MS@IW at SemEval-2022 Task 4: Patronising and Condescending Language Detection with Synthetically Generated Data

Selina Meyer, Maximilian Schmidhuber and **Udo Kruschwitz**
Chair for Information Science
University of Regensburg, Germany
(selina.meyer, udo.kruschwitz)@ur.de
maximilian.schmidhuber@stud.uni-regensburg.de

Abstract

In this description paper we outline the system architecture submitted to Task 4, Subtask 1 at SemEval-2022. We leverage the generative power of state-of-the-art generative pretrained transformer models to increase training set size and remedy class imbalance issues. Our best submitted system is trained on a synthetically enhanced dataset with 10.3 times as many positive samples as the original dataset and reaches an F1 score of 50.62%, which is 10 percentage points higher than our initial system trained on an undersampled version of the original dataset. We explore possible reasons for the comparably low score in the overall task ranking and report on experiments conducted during the post-evaluation phase.

1 Introduction

Task 4 of SemEval-2022 focuses on the detection of patronising and condescending language (PCL) in news (Pérez-Almendros et al., 2022). PCL in popular media and news sources is detrimental to an emancipated and equal society, as it is usually targeted towards minorities and socially disadvantaged communities, often in an unsuccessful attempt to show solidarity (Perez Almendros et al., 2020). PCL has the potential to strengthen existing stereotypes by representing minorities either as passive entities to be pitied and supported, thus taking away their agency and focusing on their vulnerabilities or praising members of vulnerable groups for everyday achievements simply because of their background (Nolan and Mikami, 2013). In contrast to hate speech, PCL is usually subtle, well intentioned, and free of discriminatory phrases or racial slurs, which makes it an interesting Natural Language Processing (NLP) problem.

In other domains with more discriminatory classes such as hate speech detection, generative models have recently become increasingly popular and successful as a tool to increase classification performance (Wullach et al., 2021; Anaby-Tavor et al., 2020). In our contribution to the shared task, we explored to what extent this approach is feasible for the presented use case, where classification of a text sample is less distinct and often relies on world knowledge (Perez Almendros et al., 2020). The dataset provided for the task was fairly small, with less than 10% of the data belonging to the positive class. We thus enhanced the original dataset in two ways for our system runs:

- balancing the dataset by generating only PCL samples
- increasing overall dataset size, by generating an equal amount of PCL and non-patronizing (nPCL) samples

We generally followed the approach used by Wullach et al. (2021) and initially fine-tuned a BERT classifier on the original dataset. We then fine-tuned GPT-3 (Brown et al., 2020) and generated samples of PCL and nPCL which were classified using our fine-tuned system. Samples for which the BERT classification did not correspond to the intended output were discarded. We then fine-tuned a new BERT instance with the modified dataset PCL*enhanced* including the synthetic data. Although our system only ranked middle field in the competition, both classifiers trained on the modified datasets improve our initial classifier trained on the original dataset by multiple percentage points. We conclude that this approach does add value to classification, even in cases where the distinction between the positive and the negative class relies on subtleties. The code described in the following as well as the synthetic data used for the modification of the original dataset is available on Github[1].

[1] https://github.com/khaliso/MS-IW-at-SemEval-2022-Task-4

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 363 - 368

Text	Class
Meanwhile "throughout this island, the high level of suicide is terrible and terrifying. "As Christians" we can give hope, where a person feels only darkness and hopelessnes," he said	PCL
As the house prices go up, so do rents , and the pohara poor families ca n't afford to live. Those who own houses, and are only just making it through, will be rated out of their homes	nPCL

Table 1: Examples of PCL and nPCL in the DPM.

2 Background

We participated in Subtask 1 of the competition, which entailed the binary classification of news paragraphs as either patronizing or not patronizing. Basis for the task was the *Don't Patronize Me! dataset* (DPM) (Perez Almendros et al., 2020), which contains 10,469 paragraphs of annotated data from 20 English news sources. While all paragraphs include references to potentially vulnerable groups, only 993 are examples of patronising speech. The dataset included meta-information about the country each paragraph was published in, an article id, a keyword indicating which vulnerable group is addressed, and a label ranging from 1 to 4, where 0 and 1 are treated as non-patronizing and 2 to 4 as patronizing. The task organizers define PCL as often unconscious, subtle and subjective ways in which the speaker conveys a superiority "concealed behind a friendly or compassionate approach towards the situation of vulnerable communities" (Perez Almendros et al., 2020). They explicitly exclude hate speech and discriminatory speech from PCL, making it harder to be identified not only by NLP-systems, but also by humans. We include examples of both classes in Table 1.

Transformer-based generative models such as GPT (Radford et al., 2018) and its successors have become prevalent in various NLP tasks. For instance, Liu et al. (2021a) explored the idea of synthetically constructing benchmark datasets to concur with existing benchmarks such as SQuAD, while Zhang et al. (2018) showed that a fine-tuned GPT model can accurately mimic the personal conversation style of an individual, leading to improvements in the Persona-Chat dataset.

Another increasingly popular use case is the generation of data on tasks with small labeled corpora to synthetically increase dataset size in order to train better performing classifiers. Dekker and van der Goot (2020) used synthetic data for lexical normalisation, while other researchers employed such data to train question answering models (Puri et al., 2020). Even in maths, researchers have proposed ways of creating synthetic theorems (Firoiu et al., 2021). Wullach et al. (2021) used GPT-2 (Radford et al., 2019) for their approach to hate speech detection. Their datasets were small to medium sized (6-53k labelled examples) and highly unbalanced, with as little as 1-6k hate speech samples per dataset. They created three mixed datasets containing 10k, 80k and 240k synthetic samples respectively, as well as 80% of the original datasets. The classification models trained on the largest created dataset outperformed those trained on the smaller datasets in most cases. Anaby-Tavor et al. (2020) generated data using GPT and improved sentence-level topic classification on three datasets, ranging from 4.2k to 17k entries. Wullach et al. (2021) and Anaby-Tavor et al. (2020) fine-tuned the respective GPT models on relatively small datasets, and find statistically significant improvements on classifier performance through incorporating synthetic data in the datasets used for fine-tuning classifiers.

While GPT and GPT-2 were trained on 117M and 1.5B parameters respectively, GPT-3 models were trained on up to 175B parameters (Radford et al., 2018, 2019; Brown et al., 2020). As it has been shown that an increase in model size systematically leads to improvements in text synthesis as well as common downstream tasks (Brown et al., 2020), GPT-3 is likely to produce higher quality and more natural sounding data than its predecessors. We thus expect GPT-3 generated data to have an even greater impact on performance in intricate language classification tasks such as PCL detection. We know of only few other research teams which used GPT-3 in their experiments, for instance to search for more suitable prompts for Natural Language Understanding (NLU) tasks (Liu et al., 2021b) or using prompts for few-shot generation (Yoo et al., 2021). Both achieved strong results on classification benchmarks.

While using foundation models for data generation has the potential to increase the power of language models and mitigate the data scarcity prob-

Dataset	PCL	nPCL	% PCL	$F1_{pos}$	Pre_{pos}	Rec_{pos}
DPM	993	9476	9.5	–	–	–
$DPM_{undersampled}$	804	804	50	40.74%	27.2%	**81.07%**
$DPM_{enhanced}$	10242	16937	37.7	**50.62%**	51.15%	50.10%
$DPM_{enhancedPos}$	7880	7580	49	46.76%	**54.39%**	41.01%
Official Baseline RoBERTa	–	–	–	*49.11%*	*39.36%*	*65.30%*
Post-Evaluation						
$DPM_{enhancedUnfiltered}$	24984	31886	43.93%	42.28%	43.62%	41.01%
$DPM_{enhancedPosUnfiltered}$	24984	7580	76.72%	44.07%	43.07%	45.11%

Table 2: Overview of the datasets used for fine-tuning as compared to the original dataset and test classification metrics.

lem prevalent in many NLP fields (Budzianowski and Vulić, 2019), this also bears potential risks not yet fully explored. For instance, past research showed that GPT-3 is biased in some cases, and that its defects are inherited by downstream models (Bommasani et al., 2021). Similarly, Bender et al. (2021) note, that the widespread application of foundation models carries a cost - both monetary and ethically. Thus, this approach's ethical implications should be investigated more thoroughly in future work.

3 System Overview

To generate the synthetic data, we used GPT-3's Curie model. Curie has about 13B unique parameters, while Davinci has about 175B. Although Davinci performs significantly better on a number of NLP tasks than Curie, we chose Curie, as it is more financially viable than the larger model, while retaining a comparatively strong performance (Brown et al., 2020).

For fine-tuning, we split the dataset into PCL and nPCL data and modified it to meet the API's requirements. As the API requires a prompt-completion pairing, the prompt was set to be empty (") and the completion contained the data sample. Afterwards, two GPT-3 Curie instances were fine-tuned on the PCL and nPCL data, respectively. We thus created two models, one to generate PCL and one for nPCL phrases. Following Wullach et al. (2021), we called the models with an empty (") prompt in the pipeline for synthetic data generation and the default parameters. We set *max_tokens* to the rounded mean length of the samples in the original dataset (60 for PCL and 54 for nPCL). With each iteration, we generated the maximum number samples (128), resulting in a total of 24.321 synthetic phrases by the nPCL model and 24.197 by the PCL model.

Like (Wullach et al., 2021), we classified all synthetic samples after generating the data. We used an initial baseline classifier and discarded all samples where the intended and predicted class did not match. Due to the high class imbalance of the original dataset, we randomly undersampled the negative class to the size of positive samples for training of the baseline classifier. We fine-tuned $BERT_{base}$-cased (Devlin et al., 2018) across three epochs using a learning rate of 1e-5 on the undersampled dataset. Since the synthetically generated data consisted solely of text for each label, we did not use any of the meta-information or context provided in the dataset and fine-tuned solely on text and labels. In the future, it might be useful to take meta-information into account for text generation.

39% of the generated PCL samples were classified as such by the baseline classifier, whereas 85,5% of generated nPCL samples were classified as nPCL. We explain this with the much larger sample size of nPCL in the DPM allowing the GPT-3 pipeline to generate better suited data. Based on the predictions, we created two enhanced datasets: For $DPM_{enhanced}$, we added a similar amount of synthetic PCL (9448) and nPCL (9357) samples to the DPM. For $DPM_{enhancedPos}$, we added 7086 PCL samples to balance the original dataset. For a comparison of sample sizes and share of PCL in the DPM and the different datasets used for fine-tuning see Table 2. On each of the enhanced datasets, we trained a $BERT_{base}$-cased instance the same way as our initial classifier. We submitted the classifier trained on $DPM_{enhancedPos}$ for our first and the classifier trained on $DPM_{enhanced}$ for the second run.

Intended	Text	Pred
coherent samples		
PCL	*so gao becomes emotional as he reflects on the thousands of homeless children he has come across during his decades long football career – most of them growing up without a father figure in their lives. "the kids today may be our future But there is no future for the kids today if we don't have*	PCL
PCL	*English and humanities teacher blowsy dilworth decided some kids in her georgian village needed more than a pack of cards to play baseball with – they needed an ancestral field. so she quested across the border to find some land for her students, and this week she opened a playfield on that*	nPCL
nPCL	*understandably , many sri lankans look at india with wariness, if not hostility. foster father pair of us destroyed an Eldorado of a country. thousands of families were made homeless and live on the streets today. on november medium*	PCL
nPCL	*africa has the largest block of 2017 retirements sufferance among all regions , with recent precedent of expenses course and after-inheritance taxed deaths , show disclosures by top investment funds in the united states . on the whole , fund seniors are think about leaving equities*	nPCL
incoherent samples		
PCL	*Subject : Crying Monkey Fortunetelling video 1 'sunday 's focus is on a widower , otis reigns , who recites a fortune to his 11 children while they weep , a performance that has attracted millions of views online . producer and director rebeca Ramirez says she*	PCL
PCL	*Crazy Horse 3 is aNATIVEpi agt sanctioned 51 majorityhare partnership firm jointly owned and managed by a group of indian stipendiaries and based in vancouver , b. c. agt Crazy Horse 3 is an eyaculofemoral orifice created for the purpose of*	nPCL
nPCL	*policy to homeseekers , students and the vulnerable.....transparency and public control of thebiologist!!!!!!!!!!*	nPCL
nPCL	*seems like coast is in need of some life. you could say that again about their women's Water Polo team. the t Vernons Wyr Kangas athletes recent 4ANPerformers cabinet hardwood men's schools100 result in need of some inspiring coast women*	PCL

Table 3: Examples of patronizing and non-patronizing generated data and its classification with the baseline classifier. Samples where intention and prediction matched were used for DPM*enhanced* and DPM*enhancedPos*, regardless of whether they are coherent or not. All synthetically generated data is available on github.

4 Results and Discussion

The evaluation metric used for ranking in the task was F1 over the positive class. Our baseline classifier reached an F1-score of 40.74% on the test set provided by the task organisers after the end of the competition's evaluation phase. Although it had a high recall of over 80%, precision was very low, leading to a suboptimal F1-score. The classifier trained on $\text{DPM}_{\text{enhanced}}$ scored almost 10% higher than the initial classifier, but had neither the highest recall, nor the highest precision of the three classifiers trained before the post-evaluation phase. This was suprising, as we initially expected the classifier trained on $\text{DPM}_{\text{enhancedPos}}$, which was the larger balanced dataset out of the three, to perform best. This leads to the assumption that with synthetic data, sheer amount might be more important than balancing out the dataset.

Although in the official task scoring, our system trained on $\text{DPM}_{\text{enhanced}}$ ranked in place 41 of 78 and surpassed the official baseline (fine-tuned RoBERTa) by only about 1%, we note that using both synthetically enhanced datasets led to a boost in performance compared to our initial classifier. This might seem surprising, especially considering the low performance of the initial classifier used to filter the GPT-generated data. In the post-

evaluation phase, we repeated the experiments from our two system runs without previous filtering of the GPT-output, to explore the role of the initial classifier in our system's performance. Neither $\text{DPM}_{\text{enhancedUnfiltered}}$ nor $\text{DPM}_{\text{enhancedPosUnfiltered}}$ led to better performance than $\text{DPM}_{\text{enhanced}}$. Thus, using a baseline classifier for filtering seems to be the most sensible option when working with synthetic data, regardless of its performance strength. We report on detailed classification results in Table 2. Since our baseline system did not perform very well in terms of classification, future work should first and foremost focus on improving it. The baseline system forms the basis of our approach and classification errors at this stage are likely to significantly lower the usefulness of the synthetic data.

We also looked at some of the synthetic data generated by GPT-3. Both for PCL and for nPCL, the generated samples were not always coherent on a semantic level and the occurrence of incoherent text appeared to be more common in the nPCL condition. However, it seems like coherence did not impact classification, as in both cases incoherent synthetic samples could be found in the final dataset (see Table 3).

We also found a lot of text in languages other than English, possibly because of the small size of the dataset in comparison to the vast amount of training data used to create GPT-3. We expect that filtering out such samples would increase performance further. In addition, basic data-cleaning of the synthetic data before classification might be in order. Both of this could potentially be achieved by only using data samples for which a confidence score above a certain percentage (i.e. 70%) is returned in classification. Another approach might be using an unrelated dataset to filter out all synthetic data unrelated to the task at hand. In the context of PCL detection, this could help discard generated data that is not related to vulnerable groups.

The approach of using an empty prompt (") while fine-tuning the models is debatable, because the prompt is such a powerful tool (Yoo et al., 2021; Liu et al., 2021b) and should probably be utilized. A possible approach would be to train a single model on both PCL and nPCL data, and put PCL/nPCL information in each samples' prompt. The currently unused meta-information of the dataset could also be incorporated, possibly causing additional improvements in the quality of the generated data.

5 Conclusion

We described our system submitted to Task 4, Subtask 1 of SemEval-2022. Although the system's performance did not score highly on the overall leaderboard, ranking 41st place, incorporating synthetic data in the original training set still boosted performance by up to 10% compared to our initial baseline system, which leads to the assumption that pairing this approach with more sophisticated classification systems has some potential to increase classification performance significantly. We derive some lessons learned from the presented experiments as follows:

- Using a baseline classifier to filter the synthetic data after generation seems to be essential.
- The size of the additional data seems to be more important to increase performance than balancing the data.
- Further data cleaning and filtering might be necessary to improve classification performance.
- Synthetic data leads to better performance, even if it includes a lot of incoherent samples and the baseline classifier has low performance.

In the future, we plan to improve the baseline classifier and explore different data cleaning and filtering techniques, such as using confidence scores returned by the classifier for our data selection, using unrelated datasets to filter whether a data sample fits in the task-specific domain or making use of prompts during GPT-3 fine-tuning and data generation. Exploring other augmentation strategies such as back-translation or synonym replacement of either the original data or the generated samples might further increase classification performance.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models

be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Kelly Dekker and Rob van der Goot. 2020. Synthetic data for english lexical normalization: How close can we get to manually annotated data? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6300–6309.

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*.

Vlad Firoiu, Eser Aygun, Ankit Anand, Zafarali Ahmed, Xavier Glorot, Laurent Orseau, Lei Zhang, Doina Precup, and Shibl Mourad. 2021. Training a first-order theorem prover from synthetic data. *arXiv preprint arXiv:2103.03798*.

Nelson F Liu, Tony Lee, Robin Jia, and Percy Liang. 2021a. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. *arXiv preprint arXiv:2102.01065*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.