Almost SOTA: PCL Detection Through Smart Ensembling and Calibration

Hazim Sager
Imperial College London
ms3421@ic.ac.uk

Abstract

An Optuna-based hyperparameter search is conducted using a single A100 40GB GPU to explore sampling strategies, augmentation techniques, contrastive learning, and loss functions—maximizing discriminative power before incorporating each model as a separate feature in the ensemble. Feature engineering analysis is performed using exploratory data techniques and a Random Forest classifier to assess feature importance. Finally, isotonic regression is used to calibrate model outputs to fully maximise predictive power.

The ridge regression ensemble achieves a macro F1-score of 0.78 and an F1 of 0.61 on the positive class on the provided dev set.

1 Introduction

Patronizing and Condescending Language (PCL) is subtle, challenging, and often subjective. Detecting such nuance is non-trivial, requiring models to distinguish between neutral phrasing and implicit condescension. The SemEval 2022 Shared Task (Pérez-Almendros et al., 2022) sought to address this challenge, constructing a benchmark for PCL classification. Despite advances in natural language processing as of 2022, the best-performing system in the shared task leaderboard achieved a macro F1-score of 0.65.

The winning solution (Hu et al. (Hu et al., 2022)) implemented weighted sampling to mitigate class imbalances and an advanced learning rate decay strategy to train three transformer models. These models were ensembled via majority voting, a standard technique in model ensembling.

This work uses ridge regression as an ensembling method for PCL detection, yielding a viable alternative for PCL detection. By dynamically weighting each model's contribution while suppressing collinearity, ridge regression constructs an ensemble that not only combines predictions but refines them, leveraging the strengths of each component model. An Optuna-driven hyperparameter search was conducted to explore sampling strategies, augmentation techniques, contrastive learning, and loss functions, focusing on maximising individual discriminative power before being integrated into the ensemble. A range of final ensemble configurations is evaluated, including calibration ordering, before reaching an ensemble configuration within $\approx .4$ of the SoTA F1.

2 Data Analysis

As provided, the shared task dataset includes 10,469 manually annotated pieces of text. Each text is labelled with a corresponding search keyword, country code and an aggregation of the labels provided by two annotators. The aggregated labels span the range 0 to 4, with a threshold of ≥ 2 decided on for a paragraph to be labeled with the positive class.

2.1 A Numeric Overview

The positive and negative classes are dramatically imbalanced. There are 8,528 instances of the negative class and 1968 instances of the positive class (corresponding to approx. 91% and 9% of the dataset respectively). Such imbalance is likely to produce a dramatic local minima for some learned model wherein it may achieve 91% accuracy by simply predicting the negative class for all inputs.

Figure 1 shows the distribution of the positive class per country, with a significant range (5.9% to 14.3%). Countries like Ghana and Nigeria sit at the high end, with 14.3% and 13.4% respectively. Countries like Australia and Hong Kong sit at the low end, at 6.8% and 5.9% respectively. There is broadly little correlation between other features and the labels, like text length ($r = 0.0566, p < 0.001$) and sentence count ($r = 0.0595, p <$

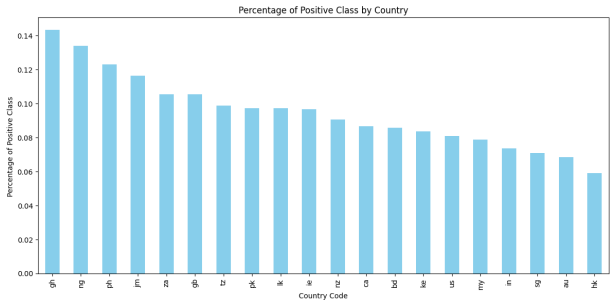

Figure 1: Percentage of Positive Class per Country

0.001). Table 1 details the correlations between numerical features engineered and the aggregated labelling.

Table 1: Spearman Correlations with Aggregated Scores

Feature	Corr.	Sig.
stopword_ratio	0.1289	***
stopword_count	0.1122	***
subjectivity	0.0833	***
word_count	0.0784	***
unique_word_count	0.0684	***
sentence_count	0.0595	***
length	0.0566	***
polarity	0.0465	***
keyword_relevance	0.0277	**
avg_sentence_length	0.0135	
lexical_diversity	-0.0910	***
keyword_count	-0.1061	***
keyword_ratio	-0.1277	***
avg_word_length	-0.1283	***
content_ratio	-0.1289	***
keyword_in_text	-0.1293	***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

There is a small but statistically significant relationship between these textual features and scores. For example, the stopword ratio in the paragraph has a correlation of $\approx$ 0.13 with scores ($p < 0.001$), explaining approximately 1.7% of the variance. However, these features in isolation do not likely constitute a powerful enough signal to determine the positive class from the negative class.

To indicate whether predictive power of the target labels could be achieved through a combination of the text and the engineered features, a random forest classifier was trained with 100 estimators. The text was vectorised using the TF-IDF algorithm, with a maximum of 1000 features. Table 2 shows the 20 most important features as determined by the random forest classifier.

The trained classifier achieved an **ROC of 0.7509** alongside a **macro-weighted F1 of 0.48** and a **positive-class F1 of 0**. These metrics suggest the random forest classifier did indeed learn some discriminative power from the signals despite the lack of conversion into correctly classified positive classes. It is imperative to note that the F1, especially the positive class F1, can be a surprisingly misleading metric. In some sense of raw strength, the classifier does produce signal amidst noise and can differentiate samples to some degree. This will motivate parts of the ensembling approach later.

Table 2: Top 5 features by importance from Random Forest Classifier.

Feature	Imp.	Type
avg_word_length	.0231	Style
stopword_ratio	.0187	Style
keyword_ratio	.0185	Style
content_ratio	.0183	Style
to	.0163	Word

2.2 A Qualitative Overview

Patronizing and Condescending language is defined in the task paper (Pérez-Almendros et al., 2022) *"language that is patronizing or condescending towards vulnerable communities (e.g. refugees, homeless people, poor families)"*. Almendros et al. (Pérez-Almendros et al., 2022) reported that the task was subtle and difficult for standard NLP tools, as corroborated. Often, language is considered condescending not based on the exact word choices but the context of the word usage and the status of the person making those choices. Unsurprisingly, this is difficult context to infer and requires picking up on subtle markers and inferences from relatively little language. Table 3 gives a qualitative analysis of the texts with varying PCL labels, and a post-hoc rationalisation to the aggregated labelling as had been given by the annotators.

The contrast in these examples shows clearly a subtle but gradual increase in patronizing tone. However, it is clear that the analysis at no point is with respect to a description in isolation, or even the description of an action in isolation. Patronizing tone, in this case, is hallmarked by a distinctive sense of excessive pity, unsuitable solutions and saviour-positioning, all of which are complex interactions between multiple people and social groups which must be 'modelled' in some way.

The example in table 3 show this effect in two separate categories - homelessness and poor fam-

ilies. In both cases, a 0-scored text generally is a factual description with little subjectivity. In 2-scored text, we generally observe a high status party positioned as a saviour, with a focus on the charitable act and the positioning of the recipients as passive. In 4-scored text, we tend to see obvious and excess shows of pity, pedantic language or entire homogenisation of a disadvantaged group.

This task is thus more difficult than usual sentiment analysis tasks or spam tasks, given often those focus on a first-order model of the language and it's implications. Here, this demands from whatever model is used, that it is capable of understanding complex social dynamics and interactions, and be able to draw decision boundaries around them. In short, this task is likely very difficult.

The gold labels are sourced from a ground truth of two annotators. Inherently then, the labelling is a subjective view. The usage of two annotators helps mitigate the annotator-bias effect, however it is not a sufficiently large number of annotators for the law of large numbers to apply and for a de-biased conclusion to be drawn. We thus expect, heavily, that we are modelling the subjective and biased opinion of two annotators whom have likely similar but different biases as shown by the many instances of annotator disagreement inferrable by the labels.

3 Modelling Choices

The previous state of the art solution used an ensemble of three models, each model a variation of what Hu et al. (Hu et al., 2022) named **BERT-PCL**. BERT-PCL is based on the RoBERTa-Large model (Liu et al., 2019), with a weighted random sampling approach used to mitigate the class imbalances and a innovative technique to vary the learning rate across different transformer layers.

This paper proposes an alternative approach. Four models were trained individually based on the **BERT-Cased** (Devlin et al., 2019), **BERT-Uncased**, **RoBERTa** (Liu et al., 2019) and **DeBERTa** (He et al., 2020) architectures. The output of each model is ensembled through a ridge-regressed regression layer and calibrated using isotonic regression. This paper motivates this choice through the initial research hypothesis that different model architectures are likely to infer a different set of linguistic features with the premise that they may be decorrelated. Naturally, decorrelated predictive signals can be ensembled to both reduce variance and increase the strength of the signal.

With all of these models being transformer-based models, we expect some correlation and collinearity in the signals nonetheless. An appropriate regularisation term is thus the L_2 ridge regression regularisation term, to both result in more explainable weights and allow for heteroskedasticity in the signals to be effectively combined.

4 Training

Each individual model was trained with varying loss functions, sampling, augmentation, batch sizes and learning rates. Both BERT-Cased and BERT-Uncased were used to explore the effect of casing on performance. Early stopping was used with a patience of 3 epochs and a maximum training time of 10 epochs.

The Optuna framework was used to search through the hyperparameters through a Tree-structured Parzen Estimator to identify better hyperparameter configurations in an extremely large hyperparameter space.

The hyperparameter search was conducted not to optimise the positive class F1 but rather the ROC AUC, to maximise discriminative power of each model while remaining threshold-agnostic.

After each instance of fine-tuning, the best model for each class was selected to form the final ensemble model. Each prediction of each model is fed into an L_2-regularised linear model. Another hyperparameter search using 5-fold cross validation, this time optimising for F1, is conducted and a regularisation strength of $\alpha = 100$ is identified as the optimal regularisation strength. We achieve a **macro F1 score of 0.78** on an internal dev set and a **positive class F1 score of 0.61** on an ensemble of all four models. This paper presents in table 5 the individual model performances, the previous state of the art, a majority voting approach on the individual models, two baselines trained on TF-IDF transformed text and the features previously discussed in section 2 and the ensemble performances.

Table 3: Qualitative analysis of texts with varying PCL scores (0: none, 2: moderate, 4: severe).

Score	Text Example	PCL Analysis
Category: Homeless		
0	"Get the homeless people shelters; assist families with health-care plans that will cover a wider range of issues and surgeries;"	Policy-focused; no condescension or pity
2	"A mass protest was held in Jaffna [...] against Tamil Nadu politicians [...] who prevented [...] Rajinikanth, visiting Jaffna to hand-over the keys of 150 houses [...] to homeless people who had been affected by the separatist war."	Positions homeless as passive recipients; focuses on charitable act
4	"Wall of kindness to provide for the homeless in city"	Homogenizes helpless group
Category: Poor Families		
0	"Computers and money were stolen [...] a crime that is extremely rare in the neighbourhood where mostly very poor families live in small subdivided units."	Factual description without judgment
2	"Later in 2008, Avril Lavigne was awarded a Certificate of Honor [...] for her work in raising funds for poor families and children with disabilities in China."	Celebrity positioned as savior; vulnerable groups as objects of charity
4	"Everyone makes a mistake, no one's perfect. Execution is very bad, no one likes it. We feel very sorry for these poor, poor families."	Excessive pity; repetition of "poor"; distancing language ("these")

model_name	learning_rate	batch_size	contrastive	contrastive_weight	focal_alpha	focal_gamma	roc_auc	scheduler_gamma	scheduler_step_size	scheduler_threshold	scheduler_T_max	scheduler_eta_min	scheduler_type
BERT-Cased	2.77917e-06	24	True	0.175234	1.26659	2.22545	0.859956	—	—	0.01	—	—	reduce_on_plateau
BERT-Cased	2.6024e-05	8	True	0.115119	1	2	0.876344	—	—	0.05	—	—	reduce_on_plateau
BERT-Cased	1.27807e-05	16	False	0.1	0.307853	1.51718	0.888419	0.3	—	—	—	—	exponential
BERT-Cased	8.79421e-05	8	False	0.1	1.13054	4.13975	0.607549	—	—	—	—	—	none
BERT-Cased	8.30643e-05	16	False	0.1	1	2	0.777544	—	—	0.05	—	—	reduce_on_plateau
RoBERTa	2.77917e-06	24	True	0.175234	1.26659	2.22545	0.907637	—	—	0.01	—	—	reduce_on_plateau
RoBERTa	2.6024e-05	8	True	0.115119	1	2	0.840094	—	—	0.05	—	—	reduce_on_plateau
RoBERTa	1.27807e-05	16	False	0.1	0.307853	1.51718	0.914239	0.3	—	—	—	—	exponential
RoBERTa	8.79421e-05	8	False	0.1	1.13054	4.13975	0.301387	—	—	—	—	—	none
RoBERTa	8.30643e-05	16	False	0.1	1	2	0.433955	—	—	0.05	—	—	reduce_on_plateau
DeBERTa	2.77917e-06	24	True	0.175234	1.26659	2.22545	0.911244	—	—	0.01	—	—	reduce_on_plateau
DeBERTa	2.6024e-05	8	True	0.115119	1	2	0.864311	—	—	0.05	—	—	reduce_on_plateau
DeBERTa	1.27807e-05	16	False	0.1	0.307853	1.51718	0.919132	0.3	—	—	—	—	exponential
DeBERTa	8.79421e-05	8	False	0.1	1.13054	4.13975	0.399687	—	—	—	—	—	none
DeBERTa	8.30643e-05	16	False	0.1	1	2	0.378937	—	—	0.05	—	—	reduce_on_plateau
BERT-Uncased	2.77917e-06	24	True	0.175234	1.26659	2.22545	0.880091	—	—	0.01	—	—	reduce_on_plateau
BERT-Uncased	2.6024e-05	8	True	0.115119	1	2	0.880208	—	—	0.05	—	—	reduce_on_plateau
BERT-Uncased	1.15104e-06	8	True	0.0518791	1	2	0.876108	0.1	3	—	—	—	step
BERT-Uncased	4.77539e-06	24	True	0.456402	1	2	0.901122	—	—	—	10	1e-07	cosine
BERT-Uncased	6.16503e-06	24	True	0.478389	1	2	0.903993	—	—	—	10	1e-07	cosine

Table 4: Optimization Results

5 Residual Analysis

5.1 To what extent is the model better at predicting examples with a higher level of patronising content?

Table 6 shows clearly a convex shape in accuracy against labels. This is not unexpected, as the extreme cases of PCL (or non-PCL) are likely easier to classify, whereas the more ambiguous labels often have annotator disagreement and thus a conflict of signal. It is therefore fair to say that the model is good at predicting the most extreme forms of PCL/non-PCL text, but struggles with ambiguity where there is annotator conflict.

5.2 How does the length of the input sequence impact the model performance?

Table 7 shows a steadily increasing F1 for larger text sizes (with exceptions for rarer samples likely due to variance). This is also not rather unexpected as we may expect that longer text sizes give the model more data points, helping it be more confident in it's predictions given more signal.

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

Table 8 shows a higher F1 for the *vulnerable*, *in-need* and *immigrant* categories, while it scores a lower F1 on the *women* and *disabled* categories. This variety may indicate to us that texts on certain topics are more obviously inferable as PCL, possibly due to language choices and biases.

5.4 A misclassification of baseline model

This paper presents a misclassification from the Bag of Words model. The following text, *" No one has the right to kill another person for a crime , " he said . " These are children coming from poor families who are engaging in theft in order to provide for their relatives .*" is misclassified as PCL

Table 5: Performance Comparison of Models

Model	Negative Class			Positive Class			Accuracy	ROC AUC
	Precision	Recall	F1	Precision	Recall	F1		
Ensemble	0.97	0.93	0.95	0.52	0.75	0.61	0.97	—
BoW	0.95	0.93	0.94	0.42	0.50	0.46	0.89	—
Naive Bayes	0.95	0.83	0.89	0.28	0.61	0.38	0.81	—
BERT-Cased	0.95	0.95	0.95	0.52	0.50	0.51	0.91	0.888
BERT-Uncased	0.92	0.99	0.96	0.73	0.18	0.29	0.92	0.901
DeBERTa	0.95	0.96	0.96	0.59	0.53	0.56	0.92	0.919
RoBERTa	0.95	0.96	0.95	0.57	0.53	0.55	0.92	0.914
State-of-the-Art	—	—	—	—	—	0.65	—	—

Table 6: Accuracy by Raw Score

Raw Score	Count	Accuracy
0	1705	0.9507
1	191	0.7120
2	18	0.3333
3	89	0.6966
4	92	0.8913

Table 7: Performance by Text Length

Length Range	Count	F1 Score	Accuracy
0-9	33	0.0000	0.9091
10-24	282	0.5652	0.9291
25-49	1021	0.5991	0.9148
50-99	650	0.6349	0.8938
100-199	108	0.7273	0.9167
200-499	1	0.0000	1.0000

Table 8: Performance by Category

Category	Count	F1 Score	Accuracy
in-need	226	0.7442	0.9027
immigrant	218	0.7273	0.9862
vulnerable	209	0.6923	0.9234
refugee	188	0.6667	0.9415
hopeless	218	0.5833	0.8624
homeless	212	0.5789	0.8491
poor-families	190	0.5682	0.8000
disabled	194	0.5000	0.9072
migrant	207	0.5000	0.9807
women	233	0.4615	0.9399

with a confidence of ≈ 0.6 whereas it has a non-PCL label. This paper believes that this misclassification is likely due to the usage of the words "children", "poor families" and general sentiment of the necessity of crime to provide for the child's family. The tone, however, is rather factual, which may be the reason for the non-PCL labelling.

6 Conclusion & Further Experiments

The results in this paper have shown that careful hyperparameter tuning, objective selection at various stages and careful usage of ensembling methods can produce near SoTA models. We present a model with 0.61 F1 and confident predictions at the extreme ends of PCL/non-PCL text. We show that the model can derive stronger signals from longer texts, and that topics such as women and disable persons are more nuanced topics regarding PCL.

A suggestion for further experimentation is to consider whether residual learnable signal exists. It would be an interesting research goal to check for whether the previously engineered features still correlate with the residuals in any meaningful manner. Indeed, this would likely take careful experimentation as the residual error on the internal dev set is almost perfectly balanced, and thus great care would need to be taken in identifying features that can predict any subtle bias in the residuals that might remain.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Dou Hu, Yihan Liang, Yinqiao Guo, and Benyou Xu. 2022. PALI-NLP at SemEval-2022 task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 330–335, Seattle, United States. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Carla Pérez-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 288–300, Seattle, United States. Association for Computational Linguistics.