Investigating Deep Learning Approaches for Hate Speech Detection in Social Media

Prashant Kapil[1], Asif Ekbal[1], Dipankar Das[2]

[1] Indian Institute of Technology Patna, India
[2] Jadavpur University Kolkata, India
{prashant.pcs17,asif}@iitp.ac.in
ddas@cse.jdvu.ac.in

Abstract. The phenomenal growth in the Internet has helped in empowering individual's expressions, but the misuse of freedom of expression has also led to the increase of various cyber crimes and anti-social activities. Hate speech is one such issue that needs to be addressed very seriously as otherwise this could pose threats to the integrity of the social fabrics. In this paper we proposed deep learning approaches for detecting various types of hate speeches in social media. Detecting hate speech from a large volume of text specially tweets which contains limited contextual information also poses a number of practical challenges. Moreover, the varieties in user generated data and presence of various forms of hate speech makes it very challenging to identify the degree and intention of the message. Our experiments on three publicly available datasets of different domains show the F1-scores in the ranges of 76-79%.

Key words: Hate speech detection, Deep Learning, Attention

1 Introduction

Social media is one platform which allows people across the globe to share their views and sentiments on various topics, but when it is intended to hurt some particular group or any individual then it is considered as hateful content. There is no such universally accepted definition of hate speech as it often varies across the different geographical regions. [1] stated that hate speech is an abusive speech with high frequency of stereotypical words. It is totally demographic dependent as some countries allow some speech to be said under *Right to speech*, whereas other countries adhere to very strict policy for the same message.

In recent times, Germany made a policy for the social media companies that they would have to face a penalty of 60$ millions if they failed to remove illegal content on time. Denmark and Canada have laws that prohibits all the speeches that contains insulting or abusive contents targeting minorities and could promote violence and social disorders. Indian government has also urged leading social media sites such as Facebook,Twitter to take necessary action against hate speech, especially those posts that hurt religious feelings and create social outrage. Setting aside legal actions our aim should be to combat with these speeches by agreeing to a set of standard definitions, guidelines and practices. [2] defined Hate speech as any communication that demean any person or any group on the basis of race, color, gender, ethnicity, sexual orientation, and nationality. Social networking sites like Twitter and Facebook are also taking preventive measures by deploying hundreds to thousands staffs to monitor and remove Offensive contents.

[3] collected messages from Whisper and Twitter to define hate speech as any offense motivated, in whole or in a part , by the offender's bias against an aspect of a group of people. They investigated main targets of hate speech in online social media and introduced new forms of hate that are not crimes but harmful. The detection can't be done manually, rather it needs a thorough investigation of the techniques and build robust techniques to accomplish this task.

The paper is structured as follows: We put discussion on the related works in Section 2. Section 3 describes our model architecture. Datasets are described in Section 4. Results along with the analysis are presented in Section 5. Section 6 presents error analysis to discuss the limitations of our proposed model. Finally, we conclude along with future work roadmaps in Section 7.

1.1 Motivation and Contribution

There has not been much research on hate speech detection because of non-availability of annotated datasets as well as lack of proper attention to this field. Its detection is challenging as these are highly contextual in nature, and poses several challenges concerning to the demographic characteristics and nature of the text. Same message can be posted in different ways, with one could be the potential candidate for hate speech, while other is not. Data imbalance also introduces challenges to build the robust machine learning model.

In this paper we propose a deep learning based approach for hate speech detection. We experimented with three publicly available benchmark datasets i.e **[4]**, **[5]** and **[6]**

2 Related Work

Most of the previous works done in this area have used different data sets. Researchers have mostly used traditional machine learning algorithms, and recently have started using deep learning. Lexical based approaches misclassifies any text containing particular slang as hate, effecting *right to freedom of speech* as the word used may have different meaning used in some different contexts.

[7] showed that Support Vector Machine (SVM) with word-n-grams employed with syntactic and semantic informations can achieve the best performance. [5] reported that using unigram, bigrams and trigrams features weighted with their TF-IDF values fed to Logistic Regression tends to perform best on their dataset by achieving 90% precision with hate class correctly predicted for 61% times. [8] classified ontological classes of harmful speech based on the degree of content, intent and affect it is creating on social media. [4] used critical race theory to annotate a dataset of 16K tweets that is made publicly available. They observed that geographic and word length distributions do not have much contributions in enhancing performance of the classifier. However, gender information combined with char-n-grams have shown a little improvement.

[9] used four types of features like n-grams, linguistic features, syntactic features and distributional semantic features to make distinction between abusive and clean data in finance and news data. [10] made use of various semantic, sentiment and linguistic features to develop a cascaded ensemble learning classifier for identifying racist and radicalized intent on Tumblr microblogging website. [11] provided solution to handle the issues of imbalanced dataset by including word knowledge in the form of Knowledge graph using text augmentation and text generation. [12] released a dataset of 10K sentences annotated into 2 labels: *Hate* and *Non Hate* extracted from Stormfront that includes the study of manual annotation and guidelines and checked their accuracy using several classification models like Support Vector Machine (SVM), Convolutional Neural Netwrok (CNN), Recurrent Neural Networ (RNN). [13] studied different forms of abusive behaviour and made public the annotated corpus of 80K Tweets categorized into 8 labels. [14] classified 2010 sentences using features like unigrams, sentiment features, semantic features and pattern based features. [15] proposed a CNN-GRU based architecture that showed promising results for 6 out of 7 datasets, outperforming other state-of-the-arts by 1-13 F1 points. They also released new dataset of 2435 Tweets focusing on refugees and muslims. [16] extracted a list of obscene words and hashtags using common patterns used in offensive and rude communications. [17] created vocabulary of *Hate* and *Non-hate* words with their best performing combination of feature groups was word2vec approach and Extended 2-grams. [18] applied bag-of-words model to learn binary classifier for the labels *racist* and *non-racist* and achieved 76% accuracy. [19] used a simple dictionary based approach to detect hate on swedish politicians. They examined different categories like anger, naughtiness, swearwords, general threats and death threats. [20] used the combinations of neural network based LSTM model with non-neural based GBDT representing words by random embedding, and achieved the best result on the dataset of [4]. The method proposed in [21] focused on detecting abusive language first and then classify into specific types of abuse. They showed that hybrid CNN i.e a combination of Char-CNN and Word-CNN perform best over Word-CNN and classical methods like Logistic Regression and SVM on dataset of 16K tweets by [4].

[22] did a comparative study on hate speech instigators and users who were targeted on twitter by studying distinctive characters and personality analysis of hate instigators. They observed that hate targets often have old accounts whereas instigators often have new accounts. [23] showed the concept of using CNN with random vectors, word vectors based on semantic information, word vectors combined with character 4-grams and compared the performance with each other. [24] provided analysis on the influence of annotator knowledge on the correct prediction. They also provided the annotation of 6909 tweets by expert and amateur.

3 Methodology

We develop multiple models based on deep learning: CNN, LSTM, Bi-LSTM, CNN-Attention, LSTM-Attention, Bi-LSTM Attention.
CNN: This model is based on the architecture by [25] that uses 5 main types of layers: **Input layer**,**Embedding Layer**, **Convolution layer**, **Pooling layer** and **Fully Con-**

nected layer.
Input Layer: All the sequences are converted to integer form where each token has been assigned unique index. The input sequences are then zero padded to have equal length as it helps in improving performance by keeping information preserved at the borders.
Embedding Layer: An Embedding is created for each sequence where each word w_i is mapped to real valued vector at the corresponding index in the embedding matrix using e(w_i), where *e* is the embedding matrix.
Convolution Layer: It is used to extract features for better representation of data using learnable filter of size i*h, where i=[2,3,4] in our experiments and h =300. Each filter is convolved through *i* words at a time and performs element-wise dot product to get a feature f_1. This process is repeated (n-h+1) times to get the feature map F = [f_1,f_2.....f_{n-h+1}]. *N* number of filters are used to get the different feature maps.
Pooling Layer: It reduces the spatial size of the representation helping in reducing overfit. Max pool takes the local maximum value from the feature map depending on the pool size whereas global max pooling takes the pool size equal to the size of the input.
Fully Connected layer: The vectorized form of features obtained from the last CNN layer is fed into the fully connected layer which has every input connected to every output by a weight. This is followed by softmax activation function that calculates the probability values for all the classes.The training of CNN is described as below.
Each sequence has actual class output A_o and predicted class output P_o.

The training of CNN proceeds by calculating *Training Error* = (A_o-P_o). The back-propagation looks for that solution of weights in the network that minimizes the training error using Delta Rule or Gradient Descent by [26]. The new weight in the network is updated as (New_weight = Old_weight - Derivative rate*learning rate). The weight will be updated till it approaches local minimum.

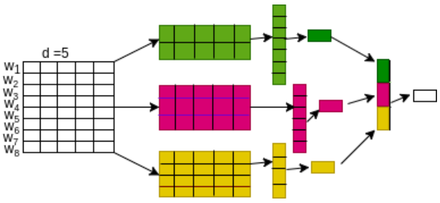

Fig. 1: **Architecture of CNN**

LSTM/BiLSTM: RNN is very suitable for sequence learning, time series but as it suffers from vanishing gradient and exploding gradient it does not perform well for the long-range dependency. So [27] introduced LSTM that is capable of learning long-range dependencies. The input sequence (i_1,i_2...i_n is transformed into its vector form of embedding size *e* which is then converted to h_1=(h_1^1,h_2^1...h_n^1 and transferred to the successive layers. It works by learning only the past information of the sequence, however

Bi-LSTM i.e a variant of LSTM comprises of 2 LSTMs to capture both past and future information. At each time step the hidden state at any time sequence is the concatenation of forward and backward states h_t=[$\overrightarrow{\mathrm{h}_t^1}$,$\overleftarrow{\mathrm{h}_t^1}$], hence the input passed to next layer is [e(w_1);h_1^1],[e(w_2);h_2^1],.....,[e(w_n);h_n^1]as the input to the next layer is the concatenation of all the previous outputs. The next layer output will be h_2 = (h_1^2,h_2^2....h_n^2). The input to the next layer will be [e(w_1);$h_1^1 h_2^1$,e(w_2);$h_1^2 h_2^2$...]. **Fig 2** shows the architure of BiLSTM.

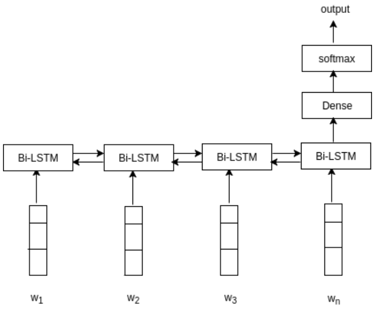

Fig. 2: **Architecture of BiLSTM**

Attention This mechanism expands the functionalities of neural network by paying attention to the specific parts of a sentence depicting human brain. We experimented with CNN, LSTM and Bi-LSTM based attention models. For CNN we used sentence level attention proposed by [28] that utilizes the concept of passing vectorized representation after pooling layer to the attention layer, which then learn the weighting for the important keywords and passes it to the softmax classifier to obtain the output. For LSTM and Bi-LSTM, our models are based on [29] that calculates the attention weight for the important words to form a representation of the sentences. Each word's hidden state representation (h_t) is passed through a learnable function a(h_t) to produce probability value α_1,α_1...α_n for each word. The sentence vector c is calculated by weighted average of h_t with weights of α. In **Fig 3** we explain the above idea.

$$e_t = tanh(Wh_t + b) \tag{1}$$

$$\alpha_t = softmax(e_t) \tag{2}$$

$$output = \sum_{t=1}^{t=n} \alpha_t h_t \tag{3}$$

4 Data sets

For the experiments, we use three types of datasets: **D1**,**D2** and **D3**. Table 1 shows the description of all the datasets with their total instances and the number of classes.

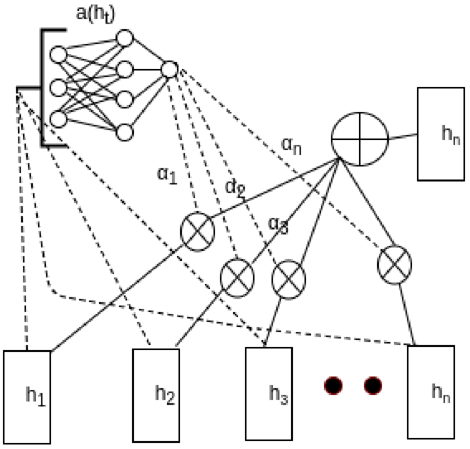

Fig. 3: **Architecture of Attention Model**

D1: This is the publicly available dataset with 16K Tweet IDs classified into three classes, *racism*, *sexism* and *neither* by [4]. As some of the tweets were deleted as well as due to account suspension of the users we were able to retrieve around 15,476 tweets.
D2:This dataset is divided into three classes *Hate*,*Offensive* and *Neither* by [5].
D3: This is the aggressive data classified into *Overtly*,*Covertly*and *Neither* by [6].
Table 2 shows some of the examples of various classes.

Table 1: **Details of the data set**

Dataset	**Total**	**Classes**	**Vocabulary Size**	**Test Data**
D1	15476	Racism(1923)Sexism(2871) Neither(10682)	12545	CV
D2	24783	Hate(1430) Offensive(19190)Neither(4163)	16362	CV
D3	12000	Overtly(2708)Covertly(4240) Neither(5052)	15830	CV

4.1 Parameter Tuning and Evaluation metrics

We use Keras with Tensorflow at the backend for our experiments. For every dataset we use 80:20 splitting with 80% for training and 20% for testing. Experiments were performed using stratified 5-fold cross validation to train all the classes according to their proportion. We report our results by standard Precision, Recall and F-score by averaging the cross-fold results. Categorical cross entropy loss function and Adam optimizer were used for training because the former is very effective on the classification task than the classification error and mean square error [30]. Hidden nodes in LSTM and Bi-LSTM layer were set to 100. For regularization dropout is applied on word embedding. In the experiment we have used publicly available Google wordvec by [31]. The batch

Table 2: **Example of a sentence of different classes**

Data Set	Original	Sentence
D1	Sexism	@Jack_McCormick1: I like my pickles like my women ,thin and cut.
D1	Racism	of course you were born in serbia...you're as fucked as A Serbian Film MKR.
D1	Neither	As long as she realizes she's not gonna look as pretty as she usually works.This character is a kind of mess.
D2	Offensive	Don't worry about the nigga you see, worry about the nigga you DON'T see... Dat's da nigga fuckin yo bitch..
D2	Hate	@Fit4LifeMike @chanelisabeth hoe don't make me put up screenshots of your texts to me hoe.
D1	Neither	Please women stay single please women when you commit to your man,commit to the gym as well.
D3	Covertly	Those who can't give justice to there women should not speak on others matter.
D3	Overtly	Thats why he always deserve slaps. ;).
D3	Neither	Yes women's must be given more powers and rights. Let's change the nation..

!

size of (16,32,64) and drop out of (0.1,0.2,0.3) were tested to build the model. The best accuracy was obtained between 4 and 10 epochs, batch size of 32 and drop out of 0.2.

4.2 Word embedding

It learns a real-valued vector representation for fixed size vocabulary from the corpus of text. Such vector representation has the advantage that different, semantically similar words may also end up having similar vectors[32]. There are two types of such embeddings: Continuous bag-of-words(CBOW) and Skip-gram model. In the CBOW architecture the model predicts the current word based on the context. In the Skip-gram model, the context words are predicted using the current word. We use the publicly available *word2vec* vectors trained on 100 billion words from Google news. All the out-of-vocabulary (OOV) words were assigned random weights in the range [-0.25, 0.25]. They were trained using CBOW architecture [31] and have dimensions of 300.

4.3 Pre-processing

As the Datasets have been crawled from social media, these contain noises and inconsistencies, such as slangs, misspelled words, acronyms etc. Hence we start by applying light pre-processing by expanding all apostrophes containing words and then removing characters like :, & ! ?. The tokens were also converted to lower-case for normalization. We also used dictionary to expand the misspelled words to its original form. All the word starting with # were normalized into words using Wordsegment in python For e.g. **#KillerBlondes** becomes *Killer blondes*, **#Feminism** becomes *feminism*, **atblackface** becomes *at black face* and **marriageequality** becomes *marriage Equality* etc. Emoticons were also replaced with tokens like Happy, sad, Disgust and Anger.

Table 3: **Results**

Model	D1			D2			D3			
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
CNN	80.73	79.22	79.70	78.19	71.30	72.09	57.04	56.15	56.06	57.43
LSTM	79.72	77.11	77.96	77.20	69.67	70.61	56.56	54.85	55.04	55.98
BILSTM	80.39	77.47	78.49	78.31	70.98	72.56	49.26	46.37	46.20	57.06
CNN-Atten.	81.36	76.47	78.37	78.88	70.98	72.56	56.77	55.79	55.90	57.93
LSTM-Attn	80.96	77.52	78.73	78.26	72.44	74.05	57.23	56.30	56.40	58.21
BILSTM-Attn	80.33	77.66	78.57	79.14	68.46	69.67	57.34	56.39	56.39	58.25
BoWV+SVM [20]	79.10	78.80	78.90	-	-	-	-	-	-	-
char-n-gram+LR [4]	72.90	77.80	75.30	-	-	-	-	-	-	-
LSTM+Random[20]	80.5	80.4	80.4	-	-	-	-	-	-	-
Logistic Regression [5]				91	90	90	-	-	-	-
Logistic Regression [33]	-	-	-	-	-	-	-	-	-	57.20
Multinomial NB [33]	-	-	-	-	-	-	-	-	-	56.86

Table 4: **Confusion Matrix for D1(CNN) and D2(LSTM Attn.)**

Class	Dataset 1			Class	Dataset 2		
	Racism	Sexism	Neither		Hate	Offensive	Neither
Racism	1509	13	401	Hate	457	849	124
Sexism	8	1921	942	offensive	354	18387	455
Neither	448	503	9731	Neither	57	318	3788

Table 5: **Confusion Matrix for D3(LSTM Attn.) and D3(Bi-LSTM Attn.)**

Class	Dataset 3			Class	Dataset 3		
	Overtly	Covertly	Neither		Overtly	Covertly	Neither
Overtly	1377	912	419	Overtly	1292	1071	345
Covertly	875	2128	1237	Covertly	805	2333	1102
Neither	362	1186	3504	Neither	353	1337	3362

Table 6: **Metric Values**

Class	True Positive	False Positive	False Negative
Racism	78.47	23.20	21.52
Sexism	66.91	21.17	23.08
Hate	31.95	47.35	68.04
Offensive	95.81	0.059	0.042
Overtly	50.84	47.32	49.15
Covertly	50.18	49.64	49.81

5 Results and Analysis

We compare our results with[4] and BoWV+Balance SVM in [20] and LSTM with Random Embedding in [20]. For D3 we are comparing our results against [33]. Of the 6 models we experimented with, CNN performed well for D1 , LSTM attention for D2 and BiLSTM Attention for D3. For D1 attention is performing good by giving good precision over previous Baseline models but our model is not at par on Recall and F-score. As the Dataset D2 do not have much comparitive study we are comparing our results with [5] in which they achieved Fscore of 90%. For D3 we have compared our results with the Best model [33] on Twitter dataset in TRAC 2018 on training data provided by the organizer. Our Deep learning model is outperforming all their baseline Linear model classifier on validation accuracy. We are reporting the Best model confusion matrix for each Dataset in Table 4 and 5. For D3 we have provided confusion matrix for both LSTM Attention and Bi-LSTM Attention.

6 Error Analysis

Error analysis was carried out to analyse the errors that were encountered in our system So we analyzed the best model confusion matrix as they were giving better performance. We perform quantitative analysis in terms of confusion matrix and qualitative analysis for analyzing the misclassified tweets.

6.1 Quantitative analysis

From Table **6** we can infer that identifying Hate, Overtly and Covertly classes posses more challenge than other subclass. Apart from data imbalance, using sarcastic phrase and racial epithets in a decietful manner makes it challenging for classifier to identify hate sentences that had 68.04% False positive Rate and with only 31.95 True Positive in D2. Due to some common obscene words between Hate and offensive classes 0.018% of Offensive instances converted to Hate. [5] reported that 76% of instances in offensive languages were choosen by 2 out of 3 annotators based on their world knowledge giving strong evidence of 95.81 True Positive value. For D1 the false negative between Racism and Sexism is only 21 instances indicating that people generally consider use of disparaging terms and abusive words as racism and sexual remarks pointing to any gender in a demeaning way is classified as Sexism . The conversion of sexism and racism to Neither class is 32.81% and 20.85% respectively. For D3 False negative between Overtly and Covertly is 912 and 875 respectively. This is because covertly and overtly class share lots of common words which are very crucial in classifying. Table **7** shows some of the highly frequency important keywords in each classes.

Table 7: **List of top occuring words in each class**

Class	High frequency words
Hate	b**ch, a**,nigger,f***ing,faggot,shit, trash, hate, kill, gay, ugly, queer, whitey
Offensive	b**ch, hoe, a**, ni**er, p***y, trash, wtf, crazy, stupid, p***s, gay, girl , hate
Racism	islam, religion, jews,women, war, christians, slave, terrorist, daesh, rape, beheaded
Sexism	sexist, women,girls,female, man, comedians, blondes, feminism, bitch , bimbos
Covertly	people,india,country,religious, party, political, muslims, fatwa, pakistan,modi,bjp
Overtly	people,india,religion,pakistan,bjp, muslims, hindu, terrorist, killed, fatwa

6.2 Qualitative analysis

For each data set we perform qualitative analysis to analyse the errors and we find that due to hate language being contextual in nature and also when the attack is directly or indirectly on women, then the model is showing poor performance. This suggests that it is indeed difficult for models to classify into fine grained labels. Table 8 contains some of the sentences converted to different classes due to system inefficiency. .

!

Data Set	Original	Predicted	Sentence
D1	Sexism	Neither	Please women stay single please women when you commit to your man,commit to the gym as well.
D1	Racism	Neither	@AdnanSadiq01 I think your goat is calling you. She is horny..
D1	Sexism	Racism	hate watch you have to be extra stupid to be a women and follow #Islam.
D1	Neither	Sexism	As long as she realizes she's not gonna look as pretty as she usually works.This character is a kind of mess.
D2	Hate	Offensive	@Fit4LifeMike @chanelisabeth hoe don't make me put up screenshots of your texts to me hoe.
D2	Hate	Offensive	@vinny2vicious faggot I knew you weren't really my friend.
D2	Hate	Neither	They should have never gave a cracker a transmitter!!!!!! @realdjTV will flip when he sees this.
D3	Covertly	Overtly	I told you wait.7 pak killed within hours of their cowardice act.Go and weep for them.
D3	Overtly	Covertly	yes we remember you are biggest terrorist country in the world you will do anything against humanity.

Table 8: **Example of a sentence predicted to different class**

7 Conclusion and Future work

In this paper we have explored the effectiveness of deep neural network for hate speech detection. The system failure on some cases highlights the subjective biases while classifying gender based message. Transfer learrning using large datasets can be very effective. Also some other linguistic features focused on gender and location will be used to improve the performance of the system. Some more other forms of Hate will also be considered.

References

1. William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.
2. John T Nockleby. Hate speech. *Encyclopedia of the American constitution*, 3:1277–79, 2000.
3. Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690, 2016.
4. Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.
5. Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.

6. Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*, 2018.
7. Hao Chen, Susan McKeever, and Sarah Jane Delany. Abusive text detection using neural networks.
8. Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. Degree based classification of harmful speech using twitter data. *arXiv preprint arXiv:1806.04197*, 2018.
9. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.
10. Swati Agarwal and Ashish Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*, 2017.
11. Borna Jafarpour, Stan Matwin, et al. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114, 2018.
12. Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.
13. Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393*, 2018.
14. Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835, 2018.
15. Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer, 2018.
16. Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, 2017.
17. Sebastian Köffer, Dennis M Riehle, Steffen Höhenberger, and Jörg Becker. Discussing the value of automatic hate speech detection in online debates.
18. Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.
19. Tim Isbister, Magnus Sahlgren, Lisa Kaati, Milan Obaidi, and Nazar Akrami. Monitoring targeted hate in online environments. *arXiv preprint arXiv:1803.04757*, 2018.
20. Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.
21. Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*, 2017.
22. Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. *arXiv preprint arXiv:1804.04649*, 2018.
23. Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

24. Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.
25. Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
26. Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.
27. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
28. Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*, 2015.
29. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
30. James D McCaffrey. Why you should use cross-entropy error instead of classification error or mean squared error for neural network classifier training. *Last accessed: Jan*, 2018.
31. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
32. Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.
33. Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vitor Beires Nogueira. 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)*, pages 28–41, 2018.