WLV-RIT at GermEval 2021: Multitask Learning with Transformers to Detect Toxic, Engaging, and Fact-Claiming Comments

Skye Morgan[1], Tharindu Ranasinghe[2], Marcos Zampieri[1]
[1]Rochester Institute of Technology, USA
[2]University of Wolverhampton, UK
sdm9815@rit.edu

Abstract

This paper addresses the identification of toxic, engaging, and fact-claiming comments on social media. We used the dataset made available by the organizers of the GermEval-2021 shared task containing over 3,000 manually annotated Facebook comments in German. Considering the relatedness of the three tasks, we approached the problem using large pre-trained transformer models and multitask learning. Our results indicate that multitask learning achieves performance superior to the more common single task learning approach in all three tasks. We submit our best systems to GermEval-2021 under the team name WLV-RIT.

1 Introduction

The popularity and accessibility associated with social media have greatly promoted user-generated content. At the same time, social media sites have increasingly become more prone to offensive content (Hada et al., 2021; Zhu and Bhat, 2021; Bucur et al., 2021). As such, identifying the toxic language in social media is a topic that has gained, and continues to gain traction. Research surrounding the problem of offensive content has centered around the application of computational models that can identify various forms of negative content such as hate speech (Malmasi and Zampieri, 2018; Nozza, 2021), abuse (Corazza et al., 2020), aggression (Kumar et al., 2018, 2020), and cyber-bullying (Rosa et al., 2019; Cheng et al., 2021; Salawu et al., 2021).

GermEval-2021 (Risch et al., 2021) focuses on identifying multiple types of comments in social media. This year's shared task is divided into three distinct classifications of comments: *i*) Toxic, *ii*) Engaging, and *iii*) Fact-Claiming. Like previous GermEval shared tasks (Struß et al., 2019), the detection of toxic content remains an integral part of GermEval-2021. Regarding engaging comments, there is an increasing desire from community managers as well as moderators to identify valuable user content (Kolhatkar and Taboada, 2017; Napoles et al., 2017). More particularly, rational comments that serve to encourage readers to engage in a discussion. In a similar light, identifying fact-claiming comments is equally important as platforms need to consistently review and verify user-generated content to uphold their responsibility as information distributors (Mihaylova et al., 2018; Shaar et al., 2020).

We pose that multitask learning (MTL) is a suitable approach for this year's GermEval as it enables what is learned from each task to aid in the learning of other tasks. The current state-of-the-art approach for offensive language identification is neural transformers modeled using single task learning (SLT) (Liu et al., 2019; Ranasinghe and Zampieri, 2020). It is well-known that training large neural transformer models often result in long processing times. As GermEval-2021 features three related tasks, from a performance standpoint, we pose that training a model jointly on three tasks is likely to be computationally more efficient than training three models in isolation. Moreover, as GermEval-2021 provides a single dataset for the three tasks, MTL can also be used to help improving performance across tasks. As such, we introduce multitask learning whereby one model can predict all three tasks as an alternative approach.

In this paper, we present the methods and results of the WLV-RIT submission to the GermEval-2021 shared task. We explore transformer architectures in two different environments, single task learning and multitask learning, and describe them in detail in Section 4. We perform several experiments using three transformer models that support German and evaluate their performance on the GermEval-2021 dataset.

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

2 Related Work

The identification of offensive language in online discussions is an extensive topic that has become popular over the past several years. The majority of the research related to this topic is centered on English data due to the availability of annotated datasets (Zampieri et al., 2019a; Rosenthal et al., 2021). Notwithstanding this, offensive language datasets are being annotated in other languages. Researchers have examined offensive content across multiple social media platforms and have both annotated and utilized data from different languages such as Greek (Pitenis et al., 2020), Marathi (Gaikwad et al., 2021), Italian (Chiril et al., 2019), Portuguese (Fortuna et al., 2019; Vargas et al., 2021), Arabic (Mubarak et al., 2021), Turkish (Çöltekin, 2020), and multiple languages of India (Ranasinghe and Zampieri, 2021a).

Past approaches to tackling the problem of offensive content on social media have relied on using a variety of computational models ranging from traditional machine learning classifiers such as Logistic Regression and SVMs (Malmasi and Zampieri, 2018), to various deep learning models (de Gibert et al., 2018). SemEval-2019 Task 5 (HatEval) (Basile et al., 2019) presented the challenge of detecting the presence of hate speech and identifying further features in hateful contents, which included two sub-tasks. For subtask A, which was the hate speech (HS) category, the best performance was achieved by training a support vector machine (SVM) model with a radial basis function (RBF) kernel. Several other high scoring teams used a convolutional neural network (CNN) which was traditionally the most popular approach to this topic (Hettiarachchi and Ranasinghe, 2019). For TRAC-1 (Kumar et al., 2018), the challenge was to develop a classifier that could discriminate between three levels of aggression in social media. The results showed that with careful consideration, classifiers like SVM and even random forest could perform at par with deep neural networks. However, in the end, more than half of the top 15 systems were trained on neural networks which demonstrates the approach's effectiveness.

The introduction of BERT (Devlin et al., 2019) spurred the use of pre-trained transformer models for classifying offensive speech (Ranasinghe and Zampieri, 2021b). As a result, neural transformer based language models have increasingly become more popular in offensive language identification. The use of pre-trained BERT models, as well as BERT-based models, was shown to be able to achieve competitive performance in popular competitions such as OffensEval (Zampieri et al., 2019b, 2020). Language-specific and multilingual models have also been introduced to assist NLP research in various languages such as GBERT for German (Chan et al., 2020), AraBERT for Arabic (Antoun et al., 2020), and the multilingual XLM-R (Conneau et al., 2019) that has been been applied to offensive language identification (Ranasinghe and Zampieri, 2020, 2021c).

3 Data

In the GermEval-2021 dataset, the focus has been extended beyond the identification of offensive comments to include two additional classes: engaging comments that can motivate readers to participate in conversations, and fact-claiming comments. The dataset for this iteration of GermEval comprises over 3,000 Facebook user comments that have been extracted from the page of a political talk show of a German television broadcaster. The training dataset has a total of 3,244 instances and comprises 1,074 instances without any toxic, engaging or fact claiming content. In Table 1, we present four different Facebook user comments along with their annotation.

Toxic	Engaging	Fact-Claiming	Training
0	0	0	1074
1	0	0	739
0	1	0	239
1	1	0	89
0	1	1	403
1	0	1	160
0	0	1	406
1	1	1	134
All			3244

Table 2: GermEval 2021 - Training Set User Comment Distribution

4 Methods

Considering the success that neural transformers have demonstrated across various natural language processing tasks (Uyangodage et al., 2021; Jauhiainen et al., 2021; Hettiarachchi and Ranasinghe, 2020a) including offensive language identification (Ranasinghe and Zampieri, 2020, 2021b; Dai et al., 2020) we used transformers to tackle this task too.

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

Comment	Sub1	Sub2	Sub3
"Die AfD sind genau so neoliberal und kapitalistische Zerstörer unserer Heimat, wie die CDU, CSU, FDP, SPD und Grüne auch."	1	0	0
"Sarazin ist ein rechtsradikaler Mensch. Ein Menschenhasser. Sie kennen nur Zerstörung. Die Geschichte hat es gezeigt."	1	0	1
"@USER, du hast das Thema im Kern nicht verstanden"	0	0	1
"Ich frage dich, verlassen Menschen gerne ihre Heimat?"	0	0	0

Table 1: Annotation examples of four different Facebook user comments. Sub1 represents toxic comments, Sub2 stands for engaging comments, and Sub3 stands for fact claiming.

Parameter	Value
learning rate‡	$1e^{-5}$
number of epochs‡	3
adam epsilon	$1e^{-8}$
warmup ratio	0.1
warmup steps	0
max grad norm	1.0
max seq. length	120
gradient accumulation steps	1

Table 3: Hyperparameter specifications. The optimised hyperparameters are marked with ‡ and their optimal values are reported. The rest of the hyperparameter values are kept as constants.

We explored transformer architectures in two different environments; single task learning and multi task learning.

Single Task Learning (STL) For the STL environment we trained three classification models based on transformers. By utilizing the hidden representation of the classification token (CLS) in the transformer model, we predict the target labels (toxic/non-toxic, engaging/non-engaging, fact-claiming, non-fact-claiming) by applying a linear transformation followed by the softmax activation (σ):

$$\hat{\mathbf{y}}_{task} = \sigma(\mathbf{W}_{[CLS]} \cdot \mathbf{h}_{[CLS]} + \mathbf{b}_{[CLS]}) \quad (1)$$

where $\cdot$ denotes matrix multiplication, $\mathbf{W}_{[CLS]} \in \mathcal{R}^{D\times 3}$, $\mathbf{b}_{[CLS]} \in \mathcal{R}^{1\times 2}$, and D is the dimension of the input activation layer $\mathbf{h}$. $\hat{\mathbf{y}}_{task}$ is the predicted value of any of the three tasks.

We construct three separate classification models minimising the cross-entropy loss for each of the three tasks as defined in the Equation 2, where y_{toxic}, y_{engage} and y_{fact} represent ground truth labels of each task. These particular losses are:

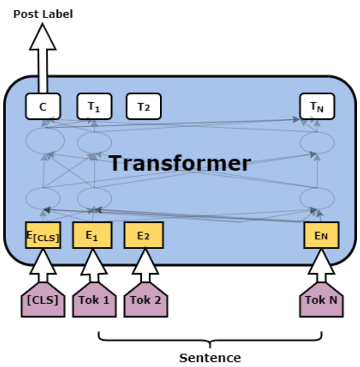
$$\mathcal{L}_{toxic} = -\sum_{i=1}^{2}\Big(\mathbf{y}_{toxic} \otimes \log(\hat{\mathbf{y}}_{toxic})\Big)[i]$$
$$\mathcal{L}_{engage} = -\sum_{i=1}^{2}\Big(\mathbf{y}_{engage} \otimes \log(\hat{\mathbf{y}}_{engage})\Big)[i]$$
$$\mathcal{L}_{fact} = -\sum_{i=1}^{2}\Big(\mathbf{y}_{fact} \otimes \log(\hat{\mathbf{y}}_{fact})\Big)[i] \quad (2)$$

where $\mathbf{v}[i]$ retrieves the ith item in a vector $\mathbf{v}$ and $\otimes$ indicates element-wise multiplication. The corresponding STL architecture is shown in Figure 1a.

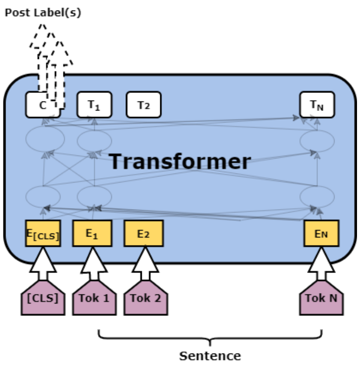
Multi Task Learning (MTL) MTL was introduced as an approach to inductive transfer (Caruana, 1997). The main goal of which was to improve generalization performance on a current task after having learned a different but related concept on a previous task. MTL is quite efficient as one model can be utilized to predict multiple tasks so long as they are related. In hate speech and offensive language detection, MTL has been shown to outperform single-task environments as well as learn task efficiently with the presence of little labelled data per-task (Djandji et al., 2020). Despite this, MTL has not been used much in the context of offensive language detection. As such, we decided to use multitask learning to compare the performance within the two different environments using different transformer models. We used the transformer as the base model for our MTL approach. Our approach will learn the three tasks jointly, i.e., Toxic comment detection, Engaging comment detection and Fact-claiming comment detection. The implemented architecture shares the hidden layers between the tasks. The shared portion includes a transformer model that learns shared information across the tasks by minimizing a combined loss.

		Toxic			Engaging			Fact-Claiming		
Model	**Environment**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	STL	0.4897	0.4421	0.4500	0.5421	0.5310	0.5380	0.5532	0.5093	0.5511
	LM + STL	0.4921	0.4432	0.4512	0.5436	0.5314	0.5398	0.5669	0.5101	0.5521
	MTL	0.5042	0.4449	0.4551	0.5472	0.5325	0.5401	0.5702	0.5113	0.5532
	LM + MTL	0.5063	0.4543	0.4665	0.5542	0.5341	0.5442	0.5732	0.5231	0.5555
gBERT	STL	0.6449	0.5801	0.6102	0.6449	0.6312	0.6342	0.6812	0.6752	0.6852
	LM + STL	0.6552	0.5841	0.6173	0.6254	0.6442	0.6354	0.6821	0.6779	0.6872
	MTL	0.7001	0.6321	0.6654	0.6777	0.6931	0.6841	0.7311	0.7211	0.7352
	LM + MTL‡	0.7124	0.6456	0.6796	0.6827	0.7027	0.6926	0.7450	0.7495	**0.7472**
gELECTRA	STL	0.6551	0.5991	0.6227	0.6391	0.6482	0.6431	0.6954	0.7002	0.7045
	LM + STL	0.6651	0.6078	0.6321	0.6422	0.6561	0.6555	0.7021	0.7102	0.7100
	MTL‡	0.7256	0.6603	0.6914	0.6895	0.6999	**0.6947**	0.7530	0.7407	0.7468
	LM + MTL‡	0.7542	0.6732	**0.7112**	0.6944	0.6924	0.6934	0.7354	0.7383	0.7369

Table 4: Results for the evaluation set in each task with Transformer models. For each model, Precision (P), Recall (R), and F1 are reported on all tasks. The best result for each task has been marked with bold considering F1. The experiments we submitted are marked with ‡



(a) STL Architecture



(b) MTL Architecture

Figure 1: The STL (top) and MTL (bottom) transformer-based architectures experimented with the GermEval-2021 dataset.

We assign equal importance to each task in our experiments. The full loss is:

$$\mathcal{L}_{multi} = \frac{\mathcal{L}_{toxic} + \mathcal{L}_{engage} + \mathcal{L}_{fact}}{3}. \qquad (3)$$

The task-specific classifiers receive input from the last hidden layer of the transformer language model and predict the output for the tasks. The corresponding MTL architecture is shown in Figure 1b

5 Experimental Setup

We performed experiments using three transformer models that support German; mBERT (Devlin et al., 2019), German BERT-large (gBERT) (Chan et al., 2020) and German Electra-large (gELECTRA) (Chan et al., 2020) transformer models available in the HuggingFace model repository (Wolf et al., 2020).

We used an Nvidia Tesla K80 GPU to train the models. We divided the input dataset into a training set and a validation set using 0.8:0.2 split. We predominantly fine-tuned the learning rate and the number of epochs of the classification model manually to obtain the best results for the validation set. We obtained $1e^{-}5$ as the best value for the learning rate and 3 as the best value for the number of epochs. We used a batch size of 8 for the training process and the model was evaluated after every 100 batches. We performed *early stopping* if the validation loss did not improve over 10 evaluation steps. The rest of the hyperparameters which we kept as constants are mentioned in the Table 3. For both STL and MTL we finetuned the considered transformer model on the GermEval 2021 training set using Masked Language Modeling (MLM) (Devlin et al., 2019) objective which we call as

		Toxic			Engaging			Fact-Claiming		
Model	**Environment**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	STL	0.5081	0.4672	0.4781	0.5689	0.5561	0.5555	0.5763	0.5286	0.5761
	LM + STL	0.5162	0.4657	0.4782	0.5698	0.5561	0.5568	0.5871	0.5389	0.5780
	MTL	0.5284	0.4672	0.4781	0.5690	0.5571	0.5678	0.5901	0.5364	0.5782
	LM + MTL	0.5243	0.4763	0.4871	0.5762	0.5590	0.5601	0.5983	0.5482	0.5782
gBERT	STL	0.6692	0.6092	0.6354	0.6678	0.6572	0.6532	0.7095	0.6982	0.7011
	LM + STL	0.6752	0.6072	0.6342	0.6453	0.6683	0.6572	0.7063	0.6982	0.7041
	MTL	0.7223	0.6532	0.6842	0.6954	0.7132	0.7041	0.7553	0.7493	0.7562
	LM + MTL‡	0.7321	0.6654	0.6941	0.7041	0.7298	0.7145	0.7653	0.7602	0.7652
gELECTRA	STL	0.6752	0.6111	0.6498	0.6531	0.6679	0.6609	0.7178	0.7285	0.7265
	LM + STL	0.6874	0.6231	0.6562	0.6666	0.6742	0.6731	0.7231	0.7303	0.7367
	MTL‡	0.7456	0.6802	0.7132	0.7001	0.7101	**0.7198**	0.7754	0.7652	**0.7653**
	LM + MTL‡	0.7853	0.6997	**0.7342**	0.7132	0.7156	0.7190	0.7542	0.7563	0.7590

Table 5: Results for the test set in each task with Transformer models. For each model, Precision (P), Recall (R), and F1 are reported on all tasks. The best result for each task has been marked with bold considering F1. The experiments we submitted are marked with ‡

Language Modeling (LM). When performing training, we trained five models with different random seeds and considered the majority-class self ensemble mentioned in Hettiarachchi and Ranasinghe (2020b) to get the final predictions.

6 Results

We show the results for the evaluation set in Table 4. In all the experimented transformer models, the MTL approach outperformed the STL approach. Furthermore in most scenarios, the systems that included a LM component outperformed those without the LM component. This corroborates the findings of previous research in offensive language identification (Ranasinghe et al., 2019). gBERT and gELECTRA models clearly outperformed mBERT in all the tasks. For the Task 1, gELECTRA model with LM and MTL achieved the best result with 0.7342 F1 score, for the Task 2 gELECTRA model with MTL, without LM achieved the best result with 0.7198 F1 score and for the Task 3 too, the same model achieved the best result with 0.7653 F1 score. Considering the overall performance we selected three best models for the submission; gELECTRA with LM+MTL, gELECTRA with MTL and gBERT with LM+MTL.

The official leaderboard of the competition was not yet released at the time of writing this paper, therefore, after the organizers released the gold labels for the test set, we calculated the Precision, Recall, and F1 values for the test set. The results are shown in Table 5. As shown in the results, the three models we selected provided the top three results for the test set too. MTL consistently outperformed STL in all the tasks with all the transformer models we experimented.

7 Conclusion and Future Work

In this paper, we presented the WLV-RIT entry to GermEval-2021. GermEval-2021 provided participants with the opportunity of testing computational models to identify toxic, engaging, and fact claiming comments. We experimented with neural transformer models in STL environment and MTL environment. MTL environment consistently outperformed STL suggesting that the use of shared learning methods improves the performance of individual tasks. Furthermore, we observed that pre-trained language-specific transformer models trained for German such as gBERT and gElectra outperform mBERT. Finally, in addition to the transformer-based MTL approach, we could observe that the use of language modelling led performance improvement in some of the tasks.

In the future, we would like to carry out an error analysis on the output of our systems to better understand the impact and limitations of MTL for these three tasks. Finally, we would like to experiment with multi-task learning in other languages, particularly low-resource languages for which only limited language resources are available.

Acknowledgments

The authors would like to thank the GermEval-2021 organizers for organizing this interesting shared task and for making the dataset available.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of LREC*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the ACL*.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28:41–75.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of COLING*.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of ACL*.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Proceedings of TALN*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. Hybrid emoji-based masked language models for zero-shot abusive language detection. In *Findings of the ACL*.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In *Proceedings of SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. 2020. Multi-task learning using AraBert for offensive language detection. In *Proceedings of OSCAT*.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A Hierarchically-labeled Portuguese Hate Speech Dataset. In *Proceedings of ALW*.

Saurabh Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of RANLP*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of ALW*.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of ACL*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2019. Emoji Powered Capsule Network to Detect Type and Target of Offensive Posts in Social Media. In *Proceedings of RANLP*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2020a. BRUMS at SemEval-2020 task 3: Contextualised embeddings for predicting the (graded) effect of context in word similarity. In *Proceedings of SemEval*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2020b. InfoMiner at WNUT-2020 task 2: Transformer-based covid-19 informative tweet extraction. In *Proceedings of W-NUT*.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing approaches to Dravidian language identification. In *Proceedings of VarDial*.

Varada Kolhatkar and Maite Taboada. 2017. Using New York Times Picks to Identify Constructive Comments. In *Proceedings of NLPJ*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of SemEval*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact Checking in Community Forums. In *Proceedings of AAAI*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic Offensive Language on Twitter: Analysis and Experiments. In *Proceedings of WANLP*.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding Good Conversations Online: The Yahoo news annotated comments corpus. In *Proceedings of LAW*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of ACL*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of LREC*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021a. An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India. *Information*, 12(8).

Tharindu Ranasinghe and Marcos Zampieri. 2021b. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Tharindu Ranasinghe and Marcos Zampieri. 2021c. Multilingual Offensive Language Identification for Low-resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In *Proceedings of FIRE*.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of GermEval*.

Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *Findings of the ACL*.

Semiu Salawu, Jo Lumsden, and Yulan He. 2021. A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *Proceedings of WOAH*.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *Proceedings of CLEF*.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings GermEval*.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Can Multilingual Transformers Fight the COVID-19 Infodemic? In *Proceedings of RANLP*.

Francielle Alves Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. 2021. Contextual lexicon-based approach for hate speech and offensive language detection. *arXiv preprint arXiv:2104.12265*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the ACL*.