amsqr at SemEval-2020 Task 12: Offensive language detection using neural networks and anti-adversarial features

Alejandro Mosquera
Symantec Enterprise Division
Broadcom Corporation
alejandro.mosquera@broadcom.com

Abstract

This paper describes a method and system to solve the problem of detecting offensive language in social media using anti-adversarial features. Our submission to the SemEval-2020 task 12 challenge was generated by an stacked ensemble of neural networks fine-tuned on the OLID dataset and additional external sources. For Task-A (English), text normalisation filters were applied at both graphical and lexical level. The normalisation step effectively mitigates not only the natural presence of lexical variants but also intentional attempts to bypass moderation by introducing out of vocabulary words. Our approach provides strong F1 scores for both 2020 (0.9134) and 2019 (0.8258) challenges.

1 Introduction

Keeping social media platforms free from unwanted publications such as spam, scam, phishing, hate speech, targeted attacks and fake news is still an active research topic nowadays. This is due not only to the relative low cost of creating fake accounts, bots (Albadi et al., 2019) and forging online identities but also to the large amount of personal information made available on the Internet which makes targeting certain groups and individuals easier than ever. While some of these threats were seen before affecting traditional messaging platforms such as email and SMS, the reach and adoption of social media applications have amplified their impact, requiring additional cost and effort to mitigate.

The use of offensive language as a vehicle to attack individuals and communities poses challenges not only for humans, which are prone to subjective and biased judgement (Sap et al., 2019) but also for automatic moderation systems. The inherent ambiguity when dealing with messages which are often short, can be written in mixed languages, contain informal words and are usually subject to adversarial modifications render naïve filtering approaches such as word lists or re-purposed spam detection models ineffective. For this reason, in order to solve this problem more sophisticated approaches such as state of the art natural language processing (NLP) is needed.

In order to keep track and measure the progress in the area on offensive language detection several English datasets with annotations for hate (Davidson et al., 2017), (Waseem and Hovy, 2016), targeted (Zampieri et al., 2019a) and personal attacks (Wulczyn et al., 2016) were released over the last years. Likewise, public evaluations such as HatEval (Basile et al., 2019) and OffenseEval (Zampieri et al., 2019b) were recently introduced highlighting the need for stronger baselines to assess the performance of more complex systems.

This paper evaluates the method and system submitted to the shared task 12 of SemEval-2020: Multilingual Offensive Language Identification in Social Media (Zampieri et al., 2020) for the subtask A (English) based on an stacked ensemble of neural networks. The rest of the document is organised as follows: In section 2, we review related work relevant for detecting abusive language. In Section 3 we describe our layered model approach including our anti-adversarial strategy based on text normalisation and stacking-based ensembling. In Section 4 we show the results obtained in the test and evaluation datasets. Finally, in Section 5 we draw our conclusions and outline future work.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1898–1905
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

Previous work on automatic hate speech and offensive language detection made use of linear models over word n-grams (Malmasi and Zampieri, 2017) and sentiment lexicons (Davidson et al., 2017). However most recent research is dominated by netural network architectures: Liu et al. (2019) and Zhu et al. (2019) applied bidirectional transformers (BERT) (Devlin et al., 2018) with success showing that pre-trained models fine-tuned for this task can outperform other approaches. On the other hand, convolutional neural networks (CNN) and bidirectional LSTMs (bi-LSTMs) provided strong results (Mahata et al., 2019) when paired with pre-trained embeddings such as FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014) or word2vec (Mikolov et al., 2013).

While adding more complexity, combining several models can effectively reduce classification bias and variance. We have seen good results using voting ensembles (Seganti et al., 2019) and stacked generalisation (Malmasi and Zampieri, 2018) when applied to this particular problem.

3 Methodology

The goal of Subtask-A is determining if a tweet is either offensive or not offensive, which conceptually translates to a binary classifier using F1-macro as scoring function. However, during the exploratory data analysis of the training set we've identified group of instances where users intentionally crafted offensive messages to bypass profanity and moderation filters. For this reason, our design choices have an anti-adversarial strategy in mind.

Best performing models in previous benchmarks (Basile et al., 2019), (Zampieri et al., 2019b) were based on popular pre-trained embeddings and architectures, either using transfer learning or leveraging these directly. While this is quite convenient in terms of computing cost it also introduces potential weaknesses which can be exploited in a black-box scenario. By guessing the base architecture the model was built upon, since there is a reduced set of high-quality pretrained models, an attacker could launch more successful black-box attacks (Wang et al., 2018). This is usually performed via input perturbation such as introducing synonyms (Jin et al., 2019), flipping characters (Pruthi et al., 2019) or including targeted keywords and typos (Shi et al., 2020). Being even possible to steal the whole model altogether (Krishna et al., 2019) in more sophisticated attacks.

3.1 Text Normalisation

Lexical normalisation techniques are particularly effective against black-box adversarial attacks (Alshemali and Kalita, 2019), while they also can increase the performance of NLP tools and applications when working with informal text (Mosquera and Moreda, 2013).

For this reason, we have applied to some of our inputs a text normalisation filter in order to reduce out-of-vocabulary words (OOV). This is not only effective against some adversarial perturbations but also replaces common typos and informal lexical variants commonly found in microblogs with their canonical version. This is performed at 2 levels: lexical and graphical. At lexical level we follow a similar modular architecture as TENOR (Mosquera et al., 2012) where a high-precision, low recall normalization dictionary is recursively combined with shortening/lengthening and re-casing rules. See table 1. Likewise, unicode homographs and near-homographs are translated to their ASCII equivalent by using a lookup table. See last entry in table 1.

Original	**Normalised**
Then these dumba$$es vote Democrat!?!!!	then these dumb asses vote democrat
@USER Again another b******* story no one is watching football because of this a******	again another bullshit story no one is watching football because of this asshole
theyre abso shite quality tho	they are absolute shit quality though
Gets Period* You are the cause of my 𝒹𝓎𝓈𝓅𝒽𝑜𝓇𝒾𝒶	gets period you are the cause of my dysphoria

Table 1: Text fragments where after normalisation a label flip was observed during validation.

3.2 Ensembling

Aiming to minimize the impact of adversarial attacks targeting popular models we have designed a 2-level classifier based on stacked generalisation as shown in Figure 1.

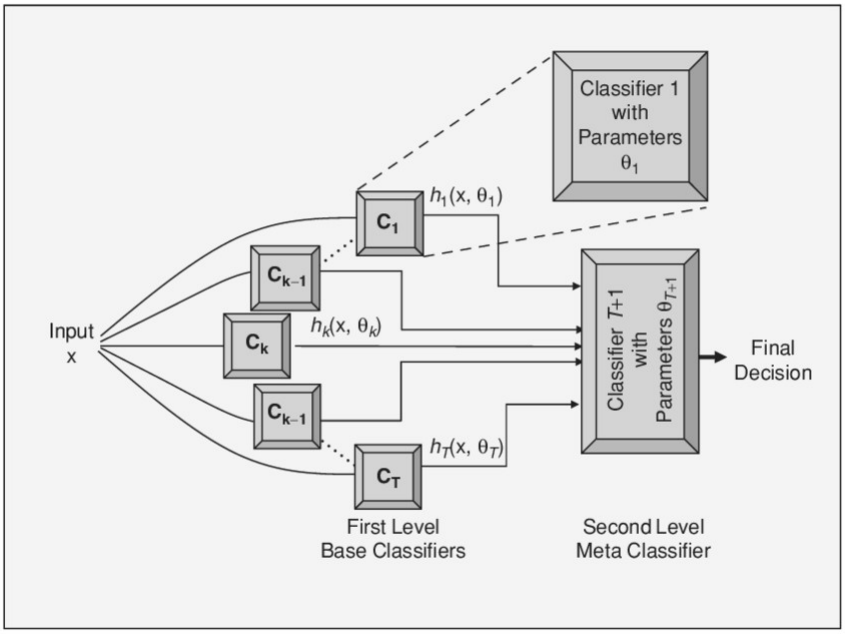

Figure 1: Stacked generalisation. Image reproduced from Polikar (2006)

The first level (L1) comprises of 42 models trained over several lexical resources using the OLID (Zampieri et al., 2019a) dataset and labels. This effectively encapsulates different models and training datasets, having more chances to thwart off-the-shelf attacks for specific architectures. Details of individual models and datasets for level 1 at can be found at Table 2.

At the second level (L2) there is a LightGBM (Ke et al., 2017) binary classifier trained over 55 boosting rounds with binary logarithmic loss. These and other model hyper-parameters were tuned against the OffenseEval 2019 evaluation set. The most important level-1 features of the final model considering both split count and gain can be seen at Figure 2. From there we can observe that both BERT (toxic, toxicnorm), and GloVe-based (capsuleglove, cnnglove) neural networks are clearly the strongest models in the ensemble.

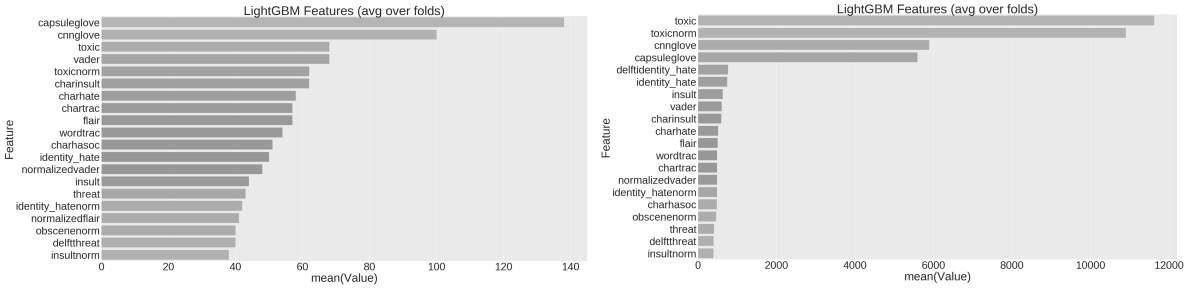

Figure 2: Feature importance (top 20) based on splits (left) and gain (right) for the LightGBM meta-classifier

[1]https://github.com/kermitt2/delft

[2]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[3]https://www.kaggle.com/c/detecting-insults-in-social-commentary/data

Level 1 model	Description	Training labels
capsuleglove	Capsule network (CapsNet) + GloVe. CapsNets (Sabour et al., 2017) have been shown as an alternative to Convolutional Neural Networks (CNNs) but more robust against white-box adversarial attacks (Frosst et al., 2018). CapsNets have been also seen outperforming CNNs in offensive text identification (Hettiarachchi and Ranasinghe, 2019)	OLID
cnnglove	CNN + GloVe	OLID
delft.*	GRU + FastText. The multi-class toxicity classifier from DeLFT [1] is leveraged to provide 6 inputs (delftidentity_hate, delftinsult, delftobscene, delftsevere_toxic, delftthreat and delfttoxic)	Wikipedia Toxic [2]
normaliseddelft.*	GRU + FastText + text normalisation produces 6 inputs (normaliseddelftidentity_hate, normaliseddelftinsult, normaliseddelftobscene, normaliseddelftsevere_toxic, normaliseddelftthreat and normaliseddelfttoxic)	Wikipedia Toxic
bert.*	BERT finetuned provides 6 inputs (toxic, severe_toxic, obscene, threat, insult and identity_hate)	Wikipedia Toxic
normalisedbert.*	BERT + text normalisation provides 6 inputs (toxicnorm, severe_toxicnorm, obscenenorm, threatnorm, insultnorm and identity_hatenorm)	Wikipedia Toxic
vader	VADER (Hutto and Gilbert, 2014) polarity score	Unsupervised
normalisedvader	VADER polarity score after text normalisation	Unsupervised
flair	Recurrent Neural Network (RNN) sentiment classifier based on Flair (Akbik et al., 2019)	IMDB (Maas et al., 2011)
normalisedflair	RNN-Flair sentiment score applied to normalised text	IMDB
qcount	Number of question marks	Unsupervised
emoji_max_max	Emoji offensive priors extracted from (Rosenthal et al., 2020). Only the highest offensive score (in case more than emoji appears in a message) of the maximum offensive score per emoji is considered	OffensEval 2020
charhasoc	Character n-gram (3-6) + logistic regression	HASOC (Mandl et al., 2019)
chartrac	Character n-gram (3-6) + logistic regression	TRAC (Kumar et al., 2018)
hateval.*	Word n-gram (1-3) + logistic regression over the 3 different labels, providing wordhate, wordtarget and wordag inputs for hate spech, targeted attack and aggression respectively at word level. The same were also trained at characted level (3-6), resulting the another 3 inputs (charhate, chartarget and charag)	HatEval (Basile et al., 2019)
charinsult	Character n-gram (3-6) + logistic regression	Kaggle insults [3]

Table 2: Summary of models and inputs.

4 Results

Our offensive text classification system obtained strong results across different datasets which are summarized in Table 3: It ranked 10th/82 at SemEval 2020 task 12/subtask A and (virtually) 2nd/104 against SemEval 2019 task 6/subtask A test set, which wasn't used as part of training nor validation.

System	Task	F1-macro
ltuhh2020 (best)	SemEval 2020 task 12 - subtask A	**0.92226**
L1 - toxic	SemEval 2020 task 12 - subtask A	0.91389
L1 - delftidentity_hate	SemEval 2020 task 12 - subtask A	0.913775
L2 - LGB	SemEval 2020 task 12 - subtask A	0.91348
L1 - toxicnorm	SemEval 2020 task 12 - subtask A	0.913211
L1 - normaliseddelftidentity_hate	SemEval 2020 task 12 - subtask A	0.912979
L1 - capsuleglove	SemEval 2020 task 12 - subtask A	0.908493
L1 - cnnglove	SemEval 2020 task 12 - subtask A	0.907237
NULI (best)	SemEval 2019 task 6 - subtask A	**0.829**
L2 - LGB	SemEval 2019 task 6 - subtask A	0.825815
L1 - cnnglove	SemEval 2019 task 6 - subtask A	0.800309
L1 - capsuleglove	SemEval 2019 task 6 - subtask A	0.787704
L1 - toxicnorm	SemEval 2019 task 6 - subtask A	0.767035
L1 - normaliseddelftidentity_hate	SemEval 2019 task 6 - subtask A	0.764185
L1 - toxic	SemEval 2019 task 6 - subtask A	0.762993
L1 - delftidentity_hate	SemEval 2019 task 6 - subtask A	0.734854

Table 3: Results of individual models (L1) and the final ensemble (L2) versus the best public scoring approach for each task.

Interestingly, BERT models fine tuned on Kaggle toxic dataset had a high correlation with the test set for this year challenge, even improving slightly final ensemble results when compared against the identity hate and toxic classes. Such correlation is not present in the previous year test set, where models trained on OLID outperformed the rest by a considerable margin.

There is another apparent trend reversal observed in normalised models, on the 2019 test set individual models with normalisation outperformed their non-normalised equivalent while in the current test set results were comparable for both normalised and not normalised.

Labelling shifts of certain keywords that caused the system to FP may be worth of further analysis: 79% of the tweets containing the pattern "sick|disgusting|sucks" were labelled as offensive in OLID, in comparison with a 55% when considering test set gold labels. Some examples of this disagreement can be found at Table 4.

Tweet	Dataset	Label
@USER That sucks {thumbs down}	OLID	OFF
@USER The game sucks	OLID	OFF
@USER man that sucks unreal	OLID	OFF
@USER Oh god, that sucks :/	Test	NOT
ldr doesn't really works it sucks	Test	NOT
Honestly they're not even pretty and the music sucks.... What do people see??	Test	NOT

Table 4: Similar tweets with different labels across OLID and test set.

5 Conclusion and Future Work

In this paper we describe our system and method for detecting offensive tweets built for SemEval-2020 Task 12 - subtask A. Our design choices had an adversarial environment in mind and therefore we've made use of anti-adversarial features such as text normalisation and ensemble learning, obtaining strong results in 2 evaluation datasets. In a future work we would like to explore different attack and defence scenarios for this particular problem.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2019. Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November.

Basemah Alshemali and Jugal Kalita. 2019. Toward mitigating adversarial texts. *International Journal of Computer Applications*, 178:1–7, 09.

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 the 13th International Workshop on Semantic Evaluation*, pages 54–63.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Hansi Hettiarachchi and Tharindu Ranasinghe. 2019. Emoji powered capsule network to detect type and target of offensive posts in social media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 474–480, Varna, Bulgaria, September. INCOMA Ltd.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT really robust? natural language attack on text classification and entailment. *CoRR*, abs/1907.11932.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2019. Thieves on sesame street! model extraction of bert-based apis.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

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 the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Ratn Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 683–690, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

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

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Alejandro Mosquera and Paloma Moreda. 2013. Improving web 2.0 opinion mining systems using text normalisation techniques. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 491–495.

Alejandro Mosquera, Elena Lloret, and Paloma Moreda. 2012. Towards facilitating the accessibility of web 2.0 texts through text normalisation. In *Proceedings of the LREC Workshop: Natural Language Processing for Improvign Textual Accessibility (NLP4ITA)*, pages 9–14.

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, October. Association for Computational Linguistics.

Robi Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *arxiv*.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3859–3869, Red Hook, NY, USA. Curran Associates Inc.

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

Alessandro Seganti, Helena Sobol, Iryna Orlova, Hannam Kim, Jakub Staniszewski, Tymoteusz Krumholc, and Krystian Koziel. 2019. NLPR@SRPOL at SemEval-2019 task 6 and task 5: Linguistically enhanced deep learning offensive sentence classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 712–721, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers.

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2018. With great training comes great vulnerability: Practical attacks against transfer learning. In *USENIX Security Symposium*.

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, June. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914.

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 (NAACL)*, 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 (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*.

Jian Zhu, Zuoyu Tian, and Sandra Kübler. 2019. UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 788–795, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.