LastResort at SemEval-2022 Task 4: Towards Patronizing and Condescending Language Detection using Pre-trained Transformer Based Model Ensembles

Samyak Agrawal **Radhika Mamidi**
International Institute of Information Technology Hyderabad
samyak.agrawal@research.iiit.ac.in,
radhika.mamidi@iiit.ac.in

Abstract

This paper presents our solutions for Task4 at SemEval2022: Patronizing and Condescending Language Detection. This shared task contains two sub-tasks. The first sub-task is a binary classification task whose goal is to predict whether a given paragraph contains any form of patronising or condescending language(PCL). For the second sub-task, given a paragraph, we have to find which PCL categories express the condescension. Here we have a total of 7 overlapping sub-categories for PCL. Our proposed solution uses BERT based ensembled models with hard voting and techniques applied to take care of class imbalances. Our paper describes the system architecture of the submitted solution and other experiments that we conducted. Our best performing models achieve an F1 score of 59.4 and 15.7 on sub-tasks 1 and 2 respectively.

1 Introduction

Patronizing and condescending attitude in language generally denotes the writer's sense of superiority over others. If someone is patronizing or condescending, it means what they write/say is accompanied by a sense of pity or compassion. Often, usage of PCL is relatively unconscious, and the intent of the writer is not to hurt a particular group or person they are referring to. So while being harmless in its intention. Usage of PCL still poses a risk of harming vulnerable people or groups by stereotyping them or normalizing specific behaviour towards them.

Task4 at SemEval-2022 (Pérez-Almendros et al. (2022)), Patronizing and Condescending Language Detection provides two sub-tasks. The goal of subtask1 is to identify if the given paragraph contains PCL. The goal of sub-task2 is to determine which subcategory of PCL expresses the condescension. The seven subcategories are Unbalanced power relations, Shallow solution, Presupposition, Authority voice, Metaphor, Compassion and The poorer, the merrier. A given paragraph can show instances of multiple subcategories.

We experimented with multiple transformer-based models. We used focal loss and Weighted Random Sampling to address the class imbalance; we also tried out ensembling models with hard voting, which improved the accuracy over the baseline models for both the sub-tasks.

The paper is structured as follows: Section 2: describes the dataset and related work. Section 3: describes our system and model architecture. Section 4 has information regarding the dataset size and splits with libraries used. Section 5 discusses the findings from our experiments, and section 6 concludes our paper.

2 Background

There has been work done to detect potentially harmful forms of language. Liu et al. (2019a) used BERT and LSTM based models to detect offensive language in the dataset, Offensive Language Identification Dataset (OLID) provided by Zampieri et al. (2019) at SemEval 2019. Indurthi et al. (2019) used InferSent (Conneau et al. (2018)) semantic sentence representations to detect Hate Speech against Immigrants and Women in the dataset provided by Basile et al. (2019) at SemEval 2019.

PCL's subtle and often unrealised nature makes its detection an arduous task for humans and Artificial Intelligence systems alike. There has been some recent work done when it comes to addressing PCL. Wang and Potts (2019) presents a dataset of social media messages annotated for condescending acts in context.

2.1 Dataset and Task Description

SemEval2022 Task 4 provides the Don't Patronize Me! dataset (Pérez-Almendros et al. (2020)). The dataset contains 10469 paragraphs. We divide the data into training and validation datasets. The paragraphs in the dataset are annotated for PCL.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 352 - 356

Paragraph	Label
Call to restore hope for homeless through inquiry	1
farooqui said women 's groups were demanding fast-track courts to deal with rape and other crimes against women .	0

Table 1: Example for Sub-Task1

Paragraph	Label
the word of god is truth that 's living and able to penetrate human souls (heb. 4:12) . consider how powerful scripture is : it can change hearts , save lives from eternal condemnation , and give hope to the hopeless	Unbalanced power relations , Compassion
these poor ladies are definitely going through some traumatic issues right now , and i am asking that they come forward so that i help them ? together with women parliamentarians - to be able to heal.	Unbalanced power relations, Shallow solution, Presupposition, Authority voice, Compassion

Table 2: Example Of Sub-Categories of PCL for Sub-Task2

The paragraphs marked positively for PCL are then annotated for seven different categories of PCL. The dataset has paragraphs in the English language and was collected by the News on Web (NoW) corpus. They queried the corpus for paragraphs using ten keywords related to vulnerable communities widely covered by the media and from the 20 English speaking countries in the corpus. Detailed information of the dataset can be found in the task description paper (Pérez-Almendros et al., 2019).

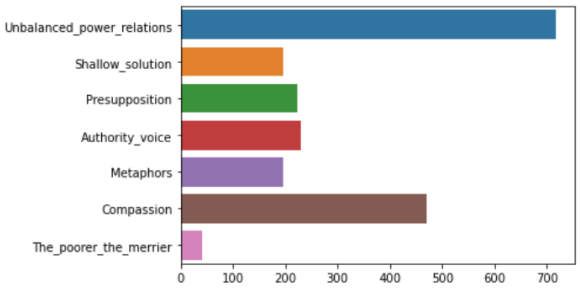

Figure 1: Frequencies of PCL subcategories

2.1.1 Sub-task1

The first sub-task is a binary classification task where given a paragraph, we have to classify whether it contains PCL. The annotation in the original paper has labels from 0 to 4. The paragraphs marked 0 and 1 are marked negative for PCL, while those marked 2 and above are marked positively. Table 1 shows positive and negative PCL examples from the dataset for sub-task 1.

2.1.2 Sub-task2

The second sub-task is a multi-label classification problem where given a paragraph, we have to classify whether it belongs to one or many of the seven subcategories of PCL. The subcategories are Unbalanced power relations, Shallow solution, Presupposition, Authority voice, Metaphor, Compassion, The poorer, the merrier. Table 2 shows examples for paragraphs marked for different categories of PCL.

3 System-Overview

For our solution, we have relied on using pre-trained transformer-based models like RoBERTa (Liu et al. (2019b)) which robustly optimizes the original Bidirectional Encoder Representations from Transformers(BERT) Devlin et al. (2019). It is pre-trained on much larger datasets, bigger batches and employs dynamic masking wherein a masking pattern is generated every time a sequence is fed to the model.

We also experimented with the newer Decoding-Enhanced BERT with Disentangled AttentionV3 (He et al. (2021a)), which is an improved version of the original DeBERTa (He et al. (2021b)). It leverages ELECTRA style (Clark et al. (2020)) pre-training by replacing DeBERTa's original mask language modelling (MLM) with a more sample-efficient pre-training task, replaced token detection (RTD), where the model is trained as a discriminator to predict whether a token in the corrupted input is either original or has been replaced by a generator.

As the dataset is highly imbalanced, we use two different ways to deal with the imbalance, focal loss and Weighted Random Sampling.

Model	UNB_POW	SHAL	PRES	AUTH	MET	COMP	POOR_MERR	AVG
RoBERTa $_{sep}$	19.4	15.9	9.3	5.3	13.9	11.3	9.1	12.0
RoBERTa $_{ens}$*	15.8	24.8	10.0	**9.3**	16	11.2	14.8	14.6
DeBERTa $_{ens}$	15.7	24.8	10.0	9.2	16	11.3	14.9	14.6
RoBERTa $_{WRS\_sep}$*	23.2	16.3	9.7	8.0	14.7	10.1	9.5	13
DeBERTa $_{WRS\_sep}$	16.3	24.9	10.3	9.1	10.0	11.6	13.8	13.7
RoBERTa $_{focal\_sep}$	10.9	**25.0**	8.1	8.8	**22.9**	**16.2**	**17.7**	**15.7**
DeBERTa $_{focal\_sep}$	15.7	24.8	10.0	9.2	16	11.2	14.8	14.5
RoBERTa $_{baseline}$	**35.35**	0.0	**16.7**	0.0	0.0	20.8	0.0	10.4

Table 3: F1 scores: Sub-Task2

3.0.1 Focal Loss

Focal loss (Lin et al. (2018)) is an improved version of Cross-Entropy Loss that tries to handle the class imbalance problem by assigning more weights to hard or easily mis-classified examples and down-weight easy examples. It results in the reduction of the contribution of easy examples. It also makes so that there is more emphasis on correcting misclassified examples.

$$L = \begin{cases} \alpha(1-p)^{\gamma}\log(p) & \text{if y} = 1 \\ (1-\alpha)p^{\gamma}\log(1-p) & \text{otherwise} \end{cases} \quad (1)$$

where p is model prediction and y is the ground truth label; α and γ are hyper-parameters, α is used to control the loss weight of positive and negative samples, and γ is used to scale the loss of difficult and easy samples. The values we take for α is 0.25 and γ is 2.0 which is the default values used in the original paper.

3.0.2 Weighted Random Sampling

Data sampling provides a way to transform a training dataset to better balance the class distribution. Data sampling techniques are helpful in cases of data imbalance as the class distribution is skewed, resulting in the models predicting the dominant class more while learning to ignore the classes with very few samples.

We use Weighted Random Sampling (WRS), which samples items from our set such that the probability of sampling item i is proportional to a given weight w_i which is equal to the class weight for the label of the i'th item.

$$w_i = 1/n_i \quad (2)$$

Here n_i is the number of items in the dataset with label i.

3.1 Sub-task1

The first sub-task is a binary classification task. We use our pre-trained BERT based transformers for this task. We experiment with either using Focal Loss or Weighted Random Sampling to deal with imbalanced data. We pass the output of the transformer model through a fully connected layer; we add a Tanh activation function with a dropout layer before passing it through our final fully connected layer, which gives us the output. We also train an ensemble of models using 5-fold cross-validation and use hard voting method to decide the final labels and combine it with Weighted Random Sampling for dealing with class imbalance

3.2 Sub-task2

For the second sub-task, we have a multi-label classification problem. We experiment with treating it as multiple binary classification tasks where for each label to be predicted, we train a separate classifier(sep). Even here, we experiment with focal loss and Weighted Random Sampling to deal with imbalanced classes.

We also train an ensemble of multi-label classifiers using 5-fold cross-validation and hard voting to decide on the final labels. As not all labels are imbalanced, we decided to use binary cross-entropy as our loss function for ensemble models for this task. We add weights to the positive samples in the loss function as done by researchers at (Gupta et al., 2021) to address the classes which do have imbalances. The formula is given below:

$$\ell(\mathbf{x}, \mathbf{y}) = -\frac{1}{Nd}\sum_{n=1}^{N}\sum_{k=1}^{d}\left[p^k y_n^k \log x_n^k + (1-y_n^k)\log(1-x_n^k)\right]$$
$$p^k = \frac{1}{f^k}(|K| - f^k) \quad (3)$$

Where N is the batch size, n index denotes n^{th} batch element, d is the number of classes, f stands for a vector of class absolute frequencies calculated

on the train set, $\mathbf{x}$ is the output vector from the last Sigmoid layer, $\mathbf{y}$ is a vector of multi-hot encoded ground truth labels and $|K|$ is the size of the train set.

Model	Precision	Recall	F1
RoBERTa $_{\text{focal}}$	64.6	45.4	53.3
DeBERTa $_{\text{focal}}$	**68.3**	34.7	46.0
RoBERTa $_{\text{WRS}}$ *	51.5	59.9	55.4
DeBERTa $_{\text{WRS}}$	50.1	65.6	56.8
RoBERTa $_{\text{ens\_WRS}}$	53.2	**67.1**	**59.4**
DeBERTa $_{\text{ens\_WRS}}$	56.2	59.6	57.9
RoBERTa $_{\text{baseline}}$	39.35	65.3	49.1

Table 4: Results: Sub-Task1

4 Experimental setup

Parameter	sub-task1	sub-task2
Dropout	0.3	0.3
BatchSize	8	8
Epochs	5	8
Learning Rate	1e-05	1e-05
Optimizer	Adam	Adam

Table 5: Hyperparameters

The dataset contains 10469 paragraphs about potentially vulnerable social groups. 9476 examples were marked negatively for PCL, while 993 were marked positively for PCL. For the second sub-task, only the 993 examples were used for training as they were marked positively for PCL.

80% of the dataset was used for training while the rest was used for validation. We only use the organisers' test set for testing out our final models. Hyper Parameters used are mentioned in table 5. Not much time was spent on hyperparameter tuning as using other previously mentioned techniques and different models gave better and more varied results.

The primary evaluation metrics used is the F1 scores. For sub-task1, precision and recall scores are also given. For sub-task2, we have individual F1 scores for each PCL subcategory along with the average F1-score. We use huggingface [1] library for our transformer models implemented in PyTorch [2].

The models were trained on 4 GeForce RTX 2080 Ti GPUs.

5 Results And Discussion

The results from all our experiments conducted for sub-task1 and sub-task2 can be seen in Tables 4 and 3, respectively. The models submitted in the evaluation phase are marked * in the tables, but we have shown results from all our experiments. We experimented with several models and techniques during the development and evaluation phases. We use the F1 score to judge our models, which is also the official metric. We ranked 23rd on the first task and 36th on the second task on our submitted models. We achieved better results on other models for both tasks, and the results are shown in their tables, respectively.

We see that all methods, namely focal loss, Weighted Random Sampling(WRS) and ensembling performed better than the baseline model. The 5-fold cross-validation, hard voting ensemble model with WRS achieves the best F1-score and Recall score for sub-task1, more than the models where only WRS is applied.

For the second task, we see the best average score from RoBERTa model trained separately (sep) for each label with focal loss to achieve the best average F1 score. Focal loss performs poorly on Unbalanced power relations, which has the highest number of positive samples (716 out of 993) and performs better on imbalanced labels having a lower number of positive samples like Metaphors and Poorer The Merrier having 197 and 40 positive samples out of 993 respectively.

6 Conclusion

This paper presents and describes our solution system for the SemEval2022 Task4: Towards Patronizing and Condescending Language Detection. We have applied BERT based pre-trained language models RoBERTa and DeBERTa with hard voting ensembling techniques along with techniques to deal with imbalanced datasets like focal loss and Weighted Random Sampling. Our submitted solutions scored F1 scores of 0.5539 and 0.1456 for the two sub-tasks, respectively.

[1] Transformers,v4.16.2,https://huggingface.co/docs/transformers/index

[2] PyTorch, v1.10.2, https://pytorch.org/

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel

Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised learning of universal sentence representations from natural language inference data.

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.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

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

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. 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. 2019b. Roberta: A robustly optimized bert pretraining approach.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2019. Cardiff University at SemEval-2019 task 4: Linguistic features for hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 929–933, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Carla Pérez-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.

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)*. Association for Computational Linguistics.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.