Machine Learning Techniques for Hate Speech Detection on Social Media

Chandradeep bhatt
CSE Department
Graphic Era Hill University
Dehradun, India
bhattchandradeep@gmail.com

Nancy Saini
CSE Department
Graphic Era Hill University
Dehradun, India
Saininancy532@gmail.com

Rahul Chauhan
CSE Department
Graphic Era Hill University
Dehradun, India
Chauhan14853@gmail.com

Ashok Kumar Sahoo
CSE Department
Graphic Era Hill University
Dehradun, India
ashoksahoo2000@yahoo.com

2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT) | 979-8-3503-0336-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CISCT57197.2023.10351228

***Abstract*— social media are computer-based and internet-based technologies that offers a platform for the concept and distribution of information, thoughts, concerns, and point of view on somethings going in the world and expression through virtual communities and networks. People use social media like Instagram, Facebook, Twitter for sharing their views and their perspective towards things happening all around the world. Some people make offensive comments and videos on these platforms that may affect the other people and can change into conflicts. Hate Speech is one of the major reasons of the conflicts in the world. Conflicts can be in the individual-level or country-level.**

Keywords- Machine Learning, hate speech, twitter, social media, support vector machine.

I. INTRODUCTION

Social media became popular in past some years. Social Media started as an approach to work together with friends and family but was later acknowledged by business that wanted to gain popularity using popular interaction technique to reach out their consumers online. It helps them to work all around the world. It has become a network or platform for public for sharing and expressing ideas, interest's despite of the socio-economic boundaries. People use social media like Instagram, Facebook, Twitter for sharing their views and their perspective towards things happening all around the world. Some people use to make offensive comments and videos on these platforms that may affect the other people and can change into conflicts. Avoiding some post could push harmful problems which could change into devastating happenings at a particular point which means it can take place as a conflict on country-level or community-level. There are many cases where violent statements have made severe fear to the society inconvenience. To recognize such issue is today's need and considerable work. The figure 1 shows the different types of hate speech.

-

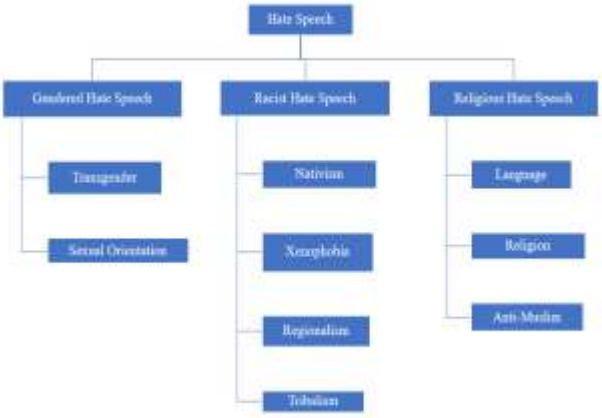

Fig. 1. Type of Hate Speech

The table I shows the percentage of speech on social media from the year 2012 to 2022.

TABLE I. HATE SPEECH ON SOCIAL MEDIA IN THE YEAR 2012, 2018 & 2022

Among 13–17-year-old social media users, percent who encounter different type of hate speech	**2012**	**2018**	**2022**
Often / Sometimes			
Racist	43%	52%	57%
Sexist	44%	52%	56%
Anti-Religion	34%	46%	54%
Homophobic	43%	52%	56%
Any of the above	37%	64%	66%
Often			
Racist	13%	12%	10%
Sexist	15%	14%	12%
Anti-Religion	10%	11%	12%
Homophobic	16%	12%	10%
Any of the above	24%	21%	19%

'Hate speech' does not mean to insulting, or foul-mouthed speech aimed at people. It is speech that can produce genuine material damage through the social,

economic, and political marginalization of society. 'Hate speech' must be recognized as linked to universal discrimination and eventual political marginalization of a society. Hate speech can be distributed offline or online using any form of representation, including images, cartoons, memes, objects, gestures, and symbols. It is abusive, bigoted, vile, or demeaning to an individual or group; it is biased; it is fanatical; it is intolerant.

II. RELATED WORK

For detecting hate speech many authors successfully proposed various methods. This section presents some work done by several authors in past years. A Linear SVC was employed by Malmasi and Zampieri [3] on word skip grams, brown clusters, and surface n-grams. Models from the SVM, bi-LSTM, and neural network are used by Arup Baruah et al. [4]. The TF-IDF properties of character and word n-grams and embeddings are used in this model. SVM and NB were employed by Asogwa D.C. The total number of 12,495 examples used in this model is 12. SVM's accuracy is 99.37%, and NB's accuracy is 50.00%. SVM offers more accurate results than NB. Aman Kharwal used machine learning to design a system. F1 score of 0.96 per cent which is generally appreciable. . R. chakravarti created a model in which hate speech is detected using three algorithms of Machine Learning i.e., SVM, Random Forest and LR. This model works on two languages i.e., English and Malayalam. The accuracy of model for SVM is 90% for English and 93.75% for Malayalam, LR is 81% for English and 90% for Malayalam, Random Forest is 86% for English and 92% for Malayalam [5]. That shows SVM gave highest accuracy among SVM, LR and Random Forest. Varsha Pathak [2] used SVC, LR, MNB, RFC and Ensemble on word n-grams, char n-gram and combine char and word n-grams. This model classified hate speech on two languages i.e., Malayalam and Tamil. The accuracy for Malayalam language of SVC is 74%, MNB is 76.08%, LR is 75%, RFC is 70% and Ensemble is 75%. The accuracy for Tamil language of SVC is 86%, MNB is 85%, LR is 86%, RFC is 81% and Ensemble is 86%.

Ching Seh Wu [6] used MNB, Linear SVM, RNN, Random Forest. They perform two experiments, one for classification into normal videos and hateful videos and another for classification into normal and racist or sexist videos. The accuracy of experiment one for MNB is 85.12, Linear SVM is 89.29, RFC is 94.64 and RNN is 80.36. The accuracy of experiment two for MNB is 79.76, Linear SVM is 82.14, RFC is 85.71 and RNN is 80.36. Mujadia et al. [9] used SVM, RF, Adaboost, voting classifiers and LSTM. The model works on dataset of Twitter and on language English, German, and Hindi. The accuracy is 69.70% for English, 47.7% of German, 80.32% for Hindi language. Sigurbergsson [10] used LSTM, bi-LSTM. The model works on dataset of Facebook and Reddit and on language English and Danish. The accuracy is 74% for English and 70% for Danish.

Altin [11] used bi-LSTM algorithm. The model works on dataset of Twitter and on language English. The accuracy is 82.9%. Umar [12] used user profiling algorithm and deep LSTM. the model works on data set of Twitter and on language English. The accuracy is 89.14% for categorizing offensive and 83.33% for recognition of user's participation. Garain [13] used Neural network algorithm. The model works on dataset of Twitter and on language English. The accuracy is 57.3% for hate speech recognition and 76.3% for aggressive recognition. MacAvaney [14] used Multiview SVM. The model works on dataset of Facebook and Stormfront on language English. The accuracy is 53.68% for Facebook and 80.33% for Stormfront. Naveen [15] used LR, NB and Linear SVM. The model works on Twitter and on Language English. The accuracy is 95.6%. Sadiq [16] used LSTM with CNN & CNN bi LSTM in Deep Neural Network. The model works on Twitter and on Language English. The accuracy is 92%. Jain [17] used bi-LSTM and CNN. The model works on Twitter and on language English and Hindi. The accuracy is 92.71% in English and 89.05% in Hindi. Kapil [18] used CNN+ Gated Recurrent Unit (GRU). The model works on Facebook and other social media and on Language English. The accuracy is 80.7% for recognition on Facebook, and 86.52% for recognition on more social media stages. Salminen [19] used SVM, NB, LR, XGBoost and Neural Networks. The model works on dataset of YouTube, Wikipedia, Twitter, and Reddit and on language English. The accuracy is 92%. Chopra [20] used CNN, bi-LSTM. The model works on dataset of Twitter and on language Hindi, English code switched. The accuracy is 73%. Modha used Linear SVM, LR and CNN. The model works on dataset of Twitter and Facebook and on language English and code-mixed Hindi. The accuracy is 64% for detection on Facebook and 58% for detection used NB, SVM, CNN, Random Forest. The accuracy is 62.33% of NB and 71.77% of SVM, 82.62% of CNN and 52.20% of Random Forest [21]. Table II shows the summary of literature review and methods used in past years to detecting the hate speech.

III. CLASSIFICATION TECHNIQUES

A. NB Classifier Algorithm

One extremely practical Bayesian learning method is the NB learner, also called the NB classifier. This is a probabilistic non-linear ML model used for classification task. Its memory requirement is very low. It handles missing feature values. It is divided into 3 types: Multinomial NB, Gaussian NB, Bernoulli NB. The NB classifier is established on the make simpler idea that the attribute principles are provisionally individual given the direct value. Figure 2 shows different types of naïve bayes algorithm.

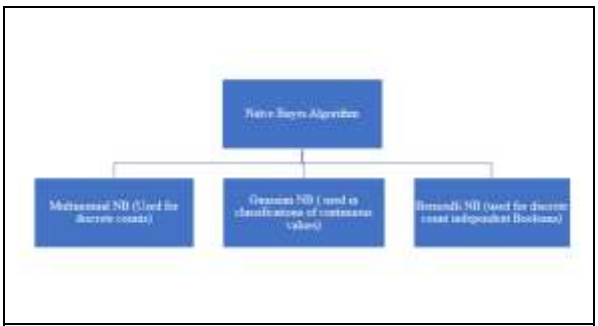

Fig. 2. Classification of Naïve Bayes Algorithm

B. SVM Algorithm

A supervised learning technique called support vector machines is used to solve classification and regression problems. However, mainly, it is used for classifications problems. It is a strong kernel method that can be used to

tackle high dimensional problems. They are also good against overfitting due to their margins. The purpose of the SVM process is to produce a line that can separate n-D space into two different classes or groups. This line is known as hyperplane. This algorithm chooses peak points that help in creating the hyperplane. Figure 3 show types of SVM.

Fig. 3. Classification of SVM

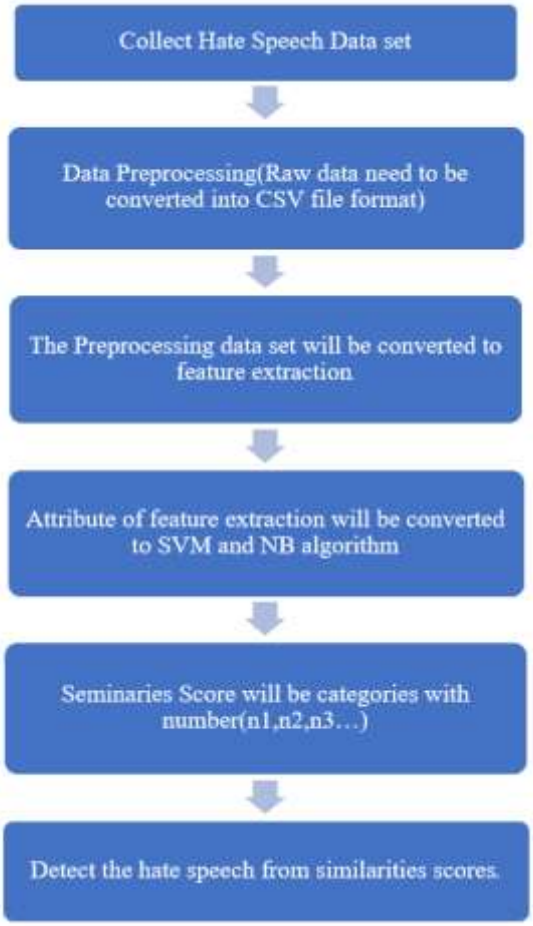

Fig. 4. Flowchart of work

Figure 4 explain all steps used for classification. The work consists of several steps and the first step is data preprocessing. Data Preprocessing is a way used to change collected data into a clean and useful dataset. The data which is collected from the different resources is not good for analysis. So, we use this technique for clearing duplicate data and useless data. In general, social media text does not obey grammatical rules and are written in more than one language. So, we must use the data preprocessing for preparing the data before training and testing of model.

IV. DISCUSSION

Support Vector Machine, Logistic Regression, and Naive Bayes were the three methods that were used in the paper. I used the Gaussian Naive Bayes method for Naive Bayes since it employs the Gaussian normal distribution and can be applied to continuous data. The Support Vector Machine method produces the best results of the three. Using SVM, Logistic Regression, and Naive Bayes, the accuracy is 94.4%, 94.02, and 47.77%, respectively. Figure 5 depicts this result as a graph.

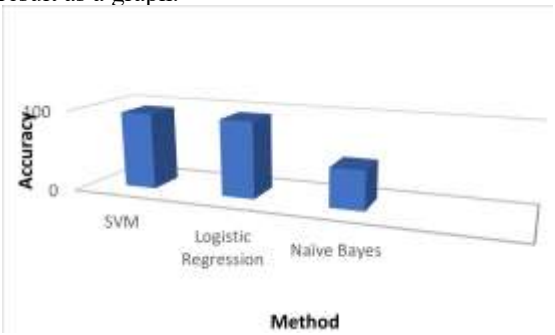

Fig. 5. Accuracy obtained by different method

V. CONCLUSION

As we see till yet, models are made using machine learning algorithm like SVM, LR, MNB. SVM is the best algorithm based on accuracy for hate speech detection. In future, we can use Deep learning algorithms like NLP. Till yet most of the researcher works on already existing dataset but later we will work on the running data. Upcoming effort should seek to apply unsupervised or semi-supervised methods on the way to create fusion DL paradigm i.e., able to automatically identify, inform, and prevent hate speech on social interacting programs. The semi-supervised application will deal with the hassles of social media displays including emojis, additional types of acronyms that does not fit the English speech specification. Suitably, for detection of hate speech chores the semi supervised learning methods can impact more because it will be able to manage huge amounts of unlabeled information for model training and checking. This paper also proposes that while there is no secure language rule, language, semantics, and spelling for transcript presentations on social media that have caused problems for models that detect hate speech, then preparing a list of profanity in mixture with other kinds of verbal structures might benefit with hate speech recognition complications. Likewise, data obtained from content other than meta-information might be beneficial in identifying existence of hate speech.

TABLE II. SUMMARY OF LITERATURE REVIEW

Author and Year	Language	Social media platform	Classification focus	Features Representation	Algorithms	Result Efficiency
Mujadia (2019)	German and English language	Twitter	Automated recognition of hate speech texts	N-grams(N=1-5), TF-IDF,	SVM, Ada-boost, RF electing classifiers & LSTM.	Approx 70%, 47.7% and 80.32%
Sigurbergsson et. al (2019)	Danish and English language	Reddit and Facebook	Automated recognition of hate speech texts	semantic elements, pre-trained word embeddings, emotion results	R, Learned bi-LSTM, Fast-Bi-LSTM, AUX-Fast-bi-LSTM.	74% and 70%
Altin et. al (2019)	English	Twitter	Automated insulting speech detection	Pre-trained word embeddings	Bi-LSTM	82.9%
Umar et. al(2019)	English	Twitter	categorization of violent speech & the people's contribution	word embedding	User profiling and LSTM	Approx 89% and 83.33%
Garain & Basu (2019)	English	Twitter	hate speech, violent behaviour	Bi-LSTM	Neural Network	57.3% a 76.3%
MacAvaney (2019)	English	Facebook And Stormfront	hate speech	TF-IDF and unigram	Multiview SVM	53.7%, 80.3%
Naveen and Kumar(2019)	English	Twitter	automatic recognition of offensive hateful and clean comments	N-grams & TF-IDF	Linear SVM, NB & LR	Approx 96%
Sadiq (2020)	English	Twitter	automatic detection of aggression	N-grams (N=1-2)	CNN-LSTM & CNN-bi-LSTM in Deep Neural Network.	92%
Jain, Kumar & Garg (2020)	English Hindi	Twitter	irony recognition	GloVe	Bi-LSTM and CNN	92.7%, 89.05%
Kapil & Ekbal (2020)	English	Facebook and other social media	Detection of abusing speech	word with character embedding	CNN + Gated Recurrent Unit (GRU)	80.6% and 86.5%
Salminen (2020)	English	Multiplatform	hate speech	bag-of-words, Word2Vec, TF-IDF, BERT, and their combinations	SVM, LR, NB, XGBoost and Neural Networks	92%
Chopra (2020)	Hindi, English code switched	Twitter	vulgarity recognition	deep graph embedding and author profiling	Node2vec + Dense CNN + bi-LSTM + DeepWalk.	73%
Modha (2020)	English and code mixed Hindi	Twitter, Facebook	online aggression	Attention-based model, BERT pre-trained language	Linear SVM, LR and CNN	64% for recognition on Facebook and 58% for detection on Twitter

REFERENCES

[1] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, G. N. Anigbgu, " Hate Speech Classification Using SVM and Naive BAYES, 2204.07057, 2022.

[2] V. Pathak, M. Joshi, P. Joshi, M. Mundada, M & T. Joshi, "Using machine learning for detection of hate speech and offensive code-mixed social media text, 2102-09866, 2020.

[3] S. Malmasi & M. Zampieri, "Detecting hate speech in social media, 1712-06427, 2017.

[4] A. Baruah, F. A. Barbhuiya & K. Dey, "Automated Hate Speech and Offensive Content Detection in English and Code-Mixed Hindi Text, HASOC In FIRE, pp. 229-236, 2019.

[5] M. K. A. Aljero & N. Dimililer, N, "A novel stacked ensemble for hate speech recognition" Applied Sciences, 11(24), 11684, 2021.

[6] C. S. Wu & U. Bhandary, "Detection of hate speech in videos using machine learning," In 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 585-590, 2020.

[7] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan & S. H. Malik, "Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques," International Journal of Information Management Data Insights, 2(2), 100120, 2022.

[8] H. Simon, B. Y. Baha & E. J. Garba, "Trends in machine learning on automatic detection of hate speech on social media platforms: A

Systematic review," FUW Trends in Science & Technology, 7(1), 001-016, 2022.

[9] V. Mujadia, P. Mishra & D. M. Sharma, "Hate Speech Detection," In FIRE, pp. 271-278, 2019.

[10] G. I. Sigurbergsson & L. Derczynski, L. (2019). Offensive language and hate speech detection for Danish," arXiv: 1908.04531, 2019.

[11] Altın, L. S. M., Serrano, À. B., & Saggion, H. (2019, June). Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bi-lstm model. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 672-677).

[12] Umar, A., Bashir, S. A., Laud, C. O., & Ibrahim, A. (2018). Profiling Inappropriate Users' Tweets Using Deep Long Short-Term Memory (LSTM) Neural Network.

[13] Garain, A., Basu, A., 2019. The Titans at SemEval2019 Task 5: Detection of hate speech against immigrants and women in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Minneapolis, Minnesota, USA, 494–497.

[14] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O., 2019. Hate speech detection: Challenges and solutions. PLoS ONE, 14(8).

[15] Naveen, K., Kumar, C. P., 2019. An effective scheme for detecting hateful and offensive expressions on twitter. Intl. Journal for Innovative Engineering and Mgt. Research. 8(8). 204-209. ISSN 2456 – 5083

[16] Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., On, B.W., 2020. Aggression detection through deep neural model on Twitter. Elsevier, Future Generation Computer Systems, 114,120-129. DOI: 10.1016/j.future.2020.07.050

[17] Jain, D., Kumar, A., & Garg, G. (2020). Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Applied Soft Computing, 91, 106198.

[18] Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems, 210, 106458.

[19] Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S., An, J., Kwak, H., Jansen, B. J., 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. Proceedings of the Twelfth International AAAI Conference on Web and social media (ICWSM). 330- 339.

[20] Chopra, S., Sawhney, R., Mathur, P., & Shah, R. R. (2020, April). Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 01, pp. 386-393)..

[21] Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. Expert Systems with Applications, 161, 113725.