ML_LTU at SemEval-2022 Task 4: T5 Towards Identifying Patronizing and Condescending Language

Tosin Adewumi, Lama Alkhaled, Hamam Mokayed, Foteini Liwicki
and Marcus Liwicki
Machine Learning Group
EISLAB, SRT
Luleå University of Technology
firstname.lastname@ltu.se

Abstract

This paper describes the system used by the Machine Learning Group of LTU in subtask 1 of the SemEval-2022 Task 4: Patronizing and Condescending Language (PCL) Detection. Our system consists of finetuning a pretrained Text-to-Text-Transfer Transformer (T5) and innovatively reducing its out-of-class predictions. The main contributions of this paper are 1) the description of the implementation details of the T5 model we used, 2) analysis of the successes & struggles of the model in this task, and 3) ablation studies beyond the official submission to ascertain the relative importance of data split. Our model achieves an F1 score of 0.5452 on the official test set.

Pérez-Almendros et al. (2020) introduced the dataset for the SemEval-2022 Task 4 (Pérez-Almendros et al., 2022)[1]. The dataset covers the English language. It is meant to support Natural Language Processing (NLP) models in identifying PCL towards vulnerable communities, such as poor families and refugees. The dataset is designed for 2 subtasks in the competition. Subtask 1 is a binary classification task of predicting the presence of PCL while subtask 2 is a multi-label classification task of predicting PCL categories. We address subtask 1 in this system paper.

PCL is an expression that depicts someone in a compassionate way or shows a superior attitude of the speaker (Pérez-Almendros et al., 2022). PCL identification is important because PCL has been shown to have harmful effects on vulnerable groups (Fox and Giles, 1996; Morris, 2007; Bell, 2013; Wang and Potts, 2019). This task of identifying and categorizing PCL is apparently more challenging than some other types of harmful language because it is subtle and generally used with good intentions (Wang and Potts, 2019; Gilda et al., 2022).

The main strategy of our system, to address the challenge, was to use a recent SoTA model (T5) in a simple, novel way to reduce out-of-class predictions. We discovered that our system achieves a relatively good performance on the task and PCL identification is a challenging task, due to its subtle nature. It achieved an F1 score of 0.5452 on the test set while the best score was 0.651. This made us rank 27 (66th percentile) out of 78 and we surpass the official RoBERTa baseline. We perform error analysis and ablation studies to evaluate the strengths and weaknesses of the model. We contribute the model checkpoint publicly on the HuggingFace hub[2] and the T5 code [3]

The rest of this paper is organized as follows. Section 1 gives a brief background of related work in PCL. Section 2 gives the system overview of what we used for the task. Section 3 describes the experimental setup for the task and the additional experiments beyond the official submission. Section 4 gives the tables of results and discusses relevant observations from the results. We share concluding remarks in section 5.

1 Background

Work on various sorts of harmful language in NLP has mostly concentrated on explicit aggressive and brazen phenomena (Pérez-Almendros et al., 2022). Scholars are striving to distinguish between harmful and unhealthy language by identifying the fundamental characteristics of unhealthy language. Price et al. (2020) proposed one of the most recent efforts in this regard. The research introduced a new dataset containing 44,000 comments with the unhealthy category sub-classified as either (1) hostile; (2) antagonistic, insulting, provocative or trolling; (3) dismissive; (4) condescending or patronizing; (5) sarcastic; and/or (6) an unfair generalisation. In their work, it is assumed that the language with a PCL tone will assume an attitude of superiority, implying that the other speak-

[1] semeval.github.io/SemEval2022/tasks

[2] huggingface.co/tosin/pcl_22

[3] github.com/tosingithub/pcl (after another competition)

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 473 - 478

ers/listeners are ignorant, naive, or unintelligent. In such scenarios, the language will usually imply that the other speaker should not be taken seriously.

Similarly, Morris (2007) explains the high likelihood of using PCL language when there is discussion between two persons with different mental health conditions. He demonstrated that patronizing language is common when a discussion occurred between a cashier with no cognitive issue and a customer who suffers from cognitive disability. Overall, PCL does not have an obvious negative or critical language and there is the challenge of limited, high-quality labelled data.

There have been different efforts at automatically detecting PCL. Wang and Potts (2019) showed that models with contextual representations are much better at identifying PCL and this bolstered the hypothesis that context is essential for PCL detection. They implemented the BERT model, which deploys a Transformer-based encoder architecture, on the TALKDOWN corpus they introduced. Both the base and large versions of the BERT model are implemented and evaluated over the new proposed corpus for balanced and imbalanced data. Price et al. (2020) added more context to their work by comparing the performance of BERT to human performance in order to better understand the model's performance. In their experiments, they observed that the BERT model detects PCL with a 78% accuracy, whereas the average over human annotators does so with a 72% accuracy. (Warholm, 2021) also finetuned a BERT model to classify the unhealthy comments in Norwegian data. This model was subjected to a variety of finetuning approaches to distinguish between condescending and non-condescending cases and in the binary classification subtask, the best accuracy was 0.862.

1.1 Data

"Don't patronize me" is an annotated dataset of PCL by (Pérez-Almendros et al., 2020) through crowdsourcing. It is a collection of texts which targets vulnerable communities. The dataset is extracted from News On Web (NoW) corpus[4], containing web articles from over 20 English-speaking countries. It contains 10,637 paragraphs. In addition to the words (*disabled, homeless, hopeless, immigrant, in need, migrant, poor families, refugee, vulnerable and women*) for identifying PCL for annotation in paragraphs, the following traits are also identified as indicators and used for acquiring the dataset:

- Words expressing feeling of pity towards the vulnerable community. For example: *god bless the victims , all those people and their poor families , and i feel so sorry but i want to tell them it was n't my son who did this , it was a different seifeddine*
- Words describing the vulnerable community as lacking certain privileges, knowledge or experience. For example: *After Vatican controversy, McDonald's helps feed homeless in Rome*
- Expressions that present members of the vulnerable community as victims. For example: *the biggest challenge is the no work policy . i think that refugees who come here , or asylum seekers , they 're unable to work and they have kids here – their kids are stateless . that 's really the cause of a lot of stress in the community*

The dataset was annotated by 3 expert annotators. It has two-level classification of PCL: binary classification used to determine if a paragraph has PCL or not, and then categorical label for those with PCL. The categorical classification has three higher-level categories: saviour, expert and poet. "Other" category is the final category to classify all paragraphs with PCL but that do not fit any of the previous categories. The saviour category represents text in which the author is in a privileged class as opposed to the target community. It has two subcategories: unbalanced power relations and Shallow solutions. The expert category is for text where the author is also in a privileged position and presents themselves as knowing better than the target group what their needs are. It also has two subcategories: presupposition and authority voice. The final category "Poet" is identified by how the author frames the community with a literary style writing. It has three subcategories: Metaphor, Compassion and The poorer the merrier.

2 System Overview

The T5 architecture (Raffel et al., 2020) is very similar to the originally proposed architecture of the Transformer by Vaswani et al. (2017). We use the pretrained base version of the model from the

[4]english-corpora.org/now/

HuggingFace hub (Wolf et al., 2020). Input sequence of tokens are mapped to embeddings and then passed to the encoder, which has alternating set of multi-head attention and feed-forward layers. The attention mechanism (Bahdanau et al., 2015) replaces each element of a sequence by a weighted average of the remaining sequence (Raffel et al., 2020). In addition to each self-attention layer of the decoder, there is the standard attention mechanism. As self-attention is order-independent, relative position embeddings are used in the architecture.

The training method (for both pretraining and finetuning) uses maximum likelihood objective (i.e. teacher forcing) and a cross entropy loss (Raffel et al., 2020). The model was pretrained on 34B tokens. Adam optimizer is used for optmization during finetuning. The model has 12 layers each in the encoder and decoder blocks and a total of 220M parameters (Raffel et al., 2020). When we refer to T5, we mean the base model, except where explicitly stated otherwise. The size of the model meant that a batch size of 64 or 32 required more memory than what is available on a single V100 GPU, so we lowered the batch size to 16. T5 takes a hyperparameter called a task prefix. We, hence, use 'classification: ' as the task prefix.

We introduced a correction to the out-of-class prediction of the model, as shown in the flow chart in Figure 1. Raffel et al. (2020) mentioned this issue as a possibility but they did not experience it. The issue appears to be because all the tasks the T5 model is trained on are framed as "text-to-text" before training. Hence, sometimes, the model might predict tokens seen during training but that do not belong to the category of classes in a classification task. This behaviour seems more common in the initial epochs of training and may not even occur sometimes. We further observed that replacing target labels with numbers and explicitly typecasting them as string reduces this occurrence, as the model becomes more stable with predictions.

We split 10% of the training set for validation (dev set) for both of our submissions to the competition. We explored different sizes, however, in further ablation studies, as explained in the next section. The 2 submissions of prediction files are based on 2 adaptive optimizers: Adam and AdamW (Loshchilov and Hutter, 2019). The predictions based on Adam had the better F1 score. Each experimental run was for 3 epochs and the model checkpoint with the lowest validation loss was saved and

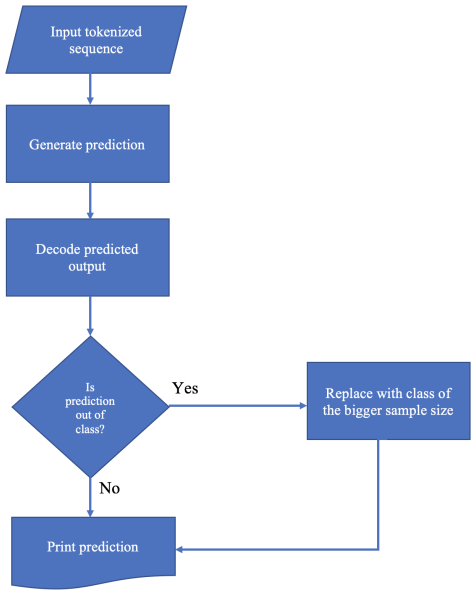

Figure 1: Flowchart of out-of-class code section for the T5 model during prediction.

used to make prediction on the test set. The initial learning rate and scheduler for both submissions are 2e-4 and linear schedule with warmup, respectively.

3 Experimental Setup

All the experiments were conducted on a shared DGX-1 cluster of 8 × 32GB Nvidia V100 GPUs. The server runs on Ubuntu 18 OS and has 80 CPU cores. The experiments were conducted in a Python (3.6.9) virtual environment with the PyTorch framework (1.8.1+cu102). We use both the training & test data provided by Pérez-Almendros et al. (2020). Besides the 2 submissions of prediction files, we perform ablation studies over the training/dev set split ratio (95%/5%, 90%/10%, 85%/15%, and 80%/20%). The training set was shuffled before splitting each dev set. We evaluate all the models using macro F1 scores, precision (P) and recall (R). In the absence of the ground truth of the test set, we perform error analysis by constructing the confusion matrix on a split of the dev set (20%). Further to that, in order to have a basis of comparison of the T5 model's strengths and struggles with the official RoBERTa baseline, we removed the 10 examples provided in Table 5 by Pérez-Almendros et al. (2022) from the training set and concatenated them with the dev set before training and evaluation. The predictions of 9 of the samples are given in Table 3.

Evaluation of the available data, by code, be-

fore and after running the script provided by Pérez-Almendros et al. (2022) to categorize the labels into 0 (neg) and 1 (pos) (for subtask 1) reveals that there are a total of 10,469 samples. The script treated paragraphs with the original labels 0 and 1 as 0 (instances not containing PCL) and paragraphs with the original labels 2, 3 and 4 as 1 (instances containing PCL). After running the script, the following are obtained: 9,476 samples classified as 0 and 993 classified as 1 in the training set. The test set has 3,832 samples. Before training, the following preprocessing steps were applied to all splits of the data:

- Emails & URLs are removed.
- All the characters are made lowercase.
- Extra spaces are removed.
- Special characters such as hashtags(#) and emojis are removed.
- Numbers & IP addresses are removed.

4 Results and Discussion

Our model performed relatively well with an F1 score of 0.5452 in the official assessment. This made it rank 27 (the 66th percentile) out of the 78 scores. All the F1 scores we report are macro scores. Our model has 11% advantage over the RoBERTa baseline, which achieved 0.4911, as shown in Table 1. Indeed, our second submission, based on the AdamW optimizer, also performs better than the baseline, achieving an F1 score of 0.5282, precision and recall of 0.5976 and 0.4732, respectively. The T5 model may have performed even better in the official rankings but for the shortcoming we described in section 2. In ablation studies, as shown in Table 2, we observe that training/dev set split ratio affects the performance of the system. All the results are based on submissions to the official evaluation system[5]. Using 5% of the training set as the dev set gave the worst F1 score but we observe improvements as the size is increased, though not linearly. We observe a sharp rise in F1 score when we increase the split from 5% to 10% but the rate of increase falls for subsequent increases.

[5]competitions.codalab.org/competitions/34344

Model	Rank	P	R	F1
best	1	0.646	0.6562	0.651
T5 (ours)	27	0.5801	0.5142	0.5452
RoBERTa baseline	43	0.3935	0.653	0.4911
worst	78	0.1059	0.0284	0.0448

Table 1: Abridged official result ranking for subtask 1.

Model (dev split)	P	R	F1
T5 (5%)	0.0725	0.8643	0.1339
T5 (10%)	0.6725	0.3628	0.4713
T5 (15%)	0.6067	0.4574	0.5216
T5 (20%)	0.5818	0.5047	0.5405

Table 2: Ablation studies results on the test set for subtask 1. Hyperparameters are the same for all model modifications. The T5 (10%) model is retrained afresh like the others, to avoid test/dev set feedback because of the samples in table 3.

4.1 Error Analysis

Since the ground truth labels of the test set are not available, we perform error analysis on the dev set. The T5 (20%) model achieves an F1 score of 0.7405 on the dev set (20%). However, the confusion matrix, as depicted in Figure 2, reveals that the model predicted 0 (neg) correctly 96.4% of the time while struggling to make the correct predictions when it came to 1 (pos), making only 47.8% of predictions correctly. This is very likely due to data imbalance, as 90.5% of the total training set contains samples labeled as 0 (neg). Ways of mitigating this may include data augmentation, possibly in a similar strategy to that used by Sabry et al. (2022), where an autoregressive model was deployed (Adewumi et al., 2022). A more careful stratification of the data split may also be helpful in this case.

Pérez-Almendros et al. (2022) report that the models they considered struggled to detect certain categories of PCL. We observe a similar challenge though our model achieves a better performance than the official baseline. For example, our T5 (20%) model's predictions for the same examples shown by Pérez-Almendros et al. (2022) for subtask 1 reveal that our model correctly predicts 5 out of the 9 displayed in Table 3, unlike the 3 correct predictions out of the 10 by the official baseline. The reason the T5 (20%) may have misclassified 2 of the samples labeled 0 (neg) in Table 3 may be because of tokens such as *vulnerable patients* and *hopelessly*, since they belong to the keywords used for annotating paragraphs with PCL, as discussed

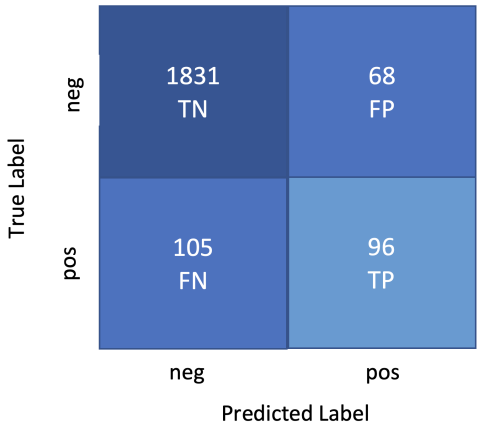

Figure 2: Confusion matrix of T5 (20%) on the dev set (20%). Macro F1 (0.7405): [0.9549 (neg) 0.5260 (pos)]

in section 1.

5 Conclusion

We describe the system involving the pretrained T5 model, which we use for our submission for the subtask 1 of the SemEval-2022 Task 4. We split 10% of the training set as dev set for hyperparameter evaluation in our official submission. Typecasting integer values, which represent classes, as string before feeding the T5 model and adjusting for out-of-class predictions improved the stability of the model in making predictions. Furthermore, in the post-competition phase, we performed ablation studies on the relative importance of dataset split by experimenting with different ratios of the training/dev set and showed what the model struggles with. Our results show that the encoder-decoder T5 model is competitive in this binary task and can obtain better performance with more hyperparameter tuning.

Acknowledgements

The authors wish to thank the anonymous reviewers for their feedback and the task organisers for their prompt attention whenever it was required.

Pred.	Paragraph	Gold
pos	from his personal story and real-life encounters with poor families , manpower correspondent toh yong chuan suggested shifting the focus from poor parents who repeatedly make bad decisions to their children (""" lifting families out of poverty : focus on the children ; last thursday) .	pos
pos	he said their efforts should not stop only at creating many graduates but also extended to students from poor families so that they could break away from the cycle of poverty	pos
neg	smyth told hkfp : """ the biggest challenge is the no work policy . i think that refugees who come here , or asylum seekers , they 're unable to work and they have kids here – their kids are stateless . that 's really the cause of a lot of stress in the community .	pos
neg	the people of khyber pakhtunkhwa are resilient . i did not see hopelessness on any face , """ he said .	pos
pos	teach kids to give back : when kang runs summer camps with kids , she includes """ contribution fridays """ – the kids work together as a team to make sandwiches for the homeless and dole out the food in shelters .	pos
pos	these shocking failures will continue to happen unless the government tackles the heart of the problem – the chronic underfunding of social care which is piling excruciating pressure on the nhs , leaving vulnerable patients without a lifeline .	neg
neg	lilly-hue : his ability to make sure our family is never in need – his sacrificial self .	neg
pos	"any kenyan small-scale farmer with such an income could not be said to be hopelessly mired in agrarian destitution . but of course , nothing in life is ever so simple as to allow for neat and precise answers ."	neg
neg	"selective kindness : in europe , some refugees are more equal than others"	neg

Table 3: Example predictions by T5 based on the dev set

References

Tosin Adewumi, Rickard Brännvall, Nosheen Abid, Maryam Pahlavan, Sana Sabah Sabry, Foteini Liwicki, and Marcus Liwicki. 2022. Småprat: Dialogpt for natural language generation of swedish dialogue by transfer learning. In *5th Northern Lights Deep Learning Workshop, Tromsø, Norway*, volume 3. Septentrio Academic Publishing.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015*.

Katherine M Bell. 2013. Raising africa?: Celebrity and the rhetoric of the white saviour. *PORTAL: Journal of Multidisciplinary International Studies*, 10(1):1–24.

Susan Anne Fox and Howard Giles. 1996. Interability communication: Evaluating patronizing encounters. *Journal of Language and Social Psychology*, 15(3):265–290.

Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. 2022. Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, 198:360–366.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Vann Morris. 2007. Patronizing speech in interability communication toward people with cognitive disabilities.

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.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovacs, Foteini Liwicki, and Marcus Liwicki. 2022. Hat5: Hate language identification using text-to-text transfer transformer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

Joakim Warholm. 2021. Detecting unhealthy comments in norwegian using bert. Master's thesis, UiT Norges arktiske universitet.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Acronyms

BERT Bidirectional Encoder Representations from Transformers. 2

LTU Luleå University of Technology. 1

NLP Natural Language Processing. 1

PCL patronizing and condescending language. 1, 2, 4

RoBERTa Robustly optimized BERT approach. 1, 3, 4

SoTA state-of-the-art. 1

T5 Text-to-Text-Transfer Transformer. 1–5