IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF CYBER BULLYINGUSING MACHINE LEARNING

Ms. Maria Kiran L[1], Divyashree S[2], Neelambika K Nadagoudra[3], and Monalisa P Naik[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Cambridge Institute of Technology (CITech), Bengaluru, India

[2,3,4] Student, Department of Computer Science and Engineering, CITech, Bengaluru, India

Abstract: Cyberbullying is a severe problem that impacts teens and adults on the internet.It has led to incidents like sadness and suicide.The demand for social media platform content regulation is expanding. To develop a model based on the use of natural language processing to identify cyberbullying in text data and machine learning, the following study uses data from two different types of cyberbullying: hate speech tweets from Twittter and comments based on personal assaults from Wikipedia forums. To determine the optimal strategy, three feature extraction techniques and four classifiers are examined. The model yields accuracy levels over 90%.

Keywords: Wikipedia, Twitter, machine learning, hate speech, personal attacks, and cyberbullying

INTRODUCTION

Following the existence of social media platforms allowing online harassment and abuse to occur on an unprecedented scale, cyberbullying is becoming an increasingly prevalent problem in today's society. Therefore, there is a critical need for efficient techniques and tools to detect instances of cyberbullying and protect victims.In recent years, machine learning has shown to be a practical technique for identifying cyberbullying.

Models like LSTM (Long Short-Term Memory) have shown significant promise in precisely identifying cyberbullying events in social media posts.The goal of this project is to employ LSTM to construct a cyberbullying detection system. The system's purpose is to determine whether or not a social media post constitutes cyberbullying based on its word sequence analysis.

There are multiple steps in the system comprising gathering data, preparing it, creating a model, and testing it. The information is gathered from social media platforms and classified as either non-cyberbullying or cyberbullying. Subsequently, the data undergoes preprocessingto eliminate superfluous information and transform the text into a numerical representation appropriate for being fed into an LSTM model.

The preprocessed data is used to construct and train the LSTM model, which helps it identify patterns and connections among words that point to cyberbullying. The model is then assessed to see how well it can identify instances of cyberbullying using a variety of criteria, including accuracy, precision, recall, and F1 score.

LITERATURE REVIEW

Md. Ashraf Uddin, Md. Manowarul Islam, Linta Islam explains In this paper, [1] With the expansion of the internet, usage of social media has increased dramatically overtime, hence making it the 21st-century's most important networking platform.But increased social networking also frequently has unfavorable effects on society, fueling a number of undesirable phenomena like cyberbullying, cyberabuse, cybertrolling, and online abuse. Cyberbullying often results in serious mental andphysical pain, and in extreme situations, it even pushes victims to attempt suicide, especially women and children. Online harassment attracts attention due to its significant negative repercussions on society. Globally, there has been an increase in instances of internet harassment in recent years, including the dissemination of sexual slurs, rumors, and private conversations. As a result, people are becoming more conscious of the telltale indicators of bullying in texts and social media posts. Online harassment attracts attention due to its significant negative repercussions on society. Globally, there has been an increase in instances of internet harassment in recent years, including the dissemination of sexual slurs, rumors, and private conversations. As a result, professionals are become more and more interested in spotting bullying texts or posts on social media. The aim ofthis research is to develop an effective way for recognizing abusive and harassing texts posted online by combining machine learning and natural language processing.

B**Haidar, M. Chamoun, and A. Serhrouch explains** In this paper, [2] Bullying has gone from classrooms and backyardsinto cyberspace because of widespread use of internet and electronic gadgets, this new kind of bullying is known as cyberbullying. Many youngsters worldwide, particularly in Arab nations, are victims of cyberbullying. Consequently, worriesabout cyberbullying are growing. To reduce cyberbullying, a great deal of research is being done. The current focus of study is on cyberbullying identification and mitigation. Previous studies focused on how cyberbullying affects both the victim and the offender psychologically. Studies have suggested ways to identify cyberbullying in English and few other languages, but none have addressed the issue in Arabic as of yet. Number of methods, namely Machine Learning (ML) and Natural Language

R. R. Dalvi, S. Baliram Chavan, and A. Halbe explains In this paper, [3] Young people are being bullied on social media in enormous numbers. The proliferation of social networking sites has led to a daily increase in cyberbullying. By using machine learning to identify word similarities in the tweets that bullies have written, an ML model that can automatically identify bullying behaviors on social media can be created. However, other techniques have been implemented to identify bullying on social media, with the majority depending on textual data. The purpose of this study is to show how software may be used to recognize tweets, posts, and other types of bullying. One suggestion is to identify and prevent bullying on Twitter by using a machine learning system. SVM and Naïve classifiers are the two used for training and testing the social media bullying content. Naive Bayes and Support Vector Machines) demonstrated a 71.25% and 52.70% accuracy rate in identifying the true positives, respectively. However, SVM performs better on the same dataset than Naive Bayes in related studies. Additionally, tweets are fetched via the Twitter API and fed into a model to determine whether or not they are bullying-related.

G. A. Leon-Paredeset al explains In this paper,[4] These days, While information and communication technologies (ICTs) continue to advance, human connections are changing, which make it possible to bring real-world experiences to a virtual platform like the Internet. In this way, issues pertaining to negative behaviors may surface even while the space-time constraints of conventional communication are broken and social bonds are reinforced. Cyberbullying is defined as any act that poses a threat to an individual's overall well-being and is conducted online, can lead to anxiety disorders, sadness, and even suicidal thoughts. It is crucial to identify this kind of behavior as soon as possible because of this.Using Twitter as the foundation for the extraction of knowledge bases or corpus, this study implements a Spanish cyberbullying prevention system(SPC), which is based on Natural Language Processing (NLP) techniques and various machine learning approaches

METHODOLOG

In this paper, we address the problem of binary categorization in relation to cyberbullying. In particular, we distinguish betweentwo primary types of cyberbullying: personal attacks on wikipedia and hate speech on twitter and we classify these

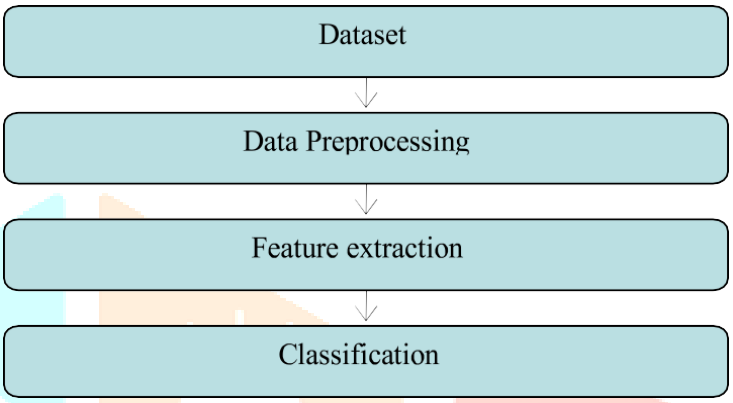

Fig: Detection and Classification Process

Twitter

Two datasets containing hate speech are combined to create the Twitter Dataset: Speech Motivated by Hate Waseem, Zeerak, and Hovy, Dirk's Twitter dataset [11] includes 17,000 tweets flagged as racist orsexist. Explainations are used to mine the tweets.5900 tweets are gone because accounts are shut down or their tweets are removed.Davidson, Thomas; Warmsley, Dana; Macy, Michael; and Weber, Ingmar, Hate Speech Language Dataset [12].

Wikipedia

The Wulczyn Thain, and Dixon dataset on Wikipedia[13] has one million comments that are classified as personal.For the examination Thirteen thousand of the forty thousand comments in the sample are classified as cyberbullying because they involve personal attacks. These remarks are taken directly from discussions between editors of wikipedia articles that have been annotated by ten people using Crowd Flower.

Data preprocessing

Initially, all text data is changed to lowercase.Afterwards, terms like "what's" and "can't" are changed to "what is" and "can not."Additionally, the string library is used to eliminate all punctuation.Next, the Natural Language Toolkit is utilized to apply the subsequent Natural Language Processing techniques: 1) Tokennization 2) Stemming 3) Stop word removal.

Feature extraction

Extraction of features is crucial to Natural Language Processing.Texta data must be translated to numerical data since classifiers cannot classify them. Every document—in this case, a tweet or comment—can be expressed as a vector, and classification can be done using those vectors. 3 feature extraction techniques are examined in the project that follows: 1)Words in a Bag 2)TF-IDF 3)Word2Vec.

Classification

The testing data is converted using the same methodology without being fitted on vectorizers or trained on the word2vec model,after the training data has been obtained by fitting it on the aforementioned feature extraction methods. The following classifierswill be trained and tested on the training set of data that follows: 1) Support vector machine 2) Logistic regrrssion 3) Randomforest 4) Multi layered percepton.

Context Diagram

The DFD is also known as a bubble chart. This simple graphical formalism can be used to show the flow of data into a system.The data flow diagram is one of the most important modeling tools (DFD). It is used to model the many components of the system. DFD shows how information moves through a system and the several ways that modifications might affect it.

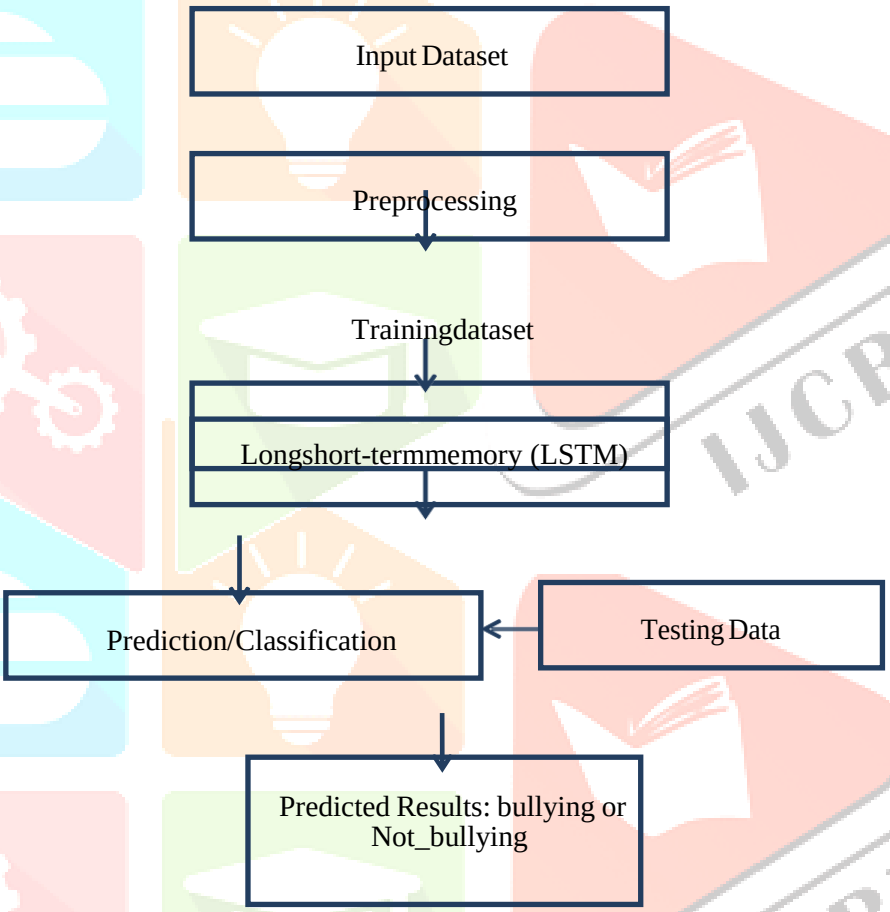

Fig: Context Diagram

RESULTS AND DISCUSSION

In the detection of cyberbullyling research, the title that reads "Detection of Cyberbullying" appears on the homepage.

Fig:HOMEPAGE

This page is designed that the user can log in to the website using it.

Fig: LOGINPAGE

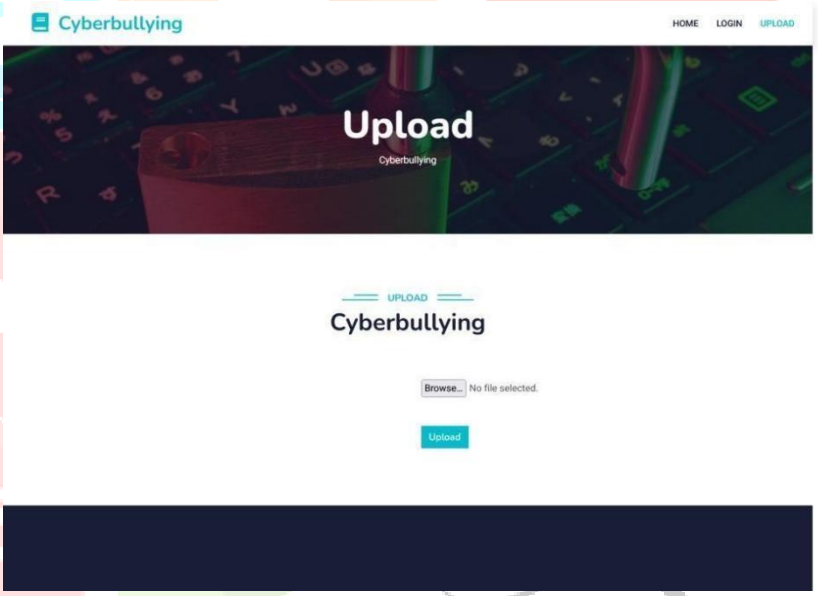

Fig: UPLOADPAGE

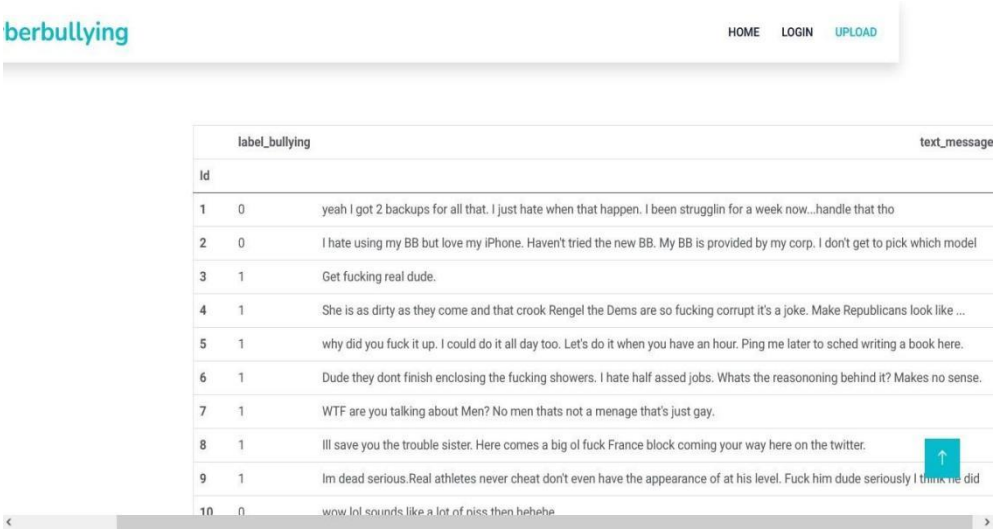

Fig: PREVIEWPAGE

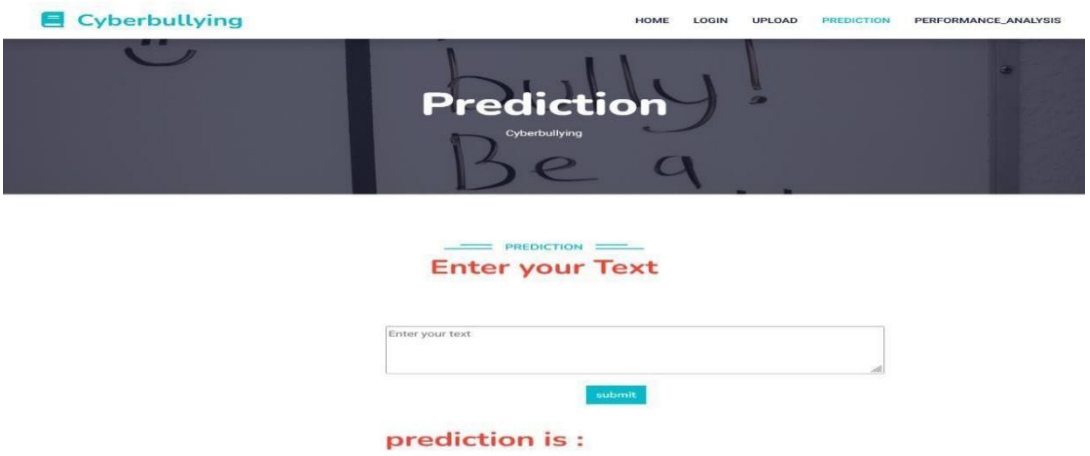

Fig:PREDICTIONPAGE

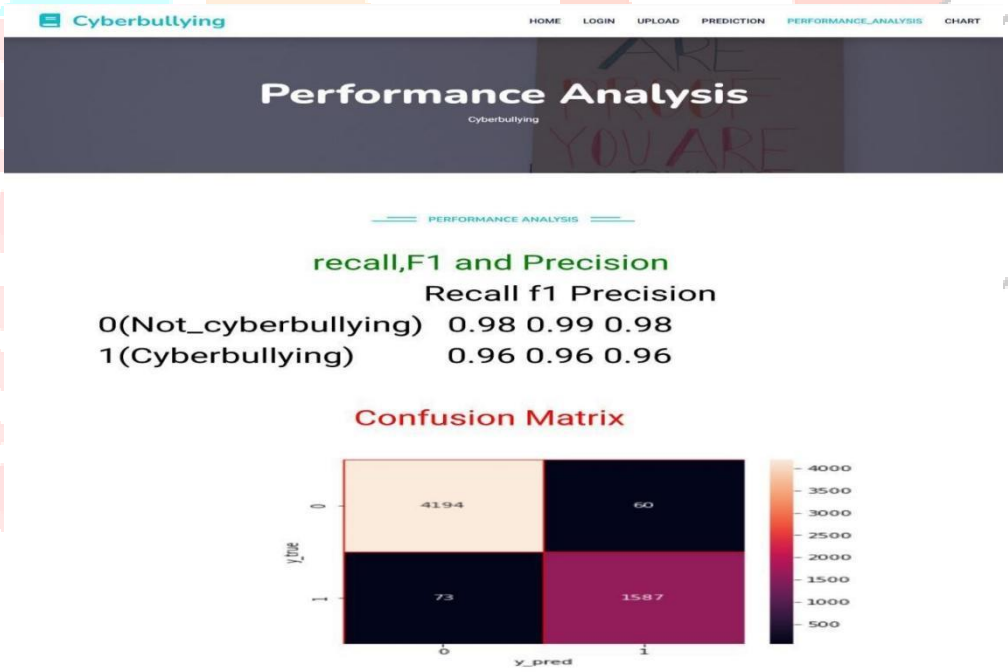

Fig: PERFORMANCE ANALYSIS PAGE

CONCLUSION

The utilization of long short term memory (LSTM) in the development of a cyberbullying detection system has yielded promising results in accurately identifying instances of cyberbullying inside social media messages. This system uses word sequence analysis to detect patterns in social media messages and decide whether or not they are examples of cyberbullying. It does this by applying machine learning and long short-term memory banks (LSTMs) to sequential data.

REFERENCE

[1] Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin Uzzal Kumar Acharjee., 2020, Cyberbullying Detection on Social Media Network using Machine Learning Approaches. In 2020 Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) IEEE.

[2] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Adv. Sci. Technol. Eng. Syst. J., vol. 2, no. 6, pp. 275–284, 2017.

[3] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, "Detecting A twitter cyberbullying using machine learning," in 2020 4th InternationalConference on Intelligent Computing and Control Systems (ICICCS), 2020.

[4] G. A. Leon-Paredeset al., "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language," in 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019.

[5] A. Ali and A. M. Syed, "Cyberbullying Detection using Machine Learning," Pakistan Journal of Engineering and Technology, vol. 3. 2, pp. 45–50, 2020.

[6] Varun Jain, Vishant Kumar, Vivek Pal, Dinesh Kumar Vishwakarma., 2021, Detection of Cyberbullying on Social Media Using Machine Learning. In 2021 Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE.

[7] Rounak Ghosh, Siddhartha Nowal and Dr. G. Manju., 2021, Social Media Cyberbullying Detection using Machine Learning in Bengali Language. In 2021 International Journal of Engineering Research Technology IJERT.

[8] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak., 2019, Detection of Cyberbullying Using Deep Neural Network. In 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS) IEEE.

[9] M. A. Al-Ajlan, and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," in International Journal of Advanced Computer Science and Applications 9.9 (2018).

[10] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," Multimed. Syst., 2020.