CS/NLP at SemEval-2022 Task 4: Effective Data Augmentation Methods for Patronizing Language Detection and Multi-label Classification with RoBERTa and GPT3

Daniel Saeedi
University of Tehran
saeedi.danial@ut.ac.ir

Sirwe Saeedi
Western Michigan University
sirwe.saeedi@wmich.edu

Aliakbar Panahi
C3 AI
ali.panahi@c3.ai

Alvis C.M. Fong
Western Michigan University
alvis.fong@wmich.edu

Abstract

This paper presents a combination of data augmentation methods to boost the performance of state-of-the-art transformer-based language models for Patronizing and Condescending Language (PCL) detection and multi-label PCL classification tasks. These tasks are inherently different from sentiment analysis because positive/negative hidden attitudes in the context will not necessarily be considered positive/negative for PCL tasks. The oblation study observes that the imbalance degree of PCL dataset is in the extreme range. This paper presents a modified version of sentence paraphrasing deep learning model (PEGASUS) to tackle the limitation of maximum sequence length. The proposed algorithm has no specific maximum input length to paraphrase sequences. Our augmented underrepresented class of annotated data achieved competitive results among top-16 SemEval-2022 participants. This paper's approaches rely on fine-tuning pretrained *RoBERTa* and *GPT3* models such as *Davinci* and *Curie* engines with extra-enriched PCL dataset. Furthermore, we discuss Few-Shot learning technique to overcome the limitation of low-resource NLP problems.[1]

Keywords: Natural Language Processing, Transformers, Data Augmentation, RoBERTa, GPT-3, Curie and Davinci Engines.

1 Introduction

Natural Language Understanding (NLU) and Interpretation (NLI) is a branch of Natural Language Processing (NLP) in Artificial Intelligence (AI), which involves understanding and analyzing human language in-depth. Recent advances in Deep Neural Networks (DNNs) have enabled NLP research scientists to achieve state-of-the-art results for tasks that were extremely difficult, if not impossible Devlin et al. (2019), Lan et al. (2020). However, understanding human emotions, reactions, and uncovering hidden insights from unstructured text data such as news stories channel is still challenging.

Language attitudes and intentions extracting in response to the support for the marginalized and vulnerable communities is one of the emergent NLP applications. Patronizing and condescending language (PCL) is a type of behavior that projects a sense of superiority to vulnerable populations Pérez-Almendros et al. (2020). Furthermore, biases and discrimination can result from patronizing attitudes, causing some people to feel unfairly treated, inadequate, unintelligent, and possibly infuriated Saeedi et al. (2021).

Since raw text data extracting from web is a common data collection method, language models can learn different forms of harmful language Heidari and Jones (2020). The PCL understanding is inherently different from sentiment analysis because positive/negative hidden attitudes in the context will not necessarily be considered positive/negative for PCL tasks. It is difficult due to the fair amount of world knowledge and commonsense reasoning required to understand this kind of language Saeedi et al. (2020). The fine-grained idea of PCL detection towards vulnerable communities was presented by Pérez-Almendros et al. (2022). They evaluated baseline results of NLP techniques to detect the presence of PCL and classify PCL types at the text span level.

In this paper, we describe systems participating in the SemEval-2022, PCL detection competition, multiple tasks of language interpretation. The competition is divided into binary classification and multi-label categorization tasks. Data quality analysis led us to explore several NLP data augmentation techniques and state-of-the-art DNN architectures for these challenging tasks. Our attempts to improve the performance of previous

[1]Our implementation is publicly available at https://github.com/daniel-saeedi/PCL_Detection_SemEval2022

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 503 - 508

Tasks	Keyword	Paragraph	labels	Combined Features
PCL Binary Classification	Homeless	Housing Minister Grant Shapps added : ' The plight of homeless people should be on our minds all year round - not just at Christmas.	1	Homeless</s>Housing Minister Grant Shapps added : ' The plight of homeless people should be on our minds all year round - not just at Christmas .
	Refugee	UNHCR gave a report on the state of refugees worldwide on Wednesday as World Refugee Day was marked.	0	Refugee</s>UNHCR gave a report on the state of refugees worldwide on Wednesday as World Refugee Day was marked.
PCL multi-label Classification	**Keyword**	**Paragraph**	**PCL Category**	**Combined Features**
	Homeless	Housing Minister Grant Shapps added : ' The plight of homeless people should be on our minds all year round - not just at Christmas.	Authority voice	Homeless</s>Housing Minister Grant Shapps added : ' The plight of homeless people should be on our minds all year round - not just at Christmas .

Figure 1: PCL data for binary and multi-label classification problems. Labels 0 and 1 are corresponding to not containing and containing PCL, respectively. "Authority voice" is the PCL category of paragraph. Training model on combined features as the concatenation of keyword and paragraph with RoBERTa separation token "</s>".

efforts ranked us 16 among 79 NLP research teams with very competitive results on the PCL detection task. Our system's performance achieved 80% and 58% F1-score on the training and test datasets, respectively. In comparison, the winning system's F1-score was 65%. Also, the in-depth dataset analysis revealed multi-label classification techniques commonly confused in the PCL categorization task.

This paper is organized as follows. In Section 2, we introduce two PCL tasks, an in-depth analysis of their datasets, and the challenges of these tasks. In Section 3, we describe our different strategies to tackle discovered challenges of data quality. Next, we explored text augmentation methods to fine-tune the Transformer-based model for each individual task. In Section 4, we discuss our applied models, the experimental setup for fine-tuning models, and their performance. Finally, we conclude the paper in Section 5.

2 Tasks Definition and Dataset Analysis

As discussed, PCL competition consists of two classification tasks, each focused on the different objectives of PCL towards underprivileged communities. Figure 1 shows samples of data, their salient features, and annotated labels in training set for both tasks. The first task aims to classify a paragraph that contains PCL as an act of appearing kind or helpful but internally feeling superior to others. The second task is the investigation of the text categorization problem, where each PCL-containing paragraph may belong to several PCL categories[2].

2.1 Data Analysis of Binary Classification

For the PCL binary classification task, we had access to 10469 human-labeled paragraphs for training our models. Two annotators consider their disagreement on borderline cases as not containing PCL. Our exploratory data analysis reveals not containing PCL paragraphs with label '0' make up a large proportion of dataset (90.4%), and target class '1' as containing PCL is the minority class.

The imbalance degree of PCL binary classification dataset can be measured in moderate to extreme range Leevy et al. (2018). The highly imbalanced data would be problematic because models are mostly trained on non-PCL data and will not learn enough from the PCL samples. In this case, a non-PCL outcome is almost always predicted by the trained model. Our experiment shows models yield inaccurate results, see Section 4.

To combat imbalanced training data and misleading classification results, we investigate several techniques in Section 3. Furthermore, this problem is highly challenging because the nature of PCL detection is different from other domains, such as hate speech, inappropriate and fake content detection.

[2]Seven categories for different traits of PCL: Unbalanced power relations (unb), Shallow solution (shal), Presupposition (pre), Authority voice (aut), Metaphor (met), Compassion (com), The poorer, The merrier (merr).

Words that might have positive connotations in sentiment analysis will not necessarily be considered positive in PCL.

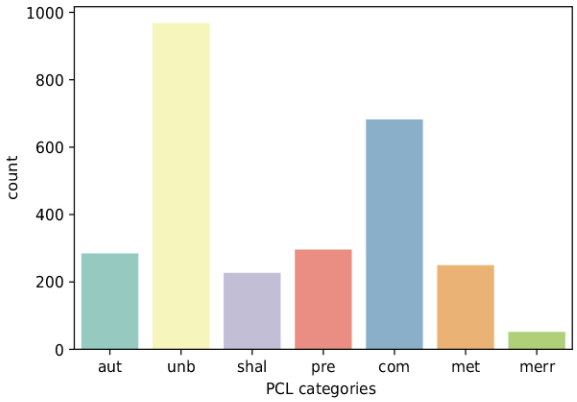

Figure 2: Imbalanced data representation. This chart illustrates the number of observations per PCL category is not equally distributed because in the first task not containing PCL class can obviously discriminate the minority class.

2.2 Data Analysis of Multi-label Classification

For identifying PCL types, the number of manually labeled samples in the datasets is 2760, including all PCL positive data from the previous task. Each text span within the containing PCL paragraph can represent one or more PCL categories.

We were challenged to build a multi-label deep learning model capable of detecting different types of PCL. The unevenly distributed labels, also in the case of multi-label classification, could be problematic. Figure 2 illustrates the number of paragraphs associated with "unb" and "com" are the dominant categories. In Sections 3, we present different methods to combat these challenges, and we describe our efforts of training model on the proper distribution to handle imbalanced dataset in Section 4.

3 Tackling Data Imbalance

Taken together, these challenges led us to approach skewed class proportion problems in the PCL dataset with various Data Augmentation (DA) techniques in NLP Wei and Zou (2019). Like many other NLP techniques, DA is not an exact science, and understanding both dataset and task is essential. We conducted an ablation study to measure the impact of DA on the performance of the system.

We aimed to enhance the size of dataset to reduce the side effect of data imbalance. Before trying text augmentation methods, we preprocessed the data by removing HTML-tags and non-alphabetic characters. Then, we expanded English language contractions, e.g., from "you've" to "you have." The following subsections explain our DA methods.

3.1 Synonym Replacement

Synonym Replacement (SR) is a simple operation that randomly chooses some non-stop words from the sentence and replaces them with one of their synonyms chosen at random. We applied *wordnet* database from *nltk* library to identify synonyms of a given word within the paragraph Miller (1995). As SR is a lightweight and efficient way of performing DA, we tried to replace 1 to 3 words at a time to create diverse PCL samples. Table 1 illustrates scores achieved by training $RoBERTa_{Large}$ model on the augmented dataset. Regardless of the approach taken, the model performance did not spike as expected. As shown later, this approach has been mixed with other text augmentation methods in training models.

3.2 Oversampling

Since containing PCL samples are underrepresented, we considered oversampling (OS) Padurariu and Breaban (2019). Oversampling randomly duplicates data in the minority class by a factor of 8 and adds them to the PCL training dataset, so the number of samples in each class becomes almost equal. The performance of the training pre-trained model with augmented training data by far exceeded the baseline result. (See Table 1)

3.3 Back Translation

We applied back translation (BT) to treat the problem of underrepresented class and boost model performance. In this case, we used a powerful augmenter method of BT in *nlpaug* library and *FairSeqMachineTranslation* Wang et al. (2020) model from *HuggingFace* [3] Transformers. The aim was to generate more PCL samples and then train model on the true distribution. BT translates all PCL samples from English to German, then translates the previously translated text back into the source language. We reused our best-performing model on OS and SR methods. However, later experiments showed that this technique led the model heavily to overfit the augmented training data (Figure 3).

[3] https://huggingface.co/docs/transformers/model_doc/fsmt

Note that we did no model validation using augmented data but did training with a mixture of OS, SR, and BT approaches. Although the improvement offered by BT is not so intelligible, statistical analysis is remarkable. The results are shown in Table 1.

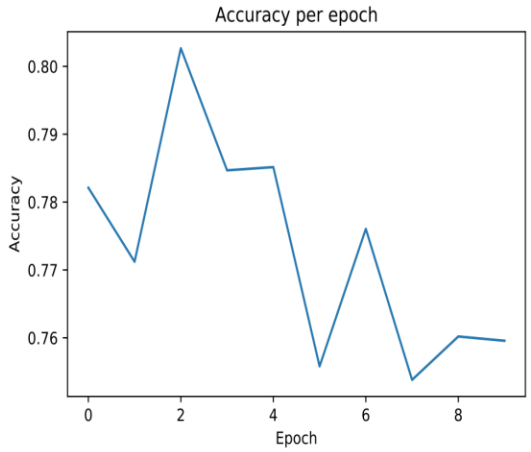

Figure 3: 80% F1-score was achieved at epoch 3. We can see a clear sign of overfitting after this epoch.

3.4 PEGASUS Paraphrasing

Paraphrase generation was the last effort in DA. Paraphrase generation models (in an encoder-decoder form) learn to reconstruct the input using different words and retaining the same meaning while paraphrasing. Paraphrasing can act as a regularizer and reduce the overfitting during the training process Fu et al. (2020).

To leverage PCL dataset efficiently, we performed paragraph paraphrasing along with SR to come up with a less imbalanced dataset. PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) Zhang et al. (2020) is a self-supervised Transformer model that masks important sentences from the input and then generates them as one output sequence from the remaining sentences.

The original PEGASUS is limited by the length of text and does truncation on long texts input. The maximum length of PCL paragraphs is 5493 tokens, while the longest input of original PEGASUS model can be 60 tokens. Therefore, we need to handle the limitation of Transformers on the size of the text while training Liu et al. (2019). We proposed an algorithm, multi-sentence PEGASUS, to modify PEGASUS model for arbitrarily long document paraphrasing. This algorithm separates each paragraph into sentences, and then multi-sentence PEGASUS generates ten paraphrased sentences from each individual sentence. The main challenge is to retrieve the original paragraph, because the number of paraphrased sentences for each paragraph was different due to different number of sentences in each sample data. This algorithm can concatenate paraphrased sentences to get the original paragraph in efficient time (The implementation is available at our GitHub). Multi-sentence PEGASUS generates a new dataset containing PCL paragraph over ten times larger than the original containing PCL training data. The following example is a containing PCL data and its corresponding paraphrased context:

Original Paragraph: Shepherding in America has always been an immigrant's job, too dirty, too cold and too lonely for anyone with options.

PEGASUS Paraphrased: In America, shepherding has always been an immigrant's job, dirty, cold, and lonely.

After multi-sentence PEGASUS paraphrasing, two words in each generated text are replaced by their respective synonyms from *wordnet* corpus. The hyper-parameters values for PEGASUS model have been selected by trial and error. We set a number of times the model searches for the most optimal follow-up word within the text to 10 and played with the parameter that regulates the chances of appearance of high/low probability words.

4 Model Description

Our system is based on pre-trained transformers models on the augmented PCL dataset. We focused on exploiting superior performance of *RoBERTa* and *GPT3* models.

4.1 Fine-tuning RoBERTa

To simulate the baseline result, we first did regular fine-tuning *RoBERTa* for each PCL task on the concatenated features of dataset (keyword and paragraph). Submitted systems on the SemEval-2022 leaderboard were evaluated on the F1-score metric. The (73%) F1-score was achieved by training the model with parameter values of $1e-5, 2, 400$ for learning rate, number of epochs, and warm-up steps, respectively while the baseline is 70.63% (See Table 1).

For the next step, we fine-tuned *RoBERTa* on the augmented datasets via each method mentioned

Models	Original Dataset	SR	OS	BT	Peg	SR/Peg	SR/OS/BT/Peg
Task1 $RoBERTa_{Large}$	73%	73%	79%	73%	76%	77%	81%
Task2 $RoBERTa_{Large}$	32%						57%

Hyper-parameters	LR	WS	Epoch
	{1e-5, 5e-6}	{400, 800, 3000, 4000}	{1, 2, 4, 8, 10}

Models	labels	precision	recall	f1-score	support	accuracy
Task1 $GPT3/Curie$	0	61%	22%	32%	50	
	1	52%	86%	65%	50	54%
Task1 $GPT3/Davinci$	0	58%	30%	39%	50	
	1	53%	78%	63%	50	54%

Table 1: Peg stands for PEGASUS paraphrasing. The training *RoBERTa* on the extra enriched dataset (SR/OS/BT/Peg) outperforms other DA methods. The learning process is controlled by setting hyper-parameters (Learning Rate (LR), Warm-up Steps (WS), and Number of Epochs) in the defined range. *GPT3* model with *Davinci* and *Curie* engines yield good performance with small subset of PCL training dataset. Support parameter indicates the number of queries which is the same for both models. 100 queries in total, and 50 queries for each label.

in Section 3, separately. Moreover, we took pre-trained *RoBERTa* and retrained on the extra enriched PCL dataset, which was boosted by a combination of three DA previously explained methods. Same as regular fine-tuning *RoBERTa*, we fed concatenation of keyword and paragraph with $RoBERTa$ special token "</s>" to the model and hyper-parameters are defined in Table 1. Augmented PCL dataset with SR and BT methods led to lower performance of our system compared to a mix of all described DA approaches. Using all DA methods together boosted the model performance to 81%. Figure 3 shows the accuracy of the model in each epoch, and it hits 81% F1-score at epoch 3, and then model start overfitting later. Our system trained and evaluated on the training dataset.

For multi-label classification, the trained *RoBERTa* model on the extra enriched dataset outperformed (57%) the same model trained on the original dataset (32%). However, the model's performance on the test dataset released in the post-evaluation phase was not the same. It is worth mentioning that the F1-average of the winning system (46%) for multi-lable classification task was not better than the random guess model.

4.2 GPT-3 Davinci and Curie

Limitation in the amount of available labeled data can be rectified with Few-Shot Learning technique by providing a few examples at inference time with a large language model. OpenAI *GPT3* Brown et al. (2020) language model uses this technique and also can be applied to PCL binary classification task. *GPT3* has been trained on a huge text dataset from the open internet with billions of parameters.

In this scheme, we considered two offered models of *GPT3* with different capabilities and price points. *Davinci* is the most capable in understanding the intent of a text, the motives of characters, and also the expensive engine. Also, *Curie* is quite faster and lower cost than *Davinci* and capable of tasks like sentiment classification.

We tried both models with Few-Shot learning technique by feeding the model a small amount of PCL training data (with an equal number of labels) as a prompt. The labeled examples were uploaded as a JSON file to OpenAI API for the purpose of classification. *Davinci* and *Curie* leverage a few labeled sets of examples without fine-tuning and enable to understand previously unseen data. We queried the model with a subset of training data to predict the most likely label for each query. In fact, *Davinci* and *Curie* engine classify specified queries using provided labeled data in a JSON file. These engines first search over the labeled data to select the most relevant for a particular query. Our implemented code is publicly available [4].

Table 1 illustrates the performance of *Davinci* and *Curie* models. OpenAI *GPT3* prices are per tokens. Therefore, we just prompted *Davinci* and *Curie* by 1000 and 200 labeled data, respectively. They were evaluated on F1-score with 100 queries of even class distribution. Surprisingly, both models perform well without hyper-parameter tuning and on just a few examples of PCL. *Davinci*'s performance was the same as *Curie*'s result but with

[4] https://github.com/daniel-saeedi/PCL_Detection_SemEval2022

five times fewer labeled examples. OpenAI API offers the ability to fine-tune their model on the desired task, which is quite costly and time-intensive. An interesting future research direction can be exploring GPT3 applications for PCL detection and multi-label classification tasks, regardless of the cost to train the model.

5 Conclusion

This paper presented a system description for PCL detection and multi-label categorization tasks. Our exploratory data analysis revealed annotated PCL dataset is highly imbalanced. We enhanced data quality with a combination of data augmentation methods. We presented a modified version of sentence paraphrasing deep learning model, Multi-sentence PEGASUS, to tackle the limitation of maximum sequence length. The proposed algorithm has no specific maximum input length to paraphrase sequences. We evaluated the performance of the large pre-trained RoBERTa model on the extra enriched PCL dataset. We boosted the baseline performance and achieved competitive results among the top-16 SemEval-2022 participants. Furthermore, we tried two models of GPT3, *Davinci* and *Curie* with Few-Shot learning technique. Our investigation showed both models perform well without hyper-parameter tuning and on just a few examples of PCL. We believe these tasks have many potentials and challenges to further improve current results.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yao Fu, Yansong Feng, and John P. Cunningham. 2020. Paraphrase generation with latent bag of words.

Maryam Heidari and James H Jones. 2020. Using bert to extract topic-independent sentiment features for social media bot detection. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0542–0547.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30.

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.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Cristian Padurariu and Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Cheionference KES2019.

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.

Sirwe Saeedi, Saraju P. Mohanty, Steve Carr, Alvis C. M. Fong, and Ajay K. Gupta. 2021. Consumer artificial intelligence mishaps and mitigation strategies. *IEEE Consumer Electronics Magazine*, pages 1–1.

Sirwe Saeedi, Aliakbar Panahi, Seyran Saeedi, and Alvis Cheuk M. Fong. 2020. CS-NLP team at semeval-2020 task 4: Evaluation of state-of-the-artnlp deep learning architectures on commonsense reasoning task. *CoRR*, abs/2006.01205.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.