Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 126–131
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Aggression and Misogyny Detection using BERT: A Multi-Task Approach

Niloofar Safi Samghabadi♠[1], Parth Patwa♢[1], Srinivas PYKL♢,
Prerana Mukherjee♢, Amitava Das♣, Thamar Solorio♠
♠ Department of Computer Science, University of Houston
♢ Indian Institute of Information Technology, Sri City
♣Wipro Research Lab
{nsafisamghabadi, tsolorio}@uh.edu
{parthprasad.p17, srinivas.p, prerana.m}@iiits.in
amitava.das2@wipro.com

Abstract

In recent times, the focus of the NLP community has increased towards offensive language, aggression, and hate-speech detection. This paper presents our system for TRAC-2 shared task on “Aggression Identification” (sub-task A) and “Misogynistic Aggression Identification” (sub-task B). The data for this shared task is provided in three different languages - English, Hindi, and Bengali. Each data instance is annotated into one of the three aggression classes - Not Aggressive, Covertly Aggressive, Overtly Aggressive, as well as one of the two misogyny classes - Gendered and Non-Gendered. We propose an end-to-end neural model using attention on top of BERT that incorporates a multi-task learning paradigm to address both sub-tasks simultaneously. Our team, “na14”, scored 0.8579 weighted F1-measure on the English sub-task B and secured 3rd rank out of 15 teams for the task. The code and the model weights are publicly available at https://github.com/NiloofarSafi/TRAC-2.

Keywords: Aggression, Misogyny, Abusive Language, Hate-Speech Detection, BERT, NLP, Neural Networks, Social Media

1. Introduction

Social media and the internet are overabundant with data. The number of users on the internet has increased by 83% from 2014 to 2019. In 2019, more than 500 million tweets and 4 billion Facebook messages were posted daily[2]. Social Media has become an important and influential means of communication as it is easily accessible and provides a lot of freedom to users. Some users misuse this by engaging in trolling, cyberbullying, or by sharing aggressive, hateful, misogynistic content. Aggressive words, abusive language, or hate-speech is used to harm the identity, status, mental health, or prestige of the victim (Beran and Li, 2005; Culpeper, 2011). This type of anti-social behavior causes disharmony in society. Hence, it is becoming quite alarming, and it is crucial to address this problem.

Aggression is a feeling of anger that results in hostile behavior and readiness to attack. According to Kumar et al. (2018c), aggression can either be expressed in a direct, explicit manner (Overtly Aggressive) or an indirect, sarcastic manner (Covertly Aggressive). Hate-speech is used to attack a person or a group of people based on their color, gender, race, sexual orientation, ethnicity, nationality, religion (Nockleby, 2000). Misogyny or Sexism is a subset of hate-speech (Waseem and Hovy, 2016) and targets the victim based on gender or sexuality (Davidson et al., 2017; Bhattacharya et al., 2020).

It is essential to identify aggression and hate-speech in social networks to protect online users against such attacks, but it is quite time-consuming to do so manually. Hence, social media companies and government agencies are focusing on building a system that can automate the identification process. However, it is difficult to draw a distinguishing line between acceptable content and aggressive/hateful content due to the subjectivity of the definitions and different perceptions of the same content by different people, which makes it harder to build an automated AI system. Facebook published its audit report[3] on civil rights, which explains its strategy to tackle abusive and hateful content. The report claims that building a complete automation system to detect hate-speech is not possible, and content moderation is unavoidable. This point brings many researchers to focus on building hate-speech/aggression detection systems since a large amount of such data is diffused in social networks. To this end, several workshops have been organized, including ‘Abusive Language Online’ (ALW) (Roberts et al., 2019), ‘Trolling, Aggression and Cyberbullying’ (TRAC) (Kumar et al., 2018b), and Semantic Evaluation (SemEval) shared task on Identifying Offensive Language in Social Media (OffensEval) (Zampieri et al., 2020).

This paper presents our system for TRAC-2 Shared Task on “Aggression Identification” (sub-task A) and “Misogynistic Aggression Identification” (sub-task B), in which we propose a BERT (Devlin et al., 2018) based architecture to detect misogyny and aggression using a multi-task approach. The proposed model uses attention mechanism over BERT to get relative importance of words, followed by Fully-Connected layers, and a final classification layer for each sub-task, which predicts the class.

2. Related Work

Hate-speech: The interest of NLP researchers in hate-speech, aggression, and sexism detection has increased recently. Kwok and Wang (2013) proposed a supervised ap-

[1] These authors contributed equally.

[2] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[3] https://www.theverge.com/interface/2019/7/2/20678231/facebook-civil-rights-audit-hate-speech-moderators

proach to detect anti-black hate-speech in social media platforms using Twitter data. They categorized the text into binary labels racist vs. non-racist and achieved a classification accuracy of 76%. Burnap and Williams (2015) utilized ensemble based classifier results to forecast cyber-hate proliferation using statistical approaches. The classifier captured the grammatical dependencies between words in Twitter data to anticipate the behavior to give antagonistic responses. Nobata et al. (2016) curated a corpus of user comments for abusive language detection and resorted to machine learning based approaches to detect subtle hate-speech. Schmidt and Wiegand (2017), give a detailed survey on hate-speech detection works. Gambäck and Sikdar (2017) used convolutional layers on word vectors to detect hate-speech. Other recent works (Zhang et al., 2018; Agrawal and Awekar, 2018; Dadvar and Eckert, 2018) also use deep learning based techniques to detect hate-speech. BERT Based approaches also have become popular recently (Nikolov and Radivchev, 2019; Mozafari et al., 2019; Risch et al., 2019).

Sexism: Recently, misogynistic and sexist comments, posts, or tweets on social media platforms have become quite predominant. Jha and Mamidi (2017) provided an analysis of sexist tweets and further categorize them as hostile, benevolent, or other. Sharifirad and Matwin (2019) also provided an in-depth analysis of sexist tweets and categorize them based on the type of harassment. Frenda et al. (2019) performed linguistic analysis to detect misogyny and sexism in tweets. Parikh et al. (2019) introduced the first work on multi-label classification for sexism detection and also provided the largest dataset on sexism categorization. They built a BERT based neural architecture with distributional and word level embeddings to perform the classification task.

Aggression: The first Shared Task on Aggression Identification (Kumar et al., 2018a) aimed to identify aggressive tweets in social media posts and provided datasets in Hindi and English. Samghabadi et al. (2018) used lexical and semantic features along with logistic regression for the task and obtained 0.59 and 0.63 F1 scores on Hindi and English Facebook datasets, respectively. Orasan (2018) utilized machine learning (SVM, random forest) on word embeddings for aggressive language identification. Raiyani et al. (2018) used fully connected layers on highly pre-processed data. Aroyehun and Gelbukh (2018) used data augmentation along with deep learning for aggression identification and achieved 0.64 F1 score on the English dataset. Risch and Krestel (2018) also employed a similar technique and got 0.60 F1 score for English.

3. Data

The datasets for this shared task are provided by (Bhattacharya et al., 2020) in three different languages: English, Hindi, and Bengali. For sub-task A, the data has been labeled with one out of three possible tags:

Not Aggressive (NAG): Texts which are not aggressive. E.g. *"hats off brother"*.

Covertly Aggressive (CAG): Texts that express aggression in an indirect, sarcastic manner. E.g., *"You are not wrong, you are just ignorant."*.

Overtly Aggressive (OAG): Texts which express aggression in a direct, straightforward, and explicit way. E.g., *"Liberals are retards"*.

For sub-task B, there are two classes:

Gendered (GEN): Texts that target a person or a group of people based on gender, sexuality, or lack of fulfillment of stereotypical gender roles. E.g., *"Homosexuality should be banned"*.

Non-gendered (NGEN): Texts that are not gendered. E.g.. *"you are absolutely true bro...but even politicians supports them"*.

Although the perception of aggression and misogyny can vary from person to person, we found some annotations that are highly improbable. The following are some examples that are mislabeled as NAG:

- *"This lady from BJP is crazy this is how u react man such a foolish and ignorant lady"*
- *"What a lousy moderator arnab is. Falthu show"*,
- *"Ha yaar bahut hi chutya movie tha.sab log keh raha tha badia movie tha isliye dekha bt bilkul jhaand tha"* (It was a stupid movie. Everyone was saying it is good so I saw but it is completely stupid)
- *"Brother puri movie bta di chutiya he kya"* (brother you spoiled the entire movie are you an idiot)

Some examples of comments mislabeled as NGEN:

- *"true feminist is Cancer"*
- *"Breif description but feminist is like urban terrorist and they will never understand"*
- *"Feminists are the next threat to our country"*
- *"chutiya hai ye feminists"* (these feminists are idiots)

Table 1 shows statistics over the train and validation data for both sub-tasks across all available languages. From this table, we can easily find out that for both sub-tasks A and B, the train and dev sets are highly skewed towards NAG and NGEN classes, respectively.

Table 2 indicates the co-occurrence of sub-task A and sub-task B labels. NAG mostly co-occurs with NGEN. The ratio of GEN to NGEN in OAG is greater than that in NAG and CAG. Overall, in all three languages, we can observe that as the directness of aggression increases (NAG<CAG<OAG), the percentage of GEN examples also increases. In Hindi and Bengali, OAG examples are more likely to be tagged as GEN than NGEN. Based on these observations, we can say that these two sub-tasks are related.

4. System Architecture

As we saw that the sub-tasks are related to each other, we create a unified deep neural architecture, following a multi-task approach. Figure 1 illustrates the overall architecture of our proposed model. Our proposed model consists of the following modules:

language	split	size	sub-task A			sub-task B	
			NAG	CAG	OAG	GEN	NGEN
English	train	4263	3375 (79.17%)	453 (10.63%)	435 (10.20%)	309 (7.25%)	3954 (92.75%)
	dev	1066	836 (78.42%)	117 (10.98%)	113 (10.60%)	73 (6.85%)	993 (93.15%)
	test	1200	690 (57.50%)	224 (18.67%)	286 (23.83%)	175 (14.58%)	1025 (85.42%)
Hindi	train	3984	2245 (56.35%)	829 (20.81%)	910 (22.84%)	661 (16.59%)	3323 (83.41%)
	dev	997	578 (57.97%)	211 (21.17%)	208 (20.86%)	152 (15.25%)	845 (84.75%)
	test	1200	325 (27.08%)	191 (15.92%)	684 (57.00%)	567 (47.25%)	633 (52.42%)
Bengali	train	3826	2078 (54.31%)	898 (23.47%)	850 (22.22%)	712 (18.61%)	3114 (81.39%)
	dev	957	522 (54.55%)	218 (22.78%)	217 (22.67%)	191 (19.96%)	766 (80.04%)
	test	1188	712 (59.93%)	225 (18.94%)	251 (21.13%)	202 (17.00%)	986 (83.00%)

Table 1: Data statistics.

language	split	total	NAG-GEN	NAG-NGEN	CAG-GEN	CAG-NGEN	OAG-GEN	OAG-NGEN
English	train	4263	134	3241	35	418	140	295
	dev	1066	38	798	9	108	26	87
Hindi	train	3984	32	2213	79	750	550	260
	dev	997	11	567	26	185	115	93
Bengali	train	3826	129	1949	129	769	454	395
	dev	957	37	485	31	187	123	94

Table 2: Co-occurrence between sub-task labels.

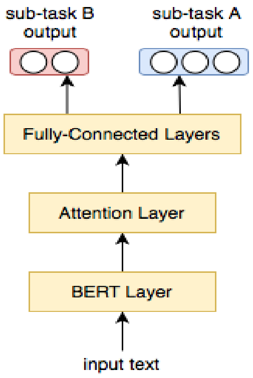


Figure 1: Overall architecture of the proposed model.

BERT Layer: We pass the input sequence of tokens to the BERT model (Devlin et al., 2018) to extract contextualized information.

Attention Layer: We feed the output of BERT layer to the attention mechanism proposed in Bahdanau et al. (2014). This layer computes the weighted sum of $r = \sum_i \alpha_i h_i$ to aggregate hidden representations (h_i) of all tokens in a sequence to a single vector. To measure the relative importance of words, we calculate the attention weights α_i as follows:

$$\alpha_i = \frac{exp(score(h_i, e))}{\Sigma_{i'} exp(score(h_{i'}, e))} \tag{1}$$

where the $score(.)$ function is defined as:

$$score(h_i, e) = v^T tanh(W_h h_i + b_h) \tag{2}$$

where W_h is the weight matrix, and v and b_h are the parameters of the network.

Fully-Connected Layers: We pass the output of the attention layer to Fully Connected (linear) layers for dimension reduction. There are two linear layers with 500 and 100 neurons, respectively.

Classification Layer: We feed the output of linear layers to two separate classification layers, one for predicting aggression class, and another for misogyny identification. For both cases, we use a linear layer with a softmax activation on top, which gives a probability score to the classes. The number of output neurons is three and two for sub-tasks A and B, respectively.

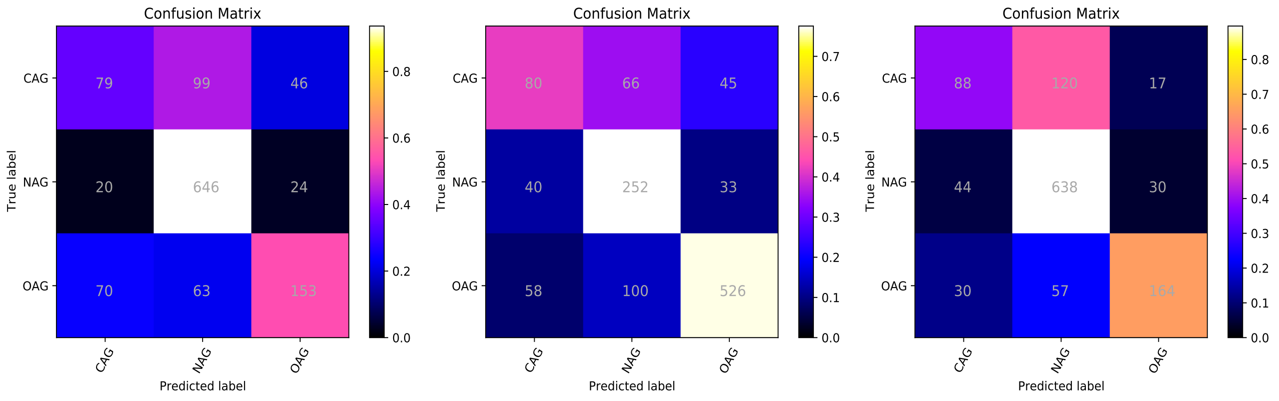
4.1. Experimental Setups

For pre-processing, we use the BERT tokenizer for text tokenization. Then, we truncate the posts to 200 tokens, and left-pad the shorter sequence with zeros. For initializing weights of the BERT layer, we use "bert_based_uncased" pre-trained weights for English and "bert_base_multilingual_cased" for Hindi and Bengali. To compute the loss between predicted and actual labels, we use Binary Cross Entropy. We calculate the sum of losses for both sub-tasks A and B. Additionally, for addressing the imbalance problem in the corpora, we add information about class weights to the loss functions for both outputs. We update the network weights using Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1e^{-5}$; however, we do not fine-tune the BERT layer. We train the model over 200 epochs using training data and save the best model based on the F1 score obtained on the validation set. We train our models on Nvidia Tesla P40 GPU having 24 GB memory, where each epoch takes around 1.5 minutes to be completed. The code and the model weights are publicly available[1].

5. Results

Table 3 shows the weighted F1 score and accuracy of our system on all the sub-tasks. Weighted F1 score is used as

[1] https://github.com/NiloofarSafi/TRAC-2

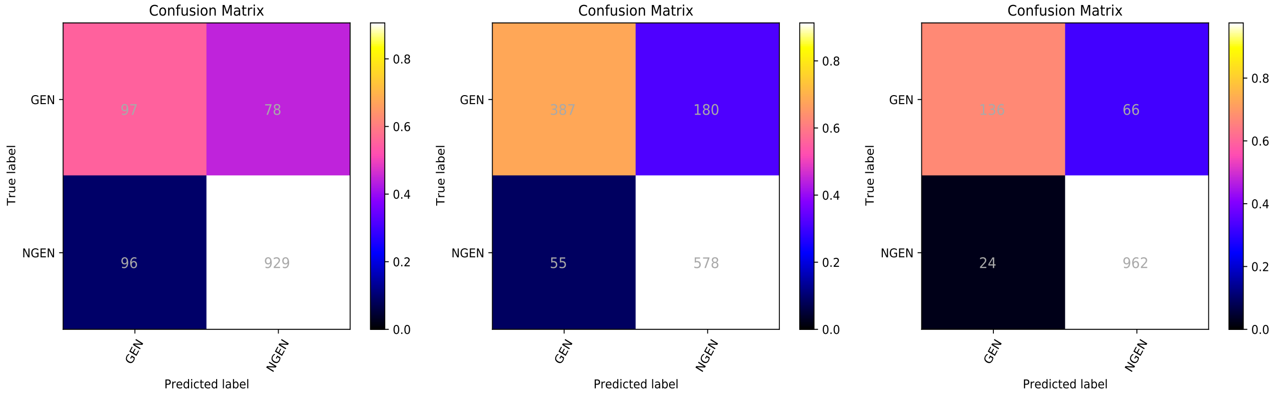


(a) English sub-task A

(b) Hindi sub-task A

(c) Bengali sub-task A

Figure 2: Heatmap of confusion matrices of our best performing systems for sub-task A across all languages.



(a) English sub-task B

(b) Hindi sub-task B

(c) Bengali sub-task B

Figure 3: Heatmap of confusion matrices of our best performing systems for sub-task B across all languages.

the official metric to rank the participants by the organizers. Based on the table, misogyny is easier to detect as compared to aggression across all available languages. The possible reason could be its binary and relatively straightforward nature as compared to sub-task A, which includes three classes. Our best score is achieved on English sub-task B, where we secured 3rd rank out of 15 teams. Our system lags behind the best performance on EN-B (0.8715 F1), and BEN-B (0.9365 F1) by 0.0136 and 0.0159, respectively, which shows our system is competitive and comparable to them.

Sub-task	F1 (weighted)	Accuracy
ENG-A	0.7143	0.7317
HIN-A	0.7183	0.7150
BEN-A	0.7369	0.7492
ENG-B	0.8579	0.8550
HIN-B	0.8008	0.8042
BEN-B	0.9206	0.9242

Table 3: Results of BERT model on all sub-tasks.

Figure 2 illustrates the confusion matrices of sub-task A for all three languages. Overall, CAG examples are more likely to be wrongly predicted as NAG than OAG. This could be due to the lack of abusive or explicit words in CAG instances. We further investigate this possibility in Section 5.1. In Hindi, OAG-NAG confusion (100) is high and is significantly more than that in English and Bengali. The reason could be that for Hindi corpus, the majority of the train instances are tagged as NAG (56.35%), whereas in its test data, the majority of the instances are labeled as OAG (57.00%).

Figure 3 shows the confusion matrices for sub-task B on all three languages. Similar to OAG-NAG, we can see that GEN-NGEN confusion for Hindi test data is higher than that in other languages. It can be explained by table 1, where we can see that for Hindi sub-task B, the distribution of classes across the test data is significantly different from the training and dev sets.

Language	Sub-task A			Sub-task B	
	NAG	CAG	OAG	GEN	NGEN
English	0.86	0.40	0.60	0.53	0.91
Hindi	0.68	0.43	0.82	0.77	0.83
Bengali	0.84	0.45	0.71	0.75	0.96

Table 4: Class-wise F1 score for both sub-tasks across all three languages.

Table 4 indicates the class-wise performance of our system

S.no	sub-task	text	actual	predicted
a	ENG-A	*Also Veere Di Wedding Fake Feminist Piece Of Shit...*	NAG	OAG
b	ENG-A	*oneitis - that's what kabir singh had with that girl in the movie ... dumb as fuck*	NAG	OAG
c	HIN-A	*Maha Chutiyapay ki film he Kabir Singh... It's totally bullshit movie...* (Kabir Singh is a very stupid film... it's totally bullshit movie...)	NAG	OAG
d	HIN-A	*Mujhe bhi jand lagi movie lakin maine chutiyo ke samne jaban nahi kholi or na hi kholuga* *(I also found this movie stupid, but I didn't open my mouth in front of idiots and won't do so.)*	NAG	OAG
e	ENG-B	*neha gupta ur are a crook if there are no evidence den how u can file a false compaint????*	GEN	NGEN
f	ENG-B	*kapil why are u listening to these chutiaasssss....give them shut up call...insane idiots*	GEN	NGEN
g	HIN-B	*Bhadwa hai rajdeep ...* (Rajdeep is an idiot.)	GEN	NGEN
h	HIN-B	*Kaunsi charas ya afeem phoonk ke aayi hai ye. Gandee aurat. Aurat ke naam pe dhabba.* *(Which weed or poppy has she smoked? Dirty lady. Blot on the name of a woman.)*	NGEN	GEN

Table 5: Instances where predicted label seems more accurate than given label.

on all the sub-tasks. For sub-task A, the performance is least for CAG across all the languages, which shows that it is the most challenging aggression class to identify. OAG and CAG scores are least for English as compared to the other two languages because the percentage of training examples for those two classes is lower in English as compared to other languages. NAG is the easiest to detect in English and Bengali, whereas OAG is the easiest to detect in Hindi. With regards to sub-task B, the performance is better on NGEN than GEN for all the three languages. The difference between the F1 score on NGEN and GEN is significantly more in English than in Hindi and Bengali. This can be attributed to the lower percentage of GEN examples in English than in the other two languages.

5.1. Error Analysis

We analyze the mistakes of our model on the validation set to see where it goes wrong. We found several instances where the actual tag is CAG, but our model classifies them as NAG. Some of those examples are listed as follows:

- *"Fat shaming is good. Why not?"*
- *"**Gay people rely on straight people to produce more gay people**"*
- *"They have no right to live"*
- *"Inko hospital bejo..ye mentally hille hue log han" (Send them to hospital, they are mentally disturbed people.)*
- *"Bhai aap na sirf review kariye baki ki baatein na hi kare toh accha h ?" (Brother you only do review, it's better of you don't talk about other things.)*

From these examples, we can see that due to the indirect/sarcastic nature and lack of profanity in CAG, it is confused with NAG. This flags CAG as the most difficult class to detect.

We also found some instances where the predicted labels seem more likely to be correct than the annotated labels. Table 5 shows such examples. In that, examples a-d are from sub-task A and are labeled as NAG, but as they include abusive and explicit words, the predicted label OAG seems more accurate. Examples e-g are labeled as GEN, but they are targeted towards a specific person not based on gender. So the model prediction NGEN is correct. Example h attacks a woman based on her gender, and hence the model predicts it as GEN.

6. Conclusion

In this paper, we present our multi-task deep neural model to identify misogyny and aggression for three different corpora - English, Hindi, and Bengali. The analysis of the label co-occurrence across the two sub-tasks shows that aggression identification and misogyny identification are related. Analysis of the results shows that CAG is often confused with NAG and is the most challenging aggression class to detect.

For future work, instead of employing BERT as a feature extractor, we plan to fine-tune it using the training data. We also plan to explore more sentiment features for better identification of the implicit forms of aggression (CAG).

7. Bibliographical References

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*. Springer.

Aroyehun, S. T. and Gelbukh, A. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Bahdanau, D., Cho, K., et al. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv: 1409.0473*.

Beran, T. and Li, Q. (2005). Cyber-harassment: A study of a new method for an old behavior. *JECR*, 32(3).

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2).

Culpeper, J. (2011). *Impoliteness: Using language to cause offence*, volume 28. Cambridge University Press.

Dadvar, M. and Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *arXiv preprint arXiv:1812.08046*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Frenda, S., Ghanem, B., Montes-y Gómez, M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5).

Gambäck, B. and Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*.

Ritesh Kumar, et al., editors. (2018b). *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018c). Aggression-annotated corpus of hindi-english code-mixed data. *CoRR*, abs/1803.09402.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting Tweets Against Blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In *Proceedings of the International Conference on Complex Networks and Their Applications*. Springer.

Nikolov, A. and Radivchev, V. (2019). Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*.

Orasan, C. (2018). Aggressive language identification using word embeddings and sentiment features. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., and Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Raiyani, K., Gonçalves, T., Quaresma, P., and Nogueira, V. B. (2018). Fully connected neural network with advance preprocessor to identify aggression over Facebook and twitter. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Risch, J. and Krestel, R. (2018). Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Risch, J., Stoll, A., Ziegele, M., and Krestel, R. (2019). hpiDEDIS at GermEval 2019: Offensive language identification using a german BERT model. In *Proceedings of the 15th Conference on Natural Language Processing*.

Ritesh Kumar, Atul Kr. Ojha, S. M. and Zampieri, M. (2020). Evaluating aggression identification in social media. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, Paris, France, may. European Language Resources Association (ELRA).

Sarah T. Roberts, et al., editors. (2019). *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.

Samghabadi, N. S., Mave, D., Kar, S., and Solorio, T. (2018). Ritual-uh at TRAC 2018 shared task: Aggression identification. *CoRR*, abs/1807.11712.

Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

Sharifirad, S. and Matwin, S. (2019). When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. *CoRR*, abs/1902.10584.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, c. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.