Taygete at SemEval-2022 Task 4: RoBERTa based models for detecting Patronising and Condescending Language

Jayant Chhillar
C2FO
Noida, India
IIIT-Delhi
Delhi, India
jayant17154@iiitd.ac.in

Abstract

This work describes the development of different models to detect patronising and condescending language within extracts of news articles as part of the SemEval 2022 competition (Task-4). This work explores different models based on the pre-trained RoBERTa language model coupled with LSTM and CNN layers. The best models achieved 15^{th} rank with an F1-score of 0.5924 for subtask-A and 12^{th} in subtask-B with a macro-F1 score of 0.3763.

1 Introduction

The use of Patronising and Condescending Language (PCL) in text or speech can affect healthy communication channels adversely. The effect of PCL on the vulnerable sections of society have been widely studied. PCL acts as a catalyst for discriminatory behaviour (Mendelsohn et al., 2020) against various vulnerable groups. It has been observed to promote exclusion and discrimination among communities and provide a conducive environment for rumour spreading and misinformation (Nolan and Mikami, 2013). These negative effects of PCL are unaffected by the intent of the writer/speaker who might have unknowingly used PCL. These reasons provide a strong argument for developing methods that can identify and prevent unwanted use of PCL in news articles, blogs, and other pieces of text.

In subtask-A of the Patronizing and Condescending Language Detection task at Semeval-2022 (Pérez-Almendros et al., 2022), the goal is to develop a model which takes a sample text as an input and outputs a label indicating the presence or absence of PCL. In subtask-B, the model was required to identify the correct set of PCL categories. The model takes in a sample text as an input and return seven separate outputs each indicating the presence or absence of the pre-defined seven categories. The dataset for the task was shared by the task organisers in the English language. To tackle these tasks, RoBERTa based models were developed. Different variations of the models involved the use of feed-forward layers, LSTMs, CNN and their combinations. For subtask-A RoBERTa with LSTM, CNN and feed-forward layers outperformed all the other variations with an F1 score of 0.5924. In subtask-B RoBERTa with feed-forward layers got the best F1 score of 0.3763 as compared to the other variations. For subtask-A, this work achieved 15th rank in the leader board and 12th rank for subtask-B details of which are discussed in section 3.2.

2 Background

Identification and analysis of PCL in text is well explored in linguistics (Aggarwal and Zhai, 2012), politics (Huckin, 2002), sociolinguistics (Thapar-Björkert et al., 2016) and other fields. However, in NLP it is still heavily unexplored and starting to gain traction. In the past topics such as sentiment analysis (Feldman, 2013), offensive speech identification (Safaya et al., 2020) and fake news identification (Shu et al., 2017) have been significantly worked upon. One major roadblock in exploring PCL in the text is the lack of well structured and labelled dataset. Recently, some new work has been developed to tackle this issue. Wang et al. (Wang and Potts, 2019) developed a model for identifying the condescending language in Reddit threads and also developed an annotated dataset for the same.

2.1 Dataset

For training and development of the model presented in this work "The Don't Patronize Me!" dataset (Perez-Almendros et al., 2020) was used. The dataset contains paragraphs in the English language extracted from the News on Web (NoW) corpus. It comprises 10469 samples out of which 993 have been classified as positive samples, i.e. they contain PCL. The dataset categorizes PCL into 7 different sub-categories, namely, Unbalanced power relations (UPR), Shallow solution

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 496 - 502
July 14-15, 2022 ©2022 Association for Computational Linguistics

Sentence	Keyword	Label
In September , Major Nottle set off on foot from Melbourne to Canberra to plead for a national solution to the homeless problem .	homeless	1
10:41am - Parents of children who died must get compensation , free medicine must be provided to poor families across UP : Ram Gopal Yadav	poor-families	1
Today , homeless women are still searching for the same thing . A place to sleep and be safe	homeless	0
For refugees begging for new life , Christmas sentiment is a luxury most of them could n't afford to expect under shadow of long-running conflicts	refugee	0

Table 1: Sample text and keyword pairs along with corresponding labels.

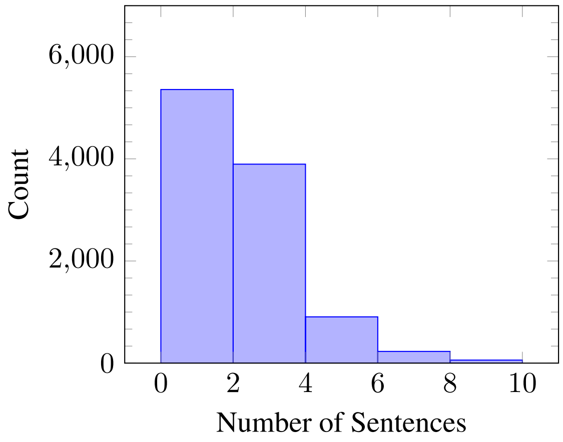

Figure 1: Distribution of the number of sentences per sample.

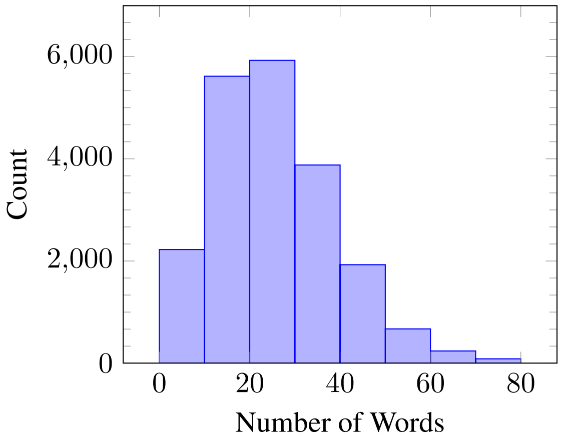

Figure 2: Distribution of the number of words per sentence in a sample.

(SSL), Presupposition (PS), Authority voice (AV), Metaphor (MTP), Compassion (CMP), The poorer - the merrier (PM). Each positive sample can belong to any combination of these categories. The distribution of each category out of all the positive samples is described in Figure 4.

For developing the models for subtask-A the dataset also provides binary labels (0 or 1) to signify the presence or absence of PCL in the text. Along with the paragraphs, the dataset includes the country of origin of the original article and keywords that occur in the paragraph under consideration. These keywords comprise the following, Disabled, Homeless, Hopeless, Immigrant, In need, Migrant, Poor Families, Refugee, Vulnerable and Women. These keywords are usually present in texts that concern the vulnerable sections of society (refer Table 1).

3 System Overview

This section describes the different model designs explored for Task A and Task B and the pre-processing techniques employed. Section 3.1 describes the pre-processing techniques and how they tackle the challenges offered by the dataset. The different models are described under section 3.2 along with a description of the different sub-components and the underlying intuition.

3.1 Data pre-processing

The "The Don't Patronize Me!" dataset offers primarily three major challenges, which are, low number of samples, high class imbalance and the low context in the textual data (smaller sentence length). To deal with high class-imbalance and lower number of samples data augmentation techniques, loss weighting strategies were adopted and to address the low context issue, keywords shared in the dataset were used to provide added context to the models.

3.1.1 Tokenisation

Each sample was tokenised using RoBERTa (Liu et al., 2019) tokenizer. To identify the optimal Tokenisation length analysis was done on the distribution of the number of a sentence per sample (Figure 1) and the distribution of the number of words for

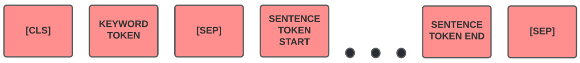

Figure 3: An example of tokenisation

each sentence (Figure 2). On analysing the two distributions length of 50 to 60 tokens seemed a viable candidate for Tokenisation operation. However, on further analysis, it was found that out of 993 positive samples, 193 (19.43 %) had more than 75 words. Thus, to prevent loss of information Tokenisation was done with a length of 100.

Each tokenised sentence was prepended with a tokenised keyword corresponding to that sample separated by the SEP token. Finally, the Tokenisation process was completed by adding a CLS and SEP token at the beginning and the end respectively (refer Figure 3).

3.1.2 Data augmentation

For data augmentation back-translation method was explored. Back-translation is the process of using a language model to translate a text from its parent language to another language, generally using a language model. The new text is then translated back to its parent language. This method introduces slight changes in the structures of the text while retaining the underlying context. This method has been shown to boost the performance of models trained over smaller datasets.(Sennrich et al., 2016). Helsinki-NLP models [1] were used to translate a sentence from English to French and back to English. Only 30 per cent (randomly sampled) of the positive samples from the dataset were back-translated.

3.1.3 Loss weighting

Initial exploratory analysis of the dataset has shown high class imbalance. To address this issue cost-sensitive re-weighting technique developed by Cui et al (Cui et al., 2019) and suggested by Jurkiewicz et al (Jurkiewicz et al., 2020) was adopted. The weighting factor for each class was identified as per the following definition:

$$(1-\beta)/(1-\beta^{n_i}) \tag{1}$$

where β is a hyper-parameter in [0,1), and n_i is the number of samples belonging to the class i. Using these weights the updated softmax cross-entropy loss is given as:

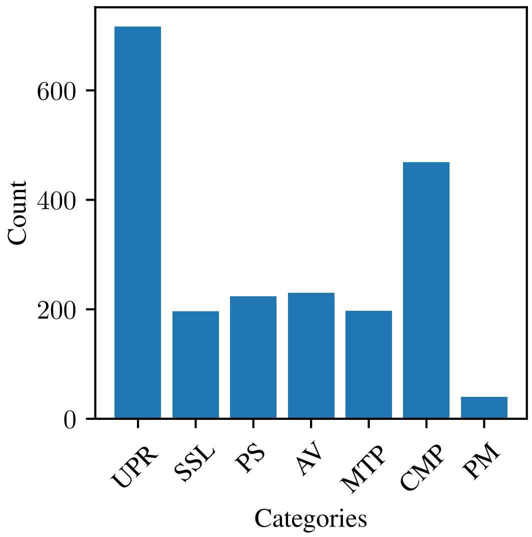

Figure 4: Number of samples for each of the seven PCL classes

Model	BASIC	AUG	WT
RB-FNN	**0.6177**	0.6301	0.6080
RB-BiLSTM	0.6140	**0.6305**	0.6258
RB-CNN	0.5879	0.5954	0.6037
RB-BLS-CNN	0.6059	0.6095	**0.6318**

Table 2: Analysis of the models trained under WT, AUG and BASIC setting for subtask-A

$$L(z,y)=\frac{1-\beta}{1-\beta^{n_i}}\log\left(\frac{\exp\left(z_y\right)}{\sum_{j=1}^{C}\exp\left(z_j\right)}\right) \tag{2}$$

where $z=[z_1,z_2,...,z_C]$ is the predicted output of the model for C classes and y being one of the possible class labels, i.e. $y\in C$

3.2 Model description

This work explores four different model designs. Each design includes RoBERTa_{LARGE} (Liu et al., 2019) as it's base layer. The output of the last hidden state (shape = 106 X 1024) is then further fed down the network to get the final prediction. For subtask-A, all the models perform binary prediction (0 = no PCL, 1 = contains PCL) to identify if the input text contains PCL, while for subtask-B each model produces 7 binary predictions, one for each possible PCL category. The design of the four models remains the same for both tasks except for the number of outputs generated by them (Figure 5). Binary Cross-Entropy (BCE) loss was used for

[1] https://huggingface.co/Helsinki-NLP

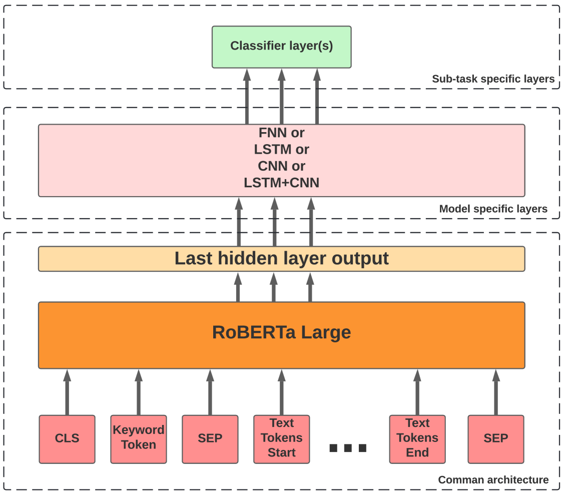

Figure 5: Generalised architecture of the models developed. For subtask-A classifier layers consist of single FNN with 2 units. For subtask-B classifier layers consist of 7 FNN layers each with 2 units.

all the outputs. Adam optimizer was utilized with a learning rate set to 1e-6 and epsilon at 1e-6.

3.2.1 RB-FNN

The model employs the use of two feed-forward layers added on top of RoBERTa$_{\text{LARGE}}$. The output of the last hidden layer is flattened and passed down the model. The initial feed-forward layer has 106 units. For subtask-A, the output of this hidden layer is passed on to a single feed-forward layer with 2 units for binary prediction, while for subtask-B the output is shared by seven feed-forward layers each with 2 units predicting the presence of each sub-category of PCL.

3.2.2 RB-BiLSTM

LSTM is a type of recurrent neural network (RNN) that allows the model to learn underlying features in temporal data without the added drawbacks of general RNN models such as exploding or vanishing gradients. LSTM allows the model to capture the long term dependencies in the data and identify the underlying temporal nature of the data(Tang et al., 2015). LSTMs have shown to achieve state of the art performance in different text classifications tasks (Tang et al., 2015) and (Li et al., 2020). Shi and Lin (Shi and Lin, 2019) also showed that using LSTM coupled with BERT can improve the performance compared to BERT by itself. For this model the output of the last hidden layer of RoBERTa$_{\text{LARGE}}$ model is fed into a Bi-Directional LSTM layer with 106 units. The output of the BiLSTM layer is then fed down to two FNN layers with 106 and 2 units respectively (subtask-A). For subtask-B, the output of the first FNN layer is fed to seven feed-forward layers each with 2 units.

3.2.3 RB-CNN

CNN based models have been shown to perform well for various text classification problems (Chen, 2015) (Safaya et al., 2020). CNN layers are able to capture the semantic relationships within the textual data and given the structured nature of the embeddings obtained from RoBERTa$_{\text{LARGE}}$ model it seemed beneficial to use CNN layers to extract the hierarchical features within the data (Rodrigues Makiuchi et al., 2019). In this model, the last layer embeddings of the RoBERTa$_{\text{LARGE}}$ model are fed to two CNN layers coupled with a max-pooling layer. The first CNN layer comprises 64 10X10 filters with stride 1 and the second layer comprises 32 5X5 filters with stride 1. After each CNN layer, a two-dimensional max-pooling operation is done with a shape of 2X2. The output of the last max-pooling operation is fed to an FNN layer with 106 units which is followed by an FNN layer with 2 units (subtask-A). For subtask-B, the output of this FNN layer is fed to seven feed-forward layers each with 2 units.

3.2.4 RB-BLS-CNN

To get the model to learn both temporal and hierarchical features within the data a hybrid model was developed employing both LSTM and CNN layers. This model is created as an amalgamation of the RB-BiLSTM and RB-CNN models. The last layer RoBERTa$_{\text{LARGE}}$ embeddings are fed to an LSTM layer with 106 units. The output of the LSTM layer is then further fed to the CNN architecture defined in the RB-CNN model. The final FNN layer with 106 units is then further fed to a single FNN layer with 2 units for subtask-A and to seven separate FNN layers each with 2 units for subtask-B.

4 Experimental setup

To gauge the effect of data augmentation and loss weighting techniques on the performance of models for each subtask four experiments were carried out (Table 6). The goal was to identify how both the techniques interacted with each other and to find the right combination for each subtask. For each experiment, the model was trained on 80 per cent of the data as the training set and 20 per cent

Model	Macro F1	UPR F1	SSL F1	PS F1	AV F1	MTP F1	CMP F1	PM F1
Baseline	0.1041	0.3535	0	0.1667	0	0	0.2087	0
RB-FNN	**0.3763**	**0.5969**	**0.4578**	**0.3333**	**0.2178**	**0.3043**	**0.536**	**0.1875**

Table 3: F1 score comparison on evaluation dataset for subtask-B between the RoBERTa baseline shared by task organisers and the RB-FNN under AUG experimental settings.

Model	F1 score	Precision	Recall
Baseline	0.4911	0.3935	0.653
RB-BLS-CNN	**0.5924**	**0.5357**	**0.6625**

Table 4: F1 score comparison on evaluation dataset for subtask-A between the RoBERTa baseline shared by task organisers and the RB-BLS-CNN under WT experimental settings.

Model	BASIC	AUG	WT
RB-FNN	**0.4054**	**0.4082**	0.3158
RB-BiLSTM	0.3643	0.3818	0.2880
RB-CNN	0.3594	0.3903	**0.3180**
RB-BLS-CNN	0.3599	0.3519	0.2871

Table 5: Analysis of the models trained under WT, AUG and BASIC setting for subtask-B

Exp	Augment	Loss Weighting
BASIC	No	No
AUG	Yes	No
WT	No	Yes
AUG+WT	Yes	Yes

Table 6: Different experiments carried out on each model.

as the validation set. The 80-20 split shared by the task organisers was used. F1 score for subtask-A and macro F1 score for subtask-B were chosen by the task organisers as the criteria to identify the best performing model, thus the same was used to evaluate the performance of different models created for the two subtasks under different experimental settings. For each experiment, training was done for 20 epochs with a batch size of 8. The best version of the model from each experiment was used to generate predictions for the evaluation dataset.

5 Results

For subtask-A RB-BLS-CNN under WT experiment achieved the highest F1 score of 0.5924 with a precision of 0.5357 and recall of 0.6625 on the evaluation dataset. While on the validation dataset the same model received an F1 score of 0.6318 with a precision of 0.5685 and recall of 0.7109.

For subtask-B RB-FNN performed best out of all the models under the AUG experimental settings. The model achieved a macro F1 score of 0.4006 on the validation dataset and 0.3763 on the evaluation dataset.

The minute difference in the F1 scores of the best models for the evaluation dataset and the validation dataset shows that the model did not overfit during the training phase despite a large number of training epochs.

The effect on class re-weighting (refer 3.1.3) and data augmentation was also explored (refer Table 2 and Table 5). It was found that for subtask-A a majority of the four models received a boost in the F1 score when class re-weighting was applied as compared to the BASIC experimental setting. However, this trend was absent for all the models of subtask-B. Rather class-weighting had a detrimental effect on the models for subtask-B as shown in Table 5. The low number of samples for each of the seven sub-classes coupled with the added complexity of the task as compared to subtask-A could have been the underlying cause behind this observation.

Similarly, the effect of data augmentation on model performance was also explored (refer to Table 2 and Table 5). For subtask-A, all the different models received a boost as compared to the models without augmented data. The same trend persisted for the majority of models trained for subtask-B.

Another interesting find in subtask-B was the significantly poor performance of the LSTM and CNN based models as compared to the vanilla RoBERTa model i.e. RB-FNN. This is not in line with the trend observed for the models in subtask-A (Table 2). The reason for this result could be similar to the unexpected trend observed for the models of subtask-B in class re-weighting experiments. Especially for LSTM based models as a large number of samples are required to train models that employ LSTMs in their design.

	UPR	SSL	PS	AV	MTP	CMP	PM
Precision	**0.6076**	0.3684	**0.5667**	**0.3428**	**0.5652**	**0.6521**	**0.6666**
Recall	0.5563	**0.3889**	0.2741	0.3157	0.25	0.4245	0.1818

Table 7: Precision and Recall values for RB-FNN model under AUG experimental settings on test data.

6 Conclusion

This work explored the design and training of different RoBERTa based models for PCL detection in text. The added benefits of using CNN and LSTM layers along with RoBERTa in boosting model performance was also shown. This work also explored the effects of using back translation as a data augmentation technique along with a class re-weighting technique to deal with low sample size and high class imbalance. Finally, the challenges offered by the models under different problem statements were explored which gives a deeper insight into the impacts of different design methodologies. The best models achieved 15th and 12th rank for subtask-A and subtask-B respectively.

Future work can include expanding the dataset with more data as the current dataset includes 10469 samples. Also, the original article can be provided against each sample which can be fed to the model as added context. This added context should significantly boost model performance.

7 Acknowledgments

The author is grateful for the support of Dr Rajiv Ratn Shah and for the guidance offered by Aradhya Mathur. The author would also like to thank the anonymous reviewers for their insightful feedback.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

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

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. Applicaai at semeval-2020 task 11: On roberta-crf, span cls and whether self-training helps them. *arXiv preprint arXiv:2005.07934*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

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*.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70.

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.

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.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Suruchi Thapar-Björkert, Lotta Samelius, and Gurchathen S Sanghera. 2016. Exploring symbolic violence in the everyday: misrecognition, condescension, consent and complicity. *Feminist review*, 112(1):144–162.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.