Team Stanford ACMLab at SemEval-2022 Task 4: Textual Analysis of PCL Using Contextual Word Embeddings

Upamanyu Dass-Vattam, Spencer Wallace, Rohan Sikand, Zach Witzel, Jillian Tang
Stanford University
{udvattam, spenwall, rsikand, zachwitz, jiltang}@stanford.edu

Abstract

We propose the use of a contextual embedding based-neural model on strictly textual inputs to detect the presence of patronizing or condescending language (PCL). We finetuned a pre-trained BERT model to detect whether or not a paragraph contained PCL (Subtask 1), and furthermore finetuned another pre-trained BERT model to identify the linguistic techniques used to convey the PCL (Subtask 2). Results show that this approach is viable for binary classification of PCL, but breaks when attempting to identify the PCL techniques. Our system placed 32/79 for subtask 1, and 40/49 for subtask 2.

1 Introduction

The goal of the task is to be able to identify whether or not a piece of text contains condescending or patronizing language, and if it contains patronizing language identity which linguistic techniques are used to express that sentiment (P'erez-Almendros et al., 2022). Studies have shown that PCL fuels discriminatory behavior, creates and feeds stereotypes, subtly blames needy individuals for their problems, and oversimplifies problems. In general, PCL makes it harder for needy communities to get the help they need and reach total inclusivity (P'erez-Almendros et al., 2020). This is obviously negative, and being able to combat it with AI may help automate the process and get past inherent biases that humans identifying PCL may have.

Our system's goal is to use contextualized word embeddings to feed the model, and thus have the model analyze the semantics of the text. We specifically focused on a purely textual analysis and did not provide the model with any metadata to see if the model could learn the PCL patterns just from the text, because although real world usage would likely include those features, the ability to learn from the text itself would be useful and more universally applicable.

In this task, we learnt that it is easier for a model to simply detect the presence of PCL than the techniques used in a piece of PCL. This is likely because of both unequal distributions of data and the fact that the context for the types of PCL likely look very similar, making the model default to a category of PCL with a higher frequency when not incredibly clear. We ranked 32/70 on subtask 1, which is detecting the presence of PCL, and ranked 40/49 on subtask 2, which involves identifying the PCL techniques used. In particular, the model struggled with the 'authority voice' and 'metaphor' categories.

2 Task Overview

The dataset provided was the Don't Patronize Me! Dataset (P'erez-Almendros et al., 2020), which is a collection of paragraphs mentioning vulnerable communities and published in media in 20 English speaking countries. Each paragraph has the country code where the paragraph was published and the keyword that was used to search for it. For subtask 1, the paragraphs are manually annotated with a label from 0-4 on how much PCL it contains; these are converted to binary labels on whether or not a paragraph contains PCL, where 2-4 indicate positive examples and 0-1 indicate negative examples. For subtask 2, all paragraphs in the dataset contain PCL and are annotated with spans containing categories of linguistic techniques that are used to express the condescension. These categories are:

- Unbalanced power relations
- Shallow solution
- Presupposition
- Authority voice
- Metaphor
- Compassion
- The poorer, the merrier

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 418 - 420

The training portion of the dataset for subtask 1 contained 10,636 paragraphs, and the corresponding dataset for subtask 2 had 993 paragraphs with 2,792 instances of PCL techniques.

3 System Overview

3.1 Data Representation

We felt that attempting to learn from the text itself and removing the contextual metadata could lead to more robust textual analysis. Therefore, we relied solely on the paragraph as our input feature. We tried two approaches: 1) using pre-trained GloVe embeddings with a dimension of 300, which track co-occurences of words in a global corpus (Pennington et al., 2014), and 2) tokenizing the paragraphs using a BERT tokenizer and inputting these into a pre-trained BERT model (Devlin et al., 2019). For subtask 2, where the focus was on shorter labeled spans rather than entire paragraphs, we still used the full text to provide more context for the detection of PCL.

3.2 Subtask 1

GloVe embeddings are good at capturing word analogies due to its global vectorization and its ability to capture sub-linear relationships in the vector space (Pennington et al., 2014). We felt that the analogy ability was specifically important to this task, because it could help capture tonal similarities between instances of PCL. Therefore, we began with a bag-of-words model where we summed the GloVe embeddings of each word in the text, then performed logistic regression to output a binary label indicating whether or not the text contained PCL. However, due to the high class imbalance present in the data, this model predicted no PCL for nearly all inputs.

Our final model was a fine-tuned BERT model. Initially, we re-labeled the training data with the final binary labels of PCL and not PCL, but this led to issues due to the high class imbalance. We decided to make the model a multiclass classifier which output the original 0-4 labels, and then convert these to binary labels. This allowed us to better adjust model weights to reflect the imbalances in class distributions, because the imbalances were not standard to all PCL, and varied dramatically based on the subtype.

3.3 Subtask 2

The most notable decision in this subtask was using the entire paragraph instead of focusing on the spans of PCL. We initially actually tried training on just the labeled spans, but these did not provide enough context for the BERT model to fine-tune to the data. Therefore, we used the entire text as the input to provide more context. Due to the small amount of data for the second subtask, we also chose to apply transfer learning and start our training from the fine-tuned model from the first subtask.

One major problem with the task is that the contexts between the types of PCL are all similar, as there can be many instances of categories within the same paragraph of text in smaller spans. This leads to the model defaulting to predicting the more frequent classes. We tried to address this by adding in the spans without context as training data; however this actually decreased performance.

4 Experimental Setup

We used BertTokenizerFast to tokenize the text and fine-tuned on the pre-trained BertForSequenceClassification model, both from the HuggingFace Transformers library (Wolf et al., 2020). We conducted hyperparameter optimization using the HyperOpt package (Bergstra et al., 2013), using population-based training (Jaderberg et al., 2017). It automatically generated sets of hyperparameters for us, and then based on the results of training with those hyper parameters updated the future hyperparameter sets. Our final model was trained using Adam optimization with a learning rate of 2.31468e-05 (Kingma and Ba, 2014); we trained the model for 6 epochs with a batch size of 8. We used a train/test split of 70/30 and evaluated based on accuracy and F1.

5 Results

Our model had precision 0.4017, recall 0.7666, and F1 0.5271 on the evaluation data for subtask 1, and has an average F1 of 0.0963 for subtask 2. We placed 32/79 for subtask 1, and 40/49 for subtask 2. Based on the results, the model had a hard time with the types of PCL that showed up less frequently in the data, and tended to perform best on the categories that were more frequent. In retrospect, using the same model architecture and setup on the two subtasks was not the optimal way

Metric	Score
Precision	0.4017
Recall	0.7666
F1	0.5271

Table 1: Results for Subtask 1.

PCL Category	Score
Unbalanced Power Relations	0.1596
Shallow Solutions	0.2694
Presupposition	0.0423
Authority Voice	0.0
Metaphor	0.0
Compassion	0.0864
The Poorer, The Merrier	0.1212
F1 Average	0.0963

Table 2: Results for Subtask 2.

to approach the task, despite some compelling reasons to approach it that way. We did not perform any quantitative analysis or ablations, but given the chance we would augment the less frequent PCL categories and see if that would fix the prediction issues for subtask 2, even if it wouldn't necessarily improve accuracy.

6 Conclusion

We developed a system which attempted to first classify whether or not a piece of text contained patronizing or condescending language, and then identify the technique used to convey the PCL. In particular, we focused on examining whether or not pure textual analysis using contextual word embeddings alone would be enough to perform the aforementioned tasks. Based on our results, this approach is only viable for the binary classification of whether or not a text contains PCL.

In the future, we would explore creating an ensemble of models, only one of which uses textual analysis, and the rest would focus on things like meta data and word frequencies which do not rely on context. Comparing the result of that ensemble to a traditional approach to this problem which uses many of those methods simultaneously would show whether or not there is a strong overlap between sources, locations, etc. and PCL, or whether only the text itself is the best indicator of a sign of PCL.

7 Acknowledgements

Thanks to the Stanford ACMLab for facilitating this entry and their instruction, particularly Ethan Chi and Patrick Liu for their reviews.

References

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Journal of the Association for Computing Machinery*, arXiv:1810.04805.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. 2017. Population based training of neural networks. arXiv:1711.09846.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Carla P'erez-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 P'erez-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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.