# Tuning and Interpreting BERT for Patronizing and Condescending Language

**Thomas Gao**
University of California, Berkeley
tgao2020@berkeley.edu

## Abstract

In support of SemEval-2022 Task 4, Patronizing and Condescending Language (PCL) Detection, we present a fine-tuned BERT model for identifying PCL embedded in news articles, which meaningfully outperforms baseline models with an F1-score of 0.52. We apply the Integrated Gradients method to attribute the model's prediction to individual words, thereby revealing the learned knowledge of the model. We found that the model encodes explicit PCL vocabulary well, but has bias for racial and religious terms. The model falls short of truly understanding PCL, struggling to differentiate factual statements and PCL and capturing PCL without explicit words.

## 1 Introduction

Patronizing and Condescending Language (PCL) is language use that shows a discourse of pity and superior attitudes towards others. PCL usage in general media, while often well-intentioned, might nevertheless routinize discrimination and lead to greater inequalities (Pérez-Almendros et al., 2020).

In SemEval-2022 Task 4: Patronizing and Condescending Language Detection, the organizers created the Don't Patronize Me! dataset, which contains over 10,000 paragraphs about vulnerable communities extracted from news stories. The dataset has been annotated to indicate the presence of PCL, categories of PCL and corresponding text spans. Our research focuses on subtask 1, which is to predict the presence of PCL as a binary classification problem.

To address this problem, we experiment with Logistic Regression with the bag-of-words model, CNN with Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), and Bidirectional Encoder Representation from Transformer (BERT) (Devlin et al., 2018). Our up-sampled BERT model yields an F1-score of 0.52, meaningfully better than the baseline models. For the BERT model, we use Integrated Gradients (Sundararajan et al., 2017) to shed light on the model knowledge and mistakes, which can be used as pointers for future research.

In the rest of this paper, we organize the content as follows: related work of harmful language modeling in section 2; section 3 introduces data description, preprocessing, and model methodologies; section 4 discusses experimental results; section 5 presents sample model predictions and their attributions; section 6 discusses interesting areas for future work. We also present the conclusion of our work at the end of the paper.

## 2 Related Work

There were similar tasks aimed at studying harmful languages. In SemEval 2019 Task 6 Identifying and Categorizing Offensive Language (Zampieri et al., 2019), models used ranged from traditional machine learning, e.g. SVM and logistic regression, to deep learning, e.g. CNN, RNN, BiLSTM, ELMo and BERT. Among the top-10 teams, seven used BERT. For the classification task, the top team experimented with different models including linear models and LSTM, and found pre-trained BERT with fine-tuning performed best. The runner-up team also used BERT model and applied techniques to address the class imbalance in the training data.

Compared to offensive language, patronizing language in news articles is more challenging to model because of the subtlety. However, the organizers show that BERT-based approaches show non-trivial results and outperform simpler models (Pérez-Almendros et al., 2020). A review of submitted papers for the task show that the top performing team PALI-NLP uses a novel Transformer-based model BERT-PCL with two discriminative fine-tuning strategies (Hu et al., 2022), and another team achieving top quartile performance ensembles five RoBERTa models that are seeded differently (Zhao and Rios, 2022). Team SATLab uses a

| Data | Train | Dev | Total |
|---|---|---|---|
| Paragraphs | 8375 | 2094 | 10469 |
| PCL(#) | 794 | 119 | 993 |
| PCL(%) | 9.48% | 9.50% | 9.49% |

Table 1: Data distribution: class imbalance with most text not including PCL.

logistic regression model only fed with characters and word n-grams and obtained performance lower than the best teams (Bestgen, 2022).

Our work therefore studies the BERT-based models. Specifically, we fine-tune the BERT model to reproduce the meaningful results and use Integrated Gradients (Sundararajan et al., 2017) method to analyse BERT's learned knowledge and predictions.

## 3 Data and Methodology

### 3.1 Data Description

The Don't Patronize Me! Dataset includes paragraphs about vulnerable communities extracted from news stories from the News on Web (NoW) corpus (Davies, 2013). Articles with at least one word from a list of selected keywords, e.g. disabled, homeless, women, vulnerable, were split into paragraphs and then randomly selected.

The data was annotated by three annotators, with two annotators annotating the whole dataset and the third one acting as a referee to provide a final label in case of disagreement. In the first step, annotators determined which paragraphs contain PCL. In the second step, the annotators indicate, in paragraphs containing PCL, the text span of PCL and its category. Out of the seven categories, the most common ones are Unbalanced Power, Compassion, and Presupposition.

The organizer created an 80/20 split of the training dataset. The dataset distribution is summarized in Table 1. They also reserved 3897 paragraphs for scoring. From the table, we observe class imbalance with the vast majority of paragraphs not containing PCL.

Regarding the text span annotation and PCL categorization, it is worth noting that between the two annotators, there are 1359 instances of agreement and 1401 instances of misalignment, demonstrating the subjective nature of PCL.

### 3.2 Preprocessing

**Data imbalance** In order to create a more balanced training dataset, we reduce the number of paragraphs without PCL to the first 1588 paragraphs, i.e. twice the number of paragraphs with PCL. This gives us 2382 training examples, of which 794 contain PCL. We also create an upsampled dataset by duplicating the minority class 5 times, creating a dataset with 11551 training examples, of which 3970 contain PCL, which is used to train the final BERT model.

### 3.3 Methodology

**Epoch** We train our deep learning models for 3-5 epochs and select the epoch that produces the highest F1-score on the dev dataset. We also monitor the validation loss and consider only epochs before validation loss rising to avoid overfitting.

**Logistic Regression** We use CountVectorizer from scikit-learn library to transform text into features. Then, we fit Logistic Regression as our baseline model to determine the baseline performance.

**CNN with GloVe** We use 300-dimensional word embeddings from GloVe (Pennington et al., 2014) to represent the text, and train single-channel CNN with 128 filters, kernel size of 5, stride of 1, 256 hidden units and L2 regularization of 1e-05.

**BERT** BERT model (Devlin et al., 2018) uses Transformer architecture (Vaswani et al., 2017) and pretrains on a massive amount of texts to learn context-aware word embeddings and sentence embeddings. We download the pre-trained *bert-base-cased* model from Hugging Face and fine-tune the model by attaching a dense layer of 256 hidden nodes to the sentence embedding vector, i.e. [CLS] token embedding. We then include a dropout layer with a dropout rate of 0.1 before making a prediction with the sigmoid activation function. We use binary cross entropy loss, Adam optimizer with a learning rate of 5e-5 and max length of 128 words and train all layers.

## 4 Experiment Results

The official evaluation metric of this task is the F1-score of the positive class. Predicting the majority class, i.e. no PCL, would produce F1-score of 0 because recall is 0. Random guess would F1-score of 0.1604.

Our machine learning models all perform better than random guesses, with the BERT models producing meaningfully better results than Logistic

Regression (BOW) and CNN (GloVe). The upsampled model produces a higher F1-score than the downsampled model, as expected given more data. We submitted our predictions on the test data reserved by the organizers, and achieved an F1-score of 0.5068, slightly lower than the score on the dev dataset. See Table 2 for model performance.

## 5 Model Interpretation

To understand how the BERT model is making the predictions, we examine the contribution of each word to the overall prediction using the Integrated Gradients method (Sundararajan et al., 2017).

For feature embedding, we use the BERT input embedding; for baseline embedding, we use zero embedding for word tokens and BERT input embedding for non-word tokens. Between the input word embedding and baseline word embedding, we create 40 equally spaced embedding steps. We record the gradients of the final prediction with respect to the input values at each embedding step, and compute the attribution of each word by summing the products of gradients and step embedding.

Figure 1 shows top examples where the model is most confident. We attribute the model predictions to words, with green and red backgrounds indicating positive and negative contribution to PCL prediction and the darkness of colors corresponding to the magnitude of the contribution. We also use the span annotations where the two annotators are in agreement as the gold span, colored in orange, and provide the PCL category. Additional examples are included in the Appendix A.

In the first example, annotators consider the sentence to contain PCL because of presupposition and compassion. Presupposition assumes a situation as certain without having all the information. Compassion presents the vulnerable people as needy, raising a feeling of pity. Our judgment is that both PCL categories refer to the word *hopeless*. From the green highlights, we see that the model rightly considers *hopeless* as the main PCL driver. We also see that the word *Africans* is assigned a green color, raising concern about racial bias in the data and the model.

In the second example, the model predicts positive class confidently based on the description of the videos. The subjects are described as being in a completely wretched and helpless state and words such as *dragged*, *worthlessness*, *hopelessness* are highlighted as main contributors. However, the annotators judged that these are not PCL, and we concur because these words show up in a factual statement about the videos. The model expertly observed relevant words but failed to understand the nuance between facts and PCL.

The third example does not contain explicit PCL words. The model predicts false on the basis of *month*, *said*, *undergraduate*, *universities*, *colleagues*. The first two are neutral vocabularies while the latter three words might be associated with capability and enablement. The annotators categorized this as Unbalanced Power most likely due to the intentional comparison between able-bodied and disabled students. We agree that the language would cause harm to the disabled community, and it is an example of implicit PCL.

In summary, our model is capable of making note of vocabularies that are likely to be associated with PCL. However, as shown in our examples, presence of such words does not automatically mean PCL, and PCL can also be implicit and not contain any explicit PCL words. In Appendix A, we also observe that vulnerable and religious vocabularies are associated with PCL, leading to false positive predictions. On the other hand, certain seemingly innocent words, e.g. *administration*, can sometimes dominate in false negative predictions. This might be due to overfitting and supports the case of ensembling.

## 6 Future Work

Due to time limitations, there are additional architectures that we would like to experiment with. One such task is to explore multi-objective learning by including an additional training objective of PCL span detection using the annotated span data. We expect that by highlighting the specific text spans that contain PCL, the model can gain more accurate understanding and achieve better prediction. The study would also reveal the usefulness of highlighting text span in addition to assigning binary labels when annotating data.

## 7 Conclusion

In this paper, we explore how pre-trained BERT models can be fine-tuned for classifying whether a paragraph contains PCL. This standard approach achieves an F1-score of 0.52, a non-trivial result compared to baselines.

We also use the Integrated Gradients method to attribute the model's decisions and gain insights

| Model | Epochs | Precision | Recall | F1 |
|---|---|---|---|---|
| Random | NA | 0.0955 | 0.5000 | 0.1604 |
| Logistic Regression (BOW) | NA | 0.2588 | 0.4824 | 0.3368 |
| CNN (GloVe) | 2 | 0.3127 | 0.4824 | 0.3794 |
| BERT (Down Sampled) | 1 | 0.3537 | 0.7286 | 0.4762 |
| BERT (Up Sampled) | 1 | 0.5464 | 0.5022 | **0.5236** |

Table 2: Model Performance for PCL Classification on Dev Data.

| Text, Attribution, Gold Span | PCL Category | Result |
|---|---|---|
| Hundreds of thousand Africans are graduating per year . Different from 1980s and early 1990s when college out ##po ##urs got immediately absorbed in the labour market , many today are job ##less and hopeless . | Compassion, Presupposition | True positive |
| The City Without Drugs organisation is still active , as is their YouTube channel . It features hundreds of videos of drug add ##ict ##s being dragged half - conscious through the street , their faces not blurred , or confess ##ing their alleged worthless ##ness , their hopeless ##ness , their shame . | NA | False positive |
| Cheung said 20 disabled undergraduate students from seven universities will start their eight - week internship in government departments this month , receiving the same salaries as able - bodied colleagues of HK $ 9 , 600 a month . | Unbalanced Power | False negative |

Figure 1: True Positive, False Positive and False Negative Examples with Highest Confidence

into the model's learned knowledge. We find that the model encodes explicit PCL vocabulary well, but falls short on differentiating between facts and PCL and capturing PCL without explicit words. We also find evidence of racial and religious bias in the model.

PCL is harmful but often unconscious language use due to a lack of perspectives of the vulnerable communities. Machine learning has a great opportunity to encode perspectives from vulnerable communities and reduce such harmful language use in our society.

## Acknowledgements

We would like to thank Joachim Rahmfeld for his guidance and feedback. We are also grateful to the W266 faculties for their wonderful teaching through live sessions, notebooks and paper reading sessions. The authors are responsible for any errors that remain.

## References

Yves Bestgen. 2022. Satlab at semeval-2022 task 4: Trying to detect patronizing and condescending language with only character and word n-grams. *arXiv preprint arXiv:2203.05355*.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. *Retrieved January*, 25:2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dou Hu, Mengyuan Zhou, Xiyang Du, Mengfei Yuan, Meizhi Jin, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. Pali-nlp at semeval-2022 task 4: Discriminative fine-tuning of deep transformers for patronizing and condescending language detection. *arXiv preprint arXiv:2203.04616*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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. *arXiv preprint arXiv:2011.08320*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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). *arXiv preprint arXiv:1903.08983*.

Xingmeng Zhao and Anthony Rios. 2022. Utsa nlp at semeval-2022 task 4: An exploration of simple ensembles of transformers, convolutional, and recurrent neural networks. *arXiv preprint arXiv:2203.14920*.

## A Appendix: Additional Examples

Figure 2 below shows additional examples of classification attribution.

| Text, Attribution, Gold Span | PCL Category | Result | Comments |
|---|---|---|---|
| " " " An ##ja Ring ##gren Love ##n I ca n ' t find a word to describe how I feel for you . . . . May God al ##mi ##ghty keep blessing you and always give you strength and sound health to continue your good work . . . . . You gave hope to the hopeless ! ! ! ! Have so much respect for you . . Stay Blessed my good fellow . . . " " says one comment ##er on Facebook . " " God b ##less you and your mission . Glad to see Hope ( and all the children ) growing up loved , well fed , happy , having fun , and going to school , | Compassion, Unbalanced power | True positive | Correctly highlighted all words in the PCL span. However, general religious words are also highlighted. |
| They sell their me ##ager belongings , they borrow money from family and friends to buy work visa ##s and airline fare ##s , they are either ill ##iter ##ate or semi - liter ##ate , they are mostly un ##ski ##lled souls from long suffering and very poor families . When they start work in Arab she ##ik ##dom ##s , despite being Muslims , they are treated as if they are sub - humans . These re ##sol ##ute and hard ##y souls , despite all the poverty and hard ##ship inflicted upon them by nature , work hard like slaves and send billion ##s of dollars to Pakistan ; with out this money the country could not survive . IF AN ##Y ONE | Compassion, Presupposition | True positive | Unclear. |
| Fast ##ing bridges the gap between rich and poor , sustained and imp ##over ##ished , fulfilled and needy . This experience should then inspire compassion and mercy , which is manifest by gene ##ros ##ity of wealth and time to help those in need . | Unbalanced power | True positive | Model focused on the right words. I would also classify it as Shallow solution. |
| Jesus begins his teaching in Matthew with the Ser ##mon on the Mount . One group he b ##less ##es is those in need of comfort , Blessed are they who m ##our ##n , for they will be comfort ##ed ( Mt 5 : 4 ) . | NA | False positive | Incorrect inference for religious text. |
| " " " Your personal leadership has been critical to addressing the p ##light of the R ##oh ##ing ##ya who fled to safety in your country . I thank you for all you have done to assist these men , women and children in need , " " he wrote in the message . " | NA | False positive | Unclear. |
| " Ad ##op ##t a Mission serves as a platform for the church and for like ##mind ##ed people to reach out to unemployed families in the communities - - whoever is in need . " " < h > The forgotten people of Brooklyn " | NA | False positive | Incorrect inference for religious and vulnerable text. |
| Famous ##ly progressive San Francisco has among America ' s most e ##co - friendly public policies and has declared itself a sanctuary to immigrants it considers per ##se ##cuted by President Donald Trump ' s administration . | Unbalanced power | False negative | Unclear why administration dominated. |
| The measures have kept the migrants living in limb ##o . The overwhelming majority have not been granted asylum and they lead a ten ##uous existence , often at the w ##him ##s of the government . | Presupposition, Compassion | False negative | Unclear. |
| Charity plans to for ##go parking so homeless can have gym and medical centre | Shallow solution | False negative | Dominated by parking. |

Figure 2: True Positive, False Positive and False Negative Examples with Highest Confidence