A Deep Learning Approach of Hate Speech and Offensive Language Detection on Twitter

Afaroz Alam
M.Tech Scholar
Department of Computer Science

Dharna Singhai
Assistant Professor
Department of Computer Science

Radharaman Engineering College
Bhadbbada Road , Ratibad ,Bhopal, MP
Bhadhbada, Road , Ratibad, Bhopal, M.P

Abstract- Recent years have seen an increase in the prevalence of hate speech, abusive language, misogyny, racism, cyberbullying, and other forms of abuse on Facebook, Twitter, and other social media platforms. People are more likely to propagate this type of action to disparage or damage someone's reputation. Such violent and offensive behaviour has grown enormously, as evidenced by. These occur as a result of people's freedom or openness to express themselves on social media platforms without fear or regard for the feelings of others. These platforms lack the capacity to effectively address the issue of online abuse, hate speech, and offensive language on their platform. Many other companies, research organizations are investing lots of money and research effort to curb this problem but they don't get much success because there is a need of great manual work to detect and remove online posts having hate speech or offensive language. The main challenge for automatic detection of hate speech on social media is to distinguish it from offensive language, cyberbullying and another form of abuses.

In our research, we introduce deep learning techniques to identify hate speech and objectionable language on Twitter. These techniques include CNN with global and average max pooling, CNN with dynamic convolution neural networks with k-max pooling, and multi-layer perceptrons. We tested these models experimentally using four publicly available Twitter hate and abusive datasets (largest twitter dataset till). On three of the four datasets, our model DCNN with k-max pooling and MLP produced state-of-the-art results. In general, our models performed better on these datasets and produced an excellent outcome when compared to earlier research on the same dataset.

Keywords— Hate speech, offensive language, sexism, Dynamic Convolution Neural networks with k-max pooling, Multi-layer perceptron, CNN.

I. INTRODUCTION

Profanity, swear words, curse words, crude, harsh, bad, and other terms are also used to describe offensive language. According to studies, an average person uses 80 to 90 offensive words daily in conversation (0.5% to 0.7% of all words). Researchers from [9][10] discovered that people use derogatory language, such as cursing, online when tweeting about the unpleasant feelings of melancholy (21.83%) and rage (16.79%).

Hate speech is any speech that targets an individual or group of individuals based on attributes such as race, religion, ethnic origin, national birthplace, sex, disability, sexual orientation, or sex personality. A few countries' laws define hate speech as any verbal, written, or visual expression that incites violence or prejudice toward a protected group or individual because of that person's membership in the group, or because it disparages or threatens that group or individual because of that membership. The law may identify a guaranteed gathering based on certain characteristics. Any encouragement of national, racial, or religious hatred that results in animosity or violence is prohibited, according to the International Covenant on Civil and Political Rights (ICCPR). ICERD forbids all provocation of racism.

On 31st May 2016, many IT companies like Facebook, Google, Microsoft, and Twitter, collectively admit to a European Union code of conduct committing them to review "the majority of valid notifications for removal of illegal hate speech" posted on their websites within 24 hours.

Before 2013, Facebook has changed in their hate speech policies when they are pressurized by more than 100 advocacy group over data released by Facebook that promote domestic sexual violence against women

Since hate speech and inflammatory language on social media platforms have a negative impact on our society and occasionally on an individual, Twitter, Facebook, and many other companies have invested heavily in research and development to address this issue. Even if they put in a lot of effort, they are nevertheless blamed for not working hard enough because it takes a lot of human labour to check online postings, identify offensive or hateful content, and then remove it.

II. OVERVIEW OF WORK

We want to automatically identify and categorise hateful and derogatory words used on Twitter in this project endeavour. We suggested using Dynamic CNN to categorise hate speech and profanity in tweets. We also suggested a deep learning method based on MLP to identify hate speech on Twitter. On the Twitter dataset, we also used a CNN-based architecture with global average pooling. In comparison to earlier trials, we obtained state-of-the-art results on each of the four datasets we used.

There are various deep learning algorithms have developed that can learn complex pattern in the dataset.

- Multilayer Perceptron (MLP)
- Convolutional Neural Network (CNN)
- Long Short-Term Memory (LSTM)
- Variational Autoencoders (VAEs)
- Generative Adversarial Networks (GANs)

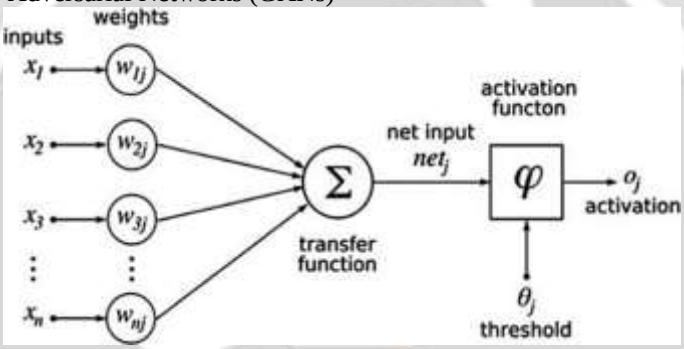

Fig-i: Multilayer Perceptron Architecture

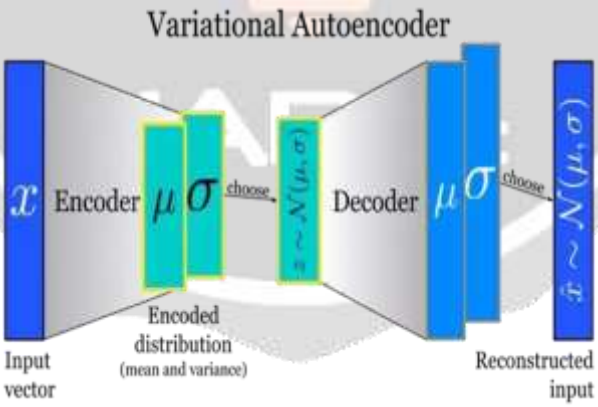

Fig.ii. VAE network architecture

III. OBJECTIVE OF THE WORK

The salient objectives of the study have been identified as follows:

The main objective of the project is to the automated approach of hate speech and offensive language detection on twitter. Automated means “using no or minimal human intervention under a controlled environment”. Automated detection corresponds to automated learning such as machine learning: supervised and unsupervised learning. We use a supervised learning method to detect hate and offensive language.

IV LITERATURE REVIEW

Hate speech, offensive language, cyberbullying and online abuse are impacted our society on a large scale in the recent time. So, there is a need of scalable, automated approach of hate speech and offensive language detection.

[Burnap P and Williams M. "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy and Internet" a supervised method of Cyber Hate speech detection on Twitter is proposed, they used extensive feature selection such as First they selected all the derogatory, expletive words used against a specific community like Muslims, blacks and made a BoW (Bag of Words) features, second, they used POS tagging of words from all the sentences which reflects sociological and common-sense reasoning shown in various instances of cyber hate speech sample. Since Rule based approach were use before that to classify cyber hate speech which was not much effective, they used support vector machines (SVM), in which the feature vector was plotted in high dimensional space and hyperplane diveded the space in such a way that all tweets belonging to "Yes" and "No" were separated. The hyperplane used was not too optimum to maximize the width of the plane and classify the tweets more accurately.

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, and Bhamidipati N. "Hate speech detection with comment embeddings" In Proceedings of the 24th International Conference on World Wide Web. they proposed distributed lower-dimension representation of comments by using neural language model like Bow (Bag of words), TF and TF-IDF, and paragraph2vec to detect Hate speech. They solved high dimensional data representation problem while classification but did not get very good accuracy of detecting hate speech.

Racist and Sexism commonly used hate speech was detected in [8] and [9]. They selected most important features by searching across character n-gram (one-gram, two-gram, tri- gram and four-gram) and performed 10-fold cross-validation to evaluate model. They also considered meta information of users like Gender of user, average length of 1-4 words per tweet, Gender + Location and Gender+Location+Length.

Davidson T, Warmsley D, Macy M, and Weber I. "Automated hate speech detection and the problem of offensive language"; a supervised method of automatic hate speech and offensive language detection is proposed, in this they used logistic regression with L2 regularization to overcome the overfitting and dimensionality reduction of data. They tested their baseline against Naïve Bayes, Decision Tree, Random-forest and Linear SVM. This was first lexicon based multi- class hate and offensive language detection method that was given very good results while automatic detection, but some-times it miss-classify offensive language as hate speech.

Lozano E, Cede˜no J, Castillo G, Layedra F, Lasso H, and Vaca C. 2017 "Requiem for online harassers: Identifying racism from political tweets"; an unsupervised method of hate speech like racism and sexism detection is proposed, they tried to find racist user as well as user who pass sexist comment on twitter during US election 2016. They used clustering to classify the racist and sexist tweet. They also clustered users who favor Donald Trump and spread racism and sexism as well as Hillary Clinton's supporter who spread racism and sexism on twitter during campaign..

Park H. J. and Fung P. "One-step and two-step classcation for abusive language detection on twitter"; They proposed a two-step method of doing classification on offensive language and a one-step method of performing one multi-class classification of detecting racism and sexism. To performed this, they used HybridCNN in one-step and Logistic regression in two-steps method. The HybridCNN made up of a combination of CharCNN and WordCNN

Zhang Z, Robinson D and Tepper J, "Detection Hate Speech on Twitter Using a Convolution-GRU based DNN" they have given a state of art technique for detecting hate speech on 7 different datasets of tweet. They employed CNN+GRU network architecture on these datasets to classify the hate speech as Racism and Sexism

against refugee Muslim in UK. They performed comparative evaluation on largest publicly available dataset and found that proposed method outperformed on all the baselines and is a state of the art among all.
We extended this work on 4 publicly available hate and offensive language dataset on twitter. In this, they used 300-dimensional GloVe word embedding, followed by 1D convolution and 1D Max-pooling, the result is further passed through a GRU then Global max-pooling is done then final soft-maxis applied to do classification.

Founta M. A., Chatzakou D, Kourtellis N, Blacknurn J, Vakali A, Leontiadis I, "A Unified Deep Learning Architecture for Abuse Detection"; they used tweets as well as meta data like user information, time of retweet etc. They applied RNN on text to do feature extraction but final classification is postponed till meta-data passed through an MLP network. The feature matrix of tweets and metadata are concatenated and then final classification was done.

Jha A, and Mamidi R. 2017. "When does acompliment become sexist? analysis and classification of ambivalent sexism using twitter data"; they used FastText classifier made by Facebook AI research team. They focused on the different form of Sexism named as Benevolent, which is very common on social media platforms. They first analyzed tweeter dataset posing sexism and classified it into three classes 'Hostile', 'Benevolent' and 'None' depending on the sexism type that it represented by using SVM. They also used sequence 2 sequence model by using TF- Sec2Sec framework for Tensorflow.

Greevy E and Smeaton A F. "Classifying racist texts using a support vector machine";they proposed a supervised method SVM to classify racist texts from different web pages. They crawled 3 million words formed a corpus. They divided it into four sets of varying size datasets with equal contribution of racist and non-racist words that can be shown in Table 1.

	Set 1	Set 2	Set 3	Set 4
# Documents in Train set	200	400	600	800
# Documents in Test set	60	100	150	200

Table 2.1: The size of train and test set in dataset

They applied BoW and Bi-grams to extract features from each of the four datasets and used SVM for classification of racist text. They found that BoW gave high precision of about 92.55% and recall of 87.00% on set-3. In Bi-grams model precision increased up to 100% on set-1 but recall decreased drastically below 75%.

V. NEED & MOTIVATIONBB

Social media like Twitter and Facebook provided us with a platform to express our views or opinion on any topic across the globe. Some people use it as a tool to defame, tarnish some one's image, spread rumor, hate and offensive language against a specific group or community. Some politicians use it to spread his propaganda, lies to influence and polarize voters towards them.

Twitter, Facebook tries to curb these but they don't get much success because it requires a lot of manual work to identify the post as hate/offensive. Automatic detection of hate/offensive post on social media is a difficult task.

VI. MODEL & DATASET PREPARATION

Dynamic Convolution neural network is a type of CNN with wide convolution [10]. It uses dynamic k-max pooling as a pooling layer. In this architecture, the size of the feature map at hidden layer changes according to the length of the sentence. Figure 1 represents Dynamic CNN

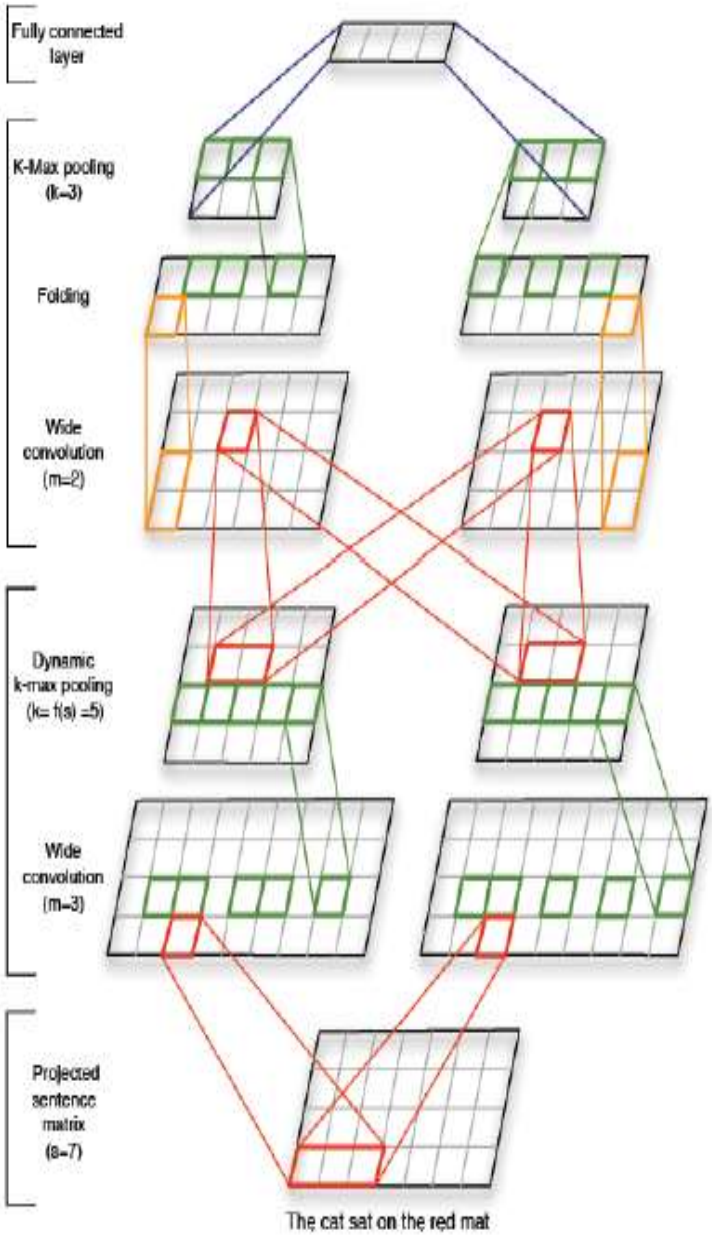

Fig.iii. A Dynamic CNN having seven-word input sentence. Word embeddings have size d=4. Two feature maps have been used for two convolution layers

We used multi-layer perceptron as a deep learning model for hate speech and offensive language detection as shown below:

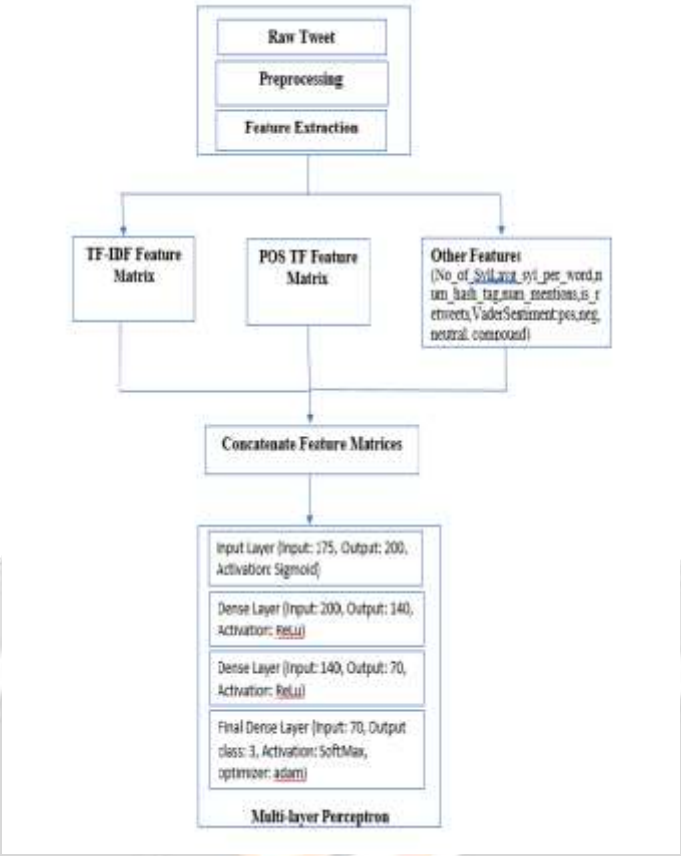

Fig. 4.2. shows the MLP based hate speech and offensive language detection model architecture

We do following preprocessing steps to clean the raw text:
1. Convert texts into lower case and remove all the stop words.
2. Remove unwanted symbols such as: & ! / \ ? & $; etc. using regular expressions.
3. Stemming and lemmatization.
4. Remove tokens having document frequency less than 5, which further removed sparse features which is less informative.
5. Further, we normalize the words like '#Hatespchaganstmslim' to 'Hate speech against Muslim', 'gooooood' to 'good' etc.

We used 4 publicly available Twitter dataset of Hate and offensive language for evaluation on our model. We crawled tweets specific to our problem using publicly available Twitter data set in the form of Tweet-id and label. We crawled tweets in the form of texts corresponding to each tweet-id and merge it to make a new dataset of tweets and label. As we know these are the most widely used dataset used in various research work

Dataset	**No of Tweets**	**Classes (%Tweets)**	**Target Class**
Hate(DT)	24,783	Hate (11.6%), offensive (76.6%), Neither (11.8%)	Hate, Offensive
WZ-LS	18,595	Racism (10.6%), Sexism	Racism, Sexism

		(20.2%), None (68.8%)	
WZ-L	16,093	Racism (12.01%), Sexism (19.56%), None (68.41%)	Racism, Sexism
WZ-S.exp	6,594	Racism (1.2%), Sexism (11.7%), both (0.53%), None (84.37%)	Racism, Sexism

VII. RESULT

on dataset DT DCNN and MLP both perform very well and given same accuracy of about 92%, while previously only Z Zhang [29] implemented CNN+GRU model and got an accuracy of 94% on this dataset and Davidson [15] got 87% accuracy by using SVM

Dataset	**SVM**	**MLP**	**CNN**	**DCNN**	**State of art**
DT	0.87	**0.92**	**0.91**	**0.92**	0.94 CNN+GRU, Zhang [29], 0.87 SVM Davidson [15].
WZ-LS	0.73	**0.82**	**0.82**	**0.83**	0.82 Park [34], WordCNN 0.81 Park [34], CharacterCNN 0.83 Park [34], HybridCNN
WZ-L	0.74	**0.82**	**0.82**	**0.83**	0.82 CNN+GRU, Zhang [29], 0.74 Waseem [19], best F1
WZ-S.exp	0.89	**0.93**	**0.90**	**0.92**	0.92 CNN+GRU, Zhang [29], 0.91 Waseem [19], `Best' features

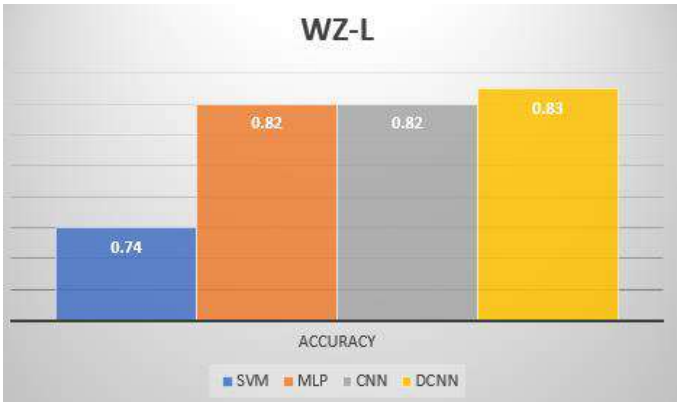

Fig. iv. Accuracy on WZ-L Dataset

Above fig shows DCNN performs well and given best result of 83% on this WZ-L dataset as compared to MLP (82.67%)and CNN* (82%) and SVM (74%.). We got the maximum accuracy of 83% as compared to previous best by Zhang [29] of about 82%.

Fig. v. Accuracy on WZ-S.exp Dataset

Above Fig shows that MLP outperformed on WZ-S.exp and gives a maximum accuracy of 93% as compared to DCNN (92%), CNN* (90%) and SVM (89%). Previously Zhang [29] got maximum accuracy on this dataset of about 92% by using CNN+GRU

illustrates the performance of our proposed model DT dataset. We got the highest precision of 0.95 for class offensive and least precision of 0.82 for class Hate. We got the highest recall of 0.96 for class offensive and least recall of 0.28 for class Hate. We got a highest F1 score of 0.96 for class offensive and a least F1 score of 0.42 for class hate. We got an average precision of 0.92, average recall of 0.92 and average F1 measure as 0.91 for dataset DT

class	Precision	Recall	F1
Hate	0.60	0.52	0.56
Offensive	0.95	0.80	0.87
Neither	0.87	0.91	0.89
Overall	0.92	0.91	0.92

Our research work proposed a new dynamic CNN based deep learning approach for detection of Hate and offensive language on Twitter. We proposed another feature concentric and deep learning MLP based approach. Dynamic CNN along with k max-pooling helped in extracting the k most active feature while preserving the order of the features. We also applied two other baseline models SVM and CNN with a combination of global max pooling and global average pooling. These two are used to compare the performance of our proposed model. DCNN outperformed on these four datasets and gives the highest accuracy of 92% on WZ-S.exp and DT. We got an accuracy of 83% on WZ-L and WZ-LS datasets.

VIII FUTURE WORK

As we used Tweets only 4-datasets for our work evaluation, we can further use metadata of tweets. If we use metadata based on networks and users like #followers and #freinds, strength and effect of friends, the effect of mentions on a user, #posts, favorite tweets etc. along with tweets. We can make a hybrid (CNN + MLP) model for classification purpose. We can pass tweets to CNN and metadata to MLP parallelly. The result of these two models will be concatenated and will be passed through a dense layer for final classification.
We can also performed the same work for other languages like Hindi, Chinese, French and Code mixed as well. Since codemixed languages are very popular on social media especially in India, Pakistan. So we can use our model for the detection of hate speech and offensive language in code mixed language (English+Roman/Urdu).

REFERENCES

[1] Nockleby, John T. (2019), "Hate Speech" in Encyclopedia of the American Constitution, ed. Leonard W. Levy and Kenneth L. Karst, vol. 3. (2nd ed.), Detroit: Macmillan Reference US, pp. 1277–79. Cited in "Library 2.0 and the Problem of Hate Speech," by Margaret Brown-Sica and Jeffrey Beall, Electronic Journal of Academic and Special Librarianship, vol. 9 no. 2 (Summer 2008)

[2] "Criminal Justice Act 2018". www.legislation.gov.uk. Retrieved 2017-01-03.

[3] An Activist's Guide to The Yogyakarta Principles (PDF) (Report). November 14, 2010. p. 125.

[4] International Covenant on Civil and Political Rights, Article 20.

[5] Convention on the Elimination of All Forms of Racial Discrimination, Article 4.

[6] "Facebook, YouTube, Twitter and Microsoft sign EU hate speech code". The Guardian. Retrieved 7 June 2016.

[7] Sara C Nelson (28 May 2013). "#FBrape: Will Facebook Heed Open Letter Protesting 'Endorsement Of Rape & Domestic Violence'?". The Huffington Post UK. Retrieved 29 May 2013.

[8] Rory Carroll (29 May 2013). "Facebook gives way to campaign against hate speech on its pages". The Guardian UK. Retrieved 29 May 2013.

[9] "#Cursing Study: 10 Lessons About How We Use Swear Words on Twitter". Retrieved 2015-01- 05.

[10] "Cursing in English on Twitter". Retrieved 2015-01-05.

[11] Iginio Gagliardone, Danit Gal, Thiago Alves, Gabriela MartinezJ," countering online hate speech", in United Nations Educational, Scientific and Cultural Organization 7, place de Fontenoy, 75352 Paris 07 SP, France.

[12] Greevy E and Smeaton A F. "Classifying racist texts using a support vector machine"; In Proceedings of the 27th Annual International ACM SIGIR Conference on Research andDevelopment in Information Retrieval SIGIR '04, pages 468–469, New York, NY, USA, 2004. ACM

[13] Burnap P and Williams M. "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy and Internet", 7(2):223–242, 2015

[14] Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, and Bhamidipati N. "Hate speech detection with

comment embeddings" In Proceedings of the 24th International Conference on World Wide Web, pages 29–30. ACM, 2015.

[15] Davidson T, Warmsley D, Macy M, and Weber I. "Automated hate speech detection and the problem of offensive language"; In Proceedings of the 11th Conference on Web and Social Media. AAAI, 2017.

[16] Kwok I and Wang Y. "Locate the hate: Detecting tweets against blacks"; In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI'13, pages 1621–1622. AAAI Press, 2013.

[17] Mehdad Y and Tetreault J. "Do characters abuse morethan words?" In Proceedings of the SIGDIAL 2016 Conference,pages 299–303, Los Angeles, USA, 2016. Association for Computational Linguistics

[18] Warner W and Hirschberg J. "Detecting hate speech on the world wide web"; In Proceedings of the Second Workshop on Language in Social Media, LSM '12, pages 19–26. Association for Computational Linguistics, 2012

[19] Waseem Z and Hovy D. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter";In Proceedings of the NAACL Student Research Workshop, pages 88–93. Association for Computational Linguistics, 2016.

[20] Waseem Z. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter"; In Proc. of the Workshop on NLP and Computational Social Science, pages 138–142. Association for Computational Linguistics, 2016.

[21] Xiang G, Fan B, Wang L, Hong J, and Rose C; "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus"; In3` Conference on Information and Knowledge Management, pages 1980–1984. ACM, 2012.

[22] Yuan S, Wu X, and Xiang Y; "A two phase deep learning model for identifying discrimination from tweets"; In Proceedings of 19th International Conference on Extending Database Technology, pages 696–697, 2016.

[23] Xiang G, Fan B, Wang L, Hong J and Rose C. 2012. "Detecting offensive tweets via topical feature discovery over a large-scale twitter corpus"; In 21st ACM CIKM,1980–1984

[24] Clarke I, and Grieve J, 2017. "Dimensions of abusive language on twitter"; In Proceedings of the First Workshop on Abusive Language Online, 1–10.

[25] Mehdad Y, and Tetreault J R. 2016. "Do characters abuse more than words?" In SIGDIAL, 299– 303.

[26] Lozano E, Cede˜no J, Castillo G, Layedra F, Lasso H, and Vaca C. 2017 "Requiem for online harassers: Identifying racism from political tweets"; In 4th IEEE Conference on eDemocracy & eGovernment (ICEDEG), 154–160.

[27] Jha A, and Mamidi R. 2017. "When does acompliment become sexist? analysis and classification of ambivalent sexism using twitter data"; In 2nd Workshop on NLP and Computational Social Science, 7–16.

[28] Park H. J. and Fung P. "One-step and two-step classcation for abusive language detection on twitter"; In ALW1: 1st Workshop on Abusive Language Online, Vancouver, Canada, 2017. Association for Computational Linguistics.

[29] Zhang Z, Robinson D and Tepper J, "Detection Hate Speech on Twitter Using a Convolution-GRU based DNN" In 15th ESWC 2018 conference on Semantic web.

[30] Joulin A, Grave E, Bojanowski P, and Mikolov T. 2016. "Bag of tricks for efficient text classification"; arXiv preprint arXiv:1607.01759

[31] Founta M. A., Chatzakou D, Kourtellis N, Blacknurn J, Vakali A, Leontiadis I, "A Unified Deep Learning Architecture for Abuse Detection"; In 32nd AAAI conference on Artificial Intelligence Hilton New Orleans Riverside, New Orleans, Louisiana, USA 2018, vol = abs/1802.00385

[32] D. Britz, A. Goldie, T. Luong, and Q. Le. 2017. Massive Exploration of Neural Machine Translation Architectures. ArXiv e-prints .

[33] Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[34] Park J.H. and Fung P. "One-step and two-step classi_cation for abusive language detection on twitter", In ALW1: 1st Workshop on Abusive Language Online, Vancouver, Canada, 2017. Association for Computational Linguistics.

[35] A. Krizhevsky, I. Sutskever, G. Hinton,"ImageNet Classification with Deep Convolutional Neural Networks," In Proceedings of NIPS 2012.

[36] https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

[37] P.W.D. Charles, Project Title, (2013), GitHub repository, https://github.com/charlespwd/project- title

[38] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In Proceedings of NIPS, pages 2672– 2680, 2014. papers.nips.cc/paper/5423-generativeadversarial-nets.pdf.

[39] Kim Y, "Bengio et al. Word vectors, wherein words are projected from a sparse,1-of-V encoding (here V is the vocabulary size", In arXiv:1408.5882v2 [cs.CL] 3 Sep 2014.

[40] Kalchbrenner N, Grefenstette E., Blunsom P. "A Convolutional Neural Network for Modelling Sentences", In arXiv:1404.2188v1 [cs.CL] 8 Apr 2014.