ASRtrans at SemEval-2022 Task 4: Ensemble of Tuned Transformer-based Models for PCL Detection

Ailneni Rakshitha Rao
Indian Institute of Technology, Gandhinagar
rao_ailneni@alumni.iitgn.ac.in

Abstract

Patronizing behavior is a subtle form of bullying and when directed towards vulnerable communities, it can arise inequalities. This paper describes our system for Task 4 of SemEval-2022: Patronizing and Condescending Language Detection (PCL). We participated in both the sub-tasks and conducted extensive experiments to analyze the effects of data augmentation and loss functions used, to tackle the problem of class imbalance. We explore whether large transformer-based models can capture the intricacies associated with PCL detection. Our solution consists of an ensemble of the RoBERTa model which is further trained on external data and other language models such as XLNeT, Ernie-2.0, and BERT. We also present the results of several problem transformation techniques such as Classifier Chains, Label Powerset, and Binary relevance for multi-label classification.

1 Introduction

In this paper, we discuss various linguistic techniques used for detecting Patronizing and condescending language (PCL). This task particularly poses a new challenge in the field of NLP because of its subjectivity, subtle usage, and requirement of world knowledge. A person is said to be condescending or patronizing when he/she uses a superior tone to talk down to people or tries to raise pity by describing their situation. Even though people often use PCL with good intentions, it encourages stereotyping, discrimination and leads to greater exclusion. Therefore, it is important to devise methods that facilitate the automatic detection of PCL. SemEval 2022 Task 4 (Pérez-Almendros et al., 2022) is the first attempt to detect the usage of PCL towards vulnerable communities. It has two subtasks: In sub-task A, we need to detect if the given paragraph contains PCL or not. Sub-task B aims to classify PCL text further among potential intersecting categories.

The two main challenges faced in this task are extreme class imbalance in the dataset and the subtle nature of PCL present in the text which makes it hard, even for humans to classify it correctly. Also, the model needs to differentiate between actual news of extremely vulnerable situations from text containing PCL.

Pre-trained language models such as BERT, XLNeT, etc., have emerged as the state-of-the-art models for many NLP tasks such as text classification, machine translation, sequence tagging, etc. However, they are trained on typical day-to-day texts. PCL text is not trivial and classifying certain classes requires some level of world knowledge and commonsense reasoning. In this paper we conduct detailed experiments using the following models: RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2020), Ernie-2.0 (Sun et al., 2019), BERT-large (Devlin et al., 2019), label specific attention network (LSAN) (Xiao et al., 2019) and their ensembles for PCL detection. Additionally, we test the Classifier Chain approach for multi-label classification.

We achieved significant improvement over the baseline RoBERTa model in both the sub-tasks. We were ranked 52^{nd} with an F1 score of 0.4421 in sub-task A and 33^{rd} with an F1 score of 0.1889 in sub-task B among 81 teams. We release the code for models and experiments via GitHub [1]

The rest of the paper is organized as follows: Section 2 describes the challenge, followed by a brief literature survey. Section 3 explains the proposed approach in detail, while section 4 presents the experimental details required to reproduce the results. Results and analysis are shown in section 5. Finally, conclusions are drawn in section 7.

[1] https://github.com/rak55/ASRtrans-semeval2022

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 344 - 351

2 Background

2.1 Problem Description

SemEval 2022 Task 4: Patronizing and Condescending Language Detection (Pérez-Almendros et al., 2022) is a paragraph-level text classification problem that consists of two sub-tasks. **Task A** is a binary classification task designed to predict if the text consists of any form of PCL. In **task B**, we need to further classify PCL text among potential categories namely *unbalanced power relations*, *shallow solution*, *presupposition*, *authority voice*, *metaphor*, *compassion*, and *the poorer the merrier*.

2.2 Related Work

Hate language detection Though the detection of patronizing and condescending language has not been studied in depth in the field of NLP, extensive work has been done in several forms of harmful language detection such as Automated hate speech detection (Davidson et al., 2017), rumor propagation (Gorrell et al., 2019), fake news detection (Conroy et al., 2015), and trust-worthiness prediction (Barrón-Cedeño et al., 2018). However, recently (Wang and Potts, 2019) introduced a labeled dataset named *TalkDown* derived from Reddit communication threads for modeling condescension. (Sap et al., 2020) introduced *Social Bias Inference* corpus to study the unbalanced power relations present in the condescending language.

Multi-label text classification There are mainly three different techniques to solve a multi-label text classification problem: Binary Relevance, Classifier Chains (Dembczyński et al., 2010) and Label Powerset method (Boutell et al., 2004). Binary Relevance treats each class independently and ignores label dependence. The Label Powerset method considers each combination of labels as a distinct class, thereby transforming a multi-label classification problem into a single-label problem. (Chen et al., 2007) propose document transformation by assigning label weights based on label entropy. (Alvares-Cherman et al., 2012) tries to incorporate label dependency into the Binary Relevance method. ReliefF and Information Gain are combined with Binary Relevance and Label Powerset approaches in (Spolaôr et al., 2013) to evaluate the importance of each label. (Wang et al., 2017) show that regularising the model during the training phase and using support inference during prediction along with F-optimizer improves the F1 score of the multi-label problem.

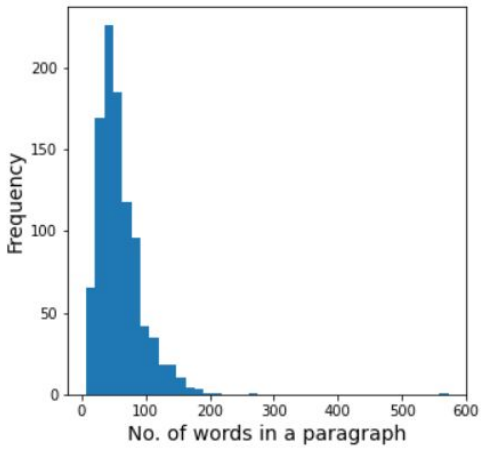

Figure 1: Frequency distribution of the text length.

Label	No. of samples
Unbalanced power relations	230
Shallow solution	716
Presupposition	196
Authority voice	224
Metaphor	469
Compassion	197
The poorer, the merrier	40

Table 1: Distribution of training data for Task B.

Transfer Learning Pre-trained transformer-based language models such as BERT, RoBERTa, etc., often outperform many traditional models trained from scratch. This success can be attributed to their rich contextual embeddings. Therefore, these models are used for many downstream tasks. (Li and Xiao, 2020) use SpanBERT for the detection of propaganda techniques in news articles. (Ranasinghe and Hettiarachchi, 2020) use BERT-based multilingual models for offensive language identification in social media.

3 System Overview

3.1 Data

The *Don't Patronize Me!* dataset (Pérez-Almendros et al., 2020) provided as training data for both the sub-tasks consists of paragraphs extracted from the News on Web (NoW) corpus. The distribution of different labels in the dataset is shown in table 1. Each training example in sub-task A consists of a *doc-id*, *keyword*, *country-code*, *paragraph*, and *label*. The text was retrieved from the news of 20 English-speaking countries based on 10 keywords belonging to vulnerable communities like disabled, homeless, etc. The frequency distribution of text

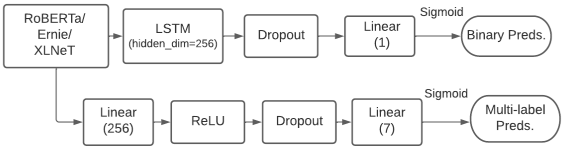

Figure 2: Flow chart for multi-task learning.

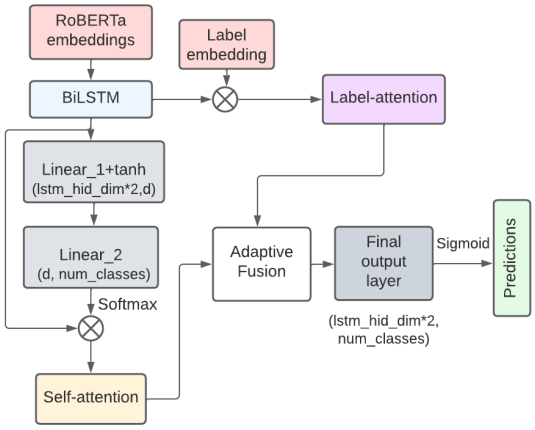

Figure 3: Architecture of LSAN.

length is shown in figure 1. The dataset for task B contains span-level annotation of PCL among seven categories mentioned earlier.

3.2 Transformer-based models

Our approach is to train several pre-trained language models including RoBERTa, Ernie-2.0, XLNet, and BERT-large individually and in an ensemble with different problem transformation approaches for both the sub-tasks. We conduct experiments with two different settings: Multi-task and single-task learning. The model used for multi-task learning is described in figure 2.

3.2.1 Using External data

A RoBERTa model is further trained using the HuggingFace library with metaphor and condescension datasets since the language and content present are similar to our training dataset. This improves the accuracy of the contextual embeddings. We extracted around 7000 text instances for training RoBERTa from the talk-down corpus (Wang and Potts, 2019), MOH corpus (Mohammad et al., 2016), VUA dataset (Steen et al., 2010) and Trofi dataset (Birke and Sarkar, 2006). We call this trained model as tuned RoBERTa (tRoBERTa) for the entirety of this paper.

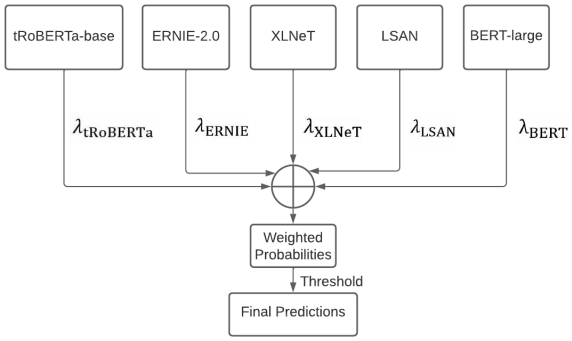

Figure 4: **Weighted-Average Ensemble**. $\lambda_{RoBERTa}$, λ_{ERNIE}, λ_{XLNeT}, λ_{LSAN}, and λ_{BERT} represent the weights of the respective models.

3.3 Label-Specific Attention Network

For sub-task B, in conjunction with transformer-based models, we used LSAN (Label-Specific Attention Network) (Xiao et al., 2019) for multi-label classification. LSAN tries to determine the label-related text from the given paragraph. It has two parts: self-attention and label-attention as shown in the figure 3. Self-attention aims to calculate the contribution of each word to a particular label. While it takes into account the context of the given text, the semantic meaning of labels themselves is captured by label-attention. The importance of these two mechanisms is determined by two trainable fully connected layers. Finally, an MLP layer with Sigmoid activation is used to get the final output.

3.4 Ensembles

Large Language models differ in the training procedures and the datasets on which they are trained. Hence, they may focus on different aspects of the input text even though they give comparable results. Therefore, it is a good practice to combine the results of these language models to get accurate word / sentence embeddings. There are several ways to combine them: we can concatenate embeddings of different models and project them to a low dimensional space for prediction, but this will require high computational power. Instead, we can tune different language models (tRoBERTa, ERNIE, XLNeT, LSAN and BERT) independently on the entire dataset and later combine their predictions as shown in the figure 4. Final results are obtained by taking a weighted average of the predictions. In this case, the weights are obtained by

Model	Task B							
	UPR	**SS**	**PS**	**AV**	**M**	**C**	**PM**	**Avg. F1**
Baseline	0.3535	0	0.1667	0	0	0.2087	0	0.1041
Ours	0.186	0.0875	0.0826	0.198	0.1324	0.2784	0.3571	0.1889
Ours*	0.1413	0.1706	0.0594	0.2086	0.2297	0.2874	0.238	0.1904

Table 2: Comparison of test results our models with baseline RoBERTa model for sub-task B. UPR: Unbalanced power relations, SS: Shallow solution, PS: Presupposition, AV: Authority voice, M: Metaphor, C: Compassion, PM: The poorer the merrier.

Model	Task A		
	Precision	**Recall**	**F1 score**
Baseline	0.3935	0.653	0.4911
Ours	0.3558	0.5836	0.4421
Ours*	0.5389	0.5678	0.5530

Table 3: Comparison of test results our models with baseline RoBERTa model for sub-task A.

* Modified System after submission and not submitted in the task.

grid search on the validation dataset. Another way to combine the predictions is the Voting Ensemble method, where the class predicted by the majority of the models is considered as the final output. We tested both approaches and found that the weighted average method yields better results than the voting ensemble method.

3.5 Classifier Chains

The classifier Chains approach connects binary classifiers in a chain such that the output of one classifier is treated as the input feature for the subsequent classifier. One of the factors that influence the performance of classifier chains is the sequence of labels used for training. The sequence of labels can be decided based on various approaches such as easiest-to-predict labels, most frequent labels first, etc. We tested the path 1–> 2 –> 6–> 7–> 4–> 5–> 3 based on the first approach. In the end we used an ensemble of the paths, 1–> 2–> 3–> 4–> 5–> 6–> 7 (P1) and 1–> 2–> 6–> 7–> 4–> 5–> 3 (P2).

4 Experimental setup

We used Pytorch (Paszke et al., 2019) and HuggingFace library (Wolf et al., 2019) for training and inference. All the models are trained on Google Colab. AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5 is used for training all transformer-based models and Adam is used for training all the other models. The maximum length of the text is limited to 200. We chose a batch size of 32 for training all the models.

4.1 Text preprocessing

We cleaned the text by removing punctuation, special characters, URLs, etc. We removed the contractions by mapping them to regular text and annotated the sentences with a [CLS] token before passing them into transformer-based models. We did not remove the stopwords as it deteriorated performance. We augmented the data based on contextual augmentation proposed in (Kobayashi, 2018) and back-translation technique.

4.2 Loss functions

We trained our model for sub-task A with binary cross-entropy loss, whereas for sub-task B we used focal loss. Focal loss (Lin et al., 2020) is a form of cross-entropy loss, but it is dynamically scaled. It is computed as:

$$FL\left(p_t\right) = (1-p_t)^{\gamma}\log\left(p_t\right)$$

By setting $\gamma > 0$, we are reducing the relative loss for easy, well-classified examples ($p_t > 0.5$). For prediction, we used a threshold of 0.35.

4.3 Training details

We split the original training data into train and development sets with 80% and 20% of the data respectively. For Sub-task A, we used *StratifiedShuffleSplit* option from *sklearn* library to split the dataset whereas for Sub-task B we used *iterative stratification* of order 1 from *skmultilearn* library. In the case of multi-task models, we used early stopping criteria on dev set independently for each task.

5 Results and Analysis

Our model for sub-task A consists of an ensemble of tuned RoBERTa (tRoBERTa), XLNet, and Ernie-2.0 models trained with binary cross-entropy

Model	Task A	Task B
Ernie-2.0 (ST)	0.535	0.498
tRoBERTa (ST)	0.537	0.503
XLNeT (ST)	0.539	0.500
LSAN	-	0.493
BERT-large	-	0.497
Avg. Ensemble (ST)	0.538	0.502
tRoBERTa (MT)	0.541	0.505
ULMFiT (ST)	0.510	0.487
tRoBERTa (CC with P1)	-	0.548
tRoBERTa (CC with P2)	-	0.550

Table 4: F1 score (Dev) of the other major models for both the sub-tasks. ST stands for single-task models, MT stands for multi-task models, CC stands for Classifier Chains with paths P1, P2 and Avg. Ensemble is the weighted average ensemble method.

Model	Task A	Task B
HAHNN (CNN)	0.530	0.493
C-BiLSTM	0.500	0.485
USE + BiLSTM	0.520	0.485
SVM (word n-grams)	0.486	0.469
SVM (ELMO)	0.492	0.475
SVM (char n-grams)	0.485	0.462
SVM (composite)	0.489	0.471
LR (ELMO)	0.485	0.465
RF (ELMO)	0.484	0.470

Table 5: Results of RNN-based and ML models.

loss. Model for sub-task B consists of a weighted average ensemble of tRoBERTa, XLNet, Ernie-2.0, BERT-large, and LSAN trained with focal binary cross-entropy explained in section 4.2. We also tested the Classifier Chain approach with tuned RoBERTa in post-evaluation phase. For this approach, we tested two sequences of labels P1, P2 (described in 3.5) and their ensemble. The official results of the models for sub-task A and sub-task B along with the baseline results on the test dataset are summarized in table 3 and 2 respectively. Apart from these models, the results of other major transformer-based models are shown in table 4.

Besides transformer-based models, we also tested Hierarchical Attentional Hybrid Neural Networks for Document Classification described in (Abreu et al., 2019), C-BiLSTM model, and a BiLSTM-attention model with USE (Cer et al., 2018) sentence embeddings. In the C-BiLSTM model, we apply convolutional and max-over-time pooling layers on the word embeddings and pass them through a bi-LSTM layer with an attention mechanism. This approach is the combination of work used in (Wang et al., 2016) and (Yang et al., 2016). We implemented traditional machine learning models like Support Vector Machine (SVM), Logistic Regression (LR), and Random Forests (RF). We used TF-IDF on words (both unigrams and bigrams), TF-IDF on character n-grams (1-5 characters), ELMO embeddings (Peters et al., 2018) and the composite features utilized in (Anzovino et al., 2018) as features for the ML models. The composite features are a combination of features based on the adjective count, length of the text, n-grams, POS tags, and doc2vec (Le and Mikolov, 2014). The results of the above RNN-based and ML models are summarized in table 5.

The comparison among three different problem transformation approaches to Multi-label classification is shown in table 6. The Label Powerset (LP) method gives the lowest F1 score as it doesn't perform well with unseen label combinations not covered in the training dataset. Classifier Chain performed better than Binary relevance as it takes label correlations into account. Incremental analysis of our system is shown in table 7. This analysis shows the importance of further training RoBERTa and focal loss.

Data augmentation is widely used to generate slightly variant larger datasets from the existing smaller ones. Since one recurring issue among all the models we trained is overfitting, three types of data augmentation techniques are tested to address this issue. We applied contextual augmentation for labeled sentences as proposed in (Kobayashi, 2018). In this method, we replace the words in a sentence with words predicted by a bi-directional language model. We also tested the back-translation approach proposed in (Sennrich et al., 2016) and easy data augmentation (EDA) techniques described in (Wei and Zou, 2019). We observed that while back-translation and contextual augmentation slightly improved the performance of the model, the use of EDA degraded it. We conclude this is because of a contextual mismatch in the text generated with EDA. Since random deletion is one of the techniques used in EDA, it may have led to the deletion of the keywords like *'them'*, *'us'*, *'poor'* etc., resulting in bad performance.

As the training dataset provided for this task is small and extremely imbalanced, we focused on analyzing the effects of data augmentation and choice

Approach	F1 Macro
Label Powerset	0.479
Classifier Chain	0.506
Binary Relevance	0.492

Table 6: Results of various problem transformation approaches on Dev set.

System	Precision	Recall	F1
RoBERTa	0.529	0.515	0.521
+ Tuned RoBERTa	0.539	0.521	0.530
+ Focal loss	0.550	0.525	0.537
+ Ensemble	0.557	0.530	0.543

Table 7: Sub-task A (Dev): Incremental analysis of our system.

of the loss function in this paper. F1 score for the *the poorer the merrier* class has greatly improved even though it is the least represented class in the entire dataset. Also, the F1 score of the *metaphor* class has improved. This improvement is the result of tuned RoBERTa which is trained on metaphor datasets. *Shallow solution* and *presupposition* are the classes with the lowest F1 score. This is because they require some form of world knowledge. Surprisingly, even though *unbalanced power relations* is the easiest class to predict, our model seemed to struggle with it.

6 Conclusion and Future Work

We have presented the results of various experiments using pre-trained language models such as RoBERTa, XLNeT, Ernie-2.0, BERT-large, and their ensembles for the detection of patronizing and condescending language. For the sake of comparison, we also presented experimental results of several RNN-based and traditional machine learning models with different features such as character n-grams, ELMO embeddings, etc. We also conducted experiments with different problem transformation approaches like Classifier Chains for multi-label classification. Since the dataset for this task is imbalanced, we also tested the effect of several data augmentation techniques and loss functions. We have achieved sizeable improvements over the baseline model by task-specific training and using techniques to mitigate class imbalance. In future work, we plan to explore other forms of Classifier Chains to more effectively model label dependence and hierarchy.

References

Jader Abreu, Luis Fred, David Macêdo, and Cleber Zanchettin. 2019. Hierarchical attentional hybrid neural networks for document classification. *Lecture Notes in Computer Science*, page 396–402.

Everton Alvares-Cherman, Jean Metz, and Maria Carolina Monard. 2012. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Syst. Appl.*, 39(2):1647–1655.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

Alberto Barrón-Cedeño, T. Elsayed, Reem Suwaileh, Lluís Màrquez i Villodre, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 2: Factuality. In *CLEF*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Weizhu Chen, Jun Yan, Benyu Zhang, Zheng Chen, and Qiang Yang. 2007. Document transformation for multi-label feature selection in text categorization. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 451–456.

Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASISamp;T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, USA. American Society for Information Science.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Krzysztof Dembczyński, Weiwei Cheng, and Eyke Hüllermeier. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 279–286, Madison, WI, USA. Omnipress.

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.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *ArXiv*, abs/1805.06201.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Jinfen Li and Lu Xiao. 2020. syrapropa at semeval-2020 task 11: Bert-based models design for propagandistic technique and span detection.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

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.

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

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tharindu Ranasinghe and Hansi Hettiarachchi. 2020. BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1906–1915, Barcelona (online). International Committee 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 *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Newton Spolaôr, Everton Alvares Cherman, Maria Carolina Monard, and Huei Diana Lee. 2013. A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science*, 292:135–151. Proceedings of the XXXVIII Latin American Conference in Informatics (CLEI).

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Bingyu Wang, Cheng Li, Virgil Pavlu, and Javed A. Aslam. 2017. Regularizing model complexity and label structure for multi-label text classification. *CoRR*, abs/1705.00740.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230, Berlin, Germany. Association for Computational Linguistics.

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.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.