Amsqr at SemEval-2022 Task 4: Towards AutoNLP via Meta-Learning and Adversarial Data Augmentation for PCL Detection

Alejandro Mosquera
Broadcom Corporation / 1320 Ridder Park Drive San Jose, 95131 California, USA
alejandro.mosquera@broadcom.com

Abstract

This paper describes the use of AutoNLP techniques applied to the detection of patronizing and condescending language (PCL) in a binary classification scenario. The proposed approach combines meta-learning, in order to identify the best performing combination of deep learning architectures, with the synthesis of adversarial training examples; thus boosting robustness and model generalization. A submission from this system was evaluated as part of the first subtask of SemEval 2022 - Task 4 and achieved an F1 score of 0.57%, which is 16 percentage points higher than the RoBERTa baseline provided by the organizers.

1 Introduction

The harmful use of language in social media can have negative and long-lasting effects such as exclusion and unfair treatment, specially when targeting vulnerable communities. For this reason, the detection of toxic, hateful and abusive comments has been the central topic of several workshops and tool evaluations, drawing a lot of attention from the Natural Language Processing (NLP) research community in the last years. However, while toxic language has a clear intent and is usually obvious to the reader, patronizing and condescending language (PCL) is more subtle and likely used in a subconscious manner even in traditional media (Perez Almendros et al., 2020). The aforementioned characteristics and its subjective nature makes PCL harder to identify than abusive comments by both humans (Sap et al., 2019) and NLP applications.

The continuously increasing taxonomies of language misuse poses new challenges to social media platforms, thus not only requiring more effort and cost in order to identify abuse across different languages and textual genres but also having to keep a balance between aggressive and conservative filtering strategies. On the one hand, users eventually devise ways of evading automatic content moderation (Gerrard, 2018), while on the other hand, policing that restricts freedom of speech can lead to distrust (Kirk and Schill, 2021). For these reasons, content filters usually rely on the latest advances in NLP research, dominated in the recent years by deep learning architectures. Despite the competitive scores achieved via transfer learning and models such as the Transformer (Vaswani et al., 2017) in this area, choosing and optimizing the right modeling framework for a given NLP task is still a non-trivial problem.

Automated Natural Language Processing (AutoNLP), the equivalent of Automated Machine Learning (AutoML) for NLP, is a relatively new field of study that aims to automate the iterative components of developing a NLP model given a specific input data and task without requiring any special domain expertise. By building upon existing concepts such as transfer learning, data augmentation and meta-learning the author hypothesizes that is possible to generate strong NLP baselines with minimal human interaction. An analysis of the results of the shared task 4 of SemEval-2022: Patronizing and Condescending Language Detection (Pérez-Almendros et al., 2022) shows that AutoNLP can be successfully applied to PCL classification, obtaining a 16% higher F1 score than the baseline provided by the task organizers.

This paper is organized as follows: In Section 2, the state of the art is reviewed. Further on, Section 3 describes the AutoNLP approach for PCL classification. Next, in Section 4, an in-depth discussion of the results obtained is described, and finally Section 5 concludes this research and outlines future work.

2 Related Work

There have been several research works on the detection of different types of harmful language, not only focused on the most explicit such as hate

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 485 - 489
July 14-15, 2022 ©2022 Association for Computational Linguistics

speech (Zampieri et al., 2019) (Garibo i Orts, 2019) but also more subtle usages such as condescending interactions (Wang and Potts, 2019) and social power implications (Sap et al., 2020). PCL towards vulnerable communities in news articles has also been characterized into 7 categories (Perez Almendros et al., 2020) used in order to label the most comprehensive PCL-annotated corpus to date: the Don't Patronize Me! (DPM) dataset, the official training resource for the shared task 4 of SemEval-2022: Patronizing and Condescending Language Detection.

3 AutoNLP for PCL

Deep neural network modeling techniques have inspired state of the art approaches in various domains, such as image classification and language modeling, thus dominating several benchmarks and shared tasks in the last years. For this reason, NLP applications relying on manually-crafted features have been less popular in comparison with deep learning (DL) architectures (Young et al., 2018), specially where extensive manual feature engineering is required to achieve a similar performance (Mosquera, 2021). However, since building a high-quality DL system for a specific task still relies on human expertise, AutoML offers a promising solution to this problem by automating most of the modeling steps (He et al., 2021).

In order to tackle an arbitrary NLP classification task, in this case PCL detection, a custom end to end AutoNLP solution has been designed and evaluated by using exclusively the DPM dataset provided by the organizers, off-the-shelf pre-trained models and without applying any special pre-processing or feature engineering besides standard tokenization. The main components of the system are described in the following section.

3.1 Adversarial Data Augmentation

Adversarial data augmentation can not only increase model robustness but also improve generalization by increasing the number of training samples (Shorten et al., 2021). This can be specially relevant when using neural networks, which tend to under-perform in a low-data regime (Antoniou et al., 2018). The different data augmentation strategies incorporated in the AutoNLP pipeline are as follows:

- **Backtranslation**: Transformation using TextAttack (Morris et al., 2020) that translates a PCL sentence into a random target language and translates it back to English.
- **Checklist**: TextAttack implementation of the Invariance Testing Method: Contraction, Extension, Changing Names, Number, Location (Ribeiro et al., 2020) applied to the positive class.
- **Wordnet**: Word swap by swapping synonyms in WordNet (Fellbaum, 1998) for PCL paragraphs.
- **Embedding**: Attack that replaces words with synonyms in the word embedding space (Mrkšić et al., 2016) for PCL texts.
- **Counterfactual**: Inspired by the concept of counterfactual augmentation (Kaushik et al., 2020), this manipulation only applies to text from the positive class which is augmented with random texts from the negative class. The resulting paragraph should still have a positive (PCL) label.
- **Shuffle**: Attack that shuffles words in a PCL paragraph.
- **Parrot**: Paraphrased PCL sentences generated with Parrot (Damodaran, 2021).
- **Pegasus**: PCL augmentation by generating paraphrases via conditional augmentation using Pegasus (Zhang et al., 2019).

3.2 Meta-learning

A common approach to meta-learning is stacked generalization (Wolpert, 1992), where a set q of base learners applied to a training set $T_{train} : \{(\tilde{X}_i, c_i)\}_{i=1}^{m}$ to produce q hypotheses $\{h_j\}_{j=1}^{q}$ is redefined into a new set T'_{train} by replacing each vector $\tilde{X}_i$ with the class predicted by each of the q hypothesis on $\tilde{X}_i$. T'_{train} is used as input to a set of meta-learners, producing a new set of hypotheses (Vilalta and Drissi, 2001).

While this approach has been successfully applied in several NLP tasks (Li and Zou, 2017) (Mosquera, 2020), an small variation that deals with skewed datasets and automatically sub-samples the majority class in each base learner (Chan and Stolfo, 1998) was considered instead for this challenge. In order to do this, a pool of 40 base learners was generated by randomly combining different data augmentation approaches, deep learning architectures via transfer learning and sub-sampling factors. Logistic regression was used as meta-learner in the second layer, with probability thresholds and hyper-parameters optimized via cross-validation.

Several pre-trained resources were used for fine-tuning with early stopping including BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), GloVe (Pennington et al., 2014) embeddings with capsule networks (Frosst et al., 2018), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019). The number of optimal training epochs was determined via cross-validation. However, for cost mitigation purposes, no model was trained for longer than 10 epochs and most hyper-parameters were left with the default values.

3.3 Model Selection

The maximum relevance and minimum redundancy (MRMR) algorithm (Zhao et al., 2019) was applied as feature selection method, reducing the final number of base learners used by the meta-model to 8.

After analyzing the cross-validation results we can observe that base models fine-tuned with ELECTRA obtained the highest F1 scores. Likewise, the most successful data augmentation was the combination of the Checklist and Backtranslation methods. The final list of base learners, including their cross validation F1 score and logistic regression coefficient is shown in Table 1.

Model	**Augmentations**	**F1**	β
BERT	Checklist	0.52	**0.31**
ELECTRA	Checklist	**0.55**	0.25
ELECTRA	Checklist Backtranslation	**0.55**	0.17
ELECTRA	Checklist Backtranslation Embedding Counterfactual Wordnet	0.54	0.13
RoBERTa	Checklist Backtranslation	0.53	0.30
RoBERTa	Parrot	0.54	0.14
RoBERTa	Checklist Backtranslation Embedding	0.54	0.09
RoBERTa	Checklist Backtranslation Embedding Counterfactual Wordnet	0.53	0.13

Table 1: Final list of base learners selected via MRMR with their cross-validation score and regression coefficient estimated during the training phase.

4 Evaluation and Results

Final test set results obtained in the PCL classification task by the AutoML system (amsqr) and the winning submission (hudou) can be found in Table 2. The official RoBERTa baseline and the development set results are also included for comparison purposes.

Model	**Precision**	**Recall**	**F1**
hudou	**0.646**	**0.656**	**0.651**
amsqr (dev)	0.587	0.578	0.582
amsqr (test)	0.547	0.599	0.572
RoBERTa baseline	0.393	0.653	0.491

Table 2: PCL classification results.

The fact that only 42 out of 78 competing teams were able to beat the RoBERTa baseline provided by the task organizers highlights the difficulty of this competition. Besides the nature of the task, other challenging factors were the strong class imbalance and the considered evaluation metric, which required careful tuning of classification thresholds via cross-validation (Lipton et al., 2014). A post-competition analysis in Table 3 shows that the automatically chosen classification threshold of 0.26 during training was also optimal for the test set.

Threshold	**Precision**	**Recall**	**F1**
0.20	0.498	**0.656**	0.566
0.22	0.516	0.634	0.569
0.24	0.532	0.621	**0.573**
0.28	0.558	0.586	0.572
0.30	**0.566**	0.574	0.570

Table 3: Post-competition classification results in the test set for different probability thresholds.

5 Conclusion and Future Work

This paper describes the system developed for the PCL detection task of SemEval 2022. The author demonstrates that the selected AutoNLP approach can produce competitive results by leveraging meta-learning, adversarial data augmentation and pre-trained resources. Automatic hyper-parameter optimization and exploring different meta-learning algorithms are left to a future work.

References

Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2018. Data augmentation generative adversarial networks.

Philip Ka-Fai Chan and S. Stolfo. 1998. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Nicholas Frosst, Sara Sabour, and Geoffrey Hinton. 2018. Darccc: Detecting adversaries by reconstruction from class conditional capsules.

Òscar Garibo i Orts. 2019. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20:4492 – 4511.

Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Rita Kirk and Dan Schill. 2021. Sophisticated hate stratagems: Unpacking the era of distrust. *American Behavioral Scientist*, page 00027642211005002.

Wen Li and Liang Zou. 2017. Classifier stacking for native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 390–397, Copenhagen, Denmark. Association for Computational Linguistics.

Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. 2014. Thresholding classifiers to maximize f1 score.

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. *CoRR*, abs/1907.11692.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Alejandro Mosquera. 2020. Amsqr at SemEval-2020 task 12: Offensive language detection using neural networks and anti-adversarial features. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1898–1905, Barcelona (online). International Committee for Computational Linguistics.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Connor Shorten, Taghi Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ricardo Vilalta and Youssef Drissi. 2001. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context.

David Wolpert. 1992. Stacked generalization. *Neural Networks*, 5:241–259.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Zhenyu Zhao, Radhika Anand, and Mallory Wang. 2019. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform.