RNRE-NLP at SemEval-2022 Task 4: Machine Learning Approaches to Detect Patronizing and Condescending Language

Rylan Yang
The Harker School
San Jose, CA
23rylany@students.harker.org

Ethan A. Chi
Stanford University
Stanford, CA
ethanchi@stanford.edu

Nathan A. Chi
De Anza College
Cupertino, CA
chinathan@student.deanza.edu

Abstract

An understanding of patronizing and condescending language detection is an important part of identifying and addressing discrimination and prejudice in various forms of communication. In this paper, we investigate several methods for detecting patronizing and condescending language in short statements as part of SemEval-2022 Task 4. For Task 1a, we investigate applying both lightweight (tree-based and linear) machine learning classification models and fine-tuned pre-trained large language models. Our final system achieves an F1-score of 0.4321, recall-score of 0.5016, and a precision-score of 0.3795 (ranked 53 / 78) on Task 1a.

1 Introduction

Patronizing and Condescending Language (PCL) is characterized by expressions that reveal a sense of compassion or superiority toward others. Research suggests that PCL can perpetuate–and even veil–discrimination toward vulnerable groups (Ng, 2007). To make matters worse, its presence is often more subtle than other offensive language (Mendelsohn et al., 2020).

Detecting PCL is a challenging task for humans annotators–and the task becomes even trickier for artificial systems. Given the varied nature of condescension, current NLP models struggle to accurately detect PCL. Part of the issue is defining what patronizing and condescending language is, exactly–what one reader considers condescending might be deemed perfectly respectful by another.

SemEval-2022 Task 4 attempts to address some of these issues (Pérez-Almendros et al., 2022). Pérez et al. classify PCL into seven distinct categories: *unbalanced power relations*, *shallow solutions*, *presupposition*, *authority voice*, *metaphor*, *compassion*, and *the poorer, the merrier*.

Task 1a seeks to determine whether the sequence of text contains any form of patronizing or condescending language. Task 1b seeks to identify the PCL category that corresponds to the sequence of patronizing or condescending text. Overall, developing systems that perform well on these tasks—-that are capable of flagging condescending language–is a critical step toward reducing discrimination toward minority groups in media. We investigate various lightweight models to determine whether such models trainable on an extremely small compute budget could effectively identify PCL, as well as larger pre-trained transformer models to identify whether performance improves as models increase in size and complexity.

2 Dataset

For Task 1a, we train and validate our models on the SemEval-2022 Task 4 training set (Pérez-Almendros et al., 2020). Each paragraph has been annotated by two annotators on a Likert-type scale from 0 to 2 as shown in Table 1. The scores from each annotator are summed together: an overall score of 0 signifies that both annotators gave scores of 0, 1 that just one annotator gave a score of 1, and 2-4 that any higher score given by both annotators was summed together. A summary of the PCL status based on the two annotators' scores is shown in Table 2.

We did not investigate Task 1b.

Rating	Description
0	no presence of PCL
1	borderline PCL
2	contains PCL

Table 1: Likert scale for annotators to describe PCL status.

2.1 Train-test split

The dataset has a total of 14366 examples, split 10469–3897 between training and testing sets. The testing set was not provided until the last phase of the competition, so we created our own validation

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 374 - 378
July 14-15, 2022 ©2022 Association for Computational Linguistics

Sum	Description
0-1	Not a PCL paragraph
2-4	A PCL paragraph

Table 2: Summary of PCL status based on two annotators' scores.

set using a 75/25 train/validation split. For this reason, our train set has 7851 examples, and our validation set has 2618 examples. In our paper, all "validation-set" performance is reported on this internal held-out set.

3 Methods

3.1 Systems Overview for Task 1a

The aim of this task is to classify a given sequence of text as patronizing and condescending or not. We implement the following lightweight machine learning classifiers in **Scikit-learn** (Pedregosa et al., 2011):

- **Logistic Regression** is a supervised classification algorithm that employs a logistic function to categorize data into discrete classes. (LaValley, 2008)

- **Support Vector Machine** is a supervised classification algorithm that maps data points in a hyperplane to maximize the width of the gaps between two or more categories. (Gold and Sollich, 2003)

- **Random Forest** is an supervised learning technique that utilizes random bagging of different bootstrap samples of data to create a prediction from uncorrelated trees that is more accurate than any one tree. (Liaw et al., 2002)

- **Multi-layer Perceptron** is a feed-forward neural network with an input layer, an output layer, and any number of hidden layers. (Gardner and Dorling, 1998)

- **Gradient Boosting** is a greedy additive algorithm that sequentially ensembles an number of weak learners (typically decision trees) (Natekin and Knoll, 2013).

- **AdaBoost (Adaptive Boosting)** is a form of gradient boosting that adds weights to each subsequent weak learner (also typically decision trees) with incorrectly classified samples until either all data points have correctly classified or the maximum iteration has been reached. (Hatwell et al., 2020)

Ensemble We also experiment with a Voting-Classifier ensemble which incorporates one Logistic Regression, one Random Forest, and one Gaussian-hybrid models. Our models were averaged with equal weights.

We also experiment with the following pre-trained language models to try and effectively classify the presence of PCL in sentences:

- **BERT** is a pre-trained masked language model. We use BERT-cased, BERT-Large-cased, BERT-uncased, and BERT-Large-uncased in our experiments. (Devlin et al., 2018)

- **RoBERTa** is an optimized BERT model that utilizes the same architecture but various changes such as larger mini-batches and learning rates. We use RoBERTa and RoBERTa-Large. (Liu et al., 2019)

3.2 Experimental Setup

Normalization We investigate standardizing the dataset (implemented with the **Scikit-learn** *StandardScaler* preprocessing function) for the lightweight models.

Pre-Trained Models Regarding the large pre-trained models, we trained with binary cross-entropy loss for 5 epochs, using a learning rate of 1×10^{-5} and batch sizes of 8 (BERT and RoBERTA-base) and 3 (BERT and RoBERTa-Large).

4 Results

The official evaluation set performances for our classifiers are listed in Table 3, while the unofficial validation set performances for our Scikit-learn and Transformer-based models are listed in Table 4.

For Task 1a (patronizing and condescending language binary classification), we submitted our two highest-performing lightweight models (Support Vector Machine and Random Forest models). Due to error, we did not submit our BERT model. Overall, we ranked 53rd out of 78 on this task, achieving a positive-class F1 score of 0.4321.

Model	positive-class F1 (1a)	recall-score (1a)	precision-score (1a)
Support Vector Machine	**0.4321**	**0.5016**	**0.3795**
Random Forest	0.3310	0.3691	0.3000

Table 3: Official validation set performance of our lightweight models on Task 1a (binary classification).

Model	Features	positive-class F1 (1a)	Accuracy (1a)	Normalize
Logistic Regression	GloVe	0.37	0.76	False
Logistic Regression	GloVe	0.37	0.76	True
Support Vector Machine	GloVe	0.37	0.73	False
Support Vector Machine	**GloVe**	**0.48**	**0.89**	**True**
Random Forest	GloVe	0.38	0.87	False
Random Forest	GloVe	0.39	0.86	True
Multi-layer Perceptron	GloVe	0.40	0.90	False
Multi-layer Perceptron	GloVe	0.34	0.88	True
AdaBoost	GloVe	0.31	0.90	False
AdaBoost	GloVe	0.31	0.90	True
VotingClassifier Ensemble	GloVe	0.42	0.87	False
VotingClassifier Ensemble	GloVe	0.42	0.87	True
RoBERTa-base	—	0.54	0.92	False
RoBERTa-large	—	**0.55**	**0.92**	**False**
BERT-cased	—	0.55	0.91	False
BERT-uncased	—	0.51	0.91	False
BERT-large-cased	—	0.56	0.93	False
BERT-large-uncased	—	0.53	0.92	False

Table 4: Unofficial validation set performances of candidate models on Task 1a (binary classification). For this task, the highest-performing lighweight models are the Support Vector Machine model and the Random Forest model, and the highest-preforming pre-trained models are BERT-cased and RoBERTa-large.

5 Conclusion

We have proposed lightweight and pre-trained systems that are able to detect PCL in text.

We find that reasonably lightweight models such as Support Vector Machine and Random Forest are effective at predicting the level of patronizing and condescending language. However, we note that ensembling these models together does not improve performance.

Additionally, normalizing the dataset had little effect for most models—-and in the case of the Multi-layer Perceptron model actually returned a lower positive-class F1 score. However, it substantially increased the F1 score with the Support Vector Machine model from 0.37 to 0.48.

Finally, we find that fine-tuning large pre-trained models like BERT and RoBERTa achieves results at least as accurate as lightweight models–if not better.

An area of interest for future work may be further experimentation with ensembles of lightweight models, as well as inquiries into adversarial and contrastive learning to improve overall accuracy.

Overall, our results show that both lightweight and fine-tuned models can achieve reasonable results at detecting patronizing and condescending language in human channels of communication.

6 Acknowledgements

The authors would like to thank Ryan A. Chi for their support and guidance throughout this project. We would also like to acknowledge Google Collaboratory for their free services.

References

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*.

Matt W Gardner and SR Dorling. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

Carl Gold and Peter Sollich. 2003. Model selection for support vector machine classification. *Neurocomputing*, 55(1-2):221–249.

Julian Hatwell, Mohamed Medhat Gaber, and R Muhammad Atif Azad. 2020. Ada-whips: explaining adaboost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making*, 20(1):1–25.

Michael P LaValley. 2008. Logistic regression. *Circulation*, 117(18):2395–2399.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

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.

Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

Sik Hung Ng. 2007. Language-based discrimination: Blatant and subtle forms. *Journal of Language and Social Psychology*, 26(2):106–122.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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.

Model	**Hyperparameter**	**Task 1a**
Logistic Regression	solver penalty class weight	lbfgs none balanced
Support Vector Machine	class weight degree	0:1, 1:13 poly
Random Forest	max depth n estimators class weight min samples leaf	10 100 balanced 10
Multi-layer Perceptron	hidden layer sizes α β learning rate	(100, 100) 0.01 0.2 adaptive
AdaBoost	learning rate n estimators	1.0 50

Table 5: Hyperparameters for lightweight supervised models.

Sentence	**label**
" Anja Ringgren Loven I ca n't find a word to describe how I feel for you .... May God almighty 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 commenter on Facebook. "God bless you and your mission . Glad to see Hope (and all the children) growing up loved , well fed , happy , having fun , and going to school , " says another .	4
We 're living in times of absolute insanity, as I 'm pretty sure most people are aware . For a while , waking up every day to check the news seemed to carry with it the same feeling of panic and dread that action heroes probably face when they 're trying to decide whether to cut the blue or green wire on a ticking bomb – except the bomb 's instructions long ago burned in a fire and imminent catastrophe seems the likeliest outcome . It 's hard to stay that on-edge for that long , though , so it 's natural for people to become inured to this constant chaos , to slump into a malaise of hopelessness and pessimism .	0

Table 6: Examples that are considered patronizing and condescending and those not considered patronizing and condescending, respectively