" feature performed the poorest due to this keyword having 79% of datapoints in class 0. Keywords with a more balanced split performed significantly better. Removing categorical data from the model resulted in a lower F1 dev. score of 0.541, indicating the importance of categorical features. These features provide additional context and allow the model to learn more complex patterns in the data, leading to better performance. While adding more features can lead to overfitting, this was prevented through the incorporation of dropouts and early stopping.

Keyword	Acc	F1 (Dev.)	Frequency
immigrant	0.972	0.400	856
vulnerable	0.943	0.647	1 140
poor-families	0.805	0.431	1 325
homeless	0.877	0.500	1 705
disabled	0.959	0.600	1 126
women	0.940	0.125	959
hopeless	0.912	0.558	1 358
migrant	0.985	0.571	921
in-need	0.907	0.712	1 689
refugee	0.931	0.480	1 242

Table 7: Influence of Categorical Data.

5 Conclusion

By introducing a combination of improvement strategies to the model, such as backtranslations (×2), upsampling (×2), including categorical data, early stopping and a learning rate scheduler, a final dev. set F1 score of 0.575 was achieved. This model achieves higher accuracy scores for text containing higher levels of patronizing content. This is due to the subjectivity associated with text that contains lower levels of PCL. The model is better at classifying text of lengths between the interquartile range of 30-58 and performs poorly on longer text. The categorical data was an important feature addition as it provided more context for the model. This analysis showed that although the overall class labels were balanced, due to the keyword level imbalances, certain keyword texts performed poorly. As a future experiment, it is suggested that the balancing of classes should be conducted such that the keywords are also balanced. Additionally, an ensemble of several RoBERTa models can be implemented. This ensemble method was introduced by D. Hu et. al. (3), which achieved a higher F1 score than traditional stand alone state of the art language models.

References

[1] A. Araujo, M. Golo, B. Viana, F. Sanches, R. Romero, and R. Marcacini. From bag-of-words to pre-trained neural language models: Improving automatic classification of app reviews for requirements engineering. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 378–389. SBC, 2020.

[2] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. 1, 2020.

[3] D. Hu, M. Zhou, and X. Du. Pali-nlp at semeval-2022 task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. 1, 2022.

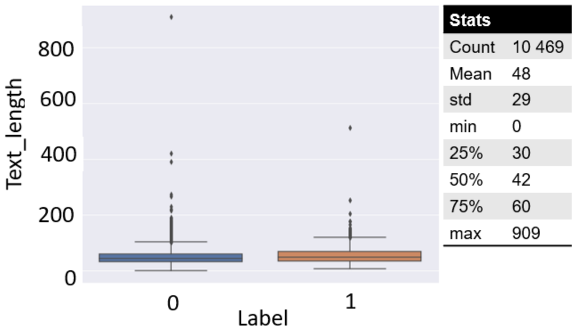
[4] Y. Liu, M. Ott, and N. Goyal. Roberta: A robustly optimized bert pretraining approach. *CoRR*, 1, 2019.

[5] C. Pérez-Almendros, L. E. Anke, and S. Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. *CoRR*, abs/2011.08320, 2020.

A Appendices

A.1 Data analysis

Figure 6 shows the training text length distribution.



Stats	
Count	10 469
Mean	48
std	29
min	0
25%	30
50%	42
75%	60
max	909

Figure 6: Distribution of text length per class label.

The below figures show the correlation of country and keyword on the presence of PCL

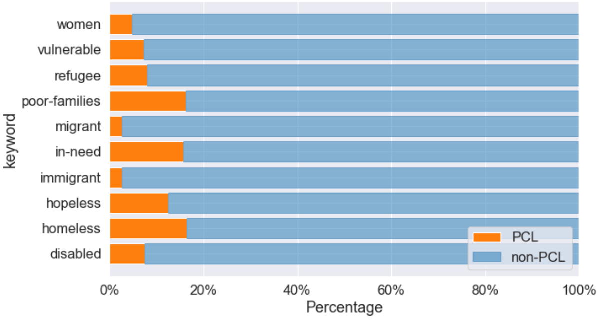


Figure 7: keyword correlation.

A.2 Further improvements

Below is a similarity analysis that was conducted to ensure that backtranslations below a similarity threshold of 0.5 was removed.

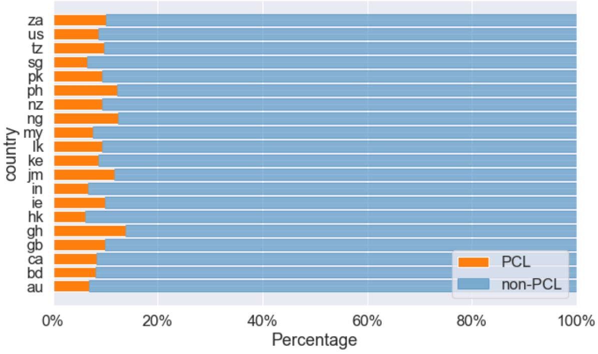


Figure 8: Country correlation.

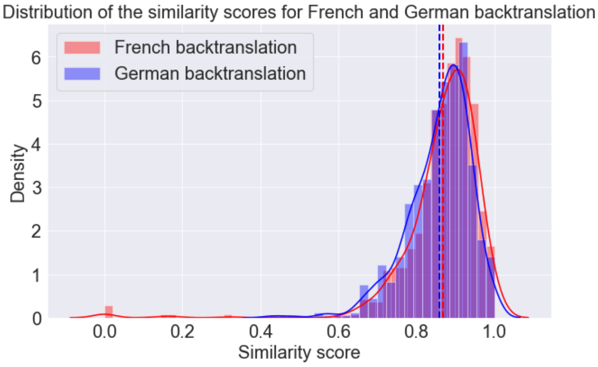


Figure 9: Similarity score (French+ German)

A.3 Modelling

Learning Rate	Val. F1
0.01	0.198
0.001	0.198
0.0001	0.237
0.00001	0.573
0.000001	0.028
0.0000001	0.0

Table 8: Initial learning rate experiment results.

A.4 Analysis

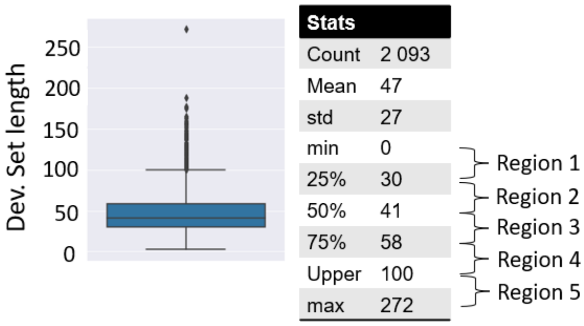
Below is an example of 2 back translations done using French and German which achieved a cosine similarity score of 0.569 and 0.56 respectively. Original : Osoyoos student excited about running across Haiti to help educate poor families

French : Osoyoos student thrilled to run through Haiti to help educate poor families

German : Osoyoo's student is looking forward to walking across Haiti to help poor families with education

No.	Att. Head	Hid. Layer	LR	F1
0	16	10	2.55E-05	0.495
1	12	10	3.91E-05	0.533
2	16	10	4.24E-05	0.524
3	16	10	6.79E-05	0.422
4	12	10	2.07E-05	0.5456
5	16	9	1.24E-05	0.500
6	16	10	4.04E-05	0.511
7	16	9	5.03E-05	0.536
8	12	11	3.15E-05	0.594
9	12	10	3.95E-05	0.540
10	12	12	2.45E-05	0.548
11	12	12	2.63E-05	0.557
12	12	12	2.97E-05	0.577
13	12	11	3.27E-05	0.574
14	12	11	1.08E-05	0.563
15	12	12	3.27E-05	0.597
16	12	11	1.76E-05	0.571
17	12	12	3.28E-05	0.577
18	12	11	1.86E-05	0.595
19	12	12	1.71E-05	0.589
20	12	11	2.05E-05	0.579
21	12	11	2.98E-05	0.587
22	12	11	1.44E-05	0.582
23	12	12	2.23E-05	0.597
24	12	12	2.12E-05	0.574
25	12	12	1.58E-05	0.589
26	12	12	1.18E-05	0.545
27	12	12	2.25E-05	0.555
29	16	12	2.55E-05	0.556

Table 9: Optuna results.



Stats	
Count	2 093
Mean	47
std	27
min	0
25%	30
50%	41
75%	58
Upper	100
max	272

Figure 10: Dev. set length distribution

Community	**Frequency**	**Class 0**	**Class 1**
immigrant	856	694	162
vulnerable	1140	696	444
poor-families	1325	518	807
homeless	1705	607	1098
disabled	1126	646	480
women	959	683	276
hopeless	1358	584	774
migrant	921	726	195
in-need	1689	603	1086
refugee	1242	705	537

Table 10: Frequency of class labels for each community.