International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

MACHINE LEARNING AND DEEP LEARNING TECHNIQUES FOR DETECTING ABUSIVE CONTENT ON TWITTER

Ajay Choudhari[1], Freya Kotak[2], Siddhant Zaveri[3], Mrs. Jyoti Bansode[4]

[1,2,3]Shah and Anchor Kutchhi Engineering College Chembur, Mumbai.
[4] Professor, Dept. Of Information Technology, Engineering, Shah and Anchor Kutchhi Engineering College, Maharashtra, India

---

Abstract - *Cyber Abuse is the use of online systems and virtual devices to inflict harm of either psychological, emotional, sexual, racist, sexist or any negative adjective in nature. Sentiment Analysis is the interpretation and classification of text data by their polarity that is positive, negative, or neutral. This is done with the help of various analysis techniques and the usage of each technique differs from author to author. Sentiment Analysis is being used by major companies in the world so that they can identify what kind of emotions are being passed around the market by the people and thus, use these emotions and make changes to the company's systems in order to grow as a whole.*

Key Words: *Twitter, Hate speech detection, BERT.*

1. INTRODUCTION

Internet being responsible for 80% communication in the world, it clearly has its own disadvantages. According to statistics presented by "Global social media statistics research summary 2022", **58.4%** of the world's population uses social media. The average daily usage is **2 hours and 27 minutes** (January 2022). It has become a part of life which also has an impact on mental health. The harm inflicted maybe physiological or emotional.

Efforts has been made by several organizations and government as well as by the companies owning the social media platforms to act upon its negative effects. Improvements are introduced but to recognize a person's intention behind a certain comment and the way it portrays is an arduous job. Finding a way to stop these abusive attacks and making a person think about the reception of the comment before they post it is idea behind this research.

1.1 LITERATURE REVIEW

The paper [6] by *W. N. Hamiza Wan Ali, M. Mohd and F. Fauzi,* basically provides an overview of the cyber bullying occuring in the social networks. It identifies the data source as text messages, instant messages, social media and online games. It compares the features used for detection such as bag-of-words and clears the latest and better version used. All the commonly used classification algorithms were summarized by considering individual social media platforms. It identifies the following challenges faced while Cyber Bullying detection: Language Challenge

Dataset Challenge

Data Representation Challenge

Zinnar Ghasem1, Ingo Frommholz1 and Carsten Maple2 talk about the urge to prevent and take strict action against cyber-crimes as they have drastically increased with the development in technology. The proposed framework [8] here is referred to as ACTS that stands for Anti Cyberstalking Text- based System. It utilizes text mining, statistical analysis, text categorization, and the use of machine learning algorithms to combat cyber bullying. The use of an attacker identification module determines the status of the attacker and his/her comments containing foul language or bad words.

[17] The paper proposed by B. Haidar, M. Chamoun, and A. Serhrouchni has provided a multilingual cyberbullying detection approach for detecting cyberbullying in messages, tweets and newspaper review for two Indian languages. Results of the experiments shows that Logistics Regression outperforms all other algorithms on these datasets. Also, generating synthesized data could help improve performance of the system. Results of our study show that our systems perform well across two languages and different domains and hence, it can be used to detect cyberbullying for other Indian languages as well.

2. METHODOLOGY

The research started with a basic machine learning approach. Choosing data, exploration, cleaning/pre-processing, applying the known algorithms and comparing the results. Later it extended to exploring deep learning mode to improve the accuracy. We examined various techniques on our dataset and chose the one for basic implementation on web interface.

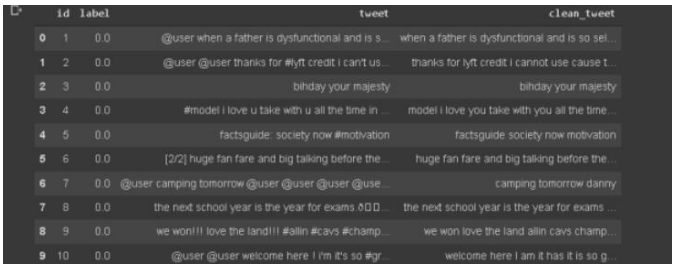
2.1 DATA DESCRIPTION AND PRE-PROCESSING

The dataset contains over 32k tweets which is then split into train and test datasets. Twitter data is complicated to

work with as it contains various symbols like at the rates and hashtags along with emoticons, slangs and more. Pre-processing is an essential part of the process to clean the data before using it. The pre-processing methods used are stated as follows:

- Html Parser
- Remove pattern
- Apostrophe removal
- Short word removal
- Emoticon Removal
- Removing brackets, hashtags and at the rates
- Tokenization
- Stemming

The image below clarifies the changes and improvements made by process of cleaning the data.



	id	label	tweet	clean_tweet
0	1	0.0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so sel...
1	2	0.0	@user @user thanks for #lyft credit i can't us...	thanks for lyft credit i cannot use cause t...
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in ...	model i love you take with you all the time...
4	5	0.0	factsguide: society now #motivation	factsguide society now motivation
5	6	0.0	[2/2] huge fan fare and big talking before the...	huge fan fare and big talking before the...
6	7	0.0	@user camping tomorrow @user @user @user @use...	camping tomorrow danny
7	8	0.0	the next school year is the year for exams.ð...	the next school year is the year for exams ...
8	9	0.0	we won!!! love the land!!! #allin #cavs #champ...	we won love the land allin cavs champ...
9	10	0.0	@user @user welcome here ! i'm it's so #gr...	welcome here i am it has it is so g...

Fig -1: Cleaned Tweets

The train test split is standard 70-30. After cleaning the tweets, next we move on to some visualizations to analyse the data trends.

2.2. FEATURE EXTRACTION

Vectorization of the tweets was carried out by count vectorizer followed by TF-IDF vectorizer. It recognizes the importance of the word in the document. Formula is given as TF_IDF = TF * IDF Where TF=Number of times word appeared in text/ number of words in text, and IDF=log (Number of documents/ Number of documents with word in it)

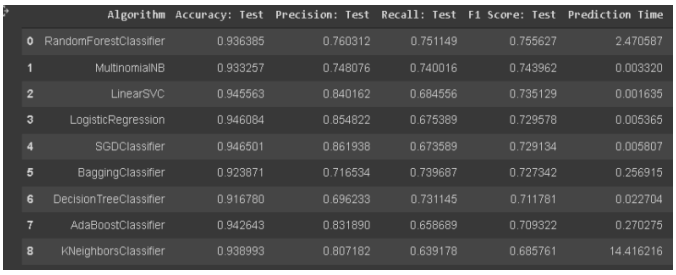
2.3. MACHINE LEARNING TECHNIQUE

Machine learning is a technique to use computer algorithms which improve over a period of time using the data. There are various pre-defined algorithms developed using the required mathematical calculations. We have chosen 9 famous machine learning techniques to apply on our data. They are as follows: 1. Random Forest Classifier 2. Multinominal NB 3. Linear SVC 4. Logistic Regression 5. SGD Classifier 6. Bagging Classifier 7. Decision Tree Classifier 8. Ada Boost Classifier 9. KNN Code has a pipeline of all these algorithms. Data is passed through it and trained.

2.4. DEEP LEARNING TECHNIQUES

Deep learning is a subset of machine learning which teaches the computer what comes to human naturally. Understanding instincts and major characteristics is a core part of deep learning. Long Short-Term Memory (LSTM) is an improvement over recurrent neural network used for sequence prediction. LSTMs make small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things.

3. RESULTS



	Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Prediction Time
0	RandomForestClassifier	0.936385	0.760312	0.751149	0.755627	2.470587
1	MultinomialNB	0.933257	0.748076	0.740016	0.743962	0.003320
2	LinearSVC	0.945563	0.840162	0.684556	0.735129	0.001635
3	LogisticRegression	0.946084	0.854822	0.675389	0.729578	0.005365
4	SGDClassifier	0.946501	0.861938	0.673589	0.729134	0.005807
5	BaggingClassifier	0.923871	0.716534	0.739687	0.727342	0.256915
6	DecisionTreeClassifier	0.916780	0.696233	0.731145	0.711781	0.022704
7	AdaBoostClassifier	0.942643	0.831890	0.658689	0.709322	0.270275
8	KNeighborsClassifier	0.938993	0.807182	0.639178	0.685761	14.416216

Fig - 2: Results

Machine Learning and deep learning give very different results as deep learning works closely on data and uses complicated algorithms and trends. The results for ML algorithms are given as follows. Figure 2. ML Results As shown, our model works best with SGD classifier, that is, Stochastic Gradient Descent. is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Accuracy and F1 score is 94.6% and 72.9% respectively. LSTM gives accuracy of 99.5% and F1 score of 74.8%.

4. CONCLUSIONS

Tackling the most important problem with today's social media is unavoidable. Keep experimenting and applying the measures is the key part. This paper focuses on trying out methods and choosing the best. ML is effective but not enough to implement. False predictions are covered well with LSTM. The main motive of this research is to cut off the abuse at the start instead of removing it after. Making the user think before they post and not let miscommunication create chaos. Implementing a suitable model at user end to stop them form coming out offensive is the aim for future.

The authors can acknowledge any person/authorities in this section. This is not mandatory.

REFERENCES

[1] International Conference on Advanced Computing Technologies and Applications (ICACTA- 2015). Online Social Network Bullying Detection Using Intelligence Techniques. By- B. Sri Nandhinia, J.I. Sheebab.

[2] Optimized Twitter Cyberbullying Detection based on deep Learning. By - Monirah A. Al-Ajlan. Information Systems Department, CCIS, King Saud University, Riyadh, Saudi Arabia, maalajlan@ksu.edu.sa and Mourad Ykhlef, Information Systems Department, CCIS.

[3] S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," 2020 Fourth International Conference.K. Elissa,

[4] A. Mody, S. Shah, R. Pimple and N. Shekokar, "Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2018, pp. 878-881, doi: 10.1109/ICEECCOT43722.2018.9001476.

[5] M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 432- 437, doi: 10.1109/ICPR.2016.7899672.

[6] W. N. Hamiza Wan Ali, M. Mohd and F. Fauzi, "Cyberbullying Detection: An Overview," 2018 Cyber Resilience Conference (CRC), 2018, pp. 1-3, doi: 10.1109/CR.2018.8626869

[7] Jezabel Molina-Gil, José A. Concepción-Sánchez and Pino Caballero-Gil, "Harassment Detection Using Machine Learning and Fuzzy Logic Techniques", Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI. Published: 20 November 2019.

[8] Zinnar Ghasem1, Ingo Frommholz1 and Carsten Maple2, "A Machine Learning Framework to Detect and Document Text-based Cyberstalking", 1 University of Bedfordshire, UK, 2 University of Warwick, UK.

[9] A. Chaudhari, A. Parseja, and A. Patyal, "CNN based Hate-o-Meter: A Hate Speech Detecting Tool," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020.

[10] N. Rai, P. Meena, and C. Agrawal, "Improving the hate speech analysis through dimensionality reduction approach," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020.

[11] D. Kumar, N. Kumar, and S. Mishra, "QUARC: Quaternion Multi-Modal Fusion Architecture for Hate Speech Classification," 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), 2021

[12] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017.

[13] S. A. E. Rahman, F. A. Alotaibi, and W. A. Alshehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019.

[14] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020.

[15] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[16] O. C. Hang and H. M. Dahlan, "Cyberbullying Lexicon for Social Media," 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), 2019.

[17] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," 2017 1st Cyber Security in Networking Conference (CSNet), 2017.

[18] Prabhu, Trisha N. Method to Stop Cyberbullying before It Occurs.

[19] Li, Bohan, et al. CYBERBULLYING DETECTION METHOD AND SYSTEM. 22 Apr. 2021.