NLP Coursework: Don't Patronize Me!

Sara Capdevila Sole
sc3719@ic.ac.uk

Alex Iordachescu
ai823@ic.ac.uk

1 Preliminaries

This team initially had three members. However, Oliver Sheridan-Methven has dropped this course, so this represents the work of 2 students.

The repository with the code can be found here. The prediction text files sit in the data folder.

2 Introduction

The *Patronizing and Condescending Language Detection Task* (Perez-Almendros et al., 2022) is a binary classification problem based on *Don't Patronize Me!* This is an annotated dataset with Patronizing and Condescending Language (PCL) towards vulnerable communities. It consists of paragraphs extracted from news articles from different English-speaking countries in which one or more vulnerable categories are mentioned. The associated paper defines PCL as a "superior attitude towards others or depicts them in a compassionate way". Moreover, it is "generally used unconsciously and with good intentions".

In the past, most efforts in Natural Language Processing (NLP) were focused on more objective and explicit tasks such as identifying intentional aggression (Potapova and Gordeev, 2016) or fact-checking (Das et al., 2023). However, there has been a notable shift in perspective, as highlighted in discussions regarding the significance and implications of this task (Sap et al., 2020), leading to a surge in related research.

The task authors outline various traits associated with PCL, including "us" versus "them" formulations and expressions of pity and saviour attitudes. Detecting this language is challenging due to nuanced expressions, requiring an understanding of linguistic cues, sarcasm, and cultural context. Another dataset for condescension detection (Talkdown corpus, (Wang and Potts, 2019)) emphasizes this complexity, underscoring the difficulty for NLP models in capturing these subtle intricacies.

3 The Don't Patronize Me! dataset

The *Don't Patronize Me!* dataset contains $10,637$ paragraphs from potentially vulnerable social communities (e.g. refugees, homeless people, poor families), from different countries. This has been annotated by three experts in communication, media and data science with labels from 0 to 4 indicating what level of PCL is present. The paragraphs have been extracted from the *News on Web* corpus (Davies, 2013).

3.1 Preliminary analysis

To convert this into a binary classification problem, we used labels 0 and 1 to refer to language with no PCL, and labels 2, 3 and 4 to refer to PCL. This resulted in an imbalanced dataset distribution; with a 9:1 no PCL to PCL ratio.

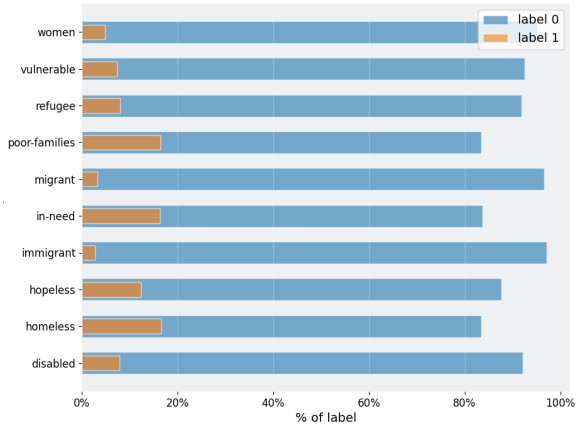

Figure 1: **Percentage of label per topic**

Our analysis suggests that certain vulnerable communities have a biased representation in the media, leading to biased levels of PCL. Notably,

Fig. 1 highlights the prevalence of 'homeless', 'in-need' and 'poor-families' for non-PCL, contrasted with 'migrant' and 'immigrant' for PCL.

We also analysed average sentence lengths, revealing variations among different communities (see Fig. 2) and countries (refer to Appendix, Fig. 4). Results showed a significant difference in average sentence length between instances labelled with PCL and those without. Notably, specific communities like 'migrants' and 'women', exhibited pronounced contrasts. Relying on these characteristics for model predictions may introduce bias into the model.

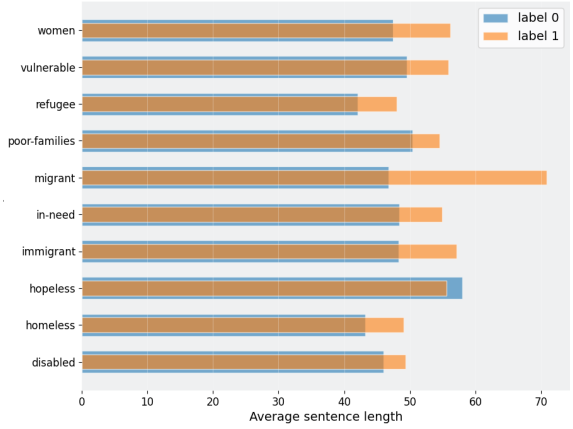

Figure 2: **Average length of sentence per topic**

3.2 Qualitative analysis

We must be aware that a high degree of subjectivity characterises this task. In (Huckin, 2002), the author illustrates that condescension is highly dependent on the reader's perception and the relative status between the reader and the writer. This is reflected by the initial inter-annotator agreement of our dataset, i.e. a moderate Cohen's Kappa coefficient of 41%.

One subjectivity factor is *speculation*. When an attitude is suggested rather than stated, the reader is forced to make assumptions about the writer's intention. The paragraph "However, that also means the French people have to live with the consequences of their system – latent racism on the one hand and resentful migrants on the other" has been labelled as borderline PCL by one of the annotators and harmless by the other. One could argue that this is an objective description of a rough reality that generates negative feelings while not being condescending. However, one could suspect that the writer considers that is part of a superior system that prevented these consequences and intends to blame the French people. Irrespective of the validity of the fact, the speculations made by the reader greatly influence the effect of the text.

Another subjectivity factor is the *impact*. As opposed to the previous discussion, some texts can be subjectively assessed in terms of outcome, regardless of the intentions of the author. The paragraph including the sentence "But displaced people were not left homeless by developers and government agencies" has been labelled as PCL level 2 out of 4. Without necessarily making assumptions about the intentions of the authors, the reader may consider the formulation to be inappropriate, as it depicts the authorities as saviours of those in need, acting from a privileged position. Even if the reader believes this is just a mistake and is not reflected in the rest of the text, this could still suffice to classify the whole paragraph as PCL.

The cultural factor makes the annotation of this dataset problematic, too. One can argue that the nuances (if not the definition) of PCL differ from country to country, so we could obtain different labels for the same paragraph depending on the country it is coming from. On the other hand, the same people annotate all of the paragraphs, so we cannot expect them to have made the decisions based on the country of origin, as they have their own backgrounds and cultural values.

We performed χ^2 tests to analyse the correlation between the labels and the other features of the paragraphs. The class label and the country of origin obtained a p-value of $1.7 \cdot 10^{-5}$, while the class label and the topic obtained a p-value of $4.07 \cdot 10^{-64}$. We can say with high certainty that these features are correlated with the assigned label. Consequently, an important question can be asked: considering the annotation methodology discussed above, do we want a model that would adapt the PCL detection mechanism to the context given by the features? Or do we want a model that treats every document equally, according to the same standards?

Nonetheless, there are plenty of examples where it is hard to agree on the nature of the paragraph. Quantitatively, this is reflected by the significant Kappa coefficient of 61% on examples that have not been classified as borderline by any of the annotators. Paragraphs such as "Lawsuit seeks release of immigrant child held in Chicago" do not raise any suspicion over the presence of

PCL. Meanwhile, in texts such as "A Deputy Governor in Zamfara State in Nigerian, Malam Ibrahim Wakkala Muhammad, has promised to sponsor the marriage ceremony of 100 vulnerable couples across the state" it is easy to identify how the governor portrays himself as a hero and how he aims to help numbers rather than people.

4 Models

We decided to implement and fine-tune RoBERTa and BERT, following the approach taken by the authors (Perez Almendros et al., 2020). We loaded the pretrained cased models, as PCL analysis is usually affected by the nuances introduced by the use of punctuation and capitalisation. To address nuances introduced by punctuation and capitalisation, we loaded the pretrained cased models. A classification head was added to each model, and subsequent fine-tuning was carried out.

BERT uses a classification head comprised of one fully connected layer, mapping 768 neurons to the two output classes. We also implemented a dropout layer with a probability of 0.2 for regularisation. We experimented with more complex classification heads with multiple layers and activation functions like ReLU but observed a significant drop in performance. Unlike BERT, RoBERTa has three fully connected layers in the classification head, with a few additions. Firstly, each linear layer is followed by a dropout layer, a LeakyReLU activation and a batch normalisation layer. Secondly, a ResNet style residual connection has been added before the output layer (He et al., 2016).

The models are optimised according to the Cross-Entropy loss. However, to further account for the imbalance in the data, we use the weighted version of Cross-Entropy to focus more on the minority class during learning (Lin et al., 2017).

Building on the prior discussion, we decided to exclusively input the paragraphs from news articles to the model, disregarding the country of origin and the topic. This choice stems from the observed correlation between these additional features and class labels. We reasoned that considering the annotation methodology, evaluating a paragraph for PCL based on these features might not be meaningful. Our aim is to prevent the model from relying heavily on these features to avoid learning shortcuts in decision-making.

We conducted a hyperparameter search for the RoBERTa model, exploring a range of values such as learning rate, batch size, epochs, dropout probability, LeakyReLU slope, early stopping patience, and positive sample weight within the loss. In addition, we tested different evaluation strategies: "every step" with a linear learning rate scheduler and "after every epoch" without a scheduler. The search ranges for each hyperparameter are detailed in Table 3 in the Appendix.

Our optimisation process involved leveraging the Optuna optimiser (Akiba et al., 2019) across a predefined hyperparameter space. Through 150 iterations, we efficiently navigated this space using the Tree-structured Parzen Estimator (TPE) algorithm. This dynamic search strategy, based on past evaluations, optimised computational time while thoroughly exploring the parameter space—a more efficient alternative to exhaustive grid search.

Throughout this process, we closely monitored the performance of each model, ultimately selecting the hyperparameter combination yielding the highest F1 score on an internal validation set. The search revealed that the most suitable hyperparameters include a learning rate of approximately $2.5{\cdot}10^{-5}$, a batch size of 32, 6 epochs of training, a dropout probability of 0.1, a LeakyReLU slope of 0.2, no early stopping patience, a weight of about 0.6 for the positive label, and the implementation of a linear learning rate scheduler.

4.1 Data Augmentation and Preprocessing

To address the class imbalance, we tried two methods; 1. Downsampling the majority class to achieve an approximate 1:1 ratio, and 2. Upsampling the minority class using Self-Supervised Manifold Based Data Augmentation (SSMBA) (Ng et al., 2020). SSMBA relies on the assumption that the underlying data can be represented in a lower-dimensional manifold. A corruption (noise) function is applied to perturb a training data point off the manifold, followed by the application of a reconstruction function (a denoising autoencoder) to reproject the data point back onto the manifold. This naturally can be used to extend the count of the minority class and possibly improve the model's learning. We adopted the recommended parameters from (Ng et al., 2020), setting the noise probability to 0.25, the augmented ratio to 2:1 (no PCL to PCL), and utilising BERT as the autoencoder.

We employed a range of data augmentation

techniques during our experiments. These included applying back-translation from German with a probability of 0.4 per text, randomly deleting and swapping words, as well as synonym replacement—all with a probability of 0.1 per text. While most of these techniques led to substantial improvements for BERT, only back-translation showed significant benefits for RoBERTa.

In comparison to the baseline models in the following subsection, no data processing was carried out for RoBERTa. This is because no data cleaning has been found to perform best for this model on PCL, as explained in (Siino et al., 2024).

We chose to implement the data augmentation techniques that demonstrated the most substantial improvement in model performance, specifically back translation and upsampling the minority class using SSMBA. The corresponding results are presented in Table 1.

	F1 Score	
Augmentation Method	**RoBERTa**	**BERT**
None	0.537	0.442
Back-Translation	0.579	0.514
Upsampling (SSMBA)	0.580	**0.519**
Downsampling	0.542	0.508
Back-Translate & SSMBA	**0.583**	N/A

Table 1: **Data Augmentation Methods**: F1 scores on the official dev set for RoBERTa and BERT using different data augmentation methods. Optimal hyperparameters from the Optuna search were employed.

The improvement in results can be attributed to the diverse set of data augmentation methods employed by SSMBA. These go beyond conventional techniques; it introduces advanced semantic transformations, perturbs the structural integrity of the data with intricate modifications, and encourages contextual variation - all adapted to the characteristics of the training set. Overall, SSMBA aims not only to diversify the dataset but also to enhance the model's robustness, making it more generalisable.

4.2 Baselines

To conduct a model comparison, we implemented three baseline models using traditional machine learning techniques with explicit feature extraction and shallow learning. The first feature extractor is Bag of Words, a method of extracting sparse embeddings from text based on the frequency of each word in the document. The second feature extractor is Bag of Bigrams which works similarly but counts the bigrams instead. Finally, the third feature extractor is based on the TfIdf statistical method and measures how important each term in each document is relative to the other documents in the corpus. We assessed the performance of these feature extractors in combination with three classifiers: SVM, Naive Bayes and Logistic Regression. In the end, we kept the best-performing pairing for each feature extractor.

As these feature extractors treat words as individual sequences of characters, we employ some preprocessing techniques to improve these representations. First, we tokenise the internet links and remove any form of referencing. We also remove all digits and special characters, because, while punctuation is important, their impact on these methods is limited and the overall effect might be negative. Next, we remove stop words, i.e. the most commonly used words in the language. According to (Luhn, 1958), these have low distinguishing power, thus they should not be considered. We omit short words (¡3 letters), considering them as noise. Finally, we run our corpus through the Porter stemmer, one of the most common algorithms for stemming English (Porter, 1980), to collapse similar forms of a word to a canonical form. This is useful when calculating frequencies. We deliberately preserve the casing, although it is generally a very useful method. Not only does it prove unhelpful empirically, but casing can also represent a meaningful feature in the context of PCL.

To address the imbalanced dataset problem, we use the Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla et al., 2002) to create new instances from the minority class. We repeatedly take the convex combination of a random sample and one of its nearest neighbours to use it as a new sample, until we achieve a completely balanced dataset. This is possible because we consider the samples to be the numerical embeddings rather than the raw tokens.

Finally, two trivial baselines have been added. One trivial model randomly assigns a label to each paragraph it is tested on. The other trivial model assigns the PCL label 1 to all paragraphs.

The results for the baseline models are presented in 2. Our best baseline misclassifies the text containing "children hailing from affluent families use dumpsites as playground" to be non-PCL, as it is not feasible to assess such expressions based on word frequencies.

Model	F1 Score	
	Baseline	Preprocessed
Bag of Words + NB	0.124	0.319
Bag of Bigrams + NB	0.083	0.286
TfIdf + LR	0.074	**0.377**
Random	0.163	N/A
All PCL	**0.174**	N/A

Table 2: **Results**: F1 Scores on the official dev set for the baseline models with and without data preprocessing & SMOTE. NB stands for Naive Bayes, and LR represents Logistic Regression.

5 Analysis

Is the model better at predicting examples with a higher level of patronising content?

The model demonstrates a noticeable performance improvement as the classified level of PCL increases, as depicted in Fig. 3. This aligns with expectations, as higher severity labels indicate reduced ambiguity, making predictions more straightforward for our model.

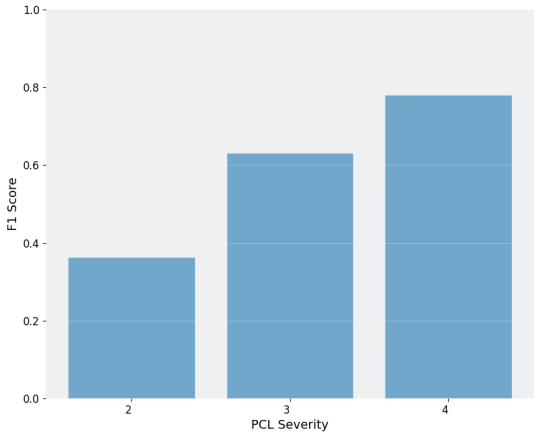

Figure 3: **F1 variation with PCL severity**

Does the length of the input sequence impact the model performance?

It is challenging to ascertain the impact of input sequence length on the model's performance, given that longer sequences are infrequent and consequently yield an unrepresentative F1 score. Appendix, Fig. 8 visually conveys these findings. The optimal F1 score is observed around the mean sentence length (shown in Fig. 2). During inference, as the sentence length extends beyond the training data's typical range, the model encounters difficulties in generalisation due to the limited representation of longer sequences. Another plausible factor contributing to the observed decrease in scores for longer sequences could be the inherent characteristics of such sequences. Longer sequences may exhibit more ambiguity, and complexity, thereby posing a harder task for the model.

Does model performance depend on the data categories?

In this analysis, we assessed the models' performance concerning country and topic. The F1 score exhibits variations across different countries and topics. This is depicted in the Appendix, Fig. 9 and 10. However, it is crucial to acknowledge that these variations could be influenced by the distribution of data counts within the dataset. Notably, the correlation between the number of PCL samples for each topic, appears to have a substantial impact on the F1 scores, as observed in the Appendix, Fig. 7 and Fig. 10. The model's ability to predict categories appears to be influenced by the representation of each category in the training set.

This aligns with the intuitive idea that increased exposure improves the model's recognition of distinctive category features. Abundant data for specific features strengthens the model's predictive accuracy. For instance, extensive 'in-need' features may not impact predicting 'vulnerable' individuals with distinct vocabularies. This analogy extends to countries, highlighting the importance of a balanced representation in the training set for diverse regional lexicons and features.

6 Conclusion

In this report, we tackle the PCL detection task (Perez-Almendros et al., 2022), together with its challenges. Employing data augmentation methods and preprocessing, we handle class imbalances and enhance the dataset. The most effective model for the classification task is RoBERTa, which achieved an F1 score of 0.583.

Building upon our insights from the previous section, we recommend improving our model by implementing a more sophisticated data-balancing approach, particularly for minority categories such as longer sequences, topics and countries. Potential strategies include exploring ensemble methods (Khan et al., 2023) or applying the upsampling technique SSMBA to each category individually, thereby generating more instances. This ensures a balanced representation, fostering improved model performance across diverse categories.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, 155:176.

Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. 2023. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness.

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, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rodmonga Potapova and Denis Gordeev. 2016. Detecting state of aggression in sentences using cnn.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context.

A Appendix

Hyperparameter	Search Range
Learning Rate	10^{-6} to 10^{-2}
Batch Size	8, 16, 32, 64
Number of Epochs	1 to 10
Dropout Probability	0.1, 0.2, 0.4, 0.5
LeakyReLU Slope	None, 0.01, 0.05, 0.1, 0.2
Patience (epochs)	None, 5
Positive Sample Weight	0.5 to 1

Table 3: **Optuna Search Ranges:** Hyperparameter ranges used for the Optuna search (Akiba et al., 2019), for the RoBERTa model.

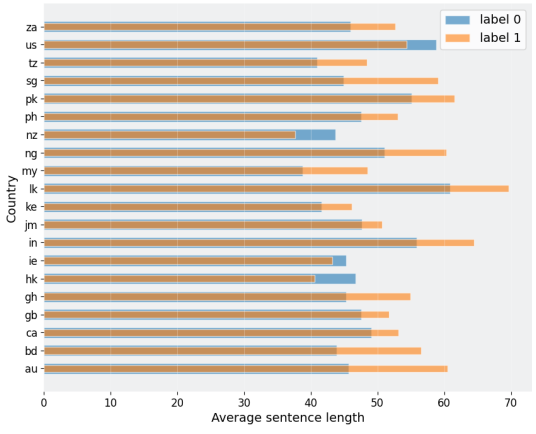

Figure 4: **Average length of sentence per country**

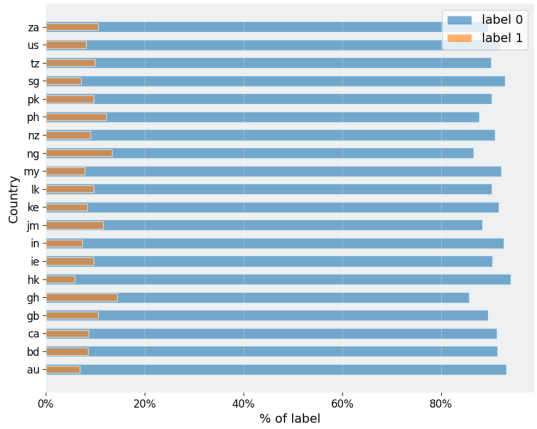

Figure 5: **Percentage of label per country**

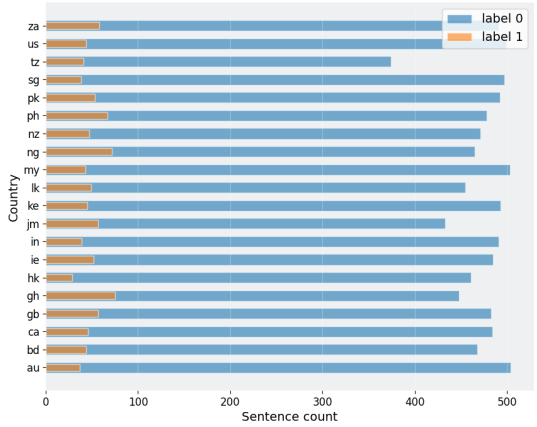

Figure 6: **Sentence count per country**

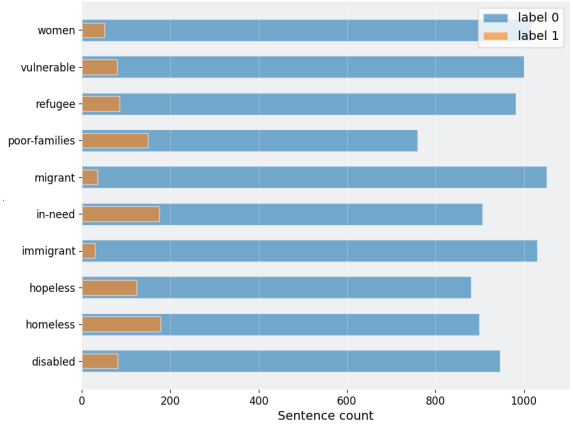

Figure 7: **Sentence count per topic**

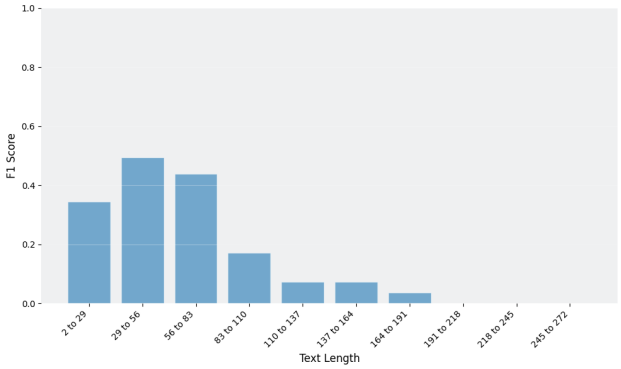

Figure 8: **F1 score variation with sentence length**

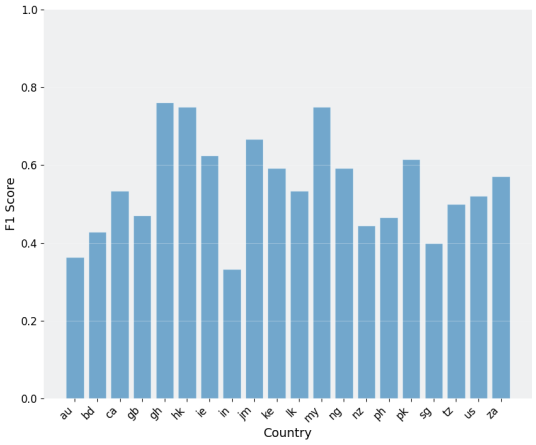

Figure 9: **F1 score variation with country**

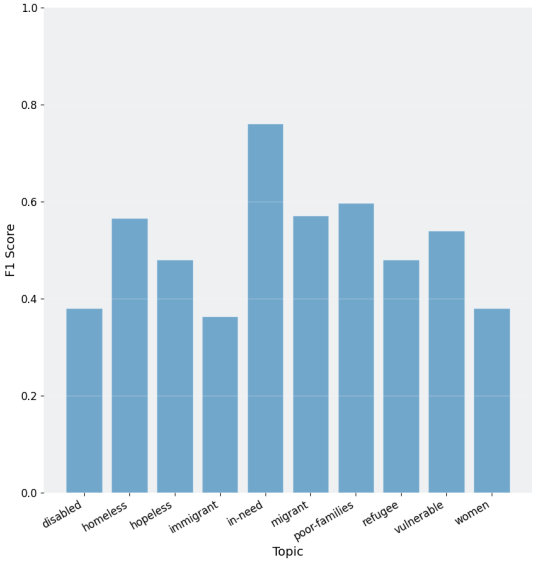

Figure 10: **F1 score variation with topic**