iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49369

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Effective Cyberbullying Detection with SparkNLP

A Rishab Vanigotha[1], M R Naveen Kumar[2], Shraddha Hiremath[3], Sujay Sukumaran Adityan[4], Dr. M John Basha[5]
Computer Science Engineering Specialization in Data Science, Jain Deemed to be University

Abstract: ***People of all ages are affected by cyberbullying, which has become a serious problem that can have negative effects like depression and even suicide. As a result, social media content regulation is becoming more and more necessary. Our research project uses SparkNLP, a potent and scalable natural language processing library, to address the problem of cyberbullying.***

A model for detecting cyberbullying was developed by us using a dataset of 48,000 tweets from Kaggle for training. We used the Universal Sentence Encoder (USE) to extract text data, and ClassifierDL, which employs deep neural networks and uses USE as an input for text classification, was used to build the classification model. Class imbalance is a problem, so we used text augmentation to address it.

Age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying are six categories into which our methodology accurately detects and categorizes cyberbullying. We evaluated the model on a held-out set of data to assess its accuracy and resilience, and it produced remarkable results of 93.34% on training data and 89% on test data. The goal of our project is to support and reinforce the ongoing initiatives that are working to stop cyberbullying and to advance a secure and positive online environment for everyone.

Keywords: cyberbullying, twitter, SparkNLP, Snscrape, Universal Sentence Encoder, Classifier DL Approach, Feature Extraction

I. INTRODUCTION

Technology has made the world smaller and brought many advantages, but it also comes with some challenges. One of them is cyberbullying, which is the deliberate use of information technology to harm others. Cyberbullying can manifest in various ways and does not only involve pretending to be someone else or hacking their online accounts. It can also mean saying nasty things about someone or spreading lies to damage their reputation. Because practically everyone uses social media, it is quite simple for anyone to misuse this access and intimidate others online.

Cyberbullying means doing things online that hurt, threaten, humiliate, or target someone for fun or with a plan. These actions are very harmful and can affect anyone quickly and severely. They happen on social media platforms, public forums, and other online information portals where people interact and share information. A cyberbully may not be a stranger; it could be someone you know who wants to hurt you or make you feel bad. Cyberbullying is a serious problem that needs to be addressed and prevented.

II. BACKGROUND

According to a report released in the year 2022 by McAfee's Chief Product Officer, Gagan Singh, cyberbullying has become a critical issue in India. The report reveals that at the age of ten or younger, over one-third of children in India face cyber racism, sexual harassment, and physical harm threats. These alarming statistics have made India the top country in the world for reported cases of cyberbullying.

On practically every social media and messaging platform, Indian youngsters witness and experience the highest cyberbullying. According to the survey, over 85 per cent of Indian children have encountered cyberbullying and they say that they are most likely to be cyberbullied by strangers compared to other children around the world, at 70 percent in India against 45 percent internationally. According to this survey, 42 percent of youngsters in India have been the target of racial cyberbullying, which is 14 percent higher than the rest of the globe at 28 percent.

The research indicates that in addition to racism, other forms of severe cyberbullying have been documented at nearly twice the global average. These include trolling at 36%, personal attacks at 29%, sexual harassment at 30%, threat of personal harm at 28%, and doxing at 23%. Notably, India reported several prevalent cyberbullying behaviors, including the spread of false stories at 39%, exclusion from groups and dialogues at 35%, and name-calling at 34%. These findings reveal the alarming prevalence of cyberbullying in India.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

III.LITERATURE SURVEY

There have been many studies conducted in the past to identify cyberbullying. Some of them, to which we specifically referenced, included:

The research paper “An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques” was published by the authors; Mitushi Raj, Samridhi singh, Kanishka Solanki and Ramani Selvanambi [1] in the year 2022. According to their study results, the CNN-BiLSTM network is the most accurate, and while the CNN alone can only train local characteristics from word n-grams, the CNN-BiLSTM can also learn global features and long-term dependencies due to its LSTM layer.

The authors Aditya Desai, Shashank Kalaskar, Omkar Kumbhar and Rashmi Dhumal [2] published “Cyber Bullying Detection on Social Media using Machine Learning” in the year 2021. The study they conducted can be summed up as a semi-supervised approach in detecting cyberbullying based on the five features utilized. Their BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models.

Muskan Patidar [3] published "Cyberbullying Detection for Twitter Using ML Classification Algorithms" in 2021. The study examined the existing literature for several machine learning algorithms and discovered the Naive bayes N-gram model, which achieved a maximum accuracy of 67%.

Andrea Pereraa and Pumudu Fernando [4] published “Accurate Cyberbullying Detection and Prevention on Social Media” in the year 2020. In this paper, the authors have presented the proposed solution which uses NLP techniques and supervised machine learning to detect cyberbullying accurately. The proposed solution resulted in 74.50% accuracy along with 74% precision, 74% recall and 74% F1 Score.

Also in the year of 2020, there was another publication “A Study of Cyberbullying Detection Using Machine Learning Techniques” by Saloni Mahesh Kargutkar and Prof. Vidya Chitre [5]. In this paper, the author employed a CNN implementation strategy with Keras, resulting in noisy labels.

González-Ibánez, Muresan, & Wacholder, Kumar [6] who published “Identifying sarcasm in twitter: a closer look” in the year 2016. The study done by them can be summarized as such that – Tweepy and Twitter4j are used to stream Twitter tweets but now Because of some limitations imposed by Twitter on the streaming API, one can only download a certain number of tweets in a particular time window.

Ennaji, El Fazziki, Sadgal, and Benslimane [7] who published “Social intelligence framework: Extracting and analyzing opinions for social CRM” in the year 2015. The study done by them can be summarized as such that – Hadoop framework for extracting and evaluating customers' opinions about a product from social networks, the provided framework extracts and analyzes social customer relationship management opinions.

Mohit Tare, Indrajit Gohokar, Jayant Sable, Devendra Paratwar, Rakhi Wajgi [8] who published “Multi-class tweet categorization using map reduce paradigm” in the year 2014. The study done by them can be summarized as such that – To classify the vast quantity of tweets, a Naive Bayes algorithm was applied. Tweets were collected using the Twitter4j library, which leverages the Twitter REST API internally.

Anjali Barskar and Ajay Phulre [9] who published “Opinion Mining of Twitter Data using Hadoop and Apache Pig” in the year 2017. The study done by them can be summarized as such that – Pig is intended for deep Hadoop analysis and also integrates with the flume its ecosystem for data retrieval and storage in HDFS. They also detected the polarity of the tweet, which tells us whether the tweet has a positive or negative meaning.

Peiling Yia, Arkaitz Zubiaga [10] who published “Session-based Cyberbullying Detection in Social Media” reviews existing approaches to cyberbullying detection, with a particular focus on session-based bullying. We examine the Social media Session-based Cyberbullying Detection framework (SSCD).

IV.METHODOLOGY

A. Proposed Methodology

This project approaches identifying cyberbullying using a multi-class classification framework that identifies six different types of cyberbullying: age-based, gender-based, ethnicity-based, religion-based, other forms of cyberbullying, and non-cyberbullying. The goal is to correctly categorize instances of online communication into one of these six groups.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

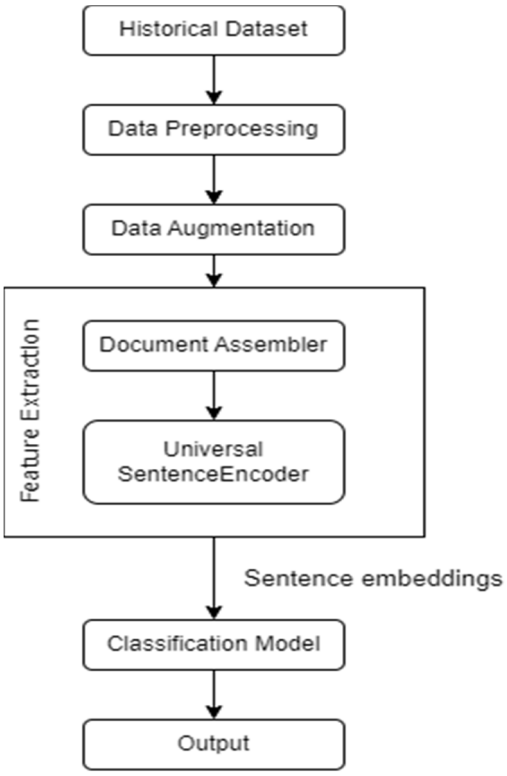


Fig. 1 Dataflow Diagram

1) First we have gathered a Twitter dataset from Kaggle.
2) After importing the data we clean the text by removing any irrelevant information,
3) We use data augmentation to solve the class imbalance problem by generating more data for the underrepresented class.
4) Then data is prepared in a form that can be processed by Spark NLP with the help of Document Assembler.
5) Universal Sentence Encoder is used to convert the document object into a dense vector of numbers that represents a sentence in the text.
6) Then we pass these sentence embeddings to a classification model built inside tensorflow.
7) Finally, snscrape is used to retrieve tweet data in order to validate the model's performance.

B. Dataset

The Kaggle dataset used in this project contains 47692 tweets, each labeled with the appropriate cyberbullying class namely, age, gender, ethnicity, religion, other-cyberbullying and not-cyberbullying. The data has been balanced so that approximately 8000 of each class are present.

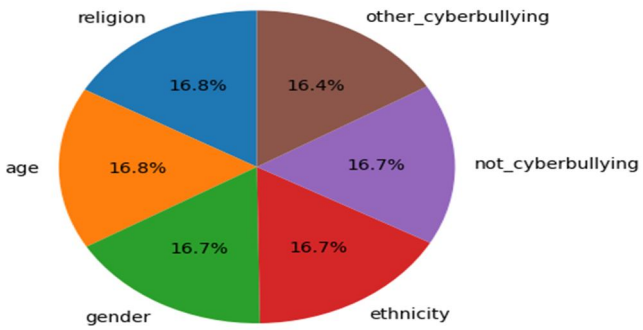


Fig. 2 Distribution of classes in the dataset

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

C. Data Preprocessing

Noisy data in the collected tweets must be cleaned in order to get useful information. We have used regular expressions to clean and remove irrelevant data from the tweets.

1) *Removing nonAscii character:* Non-ASCII characters are any characters that are not part of the ASCII character set, such as accented characters, symbols, and other international characters. Examples include é, ñ, ø, ü, ¥, and ß and emojis in tweets.
2) *Removing links:* Removing links from tweets during data preprocessing can help remove irrelevant information from the data, as well as reduce the amount of text that needs to be processed.
3) *Removing hashtags:* Hashtags can make text more difficult to process for machine learning algorithms, as they often contain symbols that the algorithm does not recognize. Additionally, hashtags typically contain information about the content of the tweet that may not be relevant to the task at hand.
4) *Removing mentions:* The removal of mention(@<name>) aids modeling by decreasing the level of noise in the data. By deleting mentions, the presence of redundant information is eliminated, which can be a distraction to the model and lead to incorrect predictions. Furthermore, by omitting mentions, the model is free to focus on more significant data that can aid in identifying the text's sentiment.
5) *Removing White-spaces:* This includes removing leading and trailing whitespaces, extra whitespaces between words, and whitespaces at the beginning and end of lines. Removing whitespaces can help reduce the size of the text and make it easier to read, as well as improve the efficiency of text processing algorithms. It can also help reduce the amount of noise in the text and make it easier to identify meaningful patterns.

Finally, the preprocessed text is converted to lowercase.

D. Data Augmentation

After data preprocessing, we encountered class imbalance for certain labels which needed to be addressed. Text data augmentation techniques are used to address the issue rather than removing text (tweets) which would lead to data loss.

Data augmentation is an effective way of dealing with class imbalance in a dataset. This technique involves generating additional data points using natural language processing techniques such as synonyms, antonyms, and paraphrasing. This increases the size of the dataset, and helps to improve the accuracy of machine learning models.

We have used “nlpaug” library for data augmentation where contextual word embeddings augmenter is used to generate newdata. We have used a pre-trained word embedding model (bert-based-uncased) to identify context-specific synonyms, which can then be used to generate new data points for training the classification model.

E. Feature extraction

Feature extraction is a process of extracting relevant information from textual data such as words, phrases, or sentences. It is used to identify patterns in large amounts of data and to reduce the dimensionality of the data. In this research, a document assembler was used to prepare the textual data into a format that is processable by Spark NLP. The output of the document assembler was then passed to the Universal Sentence Encoder, which utilizes a pre-trained model (tfhub_use_lg) to generate embeddings from the textual data. These embeddings can then be used to represent the text data for further analysis.

F. Universal Sentence Encoder

The Universal Sentence Encoder is a deep learning model that encodes text into high-dimensional vectors, allowing text categorization, semantic similarity, and other natural language tasks to be performed. The model is trained and optimized for longer text fragments, such as sentences, phrases, or short paragraphs. The input consists of variable-length English text, while the output consists of a 512-dimensional vector. The model is built on a deep average network (DAN) encoder, which, unlike typical word-level embedding models, considers the meaning of entire word sequences rather than individual words. This approach has been proven to be effective at capturing text's semantic meaning and has provided vital insight into natural language understanding challenges.

G. Deep Averaging Network

DAN encoder averages the embeddings of words and bi-grams in a sentence and then passes them through layers of feed-forward deep neural network (DNN) to produce a 512-dimensional sentence embedding as output.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

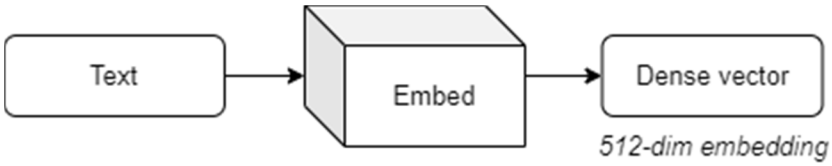


Fig. 3 Structure of Deep Averaging Network

H. Classification

Features extracted from Universal sentence encoder in the form of contextual embeddings are then passed to the ClassifierDL annotator of SparkNLP as inputs.

ClassifierDL is a deep learning-based text classification system that uses a deep neural network to automatically classify text. It is optimized to use state-of-the-art universal sentence encoders as input for text categorization. The annotator supports up to 100 classes, making it ideal for multi-class classification tasks. The model is written in TensorFlow, allowing for efficient training and deployment in the cloud. The model is designed to take in raw text and output the most probable class for the given text. This makes it ideal for finding cyberbullying, as it can quickly and accurately classify text as either cyberbullying or not. The model can also be fine-tuned to better detect cyberbullying, allowing for more accurate and robust results. For this deep learning-based text classification system, we used a batch size of 16, 42 epochs, a dropout rate of 0.4, and a learning rate of 0.004 to train the classification model.

V. EXPERIMENTAL RESULTS

We compared two deep learning models for natural language processing: SparkNLP and bidirectional LSTM. We trained and tested both models on a large corpus of text and evaluated their accuracy and speed. Our results showed that SparkNLP achieved higher testing accuracy (89%) than bidirectional LSTM (81%), but lower training accuracy (93.339%) than bidirectional LSTM (96%). SparkNLP also took less time to complete 42 epochs (14 minutes) than bidirectional LSTM to complete 5 epochs (21 minutes). These findings suggest that SparkNLP is a more efficient and robust model for natural language processing tasks than bidirectional LSTM.

Algorithm	Training Accuracy	Testing Accuracy	No of Epochs	Training Time
Sparknlp (DNNs)	93.339%	89%	42	14 mins
Bidirectional LSTM	96.60%	81.10%	5	21 mins

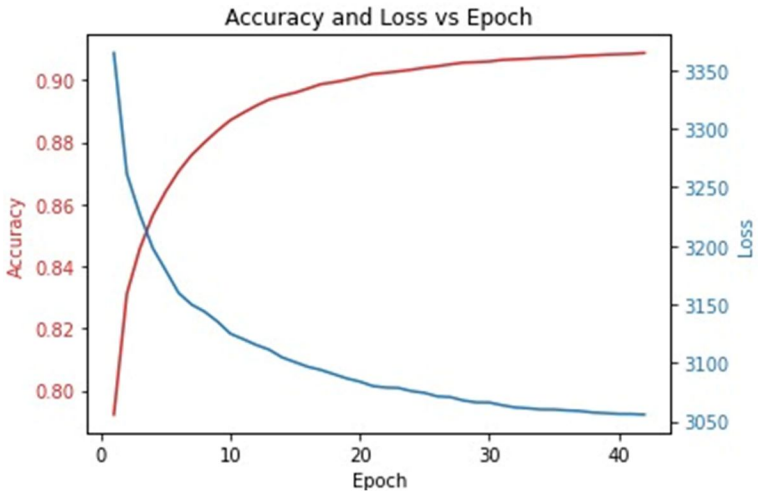


Fig. 4 Accuracy and loss versus Epoch (sparknlp)

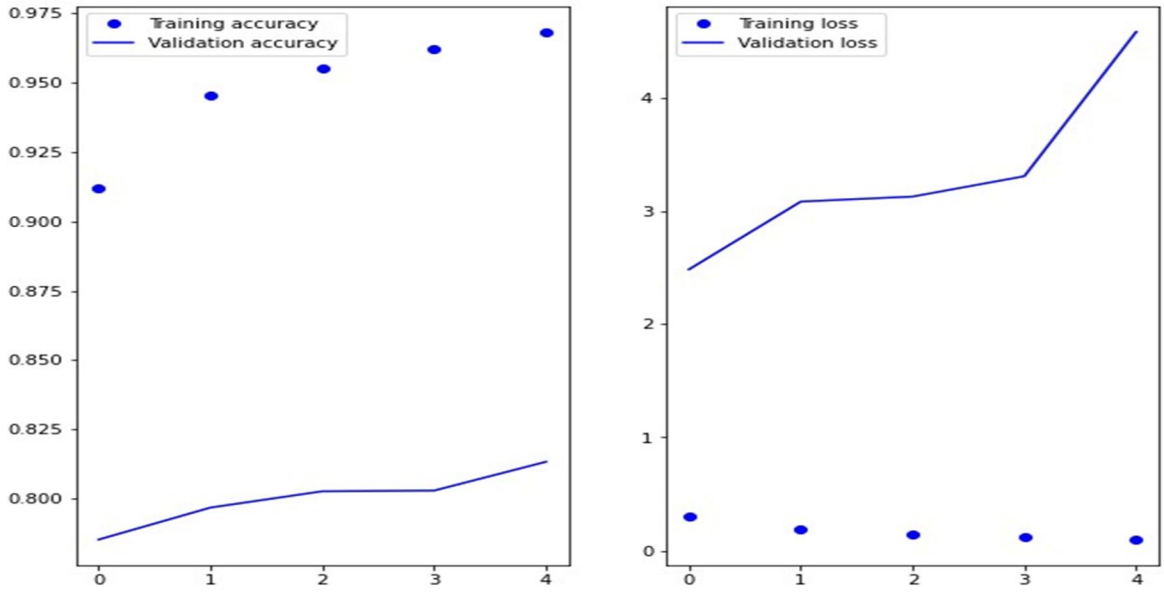


Fig. 5 Training and Validation Accuracy of Bidirectional LSTM

VI. CONCLUSION

In this paper, we have presented a novel approach to identify cyberbullying using a multi-class classification framework that distinguishes six different types of cyberbullying. We have used a Twitter dataset from Kaggle and applied various techniques such as text cleaning, data augmentation, document assembling, universal sentence encoding and tensorflow classification model to process and analyze the data. We have also used snscrape to retrieve tweet data for validating our model's performance. Our results show that our model achieved an accuracy of 89% for testing data and 93% for training data.

This paper contributes to the field of cyberbullying detection by providing a comprehensive framework that can handle multiple categories of cyberbullying and can be applied to different sources of online communication. Our work also demonstrates the potential of using deep learning methods such as universal sentence encoding and tensorflow classification model for text analysis. Furthermore, our work suggests some directions for future research such as exploring other types of cyberbullying (e.g., sexual orientation-based), improving the quality and diversity of the dataset (e.g., using more languages), and developing more robust and interpretable models (e.g., using attention mechanisms).

REFERENCES

[1] Mitushi Raj, Samridhi singh, Kanishka Solanki, Ramani Selvanambi. An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques, Published online 2022 Jul 26, doi: 10.1007/s42979-022-01308-5

[2] Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal. Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences 40, 03038 (2021), doi: https://doi.org/10.1051/itmconf/20214003038

[3] Muskan Patidar, Mahak Lathi, Manali Jain, Monika Dhakad, Prof. Yamini Barge. "Cyber Bullying Detection for Twitter Using ML Classification Algorithms," IJRASET, 2021

[4] Andrea Pereraa, Pumudu Fernando, "Accurate Cyberbullying Detection and Prevention on Social Media," ,2020, doi: https://doi.org/10.1016/j.procs.2021.01.207

[5] Saloni Mahesh Kargutkar, Prof. Vidya Chitre "A Study of Cyberbullying Detection Using Machine Learning Techniques", IEEE, 2020

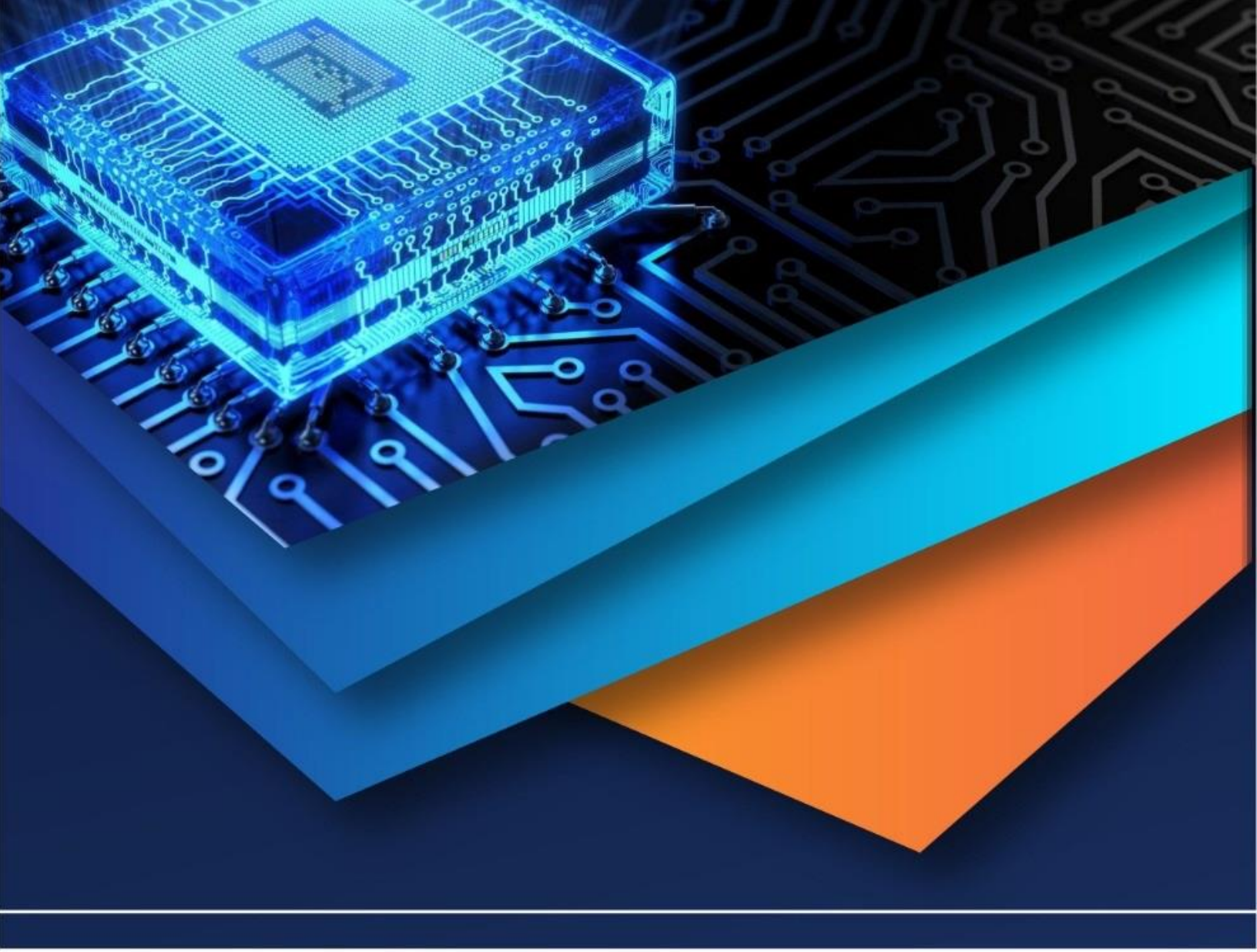
[6] González-Ibánez, R. , Muresan, S. , & Wacholder, N. (2011, June). Identifying sarcasm in twitter: A closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-volume2 (pp. 581–586), Portland, Oregon.

[7] Ennaji, F. Z. , El Fazziki, A. , Sadgal, M. , & Benslimane, D. (2015, November). Social intelligence framework: Extracting and analyzing opinions for social CRM. In Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of (pp. 1–7). Marrakech, Morocco: IEEE.

[8] Tare, M. , Gohokar, I. , Sable, J. , Paratwar, D. , & Wajgi, R. (2014). Multi-class tweet categorization using map reduce paradigm. International Journal of Computer Trends and Technology (IJCTT) , 9(2), 78–81. doi:10.14445/22312803/IJCTT-V9P117

[9] Barskar, A. , & Phulre, A. (2017). Opinion mining of twitter data using Hadoop and Apache Pig. International Journal of Computer Applications , 158, 9. doi:10.5120/ijca2017912854

[10] Peiling Yia, Arkaitz Zubiaga. Session-based Cyberbullying Detection in Social Media.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY