

# UPB at SemEval-2021 Task 5: Virtual Adversarial Training for Toxic Spans Detection

Andrei Paraschiv, Dumitru-Clementin Cercel, Mihai Dascalu

University Politehnica of Bucharest, Faculty of Automatic Control and Computers

andrei.paraschiv74@stud.acs.upb.ro  
{dumitru.cercel, mihai.dascalu}@upb.ro

## Abstract

The real-world impact of polarization and toxicity in the online sphere marked the end of 2020 and the beginning of this year in a negative way. Semeval-2021, Task 5 - Toxic Spans Detection is based on a novel annotation of a subset of the Jigsaw Unintended Bias dataset and is the first language toxicity detection task dedicated to identifying the toxicity-level spans. For this task, participants had to automatically detect character spans in short comments that render the message as toxic. Our model considers applying Virtual Adversarial Training in a semi-supervised setting during the fine-tuning process of several Transformer-based models (i.e., BERT and RoBERTa), in combination with Conditional Random Fields. Our approach leads to performance improvements and more robust models, enabling us to achieve an F1-score of 65.73% in the official submission and an F1-score of 66.13% after further tuning during post-evaluation.

## 1 Introduction

Nowadays, online engagement in social activities is at its highest levels. The lockdowns during the 2020 COVID-19 pandemic increased the overall time spent online. In Germany for instance, Lemenager et al. (2021) observed that 71% of considered subjects increased their online media consumption during this period. Unfortunately, online toxicity is present in a large part of the social and news media platforms. As such, automated early detection is necessary since toxic behavior is often contagious and leads to a spillover effect (Kwon and Gruz, 2017).

Recently, a significant effort was put into the detection of toxic and offensive language (van Aken et al., 2018; Paraschiv and Cercel, 2019; Tanase et al., 2020b,a), but the challenging nature of these problems leaves several avenues unexplored. In addition, most shared tasks focus

on the distinction between toxic/non-toxic (Wulczyn et al., 2017; van Aken et al., 2018; Juuti et al., 2020) or offensive/non-offensive posts in various languages (Struß et al., 2019; Zampieri et al., 2019a,b, 2020; Mandl et al., 2020; Aragón et al., 2020). The Semeval-2021 Task 5, namely Toxic Spans Detection (Pavlopoulos et al., 2021), tackles the problem of identifying the exact portion of the document that gives it toxicity. The provided dataset is a subset of the Jigsaw Unintended Bias in Toxicity Classification dataset<sup>1</sup>, with annotated spans that represent toxicity from a document.

In this paper, we describe our participation in the aforementioned Toxic Spans Detection task using several Transformer-based models (Vaswani et al., 2017), including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), with a Conditional Random Field (CRF) (Lafferty et al., 2001) layer on top to identify spans that include toxic language. We introduce Virtual Adversarial Training (VAT) (Miyato et al., 2015) in our training pipeline to increase the robustness of our models. Furthermore, we enhance part of our models with character embeddings based on the Jigsaw Unintended Bias dataset to improve their performance. Finally, we compare the proposed models and analyze the impact of various hyperparameters on their performance.

The rest of the paper is structured as follows. The next section introduces a review of methods related to toxic language detection, sequence labeling, and adversarial training (Kurakin et al., 2016). The third section discusses the employed models, as well as the VAT procedure. Results are presented in the fourth section, followed by discussions, conclusions, and an outline of possible future works.

<sup>1</sup><https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>

## 2 Related Work

**Toxic Language Detection.** There are several research efforts to detect toxic texts based on the Jigsaw Unintended Bias dataset, out of which most focus on the Kaggle competition task - predicting the toxicity score for a document. [Morzhov \(2020\)](#) compared models based on Convolutional Neural Networks (CNNs) ([Kim, 2014](#)) and Recurrent Neural Networks ([Cho et al., 2014](#)) with a Bidirectional Encoder Representations from Transformers (BERT) architecture ([Devlin et al., 2019](#)), obtaining the best performance from an ensemble of all used models. [Gencoglu \(2020\)](#) and [Richard and Marc-André \(2020\)](#) used the same dataset to improve on the automatic detection of cyberbullying content.

**Sequence Labeling.** Predicting the type for each token from a document rather than providing a label for the whole sequence is a task often associated with named entity recognition ([Ma and Hovy, 2016](#)), but can be performed in other Natural Language Processing pipelines, including part-of-speech tagging ([Ling et al., 2015](#)) and chunking ([Hashimoto et al., 2017](#)). A common practice in sequence tagging models ([Peters et al., 2018](#); [Avram et al., 2020](#); [Ionescu et al., 2020](#)) is to use a CRF as a final decoding layer.

**Adversarial Training.** Researched first in image classification ([Szegedy et al., 2013](#)), adversarial examples are small input perturbations that are hardly distinguishable for humans, but can dramatically shift the output of a neural network. These examples can be used in adversarial training (AT) ([Goodfellow et al., 2014](#)) as a regularization method that can increase the robustness of the model. Using the worst-case outcome from a distribution of small norm perturbations around an existing training sample, a new data point is created and inserted into the training process.

Extending AT to a semi-supervised setting, VAT ([Miyato et al., 2016](#)) does not require label information for the adversarial examples. VAT aims to increase the local distributional smoothness by adding perturbations to the embedding output. Recently, several studies ([Kumar and Singh, 2020](#); [Liu et al., 2020](#); [Si et al., 2020](#)) focused on applying VAT in Transformer-based models and obtained improvements in comparison to baseline methods on several classification tasks.

## 3 Method

### 3.1 Corpus

The dataset for the competition is a subset of the Jigsaw Unintended Bias in Toxicity Classification English language corpus, with annotated spans that make the utterance toxic. From the 8,597 trial and train records, 8,101 had at least one toxic span. By cross-referencing with the original Jigsaw dataset which contains additional information, we retrieved the toxicity scores for each text and determined that the mean toxicity score for the train and test set were very close (0.8429 versus 0.8440; see [Figure 1](#) for corresponding kernel density estimates). Moreover, only 17 out of 2,000 test data rows had a toxicity score below 0.75. Nevertheless, an off-balance was noticed between the test and train set - 80.3% entries from the test set had at least one toxic span versus a considerably higher density of 94.2% in the train set.

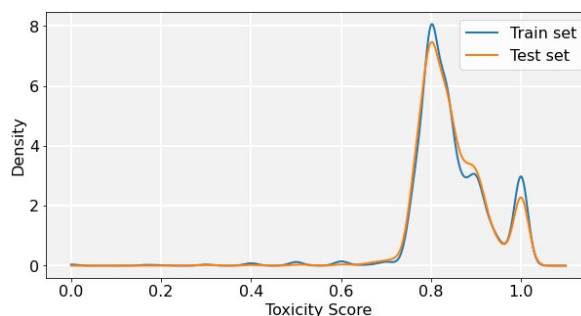


Figure 1: Kernel density estimate using Gaussian kernels for the toxicity scores in both the training and test data.

The training dataset was split into sentences while ensuring that there are no splits inside a toxic span and there are no sentences shorter than three words. Under these settings, our training dataset consists of a total of 26,589 sentences, including 10,117 records that contained toxic spans; 15% were selected for validation. Another 2,000 entries were provided by the competition organizers for testing; the labels for this dataset were made available after the competition.

For our unsupervised training samples, we selected 20,000 random records from the Jigsaw dataset, making sure there was no overlap with the Semeval-2021 training data. Additionally, we replaced all URL-s with a special token and applied lower case on all records.

Model	F1-score validation set	F1-score competition test set
LSTM-CRF-VAT	75.82%	62.49%
LSTM-CRF-VAT+chars	76.27%	63.65%
BERT-base-CRF	79.25%	62.32%
BERT-base-CRF-VAT	80.66%	64.59%
BERT-toxic-CRF-VAT*	<b>81.08%</b>	65.73%
BERT-news-CRF-VAT	80.80%	64.57%
BERT-news-CRF-VAT ( $\gamma=0.6$ )	81.01%	<b>66.13%</b>
BERT-news-CRF-VAT+chars	80.79%	64.57%
RoBERTa-large-CRF-VAT	78.13%	62.73%

Table 1: F1-scores for predictions on the validation and test set.

\* marks the model from the official submission.

### 3.2 Virtual Adversarial Training

The robustness of the model in Adversarial Training is improved through examples that are close to available training data, but the model would be likely to assign a different label than the training one, thus leading to loss increase. In VAT, Miyato et al. (2018) adapted the adversarial training from supervised to semi-supervised settings by adding an additional loss using the Kullback–Leibler divergence between the predictions of the original data and the same data with random perturbations. Since the output distributions are compared, the information about labels is not needed for the adversarial loss:

$$L_{adv} = KL(P(\hat{y}|e, \Theta) || P(\hat{y}|e + d, \Theta)) \quad (1)$$

where  $e$  is the embedding associated with the sample,  $d$  the perturbation, and  $\hat{y}$  is the predicted output.

True labels are required in general to compare the losses and find the worst case perturbations. However, this can be avoided by bounding the norm of the perturbation  $\delta$  to  $\eta$ ; thus, the value of the perturbation becomes:

$$d = \arg \max_{\delta; \|\delta\|_2 < \eta} KL(P(\hat{y}|e, \Theta) || P(\hat{y}|e + \delta, \Theta)) \quad (2)$$

Afterwards, we can estimate the perturbation  $d$  using also the gradient  $g$  and a hyperparameter  $\epsilon$  for the magnitude by applying the second-order Taylor approximation and a single iteration of the power method:

$$d = \frac{g}{\|g\|_2} \epsilon \quad (3)$$

where

$$g = \nabla_{\delta} KL(P(\hat{y}|e, \Theta) || P(\hat{y}|e + \delta, \Theta)) \quad (4)$$

In order to reduce the complexity and computation for the gradient, we ignore the dependency on  $\Theta$ . Also, the number of power iterations can be another hyperparameter for the model. The final loss function used by all models is a combination of the supervised and unsupervised adversarial loss:

$$L_{total} = \gamma L_{sup} + (1 - \gamma) L_{adv} \quad (5)$$

where  $\gamma$  is another tunable hyperparameter.

### 3.3 Implementation Details

In our experiments, pre-trained Transformer models are followed by a linear transformation of their last hidden state, and a final CRF layer. More precisely, we compare the effectiveness of several flavors of BERT models, alongside the VAT technique as follows: BERT base, a 768-dimensional model provided by Google (*BERT-base-CRF-VAT*), Unitary’s toxic BERT (Hanu and Unitary team, 2020) (*BERT-toxic-VAT*), BERT pre-trained on fake and hyperpartisan news (Paraschiv et al., 2020) (*BERT-news-CRF-VAT*), and *RoBERTa-large-CRF-VAT*, the equivalent of BERT-base-CRF-VAT that relies on RoBERTa instead of BERT.

In addition to these models, we experimented with enhancing the BERT-based representation with character embeddings (Kim et al., 2016). These character representations were trained on the entire Jigsaw dataset using a CNN-BiLSTM model (Ma and Hovy, 2016) with the next character prediction objective. We concatenated the

obtained character-level embeddings with the aforementioned Transformer’s last hidden state, and refer to this variant as *BERT-news-CRF-VAT+chars*.

As baseline systems, we design two methods: *LSTM-CRF-VAT* with GloVe embeddings (Pennington et al., 2014) and a *LSTM-CRF-VAT+chars* having character-level embeddings and VAT. In all BERT-based models, we used a maximum sequence length of 96 tokens and a sequence of 64 tokens for the LSTM baseline. Since the input words can consist of more than one token, we assign the toxicity label to a word if at least one component token is inferred as toxic.

The best hyperparameters for the BERT-base model were determined through grid search on the development set. The identified optimal values ( $\epsilon = 2$ ,  $\eta = 0.1$ , and two power iterations) were used in all other flavors;  $\gamma$  was set to 0.5 in the final loss function to balance both approaches. Furthermore, all BERT-based models were trained for one epoch in contrast with the LSTM-CRF-VAT and LSTM-CRF-VAT baselines that were trained for three epochs and four epochs, respectively.

## 4 Results

The evaluation metric for the Toxic Spans Detection task was an adapted version of the F1-score (Da San Martino et al., 2019) that takes into account the size of the overlap between prediction spans and golden labels.

Results for all developed models with the aforementioned hyperparameters (i.e.,  $\gamma = 0.5$ ,  $\epsilon = 2$ ,  $\eta = 0.1$ , and two power iterations) are presented in Table 1. Since the training data had a slightly different distribution of the span density, part of our models that performed worse on our dev set performed better on the competition test set. Adding the character embedding representation to BERT-based models did not prove to be of use in our pre-evaluation tests, but in post-evaluation, we noticed that slightly tweaking the  $\gamma$  hyperparameter for the loss from 0.5 to 0.6 brought the F1-score to 66.13%. Despite performance on the validation set was insensitive to the change in  $\gamma$  between 0.5 and 0.6, the results on the test set were more than 1.5% apart. This is mostly due to the unsupervised training that is strengthening the model’s confidence on edge cases which would lower its precision.

Figure 2 introduces the influence of the perturbation magnitude  $\epsilon$  on the overall performance of three models. The impact of  $\epsilon$  in the adversarial

training effectiveness is significant, but it is also highly dependant on the used model and can only be determined experimentally.

Our models performed well on the detection task, learning not only common toxic expressions like "moron", "stupid", "pathetic troll", "disgusting", "hang-em high", but also obfuscated expressions like "f\*cking nasty" and "b\*tchy". Nonetheless, the models fail to detect more obscured words like "you don't know s\*\*\*" or "Kill this F'n W\*ore on site". All models have the tendency to over-predict toxicity by adding words to the toxic expression - for example, "What a pile of shit" was automatically labeled as "What a pile of shit".

The character-level embeddings boosted the performance of the baseline LSTM-CRF-VAT model but did not improve any BERT model since it leads to detecting longer spans as toxic (see Table 2) which in return lowers precision.

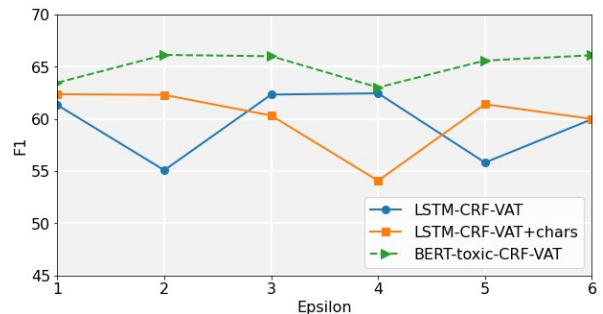


Figure 2: Impact of the perturbation magnitude  $\epsilon$  on the F1-scores for the predictions on the official test set.

## 5 Discussions and Error Analysis

In this section, we analyze the BERT-toxic-CRF-VAT performance versus the golden label values from the competition test set. The precision and recall for our best model are 65.74% and 85.54%, respectively, which are indicative of a tendency to over-predict toxic spans. As we mentioned in section 3.1, even though almost all documents in the test set had a high toxicity score in the original Jigsaw dataset, many had no annotated toxic spans. Indeed, there were 295 records where our model detected a toxic span and none were labeled in the test set. Table 3 includes examples of such detected spans those labeling is subjective and other detection errors. Words like "stupid", "dumb", and "crap" were assigned as toxic throughout the test data due to their high presence in the training data spans. There were also milder errors, spans that overlap with the golden labels, but the model omits

Annotated sample	BERT-news-CRF-VAT+chars	BERT-toxic-CRF-VAT
Greedy pig strikes again!	Greedy pig strikes again!	Greedy pig strikes again!
What's honorable is telling idiots that they're idiots (even if tardily)	What's honorable is telling idiots that they're idiots (even if tardily)	What's honorable is telling idiots that they're idiots (even if tardily)
A fool and our money are soon to be parted.	A fool and our money are soon to be parted.	A fool and our money are soon to be parted.

Table 2: Examples of model predictions with character embeddings (BERT-news-CRF-VAT+chars) and without (BERT-toxic-CRF-VAT), compared to the official annotations. The toxic spans are highlighted.

Annotated sample	Model predictions
As is Drumpf a lying gasbag.	As is Drumpf a lying gasbag .
The Liberals should listen to the economic experts and dramatically increase immigration levels (at least 500,000 per year). Stop listening to the loser racist trash.	The Liberals should listen to the economic experts and dramatically increase immigration levels (at least 500,000 per year). Stop listening to the loser racist trash .
AK Jen - Russia is no longer a communist nation. Stop spouting that - it makes you look stupid.	AK Jen - Russia is no longer a communist nation. Stop spouting that - it makes you look stupid .
The "human rights" squad need a swift kick in the derriere.	The "human rights" squad need a swift kick in the derriere .
Rabidly anti-Canadian troll .	Rabidly anti-Canadian troll .
Lord have Mercy on us, Trump is running amok	Lord have Mercy on us, Trump is running amok.
You are a nut ball	You are a nut ball
Lives in Thailand?	Lives in Thailand?
So like every other racist, he's a hypocrite	So like every other racist, he's a hypocrite
Terry Stahlman is a worthless piece of human excrement!	Terry Stahlman is a worthless piece of human excrement!
People who are anti-immigration are weak, lying, racist whiners .	People who are anti-immigration are weak, lying, racist whiners.
Some people don't need dope to be dopey.	Some people don't need dope to be dopey .
But that's what you get now for minimum wage. Increasing that to a 'livable wage' for the exact same people is stupid!	But that's what you get now for minimum wage. Increasing that to a 'livable wage' for the exact same people is stupid !
They can't, it's jammed with an overload of their hero's excrement .	They can't, it's jammed with an overload of their hero's excrement.
holy fuck you troglodytes cant even handle the simple act of meming	holy fuck you troglodytes cant even handle the simple act of meming
Meanwhile Taxed is now complaining his Hyundai Santa Fe is a piece of crap	Meanwhile Taxed is now complaining his Hyundai Santa Fe is a piece of crap
Only in that sick and twisted brain stem of yours.	Only in that sick and twisted brain stem of yours.

Table 3: Examples from the competition test dataset of differences between the annotations and the predictions from BERT-toxic-CRF-VAT model. The toxic spans are highlighted.



part of the sequence of words. Samples like "You are a nut ball" detected only as "You are a nut ball" or "So like every other racist, he's a hypocrite" marked by the model as "So like every other racist, he's a hypocrite" can be perceived as likely errors even for human annotators.

## 6 Conclusions and Future Work

In this paper, several Transformer-based models (i.e., BERT and RoBERTa) were tested together with Virtual Adversarial Training to increase their robustness for identifying toxic spans from textual information. Our experiments argue that applying VAT increases performance and that domain-specific models have higher performance when compared to larger general models.

In terms of future work, we plan to experiment with self-supervised adversarial training (Chen et al., 2020) to improve the robustness of our models. As we noticed in this dataset too, online users find clever ways to hide offensive and toxic expressions. Adversarial training can be effectively employed to detect these attempts and a study of its impact on offensive and hate speech classifiers is worth pursuing as follow-up leads.

## References

- Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42.
- ME Aragón, H Jarquín, M Montes-y Gómez, HJ Escalante, L Villaseñor-Pineda, H Gómez-Adorno, G Bel-Enguix, and JP Posadas-Durán. 2020. Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF)*, Malaga, Spain.
- Andrei-Marius Avram, Dumitru-Clementin Cercel, and Costin Chiru. 2020. Upb at semeval-2020 task 6: Pretrained language models for definition extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 737–745.
- Kejiang Chen, Yuefeng Chen, Hang Zhou, Xiaofeng Mao, Yuhong Li, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2020. Self-supervised adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2218–2222. IEEE.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 5640–5650.
- 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.
- Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Laura Hanu and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933.
- Marius Ionescu, Andrei-Marius Avram, George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb at fincausal-2020, tasks 1 & 2: Causality analysis in financial documents using pre-trained language models. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 55–59.
- Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2991–3009.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2741–2749.

- Priyanshu Kumar and Aadarsh Singh. 2020. Nutcracker at wnut-2020 task 2: Robustly identifying informative covid-19 tweets using ensembling and adversarial training. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 404–408.
- Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.
- Kyounghee Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? examining public vs interpersonal swearing in response to donald trump’s youtube campaign videos. *Internet Research*, 27(4):991–1010.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.
- Tagrid Lemenager, Miriam Neissner, Anne Koopmann, Iris Reinhard, Ekaterini Georgiadou, Astrid Müller, Falk Kiefer, and Thomas Hillemacher. 2021. Covid-19 lockdown restrictions and online media consumption in germany. *International journal of environmental research and public health*, 18(1):14.
- Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fernández, Silvio Amir, Luis Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530.
- Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. Kk2018 at semeval-2020 task 9: Adversarial training for code-mixing sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 817–823.
- 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. *arXiv preprint arXiv:1907.11692*.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.
- Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2015. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.
- Sergey Morzhov. 2020. Avoiding unintended bias in toxicity classification with neural networks. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 314–320. IEEE.
- Andrei Paraschiv and Dumitru-Clementin Cercel. 2019. Upb at germeval-2019 task 2: Bert-based offensive language classification of german tweets. In *KONVENS*.
- Andrei Paraschiv, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb at semeval-2020 task 11: Propaganda detection with domain-specific trained bert. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1853–1857.
- John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.
- 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.
- Matthew 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.
- Khoury Richard and Larochelle Marc-André. 2020. Generalisation of cyberbullying detection. *arXiv preprint arXiv:2009.01046*.
- Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. *arXiv preprint arXiv:2012.15699*.
- Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *KONVENS*.

- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Mircea-Adrian Tanase, Dumitru-Clementin Cercel, and Costin Chiru. 2020a. Upb at semeval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of bert-based models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2222–2231.
- Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020b. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN) 2020*, pages 236–245.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.
- 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.
- 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 the 13th International Workshop on Semantic Evaluation*, pages 75–86.
- 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 the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.