Amrita School of Engineering - CSE at SemEval-2019 Task 6: Manipulating Attention with Temporal Convolutional Neural Network for Offense Identification and Classification

Murali Sridharan, Swapna T R
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidhyapeetham, India
muralisridharan.10@gmail.com,tr_swapna@cb.amrita.edu

Abstract

With the proliferation and ubiquity of smart gadgets and smart devices, across the world, data generated by them has been growing at exponential rates, in particular social media platforms like Facebook, Twitter and Instagram have been generating voluminous data on a daily basis. According to Twitter's usage statistics, about 500 million tweets are generated each day. While the tweets reflect the users' opinions on several events across the world, there are tweets which are offensive in nature that need to be tagged under the hateful conduct policy of Twitter. Offensive tweets have to be identified, captured and processed further, for a variety of reasons, which include i) identifying offensive tweets in order to prevent violent/abusive behaviour in Twitter (or any social media for that matter), ii) creating and maintaining a history of offensive tweets for individual users (would be helpful in creating meta-data for user profile), iii) inferring the sentiment of the users on particular event/issue/topic . We **(CodaLab Team/User Name: murali_sr)** have employed neural network models which manipulate attention with Temporal Convolutional Neural Network for the three shared subtasks i) ATT-TCN (ATTention based Temporal Convolutional Neural Network) employed for shared sub-task A that yielded a best macro-F1 score of 0.46, ii) SAE-ATT-TCN(Self Attentive Embedding-ATTention based Temporal Convolutional Neural Network) employed for shared sub-task B and sub-task C that yielded best macro-F1 score of 0.61 and 0.51 respectively. Among the two variants ATT-TCN and SAE-ATT-TCN, the latter performed better.

1 Introduction

In the prevailing digital era, Deep Learning has penetrated almost all industry verticals and afforded several researchers an effective tool, in handling voluminous data and deriving meaningful inferences. Initially, (LeCun et al., 1998) invented Convolutional Neural Network (CNN) model for extraction of local features, which later proved to be the standard choice for Computer Vision tasks. (Hochreiter and Schmidhuber, 1997) the introduced LSTM (Long Short Term Memory) architecture, which went on to become the standard choice for Natural Language Processing (sequence) tasks due to the implicit ordering of the sequence data in words and sentences. Then several architectures, combining LSTM with CNN were introduced that went on to become successful for NLP tasks as well. Deep Learning techniques have leaped forward through multiple NLP tasks such as Modeling, Classification, Translation, Summarization, etc., and have proved to be better compared to traditional techniques.

Ever since social media has become ubiquitous there have been individuals who take gratuitous advantage of the anonymous nature of social media platforms, and engage themselves in rude and offensive communications. Such behaviour that prohibit free flow of communication and violate acceptable usage policy has necessitated to identify and capture the offensive posts, comments, etc., in order to prevent the dissemination of abusive behaviour in social media. (Zampieri et al., 2019b) focused on this aspect and organized a classification task with a particular focus on Twitter posts; unlike predictions of positive or negative sentiments, this task has three shared sub-tasks, intended to identify and capture the offense target as an entity. The task includes three shared sub-tasks that include:

i) Sub-Task A: Offensive language identification,
ii) Sub-Task B: Offense type categorization and

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 540–546
Minneapolis, Minnesota, USA, June 6–7, 2019.

iii) Sub-Task C: Offense target identification.

i) Sub-Task A: Offensive language identification in which posts are categorized into Offensive or Not Offensive. Recently (Bai et al., 2018) empirically concluded that the association between sequence modeling and recurrent neural networks should be reconsidered and established that convolutional networks are ought to be considered for sequence modeling tasks. The TCN model can be extended followed by introduction of Attention to the output of Embedding layer and TCN layer **ii) Sub-Task B: Offense type categorization** in which the Offense type is categorized into either targeted or untargeted. The objective here is to understand sentence structure by emulating the relationship between words. The sequence of words is crucial to capture the essence of sentence unlike the practice of mere focus on constituent parts of a sentence in the previous model. Based on (Lin et al., 2017), minor modifications are injected into the previous model employed in sub-task A, and introduced self-attention for embedding further to aggregate the relationship between words in a sentence and stacked attention layer, at the output of each dilated convolution blocks. **iii) Sub-Task C: Identification of target offense** in which the who, the offense is aimed at is identified and categorized into Individual, Group or Other. The same model used in the previous sub-task B is employed for this sub-task as well.

2 Related Work

In the recent past, multiple NLP tasks and papers have explored Offense identification which include (bullying, aggression, hate-speech, obscenity, insults and identity threat). (Fortuna and Nunes, 2018) has elaborately surveyed several approaches employed for automatic detection of hate speech.

(Yin et al., 2009) was one of the first to address recognition of offensive language by employing supervised classification technique along with manually developed n-gram regex matches and, contextual attributes that considered the intensity of abuse in preceding sentences. (Sood et al., 2012) indicated that certain banned words when used in appropriate manner and context, does not warrant to be categorized as abusive/offensive. Further, they showed a considerably improved scheme of profanity detection, by incorporating lists and distance metric, which enabled identification and categorization of un-normalized terms like "@$$" or "m0r0n". (Chen et al., 2012) used lexical and parser features, for detecting comments from YouTube that are offensive. Without any preset semantics of toxic content, they came up with the tool that could be manipulated through a modifiable threshold. This threshold was to be treated as a measure of toxicity, filtering the online toxic content, prior to display of contents in the client's browser. Their work incorporated Support Vector Machines (SVMs) classifiers, which included regex (manually developed), n-gram, black-lists and dependency parse features, which achieved higher precision and recall values.

(Dadvar et al., 2013) affirmed that user context was crucial in the bonafide detection of cyberbullying. (Djuric et al., 2015) highlighted the effectiveness of comment embeddings in detection of hate speech, by joint modelling comments and words using Continuous-Bag of Words (C-BOW) to generate a low dimensional embedding. The embedding is passed to binary classifier for hate speech detection. (Mehdad and Tetreault, 2016) explored the significance of features to the extent of character to word and weighed the importance of each attribute. Since the style of comments, in online forums, vary from person to person, and often includes sub-standard profane English (i.e. "f u c k e r"), learning how adjacent characters communicate with each other reveal more about the abusiveness of a comment as a whole.

(Vijayan et al., 2017) surveyed the pros and cons of several techniques of machine learning and deep learning in their comprehensive study of text classification algorithms. (Malmasi and Zampieri, 2017) employed n-gram and skip gram based SVM classifier, to detect and classify hate-speech, into three categories: Hate, Offensive and Ok. (Gambäck and Sikdar, 2017) employed multiple CNN models totaling four for Hate-Speech Classification of Twitter posts into one of the following:sexism, racism, either(sexism and racism) and not hate speech. The first model, trained on character based n-grams (4-grams), the next model trained on word vectors built using word2vec. The third model was trained on word vectors which were produced in random. The fourth model was trained on word vectors in addition to char n-gram for the classification task. The fourth model performed comparatively better in the classification task. (Waseem et al., 2017)

proposed a typology, to capture the similarity and difference between sub-tasks, and discuss their role, involved in annotating data and emulating feature construction. Their work was instrumental in identifying if the offense was targeted towards an individual or an entity, and whether the offensive language was explicit or implicit. (Zhang et al., 2018) introduced a new method, combining Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) to perform a compartive evaluation on public datasets, and set new benchmarks, in Hate Speech Detection on Twitter.

3 Methodology and Data

Besides being fast and parallel, the important aspect of TCN is causal convolution, its capability to take any arbitrary length sequence and generate an output sequence of the same arbitrary input length.

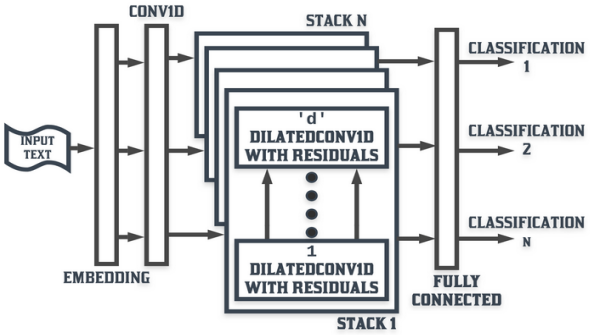

Figure 1: Temporal Convolutional Neural Network

Ever since (Bahdanau et al., 2014) introduced attention mechanism in NLP, for machine translation there have been multiple advances in memory related tasks. Further (Yin et al., 2016) established that attention based CNN performed better than attention based LSTM (Long Short Term Memory) for the answer selection task.

Here in Sub-Task A:Offense Identification, TCN was extended for sub-task A, with attention mechanism. Instead of the conventional dropout layer, a simple attention mechanism is applied at the output of embedding layer and at the output of TCN before classification layer, as illustrated in Figure 2 and Figure 3. The intention is to avoid random dropout of constituent data, which might be crucial, and introduce a mechanism with the ability, to selectively focus on input and capitalize on the crucial contributing parameters with

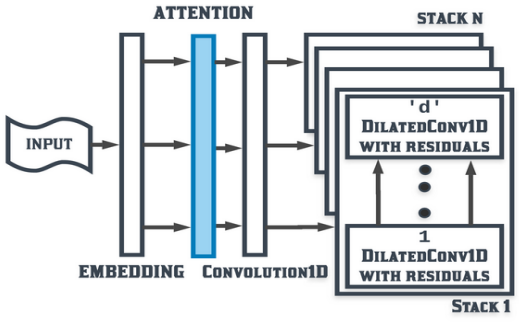

Figure 2: Temporal Convolutional Neural Network with Attention layer at the output of Embedding layer

varying attention weights and contextual vectors. ReLu (Nair and Hinton, 2010) activation is used for DilatedConvolution1D (Yu and Koltun, 2015) and softmax (Bridle, 1990) is used at the final classification layer.

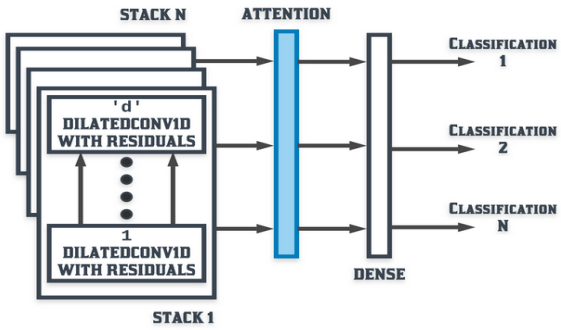

Figure 3: Temporal Convolutional Neural Network with Attention layer at the output of TCN before final Classification

Parameters	A	B	C
# Features	100*	250*	250*
# Filters	3*	5*	5*
Kernel Size	4	5	5
Dilation Range	11	11	11
Stack Count	1	1	1
Dropout Rate	0.05	0.01	0.01
Batch Size	32	32	32

Table 1: Hyper-parameters for each Sub-Task A, B and C respectively. Parameters are in numbers. *x10^2

For Sub-Task B:Automatic Categorization of Offense Type, a slight modification is introduced, to the model used for the previous sub-task, by incorporating Self Attention at the output of Embedding layer.

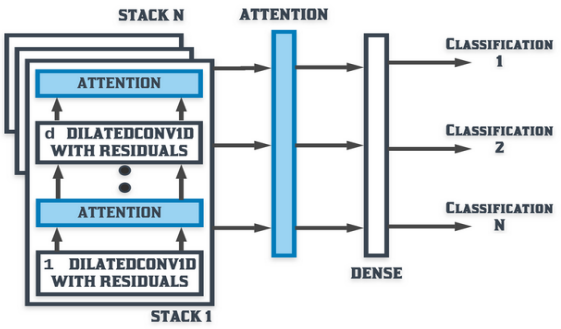

Figure 4: Temporal Convolutional Neural Network with Attention layer at the output of each Dilated-Conv1D and

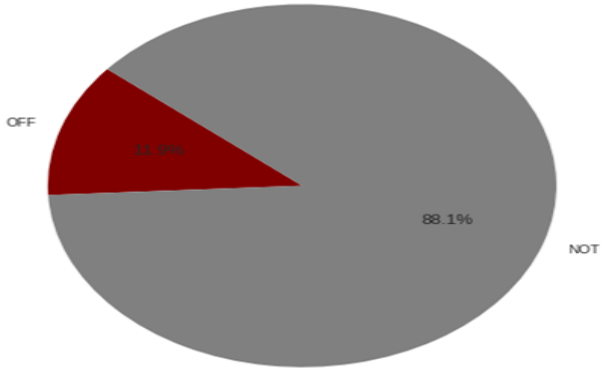

Figure 5: Shared Sub-Task A, training data instance share (OFF and NOT)

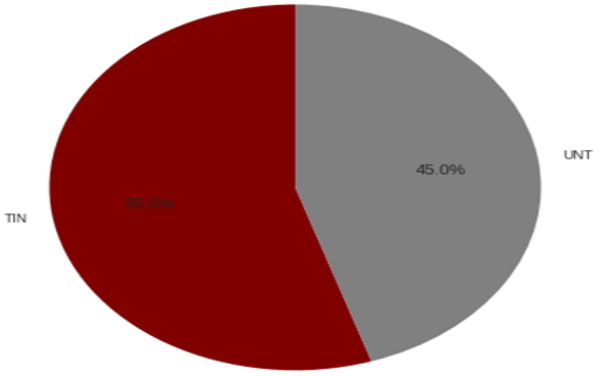

Figure 6: Shared Sub-Task B, training data instance share (TIN and UNT)

Self-Attention is introduced for characterization of multiple location of the tokens, a sentence has, in addition to extraction of semantic features. Additionally, Attention layer is stacked at the output of every 'd' dilated convolution blocks, to augment the contextual vectors, as illustrated in Figure 4. For Sub-Task C:Offense Target Identification, the model used in the sub-task B was employed to identify and categorize the target of the posts into Individual (IND), Group (GRP) and Other (OTH) classes. For all the variants, binary cross-entropy loss function is employed with the focus on categorical accuracy.

The methods employed for gathering the data, preparation and compilation of dataset, used in OffensEval shared task is described in Zampieri et al. (2019a). Two additional datasets, Kaggle Toxic Comment Classification dataset and TRAC-1 Aggression Identification in Social Media Shared Task dataset were used for sub-task A and sub-task B respectively.

In 2018, Kaggle hosted a Toxic Comment Classification competition in association with Jigsaw, which focused on classifying Wikipedia comments into one of six categories: insult, obscene, severe toxic, threat & identity hate and toxic. The instances which do not fall into one of the six categories are clean. All the six toxic categories are mapped to Offensive (OFF) class and the clean instances are mapped to Not Offensive (NOT) class.

The mapped instances were combined with the training dataset, provided for sub-task A, which produced a total of 172811 instances, of which 20625 instances were Offensive and 152186 instances were clean, as depicted in Figure 5. For the training data of Sub-Task B, TRAC-1 data (Kumar et al., 2018) was used, in addition to the provided training data, producing a total of 14174 instances containing 7799 Targeted Insults and Threats (TIN), and 6375 Untargeted (UNT) instances. No other additional training dataset was used, apart from the provided dataset for Sub-Task C. It comprised of 3876 Offensive instances, of which 1074 Offensive instances belong to Individual (IND) category, 2407 Offensive instances belong to Group (GRP) category and 395 Offensive instances target belong to (OTH) category.

4 Results

The macro averaged F1 was employed as the official metric for all the sub-tasks involved in this task accounting for the high class imbalance ratio. Our first model (ATT-TCN), employed for the shared Sub-Task A, produced an overall accuracy of 65.81% (best of 3 for evaluation test data @CodaLab). The second variant (SAE-ATT-TCN), employed for the shared Sub-Task

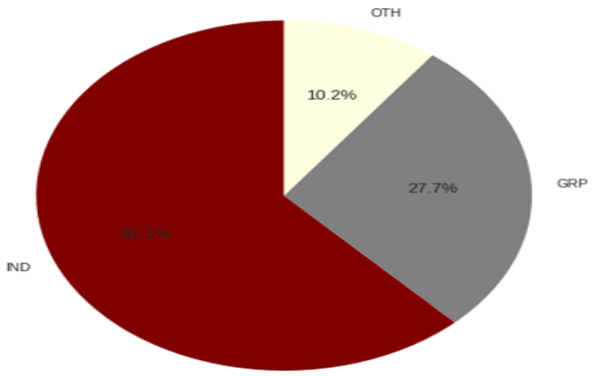

Figure 7: Shared Sub-Task C, training data instance share (IND, GRP and OTH)

B and Sub-Task C produced an overall accuracy of 75.83% and 61.5% (best of 3 for Evaluation Test data @CodaLab), respectively. The neural network model generation, fine-tuning and the evaluation test data prediction, all the activities have been executed in Google Colaboratory environment, utilizing the on hand GPU hardware accelerator. The cross validation results, and the detailed evaluation test data results have been listed in the tables accordingly.

System	Accuracy
Base ATT-TCN	0.9477
SAE-ATT-TCN[1]	0.7144
SAE-ATT-TCN[2]	0.7215

Table 2: Cross-Validation Results for Sub-Tasks A,[1]B and [2]C respectively.

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
Base ATT-TCN	**0.4682**	**0.6581**

Table 3: CodaLab Test Results for Sub-Task A.

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
SAE-ATT-TCN	**0.6164**	**0.7583**

Table 4: CodaLab Test Results for Sub-Task B.

The confusion matrices for the best performing variant of each Sub-Task have been depicted in Table 6, Table 7 & Table 8. In Table 6, for Sub-Task A, it is evident that the number of NOT Offensive instances (540) have been predicted correctly attributing to the higher count of training data instances for that class, and count of correct Offensive (OFF) instances prediction is less attributing to the less training instances for that category. The higher number of false positives for the Sub-Task A clearly indicate not so good classification performance of the variant ATT-TCN. The higher number of true positives for Targeted Insult and Threat (164 TIN instances), and the lesser true negatives (18 UNT instances), in Table 7, indicate the model has better generalization ability, and performed significantly better, compared to the previous model.

In Table 8, it is clear that the Group (GRP) offense target classification is predicted correctly compared to Other (OTH), and Individual (IND) categories respectively.

System	F1 (macro)	Accuracy
All GRP baseline	0.1787	0.3662
All IND baseline	0.2130	0.4695
All OTH baseline	0.0941	0.1643
SAE-ATT-TCN	**0.5132**	**0.615**

Table 5: CodaLab Test Results for Sub-Task C.

	NOT	OFF
NOT	540	80
OFF	214	26

Table 6: Sub-Task A, Confusion Matrix for Base ATT-TCN at threshold 0.45

	TIN	UNT
TIN	164	49
UNT	9	18

Table 7: Sub-Task B, Confusion Matrix for SAE-ATT-TCN at threshold 0.70

	GRP	IND	OTH
GRP	44	27	7
IND	10	81	9
OTH	14	15	6

Table 8: Sub-Task C, Confusion Matrix for SAE-ATT-TCN at threshold 0.55

5 Conclusion

Based on the results, it is evident that SAE-ATT-TCN has performed significantly better than the

base model ATT-TCN. From Sub-Task A, we learned that rather than going ahead with random sampling for train and test split, proceeding with a categorical split, to the best possible even ratio, would increase the generalization ability. When the dataset has class imbalance ratio, retaining even number of instances for each class as much as possible, would ensure normal distribution of the instances for each class. Such data distribution would not be skewed for a particular class which ensure better generalization capability leading to improved classification accuracy. We note that processing the sequence in both the directions (forward and backward) would further improve the classification performance attributing to better context and semantic representation learning capabilities. We are working on Bi-Directional Attention based Temporal Convolutional Network model. Our participation in the SemEval 2019: Task 6 competition has been a very good learning experience for our team, and we are eager to learn from other best performing entries.

Acknowledgement

Our team would like thank and appreciate the SemEval 2019 organizing team for affording an opportunity to participate in such classification task. We look forward to additional NLP classification tasks.

References

Dzmitry Bahdanau, Yoshua Bengio, and Kyunghyun Cho. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

John S. Bridle. 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80. IEEE.

Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 29–30. International World Wide Web Conferences Steering Committee.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hatespeech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ritesh Kumar, Shervin Malmasi, Atul Kr. Ojha, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, Santa Fe, USA.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Zhouhan Lin, Mo Yu, Cicero Nogueira dos Santos, Minwei Feng, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

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.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*.

Sara Owsley Sood, Elizabeth Churchill, and Judd Antin. 2012. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*.

V. K. Vijayan, K. R. Bindu, and L. Parameswaran. 2017. A comprehensive study of text classification algorithms. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1109–1113.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Dawei Yin, Brian D Davison, April Kontostathis, Zhenzhen Xue, Liangjie Hong, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.

Wenpeng Yin, Bowen Zhou, Bing Xiang, and Hinrich Schütze. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions.

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 The 13th International Workshop on Semantic Evaluation (SemEval)*.

Ziqi Zhang, Jonathan Tepper, and David Robinson. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.