Team LEGO at SemEval-2022 Task 4: Machine Learning Methods for PCL Detection

Abhishek Kumar Singh
GLA University, India
sabhishek.kumar166@gmail.com

Abstract

In this paper, we present our submission to the SemEval 2022 - Task 4 on Patronizing and Condescending Language (PCL) detection. We approach this problem as a traditional text classification problem with machine learning (ML) methods. We experiment and investigate the use of various ML algorithms for detecting PCL in news articles. Our best methodology achieves an F1- Score of 0.39 for subtask1 with a rank of 63 out of 80, and F1-score of 0.082 for subtask2 with a rank of 41 out of 48 on the blind dataset provided in the shared task.

1 Introduction

The explosion of social media in recent years also enables increasing the number of patronizing and condescending language (PCL). Patronizing and condescending language depicts apparently kind or helpful behavior but betraying a feeling of superiority on others. Previously, harmful behavior in language for example, hate speech (Fortuna and Nunes, 2018), offensive language (Razavi et al., 2010), fake news (Oshikawa et al., 2018), rumor propagation or misinformation (Zhou and Zafarani, 2020), and many others has been widely studied in NLP, PCL has been a neglected area of study until very recently.

Identifying PCL is hard even for humans because it is subjective and subtle. For instance, one might find condescending something which another person might consider an objective portrayal of a situation or some people might not see the harm in describing how those in a privileged position donate their remainings to those who need them. Also, we would expect a member of a so-called vulnerable community to feel more patronised than one person who does not belong to such group while reading how others refer to them.

The goal of SemEval 2022-Task 4 is to design a system to detect whether or not the text contains any form of PCL and furthermore, identify which PCL category expresses the condescension. The organizers provided two datasets, one annotated based on the intensity of PCL and other with the PCL categories. We approach this problem using various machine learning approaches using the linguistics features.

Class	Nb of Samples
PCL	9476
Non-PCL	993

Table 1: Dataset Statistics Task 1

The structure for the rest of the paper is as follows. Section 2 describes a background about the dataset. Section 3 describes the experimental setup of our experiments. It involves pre-processing, feature engineering, implementation details for all the respective ML models. Section 4 describes the results and discussion for both the subtask. And lastly, in Section 5, we concluded the paper and suggest ideas for future research.

2 Dataset

The dataset (Perez-Almendros et al., 2020) provided for this challenge was collected from the News on Web (NoW) corpus (Davies, 2013). For task1, we are provided with 10469 text paragraphs. Each paragraph instance in the dataset is provided with paragraph-level label, vulnerable community-info which includes (disabled, homeless, hopeless, immigrant, in-need, migrant, poor families, refugee, vulnerable and women), and along country of origin. There are 5 classes (0-4) based on the intensity of PCL. For task2, we are provided with 993 paragraphs. Each paragraph instance in the dataset is provided with keyword, country of origin, span-text, category label, and number of agreeing annotators. The labels comprise of 7 classes: *Unbalanced Power relations*(unb) , *Shallow solution*(shall), *Presupposition*(Pres), *Authority Voice*(Auth), *Metaphor*(Meta), *Compas-*

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 369 - 373

Class	Nb of Samples
Unbalanced power (unb)	716
Shallow solutions (shall)	196
Presupposition (Pres)	224
Authority Voice (Auth)	230
Metaphor (Meta)	197
Compassion (Comp)	469
The poorer,the merrier (Poorer)	40

Table 2: Dataset Statistics Task 2

Model	Val-F1	Test-F1
SVM	**0.91**	0.053
Logistic	0.86	0.390
SGD	0.90	0.058
MLP	0.89	0.300
AdaBoost	0.89	0.280
Ensemble	-	0.340
Roberta-baseline	-	**0.491**

Table 3: F1- score for ML Models for task1

sion(Comp), *The poorer,the merrier*(poorer). For our experiments, we perform 80-20 data split with random state 0 for both the tasks to train the models for all experimental setup.

3 Experiment Setup

3.1 Pre-processing

Task1 is a binary text classification. The dataset is annotated from 0 to 4 on the basis of PCL intensity in the text. We further re-label the dataset instances using the intensity score where 0,1 referred to Non-PCL text and (2-4) referred to PCL text. (Ref . Table 1)

Task2 is multi-label classification. Each dataset instance is annotated with different PCL category labels and the text span reflecting the PCL label is provided respectively. Many paragraph instances were annotated for more than one category of PCL over a different span of text. (Ref . Table 2)

For our experiments, we remove stopwords by using NLTK(Natural Language Toolkit) library and other non-ascii symbols from the text before performing feature engineering.

3.2 Feature Engineering

Count Vectorizer Feature extraction (Vectorization) on text, encodes the text as integers or floating point values for using as input in machine Learning algorithms. Scikit-learn's CountVectorizer is used to convert a collection of text documents to a vector of term/token counts.

Term Frequency- Inverse Document Frequency (TF-IDF) We use Sklearn TF-IDF, which is an approach to quantify words in a set of documents by computing a score for each word to signify its importance in the document or corpus.

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

TF is the ratio the frequency of a word in a document and the frequency with the total number of words in the document whereas, document frequency (DF) is the normalized count of documents in which the term is present. Inverse Document Frequency is the inverse of the document frequency which measures the informativeness of a term in the document. We used the features generated on the entire corpus and the feature length was 20244.

3.3 Models

Support Vector Machine (Burges, 1998) is an effective technique for classifying high dimensional data. It learns the optimal hyperplane that separates training examples from different classes by maximizing the classification margin. Each row vector of the word-document matrix represents the vectorization of text that are mapped to a latent semantic space in this module by LSA vector space model, to generate representation vectors and further classify them. We perform Principal Component Analysis (PCA) (`number of components`=500) to perform dimension reduction over text features before inputting to this model.

Logistic Regression (Cramer, 2002) is a classification algorithm used to solve binary and multi-label classification. The logistic regression classifier uses the weighted combination of the input features and passes them through a sigmoid function.

Stochastic Gradient Descent (Ruder, 2016) is an iterative algorithm that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function. We perform feature scaling before inputting the features to the model.

AdaBoost (Freund and Schapire, 1997) Adaptive Boosting is very popular boosting technique that combines multiple local weak classifiers into

Model	Unb	shall	Pres	Auth	Meta	Comp	poorer	Average F1
SVM	0.87	0.2	0.27	0.16	0.05	0.75	0	0.32
Logistic	0.84	0.56	0.53	0.37	0.27	0.76	0	**0.47**
SGD	0.84	0.25	0.35	0.08	0.05	0.68	0	0.32
MLP	0.84	0.42	0.47	0.27	0.26	0.71	0	0.42
AdaBoost	0.82	0.3	0.36	0.25	0.28	0.6	0	0.37

Table 4: F1-Score for ML Models for task2 on validation set

Model	Unb	shall	Pres	Auth	Meta	Comp	poorer	Average F1
SVM	0.11	0.18	0.02	0.094	0.10	0.10	0	0.089
Logistic	0.13	0.14	0.06	0.08	0.04	0.099	0	0.082
SGD	0.11	0.14	0.029	0.046	0.023	0.068	0	0.060
MLP	0.13	0.12	0.029	0.045	0.028	0.092	0	0.063
AdaBoost	0.12	0.089	0.027	0.039	0.054	0.101	0.045	0.068
Ensemble	0.1180	0.2050	0.0192	0.0643	0.0645	0.1018	0	0.082
Roberta-baseline	0.3535	0	0.1667	0	0	0.2087	0	**0.104**

Table 5: F1-Score for ML Models for task2 on test set

a single strong classifier. It can be used to significantly reduce the error of any learning algorithm that consistently generates classifiers whose performance is a little better than random guessing. We initiate the model with `number of estimators` = 400, `learning rate` = 1, and `base classifier` = DecisionTreeClassifier with `criterion` ='entropy'.

Multilayer Perceptron (MLP) (Rosenblatt, 1961) is a classical type of neural network. They are composed of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer, also called the visible layer. MLPs are suitable for classification prediction problems because they are known to be capable of modelling complex functions. We used number of `hidden layers`= 30(for task1),and 1000 (for task2), and `maximum iteration` = 2000 to initiate the MLP model. Activation functioned used for the hidden layer is by default i.e ReLU (Nair and Hinton, 2010).

3.4 Ensemble

For our submitted system, after predictions were extracted from different models, we calculate the ensemble (Kuncheva and Whitaker, 2003) using the mode to find the most frequently occurring label. In the presence of a tie-breaker scenario, we select the label predicted by the best performing model.

3.5 Implementation

For all the models, we used Scikit learn library for our (Pedregosa et al., 2011) implementation. For task1, our validation set has 2094 examples and for task2, we have 199 examples from the training data according to the initial data split. All the models were initiated with `class_weight` ='balanced' setting , `maximum iteration`=1000. For the rest of the hyper-parameters we use the default setting otherwise specified in their sections. The github repository containing all the details of our experiments is made publicly available[1].

4 Results & Discussion

We evaluate and report the performance of different models on our validation set and the blind-test set (Table 3, 4, 5). We can see that logistic regression model performs the best among all the models with TF-IDF features as the input for task1 and SVM works better than of LR classfier for task2.

We see a steep drop in the performance of model on the test-set. This can be attributed to the difference distribution in the training and testing data which reduces the effectiveness of the TF-IDF features. The imbalance in the dataset also can be a reason for the model to perform badly on the 'Non-PCL' class in task1 (1) and 'poorer' class in task2 (1) . The machine learning models are not effective compared to the roberta-baseline as we see that

[1]https://github.com/Abhi-020/PCL

Gold	Paragraph	Pred
non-pcl	On the other hand,in Europe and North America,educated and young Muslims are surprisingly found to be **vulnerable** to such extremism	*pcl*
pcl	Many celebrities wore blue ribbons to support the American Civil Liberties Union, which is seeking to shed light on the plight of young **immigrants** facing the potential of being deported .	*non-pcl*

Table 6: Examples of incorrect predictions made by Logistic Regression Classifier in Task 1.

Gold	Paragraph	Pred
unb	Fast food employee who fed **disabled** man becomes internet sensation	*non-unb*
non-unb	When I was born , this was a nightmare town for **disabled** children , he said to me then	*unb*
shall	After a big casino win , Mario Balotelli gave a **homeless** man? 1,000 (PA)	*non-shall*
non-shall	I rather donate to the less privilege in the church or **homeless** than to pour a cup of water into Nigeria sea of wealth so that the thieves can grab my little contribution .	*shall*
Pres	Once again the stateless Rohingya are on the run – **homeless** and increasingly hopeless .	*non-Pres*
non-Press	Antidote for **hopelessness** Pulitzer Prize-winning journalist Roy Gutman , author of How We Missed the Story , argued that journalism in conflict zones provides change-makers and hope as an antidote for **hopelessness** .	*Press*
Auth	Every family which qualifies for the program should be covered . Every child in **poor families** must be placed and kept in school , and they should enjoy health and nutrition assistance , Romualdez said	*non-Auth*
non-Auth	The government is implementing several schemes would change the economic position of **poor families** , " she added	*Auth*
Meta	It is the supreme task of this generation to give **hope** to the **hopeless** strength to the weak and protection to the defenceless	*non-Meta*
non-Meta	They discounted and denied every conceivable poll which, showed Jonathan losing the election ,preaching that Nigerians wanted continuity ,not the change the opposition advocated . he people of Nigeria were portrayed as somehow loving their poverty and insecurity , their darkness and weakness , **hopelessness** and joblessness.	*Meta*
Comp	Today , **homeless** women are still searching for the same thing . A place to sleep and be safe .	*non-Comp*
non-Comp	Housing Minister Grant Shapps added :'The plight of **homeless** people should be on our minds all year round - not just at Christmas . families and be symbols of hope and possibility , of never giving up .	*Comp*
Poorer	A lot of my disabled patients over the years have gained strength and hope from me when they see that I also have a **disability** , but that I 'm coping . Sometimes the biggest gift I can give other people with **disabilities** is to show them that you can get a job .	*non-Poorer*
non-Poorer	One of her proudest achievements as an MP is challenging how the **disabled** are treated She became the first **disability** issues spokesperson and later minister .	*Poorer*

Table 7: Examples of incorrect predictions made by Logistic Regression Classifier in Task 2.

static text features are less helpful in the detection of PCL.

The table 6, 7 contains examples from the task1 and task2 validation set respectively, where the model failed to label the paragraph instances correctly. We can see that, the presence of vulnerable community keywords (Highlighted Table 6,7) often confuses the model leading it to mislabel the instances. We observe that TF-IDF features are not able to capture contextual information as they rely only on the presence of the word indicators . We believe that this is the reason behind the inefficiency of the ML models.

5 Conclusion

This paper presents our study of machine learning models for the binary and multi-label text classification on the PCL detection shared task. We find that tf-idf features can be effective in cases where train and testing data are from the same distribution but it may fail otherwise. For future work, we plan to experiment with contextual embeddings from BERT, and other transformer-based models. We also would like to look into bootstrapping and data augmentation techniques to solve the class imbalance problem more effectively.

References

Christopher JC Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Jan Salomon Cramer. 2002. The origins of logistic regression.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Ludmila I Kuncheva and Christopher J Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Frank Rosenblatt. 1961. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.