e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

Miss. Jafri Sayeedaaliza Abutorab[*1], Miss. Wagh Roshani Balasaheb[*2], Miss. Gaikwad Vaishnavi Subodh[*3], Miss. Sonawane Ujjwala Dattu[*4], Professor Waghmare. A .I.[*5]

[*1,2,3,4,5]Dept Of Information Technology Engineering, SND College Of Engineering & Research Center, Nashik, Maharashtra, India.

ABSTRACT

Increasing internet use and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyberbullying, a new form of bullying that emerged recently with the development of social networks, means sending messages that include slanderous statements, or verbally bullying other people in front of rest of the online community. The characteristics of online social networks enable cyberbullies to access places and countries that were previously unattainable for using SVM we are going to identify cyberbullying in twitter. Objectives of this implementation written in objective section. Image character with the help of OCR will be done by us to find image - based cyberbullying the impact on individual basis thus will be checked on dummy system. Machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching textual data to the identified traits. On the basis of our extensive literature review, we categorise existing approaches into 4 main classes, namely supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning-based approaches typically use classifiers such as SVM and Naïve Bayes to develop predictive models for cyberbullying detection. We are using natural language processing techniques and machine learning methods namely, Bayesian logistic regression, random forest algorithm, support vector machines have been used to determine cyberbullying.

Keywords: Machine Learning, Cyberbullying, Social Media, Twitter.

I. INTRODUCTION

Modern young people ("digital natives") have grown in an era dominated by new technologies where communications are pushed to quite a real-time level, and pose no limits in establishing relationships with other people or communities. The fast growing use of social networking sites among the teens have made them vulnerable to get exposed to bullying. Comments containing abusive words effect psychology of teens and demoralizes them. In this work we have devised methods to detect cyberbullying using supervised learning techniques. Cyberbullying is the use of technology as a medium to bully someone. Although it has been an issue for many years, the recognition of its impact on young people has recently increased. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. Moreover, they are a place where people engage in social interaction, offering the existing friendships. On the negative side however, social establish new relationships and maintain existing friendships. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user for such a message to be labelled as cyberbullying. The growth of cyberbullying activities is increasing as equally as the growth of social networks. Cyberbullying activities poses a significant threat to mental and physical health of the victims. Project about detection of bullying is present but implementation for

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

monitoring social network to detect cyberbullying activities is less. Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using natural processing language.

II. LITERATURE SURVEY

1. M. Di Capua, et al. [1] raises the unsupervised how to develop an online bullying model based on a combination of features, based on traditional textual elements and other "social features". Features were divided into 4 categories as Syntactic features, Semantic features, Sentiment features, and Community features. The author has used the Growing Hierarchical Self Map editing network (GHSOM), with 50 x grid 50 neurons and 20 elements as the insertion layer. M. Di Capua, and others used an integration algorithm k-means to separate the input database and GHSOM in the Formspring database. The effects of this hybrid the unsupervised way surpasses the previous one results. The author then checked the youtube database at 3 p.m. Different Machine Learning Models: Naive Bayes Classifier, Decision Tree Classifier (C4.5), and Support Vector Machine (SVM) with Linear Kernel. It was saw that the combined effects of hate speech turned around so that we have lower accuracy in the youtube database compared to FormSpring tests, as in the text analysis and syntactical features work differently in on both sides. When this hybrid method is used in Twitter Database, led to weak memory and F1 Score. The model proposed by the authors can also be improved used in building constructive mitigation applications cyberbullying problems.

2. J. Yadav, et al. [2] suggests a new approach to this the discovery of internet cyberbullying on social media in using a BERT model with a single line neural and was tested on the Formspring forum and Wikipedia Database. The proposed model provided performance 98% accuracy of spring Form databases and 96% accuracy in a relatively comprehensive Wikipedia database previously used models. The proposed model provided better Wikipedia database results due to its size g without the need for excessive sampling while I The spring data form requires multiple samples.

3. R. R. Dalvi, et al. [3] suggests a way to do this detect and prevent online exploitation on Twitter using Classified supervised machine learning algorithms. In this study, the live Twitter API is used for compilation tweets and data sets. The proposed model tests both Support Vector Machine and Naive Bayes on the data sets are collected. To remove a feature, use it TFIDF vectorizer. The results show that it is accurate of an online bullying model based on Vector Support The machine is about 71.25% better than Naive Bayes was almost 52.75%.

4. Trana R.E., et al. [4] The goal was to design a machine learning model to minimize special events including text extracted from image memes. Author include a site that contains approximately 19,000 text views have been published on YouTube. This study discusses the operation of three learning machines equipment, Uninformed Bayes, Support Vector The machine, as well as the convolutional neural network used in on the YouTube website, and compare the results with existing details of the Form. They do not write continuously investigate cyber bullying algorithms sub-sections within the YouTube website. Naive Bayes beat SVM and CNN in the next four categories: race, nationality, politics, and general. SVM passed well with the inexperienced Naïve Bayes again CNN is in the same gender group, with all three algorithms show equal performance with the middle body group accuracy. The results of this study provided inaccurate data used to distinguish between incidents of abuse and non-violence. Future work can focus on the construction of a two-part separation system used to test text taken from photos to see that the YouTube website provides a better context for aggression-related collections.

5. N. Tsapatsoulis, et al. [5] detailed review of introduced cyberbullying on Twitter. The importance of identifying the various abusers on Twitter is provided. Kwe paper, various practical steps required the development of an effective and efficient application for Internet traffic detection is well defined. I styles involved in data classification and recording platforms, machine learning models and feature types, and model studies using such tools explained. This paper will serve as the first step in the process project on acquisition of cyberbullying technology using the machine reading.

6. G. A. León-Paredes et al. [6] they described I the development of an online bullying detection model is used Native Language Processing (NLP) and Mechanics Reading (ML). Spanish Cyberbullying Prevention The system (SPC) was developed by installing the machine learning strategies Naïve Bayes, Support Vector Machine, and Logistic Regression. Database used this study was posted on Twitter. Plurality 93% accuracy was achieved with

the help of third parties techniques used. Charges of online exploitation were found with the help of this system presented the accuracy of 80% to 91% on average. Stemming and lemmatization techniques in NLP can be used continuously increasing system accuracy. Such a model can and used for adoption in English and local languages if possible.

7. P. K. Roy, et al. [7] details about creating a request for hate speech on Twitter via the help of a deep neural convolutional network. Machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) have it used to identify tweets related to hate speech in Twitter and features have been removed using tf-idf process. The leading ML model was SVM but it managed predict 53% hate speech tweets on a 3: 1 database to be tested train. The reason behind the low forecast scale was unequal data. The model is based on the prediction of hate speech tweets. Advanced learning based methods on the Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and its Content LSTM combinations (CLSTM) have the same effects as separate distributed database. 10 times cross confirmation was used in conjunction with the proposed DCNN model once you got a very good rate of memory. It was 0.88 hate speech and 0.99 non-hate speech. Test results confirmed that the k-fold opposite verification process is a better resolution with unequal data. In the future, the the current database can be expanded for better performance accuracy.

8. S. M. Kargutkar, et al. [8] had proposed a program to give the characters twice for cyberbullying. The system uses Convolutional Neural Network (CNN) and Keras content testing as appropriate strategies then providing them.

III. PROJECT AIM

The main aim of the detecting the cyberbullying model will help to improve manual monitoring for cyberbullying on social networks. In this project we fetch the tweets from twitter accounts and preprocess the twits and images and applying generated model will detect the cyberbullying or not. The objectives of the systems development and event management are: Collect the data set of bullying words and preprocess it and apply natural language processing and then machine learning algorithms Generate different machine learning algorithm model. Fetch the tweets from twitter account and preprocess it. Apply generated model on the fetched tweets and get final output cyberbullying or not.

IV. PROJECT SCOPE

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So overcome these issues detecting the cyberbullying is very important in now a days which will help to stop cyberbullying on social media networks. search and find the the dataset and download it for train the model. After downloading first we will pre process the data and then transferred to Tf Idf. It then trains the dataset using Naive Bayes, Support Vector Machine (SVM), and DNN algorithms and creates models separately. Next, we will develop a web application using the FLASK framework. It imports live tweets from Twitter, then applies the generated model to the imported tweets and checks whether text or images are cyberbullying. For this purpose, we use python as backend, Mysql as database and html, css, javascript etc as frontend.

V. METHODOLOGY

We will develop this project using Python and web technologies. We train the model by first searching, finding, and loading the dataset. After loading, I first preprocess the data and then pass it to the Tf-Idf. It then uses Naive Bayes, SVM (Support Vector Machine), and DNN algorithms to train the data set and create the model separately. Next, we will develop a web application using the FLASK framework. It imports live tweets from Twitter, then applies the generated model to the imported tweets and checks whether text or images are cyberbullying. For this purpose, we use python as backend, Mysql as database and html, css, javascript etc as frontend. threatening situations including grooming or sexually transgressive behaviour, signals of depression

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be needed to prove that one offensive message is part of a series of online bullying directed at users. The rise of cyberbullying and the rise of social media are also on the rise.

We have divided all the features we have extracted into five categories:

- Sentimental Features
- Sarcastic Features
- Syntactic Features
- Semantic Features
- Social Features

Cyberbullying poses a significant threat to the mental and physical health of victims. Although there are projects to detect bullying, there are few implementations of social media monitoring to detect cyberbullying. Therefore, the proposed system is suitable for natural language processing and Automated solutions related to cyberbullying detection have not been adequately studied in the past. This is one of the main causes of errors due to insufficient training data set available. Some data sets of can be used instead of general sentiment analysis, and all are used in a controlled approach. While reports of bullying are published daily, reports of bullying are only a fraction of that compared to hundreds of thousands of messages per second. Random sampling only generates a few aggressive messages, so collecting enough training data is a real big deal.

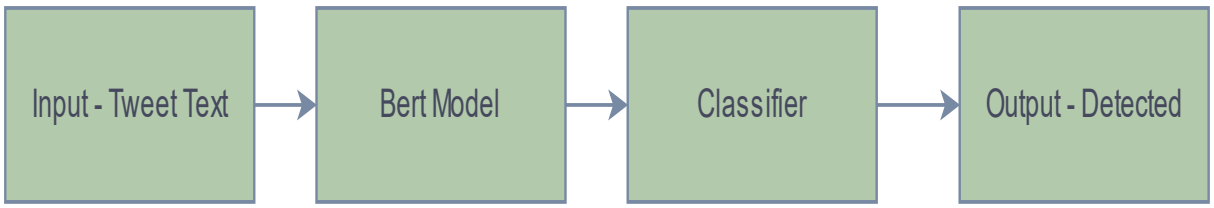

Fig 1: Flowchart of Detection System

VI. PROBLEM STATEMENT

Social networks Networks give us great opportunities to communicate, and also increase the vulnerability of young people to threatening situations on the Internet. Cyberbullying on social media is a global phenomenon due to its large number of active users. The trend shows that social network cyberbullying is increasing rapidly day by day. Recent research shows that cyberbullying is a growing problem among young people. Successful prevention depends on the proper detection of potentially harmful messages, and the information overload of the Internet requires intelligent systems to automatically detect potential hazards. Therefore, in this project, we will focus on creating a model to automatically detect cyberbullying in social media text by simulating messages created by social media bullying.

VII. DETERMINING CYBERBULLIED PERSON BY FOLLOWING DIAGRAMS

Victims may change their profile data, post content with negative emotions, or abruptly leave the network after these interactions. Such provocative interactions could be flagged for later review by someone who could take appropriate action where Twitter disallows user profile entry. To do this, we can create our own systems that can identify such changes and determine how the bullying has affected them. person.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

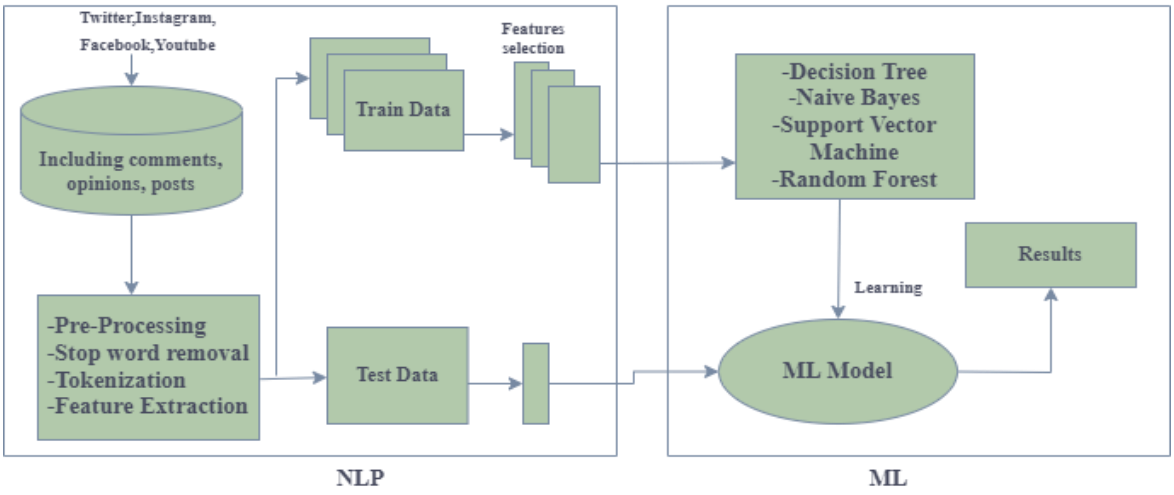

Fig 2: Framework of Cyberbullying Detection System

VIII. OUTPUT OF CYBERBULLYING DETECTION SYSTEM

In this screenshots we have shown the detection plane shows an example of bullying content that is detected as bullying in the result," This is a type of bullying content", and if the text is not the content of cyberbullying there will be a output is "This is not a type of bullying content".

As the framework is done in the Django show that " NAME ,USERNAME ,PASSWORD " for register option and if the user have the account then they can login with already created account with " LOGIN" option.

Cyberbullying research has often focused on detecting cyberbullying 'attacks' and hence overlook other or more implicit forms of cyberbullying and posts written by victims and bystanders. However, these posts could just as well indicate that cyberbullying is going on.The main contribution of this paper is that it presents a system to automatically detect signals of cyberbullying on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders. We evaluated our system on a manually annotated cyberbullying corpus for English and Dutch and hereby demonstrated that our approach can easily be applied to different languages, provided that annotated data for these languages are available.

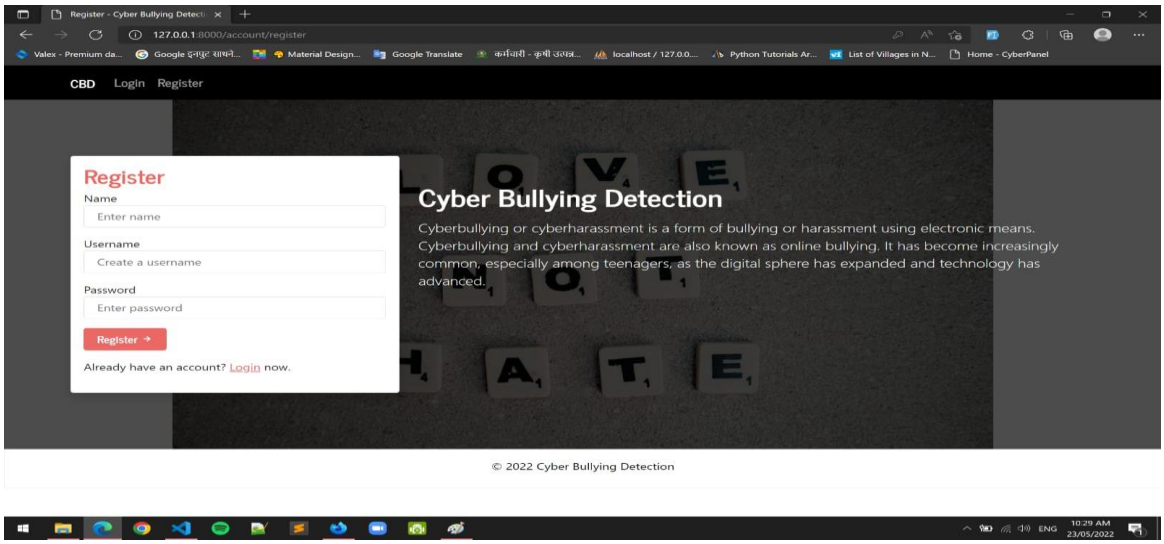

Fig 3: Page of Registration for Cyberbullying Detection

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

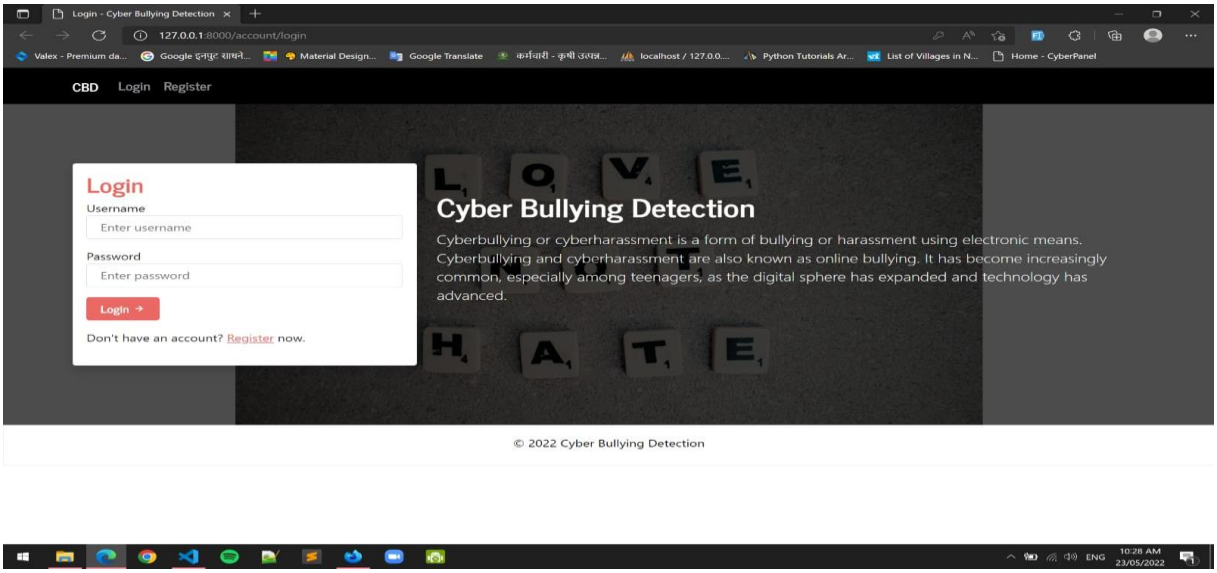

Fig 4: Login Page

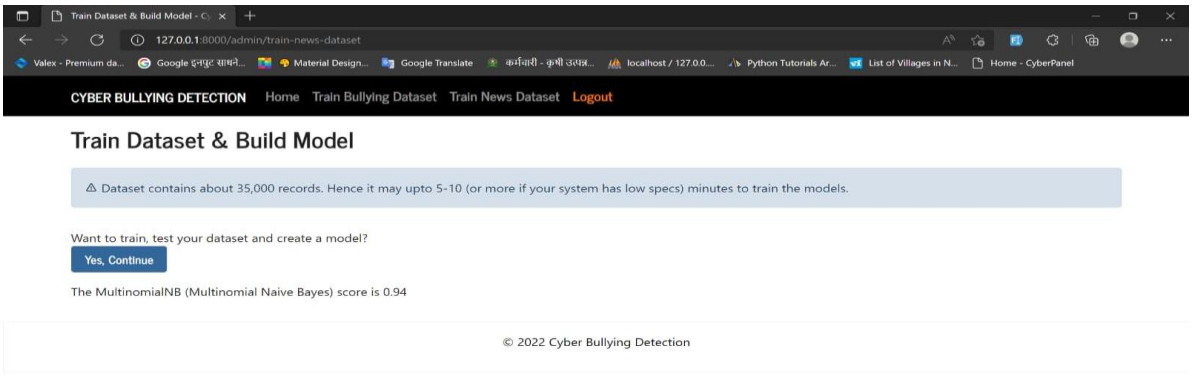

Fig 5: Data set Building

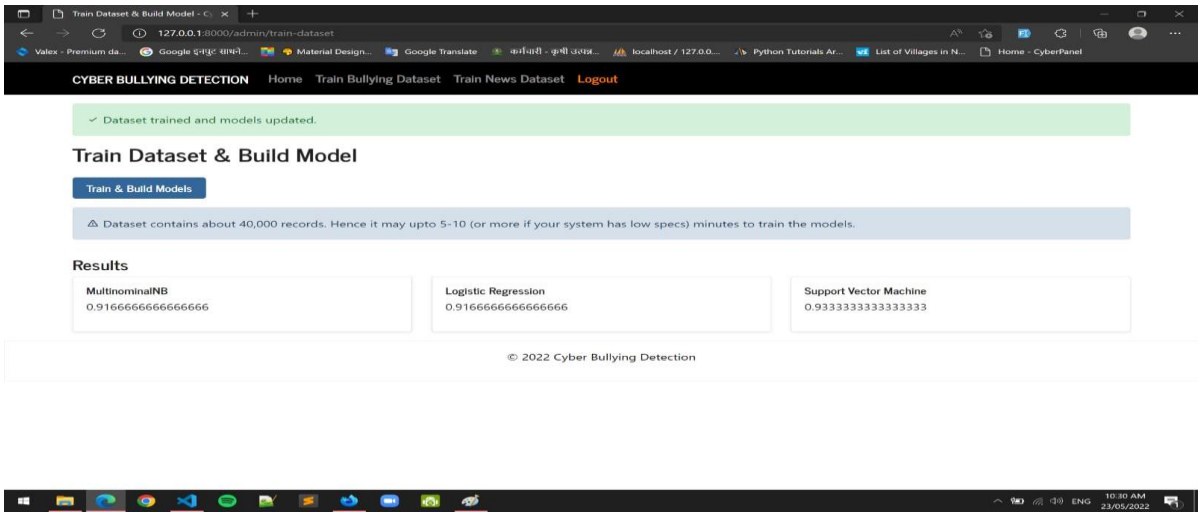

Fig 6: Results of Model

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

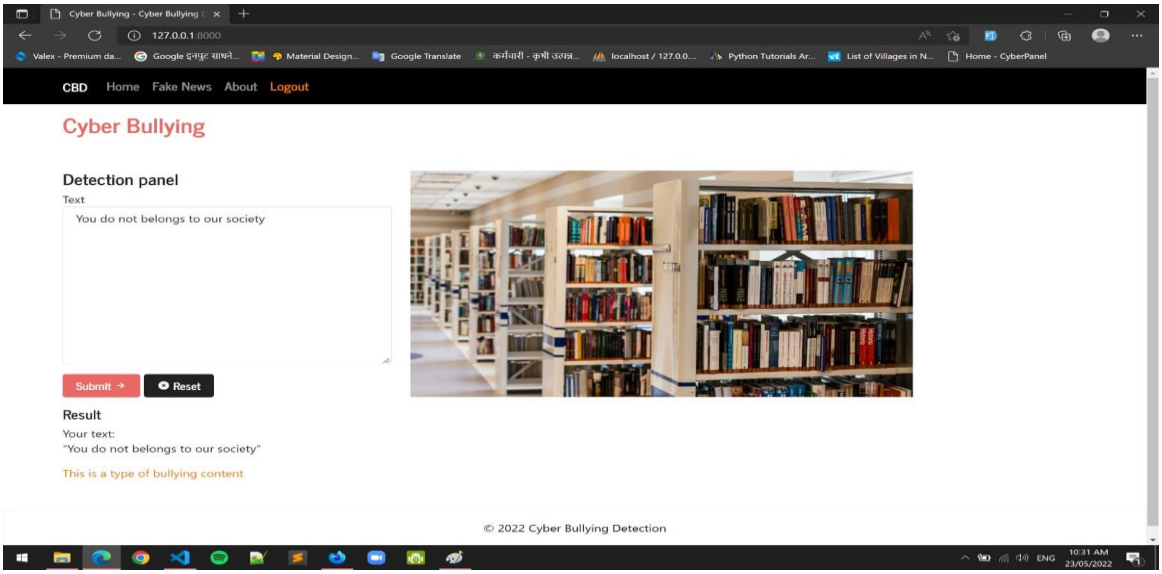

Fig 7: Output of Detection Panel

IX. ADVANTAGES

1. Cyberbullying detection process is automatic and time taken for detection is less.
2. It works on live environment .
3. The latest machine learning models are used for training models that are accurate.

X. APPLICATIONS

1. This software application would be capable of accurately classifying Twitter message negative or positive with respect to some commonly used terms .
2. Mainly focused on gender bullying by using four words with different polarity.

XI. CONCLUSION

We proposed a semisupervised approach in detecting cyberbullying based on the five features that can be used to define a cyberbullying post or message using the BERT model. While considering just one of the features which was sentimental features the BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models. The BERT model can achieve more accurate results if provided with a large data set. We can try to achieve even better results in the cyberbullying detection process if we consider all the features that we have proposed in this research paper. Based on all the features an application can be created to detect the bullying traces and thus help in detecting and reporting such posts. A combination of other models on top of the BERT model may be used in the future to generate a heart condition model for a specific NLP cyberbullying detection task.

XII. REFERENCES

[1] M. Di Capua, E. Di Nardo and A. Petrosino, Unsupervised cyberbullying detection in social networks, ICPR, pp. 432-437, doi: 10.1109/ICPR.2016.7899672. (2016)

[2] D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online].Available: http://www.pcmag.com/article2/0,2817,2388540,00.asp

[3] J. W. Patchin and S. Hinduja, "Bullies move Beyond the Schoolyard; a Preliminary Look at Cyberbullying," Youth Violence and Juvenile Justice, vol. 4, no. 2, pp. 148–169,2006

[4] Anti Defamation League. (2011) Glossary of Cyberbullying Terms.adl.org.[Online].Available: http://www.adl.org/educati on/curriculum connections/cyberbullying /glossary.pdf

[5] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, 2007.

[6] D. Maher, "Cyberbullying: an Ethnographic Case Study of one Australian Upper Primary School Class," Youth Studies Australia, vol. 27, no. 4, pp. 50–57, 2008.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in Proc. Content Analysis of Web 2.0 Workshop (CAW 2.0), Madrid, Spain, 2009.

[8] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, 2011.

[9] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. San Francisco, CA: Morgan Kauffman, 2005.

[10] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kauffman, 1993.

[11] W. W. Cohen, "Fast Effective Rule Induction," in Proc. Twelfth International Conference on Machine Learning (ICML'95), Tahoe City, CA, 1995, pp. 115–123.

[12] D. W. Aha and D. Kibler, "Instance-based Learning Algorithms," Machine Learning, vol. 6, pp. 37–66, 1991.

[13] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," Advances in Kernel Methods, pp. 185–208, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?id=299094.299105

[14] https://www.sciencedirect.com/topics/computer science/deep-neural-network

[15] An Effective Approach for Cyberbullying Detection and avoidance ieee paper

[16] Approaches to Automated Detection of Cyberbullying: A Survey ieee paper

[17] Cyberbullying Detection System on Twitter ieee paper

[18] Methods for Detection of Cyberbullying: A Survey ieee paper

[19] Using Machine Learning to Detect Cyberbullying ieee paper

[20] Deep Learning Algorithm for Cyberbullying Detection ieee paper

[21] Online Social Network Bullying Detection Using Intelligence Techniques ieee paper

[22] Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: The perspective of European children. Initial Findings. London: EU Kids Online; 2010.

[23] Tokunaga RS. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. Computers in Human Behavior. 2010;26(3):277–287.

[24] Mckenna KY, Bargh JA. Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology. Personality & Social Psychology Review. 1999;4(1):57–75.

[25] Gross EF, Juvonen J, Gable SL. Internet Use and Well-Being in Adolescence. Journal of Social Issues. 2002;58(1):75–90.

[26] Juvonen J, Gross EF. Extending the school grounds?—Bullying experiences in cyberspace. Journal of School Health. 2008;78(9):496–505. pmid:18786042

[27] Hinduja S, Patchin JW. Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. Youth Violence And Juvenile Justice. 2006;4(2):148–169.

[28] Van Cleemput K, Bastiaensens S, Vandebosch H, Poels K, Deboutte G, DeSmet A, et al. Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings.) (White Paper). University of Antwerp & Ghent University; 2013.

[29] Livingstone S, Haddon L, Vincent J, Giovanna M, Ólafsson K. Net Children Go Mobile: The Uk report; 2014. Available from: http://netchildrengomobile.eu/reports. [Accessed 30th March 2018].

[30] O'Moore M, Kirkham C. Self-esteem and its relationship to bullying behaviour. Aggressive Behavior. 2001;27(4):269–283.

[31] Fekkes M, Pijpers FIM, Fredriks AM, Vogels T, Verloove-Vanhorick SP. Do Bullied Children Get Ill, or Do Ill Children Get Bullied? A Prospective Cohort Study on the Relationship Between Bullying and Health-Related Symptoms. Pediatrics. 2006;117(5):1568–1574. pmid:16651310

[32] Cowie H. Cyberbullying and its impact on young people's emotional health and well-being. The Psychiatrist. 2013;37(5):167–170.

[33] Price M, Dalgleish J. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. Youth Studies Australia. 2010;29(2):51–59.

[34] Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics. 2014;.

[35] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273–297.

[36] Vandebosch H, Van Cleemput K. Cyberbullying among youngsters: profiles of bullies and victims. New Media & Society. 2009;11(8):1349–1371.

[37] Vandebosch H, Van Cleemput K. Defining cyberbullying: a qualitative research into the perceptions of youngsters. Cyberpsychology and behavior: the impact of the Internet, multimedia and virtual reality on behavior and society. 2008;11(4):499–503.

[38] Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In: ADC.Databases Theory and Applications. Springer International Publishing; 2014. p. 160–171.

[39] Galán-García P, Puerta JGdl, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying

[40] Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: The LREC 2010 Workshop on new Challenges for NLP Frameworks. University of Malta; 2010. p. 45–50.