BpHigh@TamilNLP-ACL2022: Effects of Data Augmentation on Indic-Transformer based classifier for Abusive Comments Detection in Tamil

Bhavish Pahwa
Bits Pilani Hyderabad Campus
bhavishpahwa@gmail.com

Abstract

Social Media platforms have grown their reach worldwide. As an effect of this growth, many vernacular social media platforms have also emerged, focusing more on the diverse languages in the specific regions. Tamil has also emerged as a popular language for use on social media platforms due to the increasing penetration of vernacular media like ShareChat and Moj, which focus more on local Indian languages than English and encourage their users to converse in Indic languages. Abusive language remains a significant challenge in the social media framework and more so when we consider languages like Tamil, which are low-resource languages and have poor performance on multilingual models and lack language-specific models. Based on this shared task, "Abusive Comment detection in Tamil@DravidianLangTech-ACL 2022", we present an exploration of different techniques used to tackle and increase the accuracy of our models using data augmentation in NLP. We also show the results of these techniques.

1 Introduction

The growth of social media platforms has been a significant factor in increasing awareness and connecting the world. Social media has changed the conventional way of communication and has introduced certain short forms and slang that are not present in the traditional vocabulary of any language.[1] At the same time, social media platforms have given rise to a new dynamic of cyber harassment utilizing the veil of anonymity that most platforms provide. Abusive language is a broad term often used to describe the posts and comments on social media platforms written to cyberbully, spread toxicity, spread hate, hurt others based on sex, caste, or creed(Pamungkas et al., 2021).

In the recent past, many social media platforms have updated their guidelines and added moderation policies to curb the spread of abusive language on them. Platforms like Facebook, YouTube, and Twitter have added features to report several posts/videos and comments. Many social media platforms also employ content moderators to clamp down on abusive language on their platforms. Still, this strategy is not sustainable for the long term as social media users continue to grow, and this approach cannot scale (Saha et al., 2021). Many content moderators feel the mental and psychological effects of viewing and moderating several extreme contents and are profoundly affected by such content.[2] Hence many platforms have started building automated abusive language detection and classification systems to improve their moderation capabilities.

In India, vernacular media faces more challenging problems dealing with more diverse languages and code-mixed data. For example, dealing with a vernacular language like Tamil is challenging as it is a low-resource language. Hence, it has insufficient datasets and code-mixed data, and data belonging to both Tamil script and transliterated data. Sometimes these challenges can lead to difficulty for the social media platforms to detect and remove the abusive language, leading to skirmishes with the government.[3]

Sharechat and Moj also organized a challenge recently to improve the abusive language detection systems in Indic languages and released their proprietary dataset for further research.[4] A shared task on "Offensive language detection in Dravidian Languages" was also introduced in the "First Workshop on Speech and Language Technologies for Dravidian Languages at EACL 2021". This shared task consisted of a large cor-

[1] https://www.languageservicesdirect.co.uk/social-media-changing-english-language/

[2] https://www.theverge.com/

[3] https://www.npr.org

[4] https://www.kaggle.com

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 138 - 144
May 26, 2022 ©2022 Association for Computational Linguistics

pus of comments/posts in code-mixed languages Tamil-English, Kannada-English, and Malayalam-English(Chakravarthi et al., 2020b,a, 2021; Hande et al., 2020). Further extending this shared task, the organizers of "The Second Workshop on Speech and Language Technologies for Dravidian Languages at ACL 2022" have released a shared task on "Abusive Comment Detection in Tamil" this task focuses specifically on abusive language detection in Tamil. It consists of datasets of both Tamil as well as code-mixed Tamil-English(Priyadharshini et al., 2022; Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021; Hande et al., 2021).

Our paper makes a two-fold contribution to the shared task. First, we experiment with the state-of-the-art transformer models pre-trained on Indian languages. Secondly, we show how data augmentation techniques in NLP perform in this task and how training with word-level augmented sentences affect our model accuracy. We also provide the trained model weights and the implementation code.[5]

2 Related Work

Steimel et al. (2019) investigated multilingual abusive comment detection focusing on English and German languages. They used a publicly available dataset of Twitter hate speech made by Waseem and Hovy (2016) for English and for German they used the 2018 GermEval shared task data set(Wiegand et al., 2018). They experimented with different text classification algorithms and found that there was no single algorithm which gave the best results on both the languages. They got best results on the German dataset by using SVM , 72.01 F-score and best results on English dataset were obtained using XGBoost, 80.49 F-score.

Pamungkas et al. (2021) wrote a summary paper on multilingual and multi domain abusive language detection, in the paper the authors highlighted various different techniques and datasets created and used by different researchers to properly define and solve the multilingual challenges related to abusive language detection. In the papers the author mentions several Transformer architecture based models like Multilingual BERT and XLM RoBERTa which are pre trained on corpus of several languages and can be finetuned for various tasks. They also mention various datasets in Indic languages like Hindi and code mixed hindi-english which are created specifically for the purpose of Hate speech/Abusive language detection, like HASOC, 2019.[6]

Khanuja et al. (2021) published a research paper along with a new transformer language model ‘MURIL’ based on the BERT architecture which is specifically designed for Indian languages. In the paper the authors also compared performance of both Multilingual BERT and MURIL on various tasks in Indian languages.The model currently has support for 17 Indian languages. In the papers the author shows that MURIL beats Multilingual BERT across all tasks and benchmarks in Indian languages. On the famous XTREME benchmark, Multilingual BERT gives an average performance of 59.1 whereas MURIL gives a 68.6 average performance.

Feng et al. (2021) published a survey paper on data augmentation approaches utilized in NLP. The paper’s authors discussed how data augmentation techniques could help fix the class imbalance. They also discussed various data augmentation methods like BACKTRANSLATION (Sennrich et al., 2016) and EASY DATA AUGMENTATION (EDA)(Wei and Zou, 2019).

Kobayashi (2018) presented a novel data augmentation technique called contextual augmentation. In this technique the authors used a bi-directional language model to replace words in the given input with other words according to the context. They further showed how this technique helps improve the accuracy of classifiers based on Recurrent Neural Networks(RNN) and Convolutional Neural Networks(CNN).

3 Dataset Description

The shared task on Abusive comment detection in Tamil-ACL 2022(Priyadharshini et al., 2022) is a comment classification problem that can be further described as a multi-class text classification problem in Tamil native script and Tamil-English code-mixed. The main objective is to build two separate systems that can classify comments, one for Tamil native script and another for code-mixed Tamil-English.

The purpose can also be redefined as developing a common system that can classify comments of both Tamil and Tamil-English languages. This paper treats the objective as building a common

[5] https://github.com/bp-high

[6] https://hasocfire.github.io

system for both languages rather than separate, to have a more standard approach.

The dataset was generated by scraping Youtube comments belonging to Tamil and Tamil-English code-mixed languages and annotated on comment level by linguists/annotators based on the platform's set guidelines and code of conduct. The dataset is split into two datasets based on the language. The labels used for annotation of the dataset are Misogyny, Misandry, Homophobia, Transphobia, Xenophobia, Counter Speech, Hope Speech, and None of the above. The dataset is further split into the training and development sets. The dataset consists of rows that contain the comment text and the label assigned to that comment.

4 Methodology

This section discusses the experiments and approaches undertaken to build a system for abusive comment detection. As explained earlier, we create a common system for both Tamil and Tamil-English languages, and hence for this purpose, we combine the dataset of Tamil and Tamil-English languages to make a combined dataset. Figure 1 shows a flowchart of the different approaches that we explored.

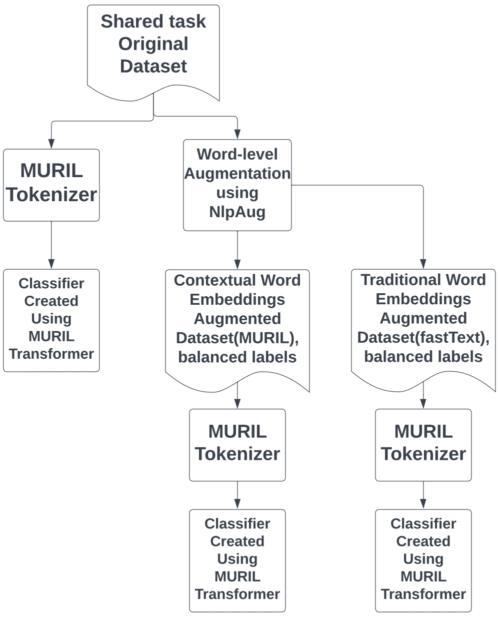
4.1 Transformer Model

Recently Transformer models have become quite widely used in NLP due to their property to capture context and the attention mechanism. In many downstream tasks in NLP, Transformers based models are state of the art, and due to organizations like Hugging Face, their implementation and Fine-tuning have become quite accessible.

We build a classifier using the **MURIL Transformer**(Khanuja et al., 2021) as our embedding layer(all layers frozen) and attach a classifier head by adding subsequent convolution and dense layers. The final output dense layer has softmax activation, which gives us the final predictions. The details of the model structure are present in Figure 2.

We used the MURIL Transformer as our embedding layer as it supports both Tamil and Tamil-English code-mixed as it was trained on both translated and transliterated document pairs.

Also, pre-processing of the comments is done using the MURIL tokenizer, also from Khanuja et al. (2021) we can see that MURIL produces lesser sub-words per word when compared to other multilingual models trained for Indian languages and has higher preservation of semantic meaning for Indian languages.



Shared task Original Dataset
MURIL Tokenizer
Word-level Augmentation using NlpAug
Classifier Created Using MURIL Transformer
Contextual Word Embeddings Augmented Dataset(MURIL), balanced labels
Traditional Word Embeddings Augmented Dataset(fastText), balanced labels
MURIL Tokenizer
MURIL Tokenizer
Classifier Created Using MURIL Transformer
Classifier Created Using MURIL Transformer

Figure 1: Flowchart of the different approaches

4.2 Data Augmentation in NLP

The labels in the initial dataset were unbalanced, with an overwhelming number of labels belonging to the "None of the above" class. We use data augmentation techniques in NLP to balance the dataset by performing word-level augmentation on the sentences belonging to the classes with lower representation in the dataset to reach a net balanced representation of all classes. We take the help of the NlpAug library[7](Ma, 2019), which provides the methods to perform word-level augmentation using contextual models as well as non-contextual word embeddings like Word2vec(Mikolov et al., 2013), fastText(Bojanowski et al., 2017), and Glove(Pennington et al., 2014).

$$M(i) = \lfloor (maximum_{\mathrm{j \in L}} (\mathrm{N_j})) / \mathrm{N_i} \rfloor$$

The above equation shows us the multiplier value M, used while generating the augmented sentences. M refers to the value by which the number of occurrences of a label should change, and N is the number of occurrences of a label, also called the value count of a label. L refers to the set of class labels. In terms of words, the above equation conveys that the multiplier value M(i) for label i is equal to the floor division of the value count for the

[7] https://github.com/makcedward/nlpaug

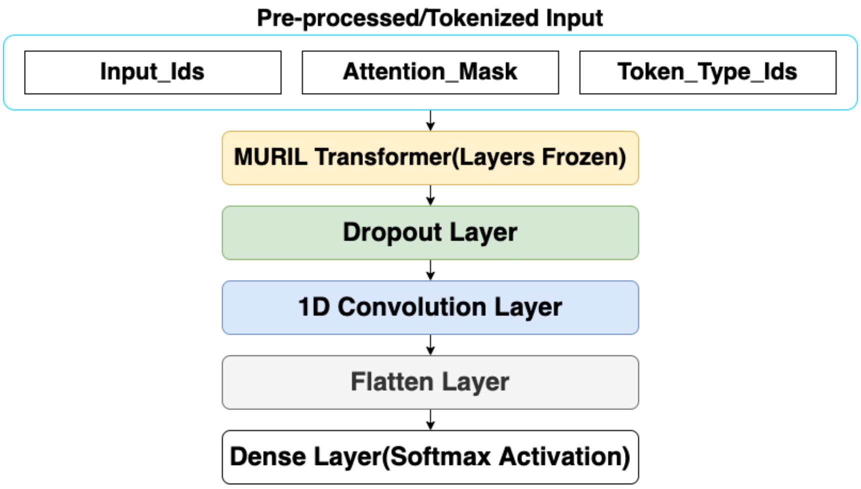

Pre-processed/Tokenized Input
Input_Ids
Attention_Mask
Token_Type_Ids
MURIL Transformer(Layers Frozen)
Dropout Layer
1D Convolution Layer
Flatten Layer
Dense Layer(Softmax Activation)

Figure 2: Structure of the classifier based on MURIL Transformer

Approach	Language	Accuracy	Macro-avg F1-score	Weighted-avg F1-score
No Augmentation	Tamil	0.64	0.17	0.56
No Augmentation	Tamil-English	0.67	0.19	0.59
Augmentation(MURIL)	Tamil	0.55	0.16	0.48
Augmentation(MURIL)	Tamil-English	0.59	0.13	0.50
Augmentation(Tamil fastText)	Tamil	0.49	**0.25**	0.52
Augmentation(Tamil fastText)	Tamil-English	0.52	**0.27**	0.56

Table 1: Results of all the approaches on test dataset

label having maximum count and the value count for label i. Using the mentioned equation we apply two word-level augmentation approaches on our train dataset. One using the contextual model and the other using the traditional non-contextual word embedding. Do note that no changes are made to the development/validation dataset.

4.2.1 Data Augmentation using Contextual Model

We use the **MURIL Transformer**(Khanuja et al., 2021) again as a "Contextual Word Embedding Augmenter" to generate word-level augmented sentences(Kumar et al., 2020). Then we train our classifier using this new balanced version of the train dataset.

4.2.2 Data Augmentation using Non-Contextual Word Embedding

We use the IndicNLP tokenizer for Indian languages[8] for pre-processing the input sentences and the **Tamil fastText model** from the IndicNLP suite(Kakwani et al., 2020) as a 'Word Embeddings Augmenter' to generate word-level augmented sentences. Then we train our classifier using this new balanced version of the train dataset.

MODEL PARAMETERS	VALUE
Fixed Parameters	
Batch Size	32
Optimizer	Adam
Learning Rate Schedule	Exponential Decay
Max Sequence Length	64
Tuned Parameters	
No Augmentation	
Num Epochs	50
Dropout	0.5
Learning Rate	0.01
Augmentation(MURIL)	
Num Epochs	60
Dropout	0.5
Learning Rate	0.001
Augmentation(Tamil fastText)	
Num Epochs	50
Dropout	0.3
Learning Rate	0.01

Table 2: Hyperparameters optimized for different approaches

[8] Indic NLP library

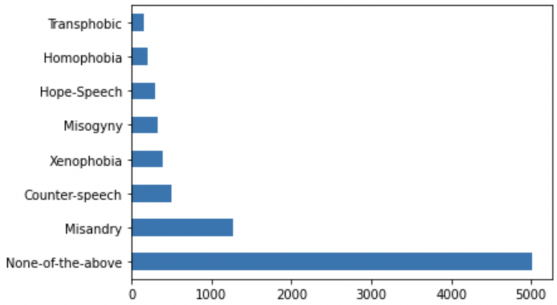


Figure 3: Bar graph of the number of occurrences of each label in the original train dataset

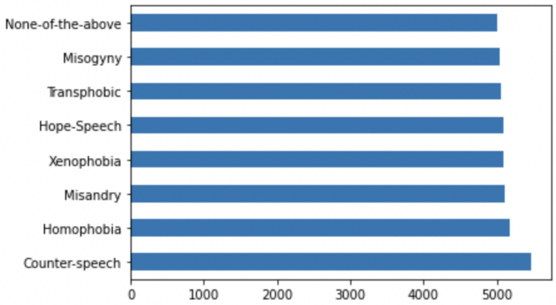


Figure 4: Bar graph of the number of occurrences of each label in the augmented train datasets generated

5 Results

We optimize the hyperparameters of the transformer-based classifier in all of our approaches on the training and development set and then get the predictions on the test set to observe the results of our approaches. The test dataset results on both languages are present in Table-1. In this task the approaches are evaluated with Macro-avg F1-score and the best performing approach for each language has been highlighted. For both the languages Tamil and Tamil-English, we observe that using the original dataset and training it with our transformer-based classifier yields better results than the data augmentation approach using **MURIL** and then training it with the transformer-classifier. However we observe that the results for the data augmentation approach using **Tamil fastText** produced better results for both the languages. See Table-2 for details in our training setup for the transformer-based classifier for all our approaches.

6 Conclusion

We explored the effects of data augmentation techniques on the Indic-Transformer based classifier created using MURIL Transformer on the task of Abusive Comment Detection in Tamil. We observe a negative result in the case of word-level augmentation using Contextual Models(**MURIL**) and an improvement in performance in the case of augmentation using Non-Contextual Word Embeddings(**Tamil fastText**).

As we further try to speculate why our augmentation technique based on Contextual Models failed to yield a better result, we consider the reasons stated in Longpre et al. (2020), which show that data augmentation techniques help improve performance on the task only when the approaches provide a language pattern that is not seen before during pretraining of the Transformer model. As both the Contextual Model for augmentation and the Indic-Transformer used to create the classifier is MURIL transformer, we cannot observe new linguistic patterns.

Also, in Kobayashi (2018), the authors observe that augmentation based on Contextual Models might not be able to remain compatible with the annotated labels of the original input and thus might harm the training process. They suggest using information from both label and context to generate word-level augmentations to control this incompatibility.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclu-*

sion, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language. *arXiv preprint arXiv:2108.04616*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online. Association for Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. Towards multidomain and multilingual abusive language detection: a survey.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

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.

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1151–1160, Varna, Bulgaria. INCOMA Ltd.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language*.