DETECTION OF CYBER BULLYING

Mini Project 1-B
SEM IV

Abstract

Cyberbullying can manifest in various forms like cyber-stalking, identity theft, etc. Teenagers are more prone to be victims of cyberbullying since they spend a significant amount of their time daily engaging with various social media entities. UNESCO drew attention to this issue by producing case studies from several geographically distant countries. These studies show that the urgency of countering cyber bullying cannot be gainsaid and is a subject of human rights. Similar to bullying, cyberbullying affects a person’s mental well-being, often leading to low self-esteem, self-harm, anxiety, depression, suicidal thoughts, etc. Due to the widespread instances of cyberbullying and harassment, combating cyberbullying is a pressing issue.We have completed a thorough literature review of Cyberbullying. The literature study concluded that social media spaces are rife with instances of harassment and bullying. There is a paucity of mere documentation of cyberbullying, let alone actions to prevent it.

Of late, researchers are in favor of countering cyberbullying through automation[2]. Machine Learning models have been designed to detect bullying and harassment through text in comments. But these systems are lacking in terms of texts/comments that are available in regional scripts. Training ML models with regional languages shall make them more efficient in countries like India, where various regional languages, differing in scripts, are used. Hindi has around 73 million native speakers (1.1% of the world's population). Thus, designing automated systems for Hindi will significantly impact the safety of online communication spaces.

INTRODUCTION

Cyberbullying is the act of humiliating, scaring, or causing danger to other people through electronic devices and online networks. Bullies, if left unquestioned of their behavior, can normalize such behavior leading to rampant abuse. Automatic detection and cyberbullying reporting can help keep a check on individuals occupying online space and induce a sense of responsibility towards their behavior. Implementation of this method of countering cyberbullying shall require automation of classification of individuals' various activities as offensive or not offensive. Since the most common form of bullying is mean comments, classifying text according to the conveyed sentiments can help bring bullies to book. For this purpose, ML models can be trained and employed.

CYBERBULLYING

Cyberbullying is prevailing in most countries across the world. The paradigm shift within the IoT around ten years back, resulted in tremendous growth in the number of active social media users. Nowadays this number has exceeded over 3 billion. Cyberbullying takes a toll on a person's mental and emotional health, brought to light in Hinduja et al.

Traditionally, guidelines and protocols have been drafted, but they involve the victim reporting the abuse to their parents or guardians. This approach of dealing with cyberbullying has not proved very efficient since most victims do not report the incidents. Therefore, automated cyberbullying detection is of immense importance.

DATA AVAILABILITY

The increase in the consumption of internet resources in our country might even be attributed to the "Jio effect." Reliance Jio which arrived in 2016 with almost negligible cost information plans in comparison to the already existent plans. The network gained a lot of customers and turned into the most important mobile operator globally with over three hundred million subscribers. Reliance Jio's client base is regarded as massive as the entire population of the United States of America to place this in context. This also means that the number of engagements people have in social media spaces has increased manifold(Big Data. Era)[4], and many scalable data can be employed to train Machine Learning models.

MACHINE LEARNING

"ML is the science of getting computers to learn and function like humans do, and enhance their learning over time in an independent fashion, by providing them information and data in the form of observations and real-world interactions."[5]. Computational methodology such as graph mining and relational learning has been in use for business-centered Data analytics. Reorientation of these methods can be instrumental in mitigating cyberbullying, if not cease it.

Most of the time, the focus while learning Machine Learning is placed on choosing the best algorithm, but the equal emphasis should be placed on the data we use to train the models.

We tend to propose a possible answer for detection of these cyberbullying instances which contain detrimental and disparaging text or sentences.

RELATED WORKS

Owing to the importance of this topic, this is a widely covered topic both online and offline. Many research studies are done in psychology, social sciences, and research to detect cyberbullying using current technologies, and testament to the above statement is the abundance of research papers available regarding this topic. This section includes a discussion of the prevalence and seriousness of cyberbullying.

Kontostathis et al. [6] reviewed data from Formspring.me, and manually labeled data using Amazon's Mechanical Turk. They implemented text mining to detect vandalism, spam, internet abuse, and cyberterrorism.

Dinakar et al. [7] at MIT conducted a supervised machine learning approach to develop a model to identify cyber bullying about sexuality, race & culture, intelligence. They collected Youtube comments and manually labeled them, and implemented various binary and multiclass classifiers. Their study revealed that binary classifiers outdid the multiclass classifier. The authors used SVM learner and reached an accuracy of 66.7%.

Yin et al. [8] used supervised learning to develop a model to detect harassment. They employed a bag of words approach based on parameters such as content, sentiment and contextual attributes of data to train an SVM classifier and reached a recall level of 61.9%

Dadavar et al.[9] analyzed the data from MySpace, a platform that offers an interactive, user-submitted community of friends with blogs, personal profiles, etc., and specifically focused on going down the route of a gender-based approach. TO train a gender-based classifier, they used an SVM model. The dataset consists of 381.000 posts. The result obtained by the gender-based approved baseline by 15% in F-measure, 39% in precision and 6% in the recall.

Nandhini et al.[10] used Naive Bayes machine learning algorithm. They used the dataset from MySpace, and they achieved an accuracy of 91%. They then proposed another model[11], Naive Bayes Classifier, and genetic operation(FuzGen), and they reached 87% accuracy.

Isa et al.[12] retrieved dataset from Kaggle. They employed two classifiers, namely., SVM and Naive Bayes. They scored an accuracy of 92.81% with Naive Bayes, while with the SVM poly kernel, they achieved an accuracy of 97.11%.

Haidar et al.[13] put forward a model to discern cyberbullying by using the Arabic language. They used SVM achieved 94.1% precision and Naive Bayes and achieved 90.85% precision.

Chavan et al.[14] implemented Support vector machines (SVM) and logistic regression. With logistic regression, they attained 60% recall and 73.76% accuracy, and 64.4% precision. While with SVM, they achieved 70% precision, 77.65% accuracy and 58% recall.

Eckert et al.[15] presented the reference literature for detection of cyberbullying incidents in social media spaces using DNN based models.

H. Rosa et al.[16] produced an in-depth study of research that focused on automatic detection of cyberbullying through a quantitative systematic review approach. We also inspected this approach with a detailed experiment to examine current implementation by using feature engineering.

Some papers also mentioned deep learning and neural networks as a way to detect cyberbullying. Zhang et al.[17] used innovative pronunciation-based convolution neural networks(PCNN). They used 1313messages off twitter and 1300 messages from Formspring.me. They achieved a precision of 56%,78% recall, and 96% accuracy.

Parime et al.[18] used MySpace as their dataset source and manually marked them. They employed SVM as their classifier.

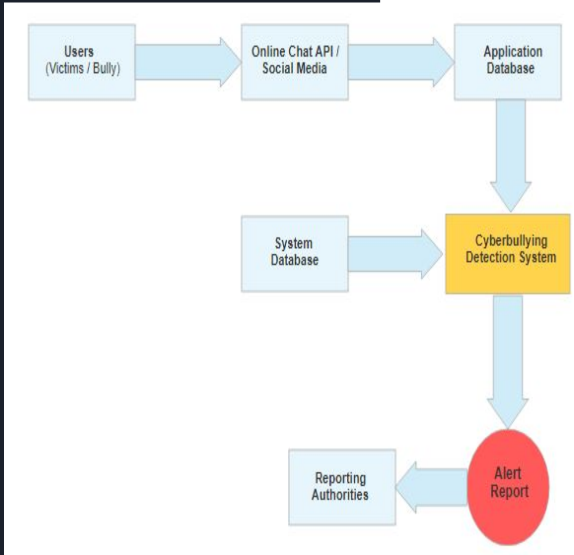
Capua et al. [19] used GHSOM networks to be well suited for extensive collections of documents that need to be classified. They used a dataset from Kontostathis et al. [6]. They achieved 72% precision. 73% accuracy and 71% on F1 using GHSOM. Using C4.5, they earned 60% precision and 67% recall. Using SVM, they received 67% recall on this dataset. They also used a dataset from Dadavar et al. [9]. Using GHSOM, they achieved 60%precision, 69% accuracy, 94% recall.

Salawu et al.[20] conducted a unique survey under the umbrella of cyberbullying detection.

The studies mentioned above highlighted a significant gap in detecting cyberbullying. Most of them are likely to stick to a dataset of one language, i.e., English.

PROPOSED TECHNOLOGY

In this model, three algorithms are used to accomplish the detection of Cyberbullying. They are as follows: Random Forest, Naive Bayes, and SVM.



The proposed model has been trained and tested for various English and Hindi languages datasets to attain the below-mentioned performance.

DATA COLLECTION

We have collected datasets from dhavapotdar/cyberbullying-detection, which provided datasets in the English Language, consisting of around 30,000 facebook comments and kmi-linguistics/track-1 datasets in Hindi Language consisting of approximately 15,000 tweets on github.com.

The Hindi dataset consists of data in the Hindi language and Hindi words transliterated using the English language. The datasets consisted of various comments classified as either cyberbullying or not. The classification classes were skewed and distributed such that 12384 comment entities out of 26133 fell under the label of cyberbullying, while 13749 belonged to the other type for one English dataset. The dataset consists of 2 columns. The first column consists of text and the second column consists of labels. The dataset was labeled 1: offensive texts and 0 for non-offensive texts. 80% of the dataset was used for training the model, and the remaining 20% was used for testing the model.

ALGORITHMS

1. ***Naive Bayes*:** This algorithm is based on Bayes' Theorem with an assumption of independence among predictors.

   Bayes theorem formulated by Thomas Bayes, calculates the probability of an event occurring based on prior knowledge conditions related to the event. It is based on the formula

   P(A|B)=P(A)*P(B|A)/P(B)

   Multinomial Naïve Bayes (Multinomial NB) is popularly used for Natural Language Processing(NLP).NB was used to detect instances of cyberbullying in many cases. This model assumes that a parametric model produces the text and uses training data to determine Bayes-optimal parameter estimates of the model.

2. ***Random Forest*:** Random Forest is an ensemble of decision trees generally trained via the bagging method . It can be used for both classification and regression. RF gives good accuracy in comparison with other classification methods with overfitting not being a concern while using Random Forest.

ALGORITHMS

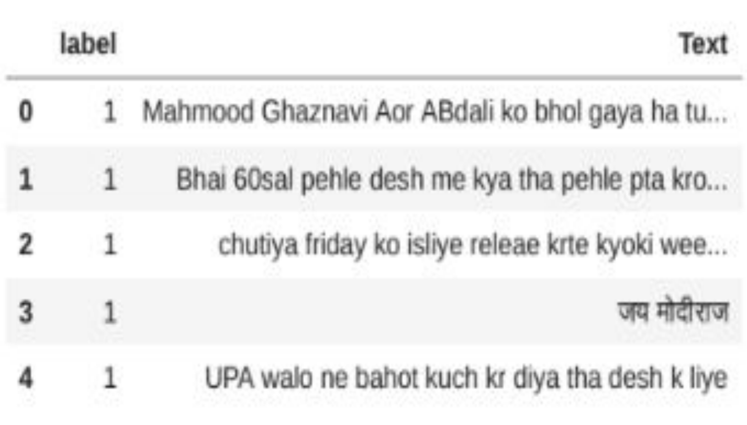
<u>3. Support Vector Machine:</u> SVM may also be referred to as a supervised machine learning algorithm that can be used for classification or regression challenges that provide higher accuracy in the cyberbullying detection project.

Support Vector Machine is a versatile machine learning model capable of performing linear and non linear classification.

The main idea behind SVM is that the two classes are linearly separable. We can think of SVM as fitting the widest possible street between the two classes

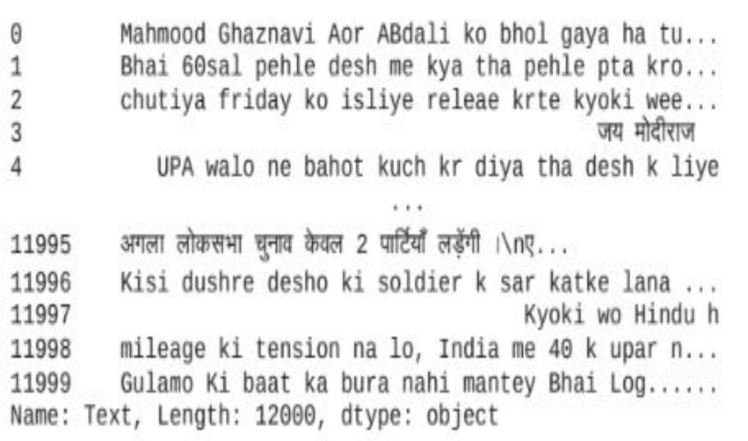
SVM classifiers work efficiently even if the classes can be separated linearly.Specific kernels are implemented in the SVM to turn the function space appropriately. Kernelized SVM models help us to achieve this.

IMPLEMENTATION



	label	Text
0	1	Mahmood Ghaznavi Aor ABdali ko bhol gaya ha tu...
1	1	Bhai 60sal pehle desh me kya tha pehle pta kro...
2	1	chutiya friday ko isliye releae krte kyoki wee...
3	1	जय मोदीराज
4	1	UPA walo ne bahot kuch kr diya tha desh k liye

Figure 2. Annotated Dataset in Hindi



```
0        Mahmood Ghaznavi Aor ABdali ko bhol gaya ha tu...
1        Bhai 60sal pehle desh me kya tha pehle pta kro...
2        chutiya friday ko isliye releae krte kyoki wee...
3                                              जय मोदीराज
4           UPA walo ne bahot kuch kr diya tha desh k liye
                               ...
11995    अगला लोकसभा चुनाव केवल 2 पार्टियाँ लड़ेंगी ।\nए...
11996    Kisi dushre desho ki soldier k sar katke lana ...
11997                                      Kyoki wo Hindu h
11998    mileage ki tension na lo, India me 40 k upar n...
11999    Gulamo Ki baat ka bura nahi mantey Bhai Log......
Name: Text, Length: 12000, dtype: object
```

Figure 3. Comments from Dataset in Hindi

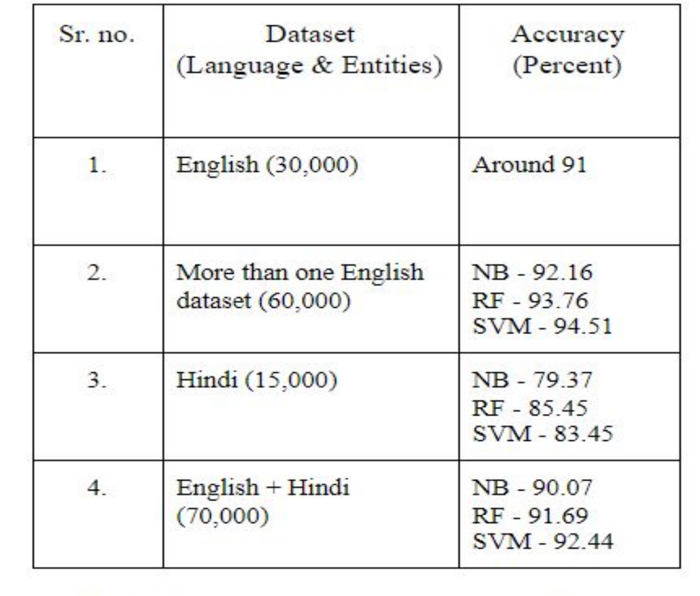
Results

Primarily, the model was trained, thus tested, for two different datasets in the English language, each consisting of entries around 30,000 tweets. This resulted in the accuracies of around 91%.

Later, to extricate the drawback of an insufficient dataset, we merged the existing datasets in English (total entries around 60,000 tweets), which realized the accuracies as 92.16% through Naive Bayes, 93.76% through Random Forest and 94.51% through SVM.

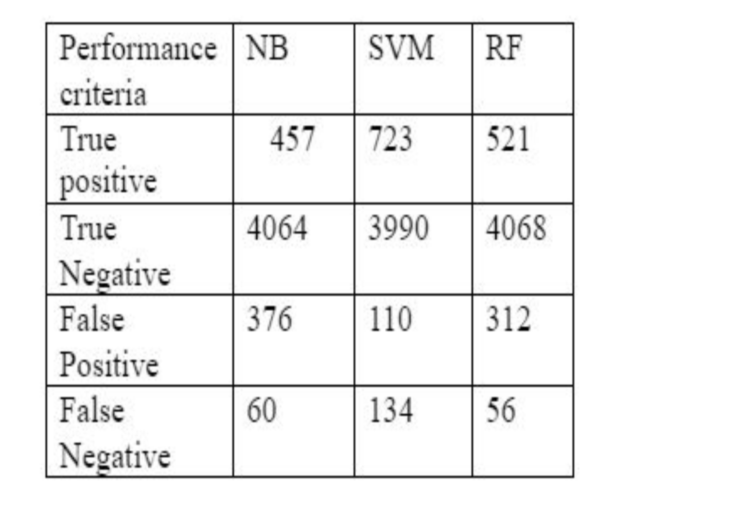
Furthermore, to upgrade the model and outperform the existing systems, the model was trained and tested for a Hindi language dataset (around 15,000 tweets). Consequently, the following accuracies were achieved: 79.37% through Naive Bayes, 83.45% through SVM, and 85.45% through Random Forest. However, the existing models conferred the accuracy between 50-60% only.

In the final analysis, as shown in Table1, the model was trained and tested for multilingual datasets; wherein, all the datasets as mentioned above were combined, emanating a solid dataset consisting of tweet entries of around 70,000. The accuracy obtained for the same was 90.07% through Naive Bayes, 91.69% through Random Forest, and 92.44% through SVM. Thus, the existing models with insufficient datasets provide an accuracy of around 80% for multilingual purposes.



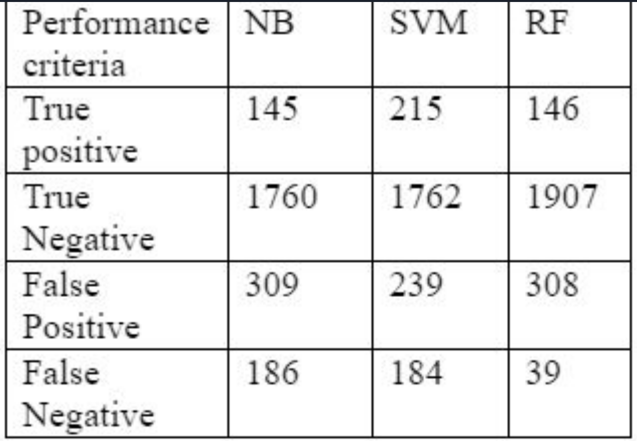
Sr. no.	Dataset (Language & Entities)	Accuracy (Percent)
1.	English (30,000)	Around 91
2.	More than one English dataset (60,000)	NB - 92.16 RF - 93.76 SVM - 94.51
3.	Hindi (15,000)	NB - 79.37 RF - 85.45 SVM - 83.45
4.	English + Hindi (70,000)	NB - 90.07 RF - 91.69 SVM - 92.44

Table 1. Performance of the proposed methodology



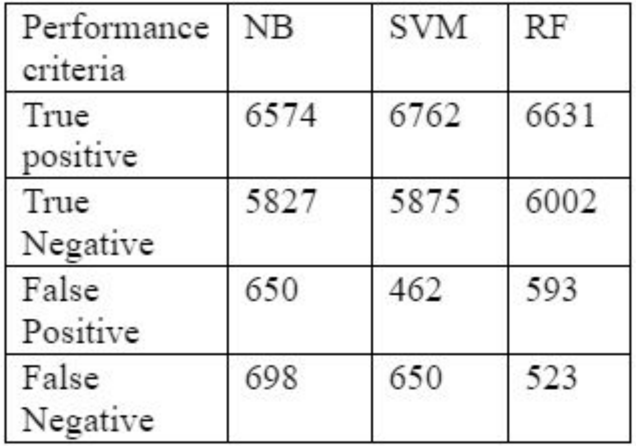
Performance criteria	NB	SVM	RF
True positive	457	723	521
True Negative	4064	3990	4068
False Positive	376	110	312
False Negative	60	134	56

Table 2: Confusion matrix of various algorithms for English language dataset



Performance criteria	NB	SVM	RF
True positive	145	215	146
True Negative	1760	1762	1907
False Positive	309	239	308
False Negative	186	184	39

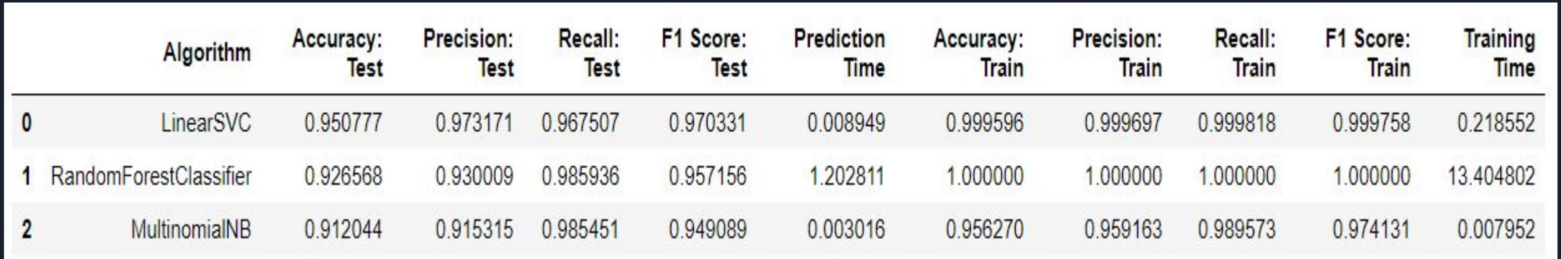
Table 3. Confusion matrix of various algorithms for Hindi language dataset



Performance criteria	NB	SVM	RF
True positive	6574	6762	6631
True Negative	5827	5875	6002
False Positive	650	462	593
False Negative	698	650	523

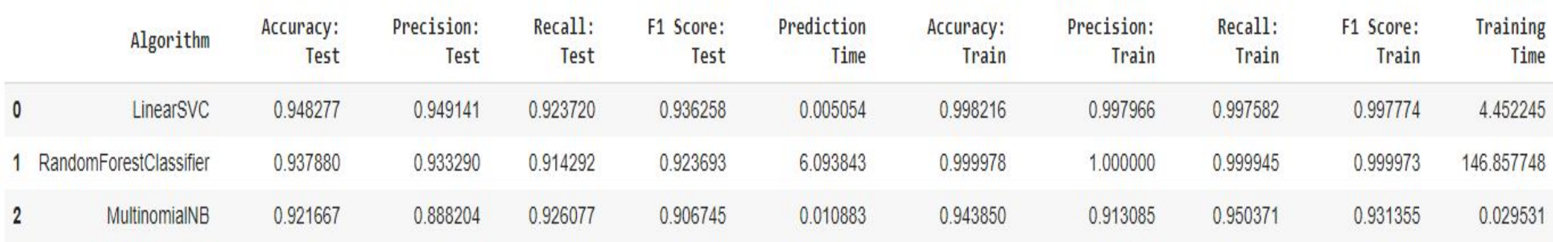
Table 4. Confusion matrix of various algorithms for Hindi and English language dataset.

Comparisons of various Algorithmic Performances:



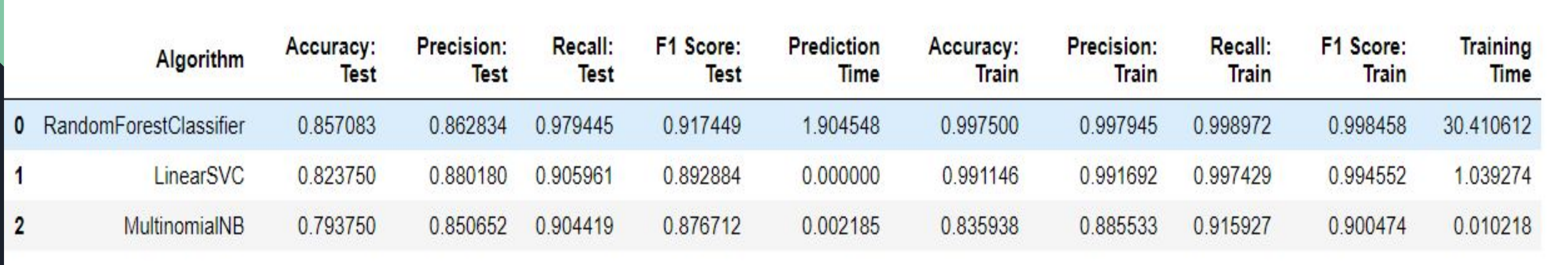
	Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Prediction Time	Accuracy: Train	Precision: Train	Recall: Train	F1 Score: Train	Training Time
0	LinearSVC	0.950777	0.973171	0.967507	0.970331	0.008949	0.999596	0.999697	0.999818	0.999758	0.218552
1	RandomForestClassifier	0.926568	0.930009	0.985936	0.957156	1.202811	1.000000	1.000000	1.000000	1.000000	13.404802
2	MultinomialNB	0.912044	0.915315	0.985451	0.949089	0.003016	0.956270	0.959163	0.989573	0.974131	0.007952

Performance of algorithms with English dataset



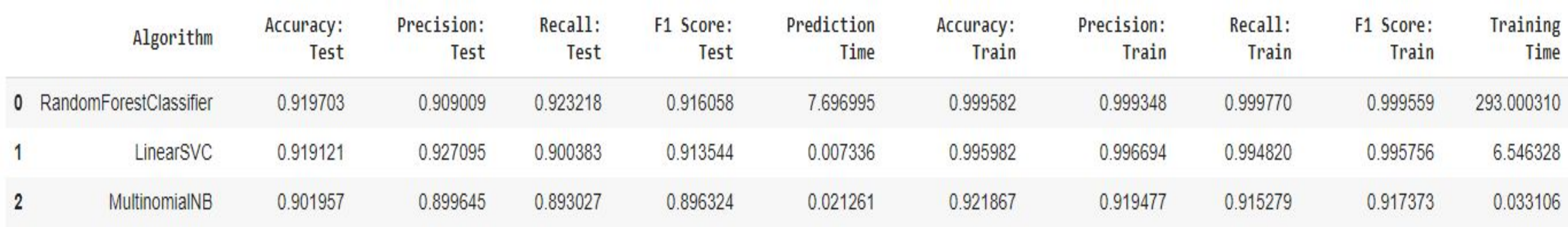
	Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Prediction Time	Accuracy: Train	Precision: Train	Recall: Train	F1 Score: Train	Training Time
0	LinearSVC	0.948277	0.949141	0.923720	0.936258	0.005054	0.998216	0.997966	0.997582	0.997774	4.452245
1	RandomForestClassifier	0.937880	0.933290	0.914292	0.923693	6.093843	0.999978	1.000000	0.999945	0.999973	146.857748
2	MultinomialNB	0.921667	0.888204	0.926077	0.906745	0.010883	0.943850	0.913085	0.950371	0.931355	0.029531

Performance of algorithms with more than one English dataset



	Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Prediction Time	Accuracy: Train	Precision: Train	Recall: Train	F1 Score: Train	Training Time
0	RandomForestClassifier	0.857083	0.862834	0.979445	0.917449	1.904548	0.997500	0.997945	0.998972	0.998458	30.410612
1	LinearSVC	0.823750	0.880180	0.905961	0.892884	0.000000	0.991146	0.991692	0.997429	0.994552	1.039274
2	MultinomialNB	0.793750	0.850652	0.904419	0.876712	0.002185	0.835938	0.885533	0.915927	0.900474	0.010218

Performance of algorithm with hindi dataset



	Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Prediction Time	Accuracy: Train	Precision: Train	Recall: Train	F1 Score: Train	Training Time
0	RandomForestClassifier	0.919703	0.909009	0.923218	0.916058	7.696995	0.999582	0.999348	0.999770	0.999559	293.000310
1	LinearSVC	0.919121	0.927095	0.900383	0.913544	0.007336	0.995982	0.996694	0.994820	0.995756	6.546328
2	MultinomialNB	0.901957	0.899645	0.893027	0.896324	0.021261	0.921867	0.919477	0.915279	0.917373	0.033106

Performance of dataset with English + Hindi Dataset

CONCLUSION

Although this is a very well-researched area that is evident from the abundance of research papers in the field: Detection of cyberbullying, existing research has not attained satisfactory accuracy. This model has utilized various algorithms namely, Random Forest, Naive Bayes, and supervised machine learning algorithm SVM.

To maintain the efficiency of these systems, one must consider rapidly evolving vocabulary and behavioral patterns. Attention needs to be paid to cultural nuances to classify a particular behavior as offensive or not offensive successfully; tested most of the work done in this field for datasets in the English language only. Therefore we have also tried to expand this research by detecting cyberbullying in other native languages such as Hindi. But a significant hindrance while doing so was the unavailability of a sufficient dataset. Hence, we have successfully attained utmost accuracies for the combination of datasets in English and Hindi.

References

[1]. D. Richardson and C. F. Hiu, Ending the torment: tackling bullying from the schoolyard to cyberspace. 2016.

[2]. Van Royen, K., et al. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics, 2015. 32(1): p. 89-97.

[3]. S. Hinduja, and J. Patchin (2007). Offline consequences of online victimization: School violence and delinquency. Journal of school violence, 6(3), 89-112

.

[4]. Wu, X., et al., Data mining with big data. IEEE transactions on knowledge and data engineering, 2014. 26(1): p. 97-107

[5]. https://emerj.com/ai-glossary-terms/what-is-machine-learning/

[6]. A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. 2013, “Detecting cyberbullying: query terms and techniques," In Proceedings of the 5th WebSci 2013. ACM, New York.

[7]. K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," MIT, International Conference on Weblog and Social Media. Barcelona, Spain, 2011.

[8]. D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0”, In Proceedings of CAW2.0 Workshop, 2009.

References

[9]. M. Dadvar, R.B. Trieschnigg and F.M.G. de Jong. “Experts and Machines Against Bullies: A Hybrid Approach to Detect Cyberbullies''. In 27th Canadian Conference on Artificial Intelligence, University of Waterloo, Montral, Canada, 2014.

[10]. B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithms. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015

[11]. B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45:485–492, 2015

[12.]Sani Muhamad Isa, Livia Ashianti, et al. Cyberbullying classification using text mining. In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.

[13]. Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6):275–284, 2017.

[14]. Vikas S Chavan and SS Shylaja. Machine learning approach for detection of cyber-aggressive comments by peers on social media networks. In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on, pages 2354–2358. IEEE, 2015

References

[15]. Maral Dadvar and Kai Eckert. Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. Web-based Information Systems and Services, Stuttgart Media University Nebenstrasse 8, 70569 Stuttgart, Germany {dadvar , eckert}@hdm-stuttgart.de

[16]Automatic cyberbullying detection: A systematic reviewH. Rosaa,b, N. Pereirac,d,∗, R. Ribeiroa,e, P.C. Ferreiraa,c,d, J.P. Carvalhoa,b, S. Oliveirac,d, L. Coheura,b, P. Paulinod,f, A.M. Veiga Simãoc,d, I. Trancosoa,b. Computers in Human Behavior.

[17]Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection with a pronunciation-based convolutional neural network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 740–745. IEEE, 2016

[18]Sourabh Parime and Vaibhav Suri. Cyberbullying detection and prevention: Data mining and psychological perspective. In-Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on, pages 1541–1547. IEEE, 2014.

[19]Michele Di Capua,Emanuel Di Nardo, Alfredo Petrosino : Unsupervised Cyber Bullying Detection in Social Networks from 2016 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016

[20]Approaches to Automated Detection of Cyberbullying: A Survey. Semiu Salawu, Yulan He, and Joanna Lumsden. IEEE.

Thank you!