Felix&Julia at SemEval-2022 Task 4: Patronizing and Condescending Language Detection

Felix Herrmann & Julia Krebs
University of Regensburg
Regensburg, Germany
felix-georg.herrmann@stud.uni-regensburg.de, julia.krebs@stud.uni-regensburg.de

Abstract

This paper describes the authors' submission to the SemEval-2022 task 4: Patronizing and Condescending Language (PCL) Detection. The aim of the task is the detection and classification of PCL in an annotated dataset. The authors of this paper worked on two different models with finetuned hyperparameters focusing on number of epochs, training batch size, evaluation batch size, gradient accumulation steps and learning rate. The authors submitted one RoBERTa model and one DistilBERT model. Both systems performed better than the random and RoBERTA baseline given by the task organizers. The RoBERTA model finetuned by the authors performed better in both subtasks than the DistilBERT model.

1 Introduction

With the rise of social media, online hate speech has skyrocketed and has posed a problem for social media platforms: How can the giant number of messages on social platforms be surveilled, so that hateful comments can be reported and deleted? The answer to this is hate speech detection, a discipline within Natural Language Processing that has become increasingly popular and successful in recent years.

However, apart from hate speech, there is also other harmful language that should be studied. This is what the research by Perez-Almendros, Espinosa-Anke and Schokaert dives into. They collected an annotated dataset that focuses on a type of harmful language that is not so easily detected: patronizing and condescending language (PCL). The researchers describe PCL as follows: "An entity engages in PCL when its language use shows a superior attitude towards others or depicts them in a compassionate way." (Perez-Almendros et. al., 2020)

One of the difficulties in detecting PCL as opposed to openly hateful speech, is that persons who use PCL often do not intend to do harm, but instead want to support the groups that they name. PCL is usually aimed at vulnerable communities which makes the detection of PCL even more important.

The SemEval 2022 task 4 competition is based on this research by Perez-Almendros et. al: the detection of PCL. In two subtasks, one a binary classification and one a multi-label classification, the participants of the SemEval competition were tasked to submit up to two models.

The authors of this paper participated in both subtasks and propose a finetuning based approach on a pre-trained RoBERTa language model. Two models were submitted to the task organisers, one RoBERTa model and one DistilBERT model. The RoBERTa model performed better in both subtasks.[1]

In this paper, the authors will first describe the task set out by the SemEval 2022 competition in sections 2. In section 3, the authors explain the relevance of related work to this topic. Section 4 gives an overview over the BERT model. In the subsequent section, the two pre-trained models are presented. In the section experimental setup, the two different finetuned models are described in detail. Section 7 presents the results of the competition and section 8 the conclusion.

[1] The code can be found here: https://github.com/julia-ecrevisse/SemEval2022

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 357 - 362

2 Task description

This research paper contributes to task 4 of the SemEval 2022 competition[2]: PCL detection. The starting point of the research task is the paper "Don't Patronize Me! An annotated Dataset with Patronizing and Condescending Language Towards Vulnerable Communities" (Perez-Almendros et al., 2020). The researchers provide an annotated dataset with paragraphs taken out of news articles from English speaking countries[3] where vulnerable communities are mentioned. The following communities are included: disabled, homeless, hopeless, immigrant, in need, migrant, poor families, refugee, vulnerable and women. The dataset includes a total of 10,637 paragraphs extracted from the News on Web corpus. 3,554 of the paragraphs had been labeled as PCL by the annotators.

The aim of the SemEval task is 1) to identify which paragraphs include PCL and 2) if the paragraph includes PCL which category or categories it belongs to. The two subtasks are described as follows:

Subtask 1: Binary classification. Given a paragraph, a system must predict whether or not it contains any form of PCL.

Subtask 2: Multi-label classification. Given a paragraph, a system must identify which PCL categories express the condescension. There are seven different categories: Unbalanced power relations, shallow solution, presupposition, authority voice, metaphor, compassion and the poorer, the merrier.

3 Related Work

While other areas related to hate speech and hateful language have been a focus of NLP research in recent years, the research into PCL is still limited. However, there are a few researchers that have delved into this discipline from different angles.

Wang and Potts (2019) discuss condescending language and its detection in their research and also point out that high quality data for this kind of language detection is still limited. They introduce the dataset "Talkdown" which is a data set from the social media platform Reddit. Their approach concentrated on BERT baseline models and the researchers concluded that condescending language is highly connected to context (Wang and Potts, 2019).

Taking a different angle, Sap et al. (2020) write about the social and power implications of language. They claim that by using language with, e.g. social bias, stereotypes and prejudices are reinforced. Their collected data stems from different social media sites. They state in their results that previously successful models can categorize social bias relatively well, but they have difficulties classifying the social bias frames which were developed by the authors (Sap et al., 2020).

BERT

The state of the art in natural language processing is significantly characterized by architecture-based transformer-encoder models called BERT (Bidirectional Encoder Representation from Transformers) (Peters et al., 2018).

BERT's architecture relies on a two-part training process, a pretraining using unlabeled text corpora and a subsequent test run using labeled data (Develin et al, 2019). Here, BERT models do not use any decoder layers (Rothman, 2021) and fully rely on the encoder structures developed by Devlin et al. (2018). Here, these models are divided into base and large.

In order not to have to perform the training process, which is time-consuming, for each new task, especially pre-trained models are made available on platforms such as Huggingface. These can be adapted to the respective requirements by means of fine-tuning.

This previous research into hate speech and hateful language as well as the research into BERT model provides the basis of this paper. The challenge of condescending or patronizing language as opposed to hate speech is that condescending or patronizing language is often more difficult to detect as it is often not on purpose or not as obvious. Still, the steps taken by previous authors show a way on how to tackle these challenges.

4 System overview of pretrained models

In the paper "Don't Patronize me!" which is the starting point of the SemEval Task, the researchers

[2] https://sites.google.com/view/pcl-detection-semeval2022/home?authuser=0

[3] News on Web Corpus: https://www.english-corpora.org/now/

point out that an NLP model with BERT achieves the best results. More specifically RoBERTa performs slightly better than DistilBERT and BERT-base (Perez-Almendros et. al., 2020).

4.1 RoBERTa model

The RoBERTa model stands for Robustly Optimized BERT Pretraining Approach and is based on the basic architecture of BERT, which, however, has been insufficiently trained for the subtasks to be handled. RoBERTa increases performance by improving the pretraining processes. To generate this progress, the number of pretraining transformers is increased. An example of this is the omission of WordPiece tokenization (Song et al., 2021) and the associated structuring at the byte-level byte pair encoding level (Rothman 2021).

For both subtasks, the RoBERTa base model was used, which consists of 12 encoders. In contrast, the RoBERTa large consists of 24 (Nester, 2022). The decisive factor for the choice of the model was the improved modification of the hyperparameters in the fine-tuning, which were our primary focus. In addition, the large amount of data used to pre-train RoBERTa was a reason for the choice. From this, we aimed to improve the generalization of the responding capability compared to the conventional BERT model (Delobelle et al., 2020).

Apart from the RoBERTa baseline, another RoBERTa model from the transformers library was tried (Devlin et al, 2019): *xlm-roberta-base* (Huggingface), with the following results: Subtask 1 F1 score: 0.466, Subtask 2 average F1 score: 2: 0.110. As this baseline performed worse than the original baseline, the authors decided to continue with the original RoBERTa baseline.

4.2 DistilBERT model

DistilBERT is often called the "faster and cheaper" version of BERT. Using up massive amounts of data is expensive and also wasteful. In their research Sanh, Debut, Chaumond and Wolf describe that they were able to "reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster" (Sanh et. al., 2020). That is why it makes sense to see if there are promising results with a DistilBERT model also for this SemEval task.

Two different DistilBERT models from transformers were tried for the two subtasks: *distilbert-base-uncased* and *distilbert-base-uncased-finetuned-sst-2-english*. Both were first tried as a baseline. Comparing both models, the *distilbert-base-uncased* showed better results for the baseline, especially for subtask 2 (see table 1),

Baseline	***distilbert-base-uncased***	***distilbert-base-uncased-finetuned-sst-2-english***
Subtask 1 F1 score	0.488	0.433
Subtask 2 average F1 score	0.121	0.060

Table 1: DistilBERT baselines F1 scores

that is why it was decided to continue with this model in the following fintetuning stage. For the multilabel classification of subtask 2 the most promising predictions of the *distilbert-base-uncased* baseline were in the categories unbalanced power relations, presupposition, and compassion.

The *distilbert-base-uncased-finetuned-sst-2-english* model only showed a good F1 score for the category unbalanced power relations, whereas all other categories had a F1 score of 0.0.

5 Experimental setup

Fine-tuning was performed on Google Colab. The authors used the train and test set as it was provided by the task organizers.

5.1 RoBERTA model

Subtask 1

The experimental setup used was the RoBERTa baseline, which was provided by the Semeval 2022 team for the task. For fine-tuning, 9 different hyperparameters were run in different combinations with the model. For this, the F1 score was set as a benchmark and depending on the development, the hyperparameter was pushed to its limit. Using the example of the epochs, a steady improvement of the F1 score could be observed up to level 5, before it deteriorated again. If such a limit of a parameter was reached, it could be set as default and the tuning could be supplemented by another one. This resulted in a combination of 5 epochs, a learning rate of 3e-5, an evaluation batch

size of 32, a training batch size of 16 and gradient accumulation steps of 2.

In addition, the weight_decay, adam_epsilon, max_grad_norm and the warmup_steps were implemented at different levels. However, these led to a deterioration of the precision and the F1 score in all variants. However, it is worth mentioning that there was a significant improvement of the recall by using warmup_steps= 3500. Without further illuminating this direction, a recall of 0.804 was achieved, which cannot yet be called a maximum.

Subtask 2

The findings from Subtask 1 were to be applied to multi-labeling. For this purpose, both the RoBERTa-baseline model and the same hyperparameters were adopted. Again, the different hyperparameters were pushed to maximum improvement with the goal of obtaining the highest possible mean value of the individual F1 scores of the labels. Epochs could be increased to 20 until the peak was reached. All other efforts to obtain an optimized result away from the epochs or in combination with them were unsuccessful. The mean value reached its possible optimum after 20 epochs with a result of 0.258.

5.2 DistilBERT model

For the DistilBERT model, three hyperparameters were tried by the authors: number of trained epochs, the evaluation batch size and the training batch size. The number of trained epochs has been adapted from 1 to 20 epochs, while for the evaluation and training batch size 16, 32 and 64 was used respectively.

	Subtask 1	Subtask 2
Number of trained epochs	5	20
Evaluation batch size	32	64
Training batch size	32	16

Table 2: Hyperparameters used for DistilBERT model

The best results were achieved with the hyperparameters described in Table 2.

6 Results

An overview of the results can be found in table 3 and 4.

	RoBERTa model	DistilBERT model	Baseline RoBERTa
Precision	0.401	0.357	0.394
Recall	0.773	0.640	0.653
F1	0.528	0.459	0.491

Table 3: Results subtask 1

F1 scores	RoBERTa model	DistilBERT model	RoBERTa baseline
unbalanced power relations	0.366	0.352	0.354
shallow solution	0.351	0.345	0
presupposition	0.176	0.2	0.167
authority voice	0.221	0.163	0
metaphor	0.211	0.095	0
compassion	0.285	0.271	0.209
the poorer, the merrier	0.167	0.0	0
average	0.254	0.204	0.104

Table 4: Results for subtask 2

6.1 RoBERTa model

In Subtask 1 as well as in Subtask 2 significant improvements could be achieved compared to the baselines. This is reflected in the results of the test data set as well as the final data set. The finetuning improved the initial value (F1: 0.483) of the RoBERTa baseline in Subtask 1 by about 0.05 to 0.547. A significant increase of about 0,10 (0.441 to 0.350) was observed in the Precision section. In the Competition a similar value was achieved with, so that the robustness of the system and its configuration is given. An overfitting could not be determined. The results were ranked 31st with a Precision of 0.401, a Recall of 0.773 and the F1 0.528.

In Subtask 2, RoBERTa nearly doubled the mean F1 score from baseline 0.134 to 0.258. On the official leaderboard, an almost identical value of 0.2536 was achieved, so that here, as in Subtask 1, no overfitting prevailed and the robustness of the system was confirmed. The RoBERTA model performed best in the categories “unbalanced power relations” and “shallow solution” (see table 4). The worst results were reached for the

categories “presupposition” and “the poorer the merrier”.

6.2 DistilBERT model

A comparable scenario occurred when running the DistilBERT model with the test data. Compared to the Random as well as the RoBERTa Baseline, the results significantly improved in both subtasks. In the comparison, however, the Random Baseline is used first. An increase from 0.174 (Random) to 0.512 could be achieved by applying the above mentioned hyperparameters. The RoBERTA baseline was thus also exceeded by approximately 0,30. However, in the official results, the score was not repeated and was 0.459 (see table 3).

In Subtask 2, the Random Baseline result with the test data of 0.055 for the mean of the F1 scores was almost tripled to 0.146. However, compared to the RoBERTa baseline of 0.134, only a small improvement was observed.

Surprisingly, the model performed much better with the official data and thus achieved its maximum value of 0.203. Also, for the DistilBERT model, the categories with the best results were “unbalanced power relations” and “shallow solutions” (see table 4). The DistilBERT model performed worse than the RoBERTA model in most categories, except “presupposition”. The system performed worst for the category “the poorer, the merrier” compared with the other categories.

7 Conclusion

The purpose of this work was to identify PCL in texts using NLE. For this purpose, existing text classification models were used and adapted to the task by fine tuning. The goal was to achieve the highest possible F1 scores by the model, which is equivalent to the percentage detection of condescending terms. A complete detection of all these terms was not achieved, but in the binary classification with RoBERTa a score above 0.52 could be obtained. The multi classification was the bigger challenge and could only finish with a score around 0.250.

Despite these results, a basis, if not an improvement, has been created for future work. The hyperparameters which were used can already be set to default in the next works at the beginning and thus increase the speed of the development process for researches in PCL detection by BERT models. More research could also be done into the different categories of PCL. There were categories that consistently performed better than others (unbalanced power relations and shallow solution), while the systems had more difficulties with other categories (especially the poorer, the merrier).

There are still a lot of directions that the research into PCL detection can continue. As stated above, PCL detection in NLE is an emerging field that would benefit from further research. The task is more difficult than traditional hate speech detection as PCL is often more subtle. However, PCL can contribute to repeating and furthering stereotypes and discrimination, especially among vulnerable communities and therefore the research into PCL detection systems is a vital endeavor.

Acknowledgments

We would like to thank our professor Dr. Udo Kruschwitz at the University of Regensburg for encouraging us to take part in the SemEval 2022 evaluation and supporting us during the process. We would also like to thank our study colleagues Aenne and Amela with whom we exchanged ideas and approaches throughout the timeframe of the SemEval 2022 competition.

References

Delobelle P., Winters T., and Berendt B.. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics

Devlin J., Chang M., Lee K., and Toutanova K.. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nester, J.. 2022. Rassistische Sprache mit BERT erkennen: Eine Untersuchung am Beispiel deutscher Plenarprotokolle. Universität zu Köln.

Perez-Almendros, C., Espinosa-Anke, L., Schockaert, S.. 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.

Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L.. 2018. Deep contextualized word representations. In *NAACL 2018*.

Rothman, D.. 2021. Transformers for natural language processing: build innovative deep neural network architectures for NLP with Python, Pytorch, TensorFlow, BERT, RoBERTa, and more. Birmingham, Mumbai 2021

Sanh V., Debut L., Chaumond J., Wolf T.. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sap M., Gabriel S., Qin L., Jurafsky D, Smith N., and Choi Y.. 2020. Social bias frames: Reasoning about social and power implications of language. arXiv preprint arXiv:1911.03891.

Song X., Salcianu A., Song Y., Dopson D., and Zhou D.. 2021. Fast WordPiece Tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang Z., Potts C.. 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*.