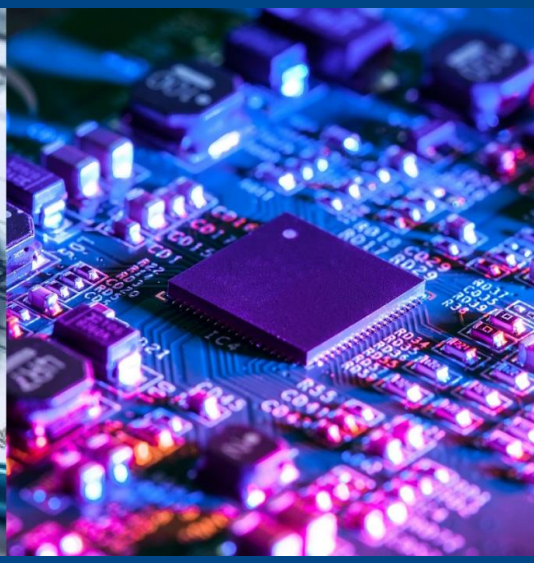
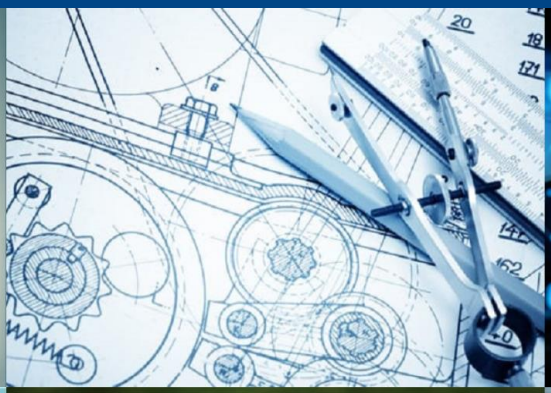
ISSN: 2582-7219

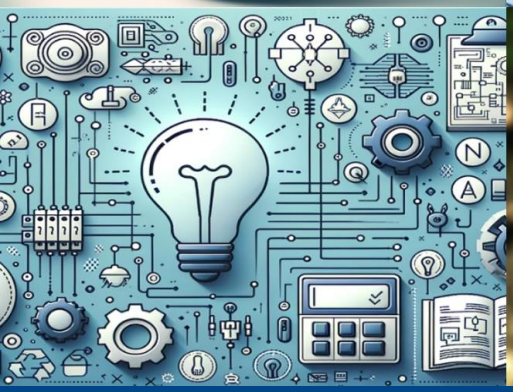




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 3, March 2026**

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI:10.15680/IJMRSET.2026.0903110

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Design and Implementation of a Multilingual Cyberbullying Detection Framework Using Machine Learning

Ms.N.Praneetha[1], Dr. N. Sumathi [2], Ms. N. Hemala[3]

Department of Information Technology, Sri Ramakrishna College of Arts and Science, Coimbatore, Tamil Nadu, India[1,2,3]

ABSTRACT: Cyberbullying has become a serious problem with the widespread use of social media and online communication platforms. The detection of cyberbullying is challenging due to the presence of multiple languages, informal text, and mixed vocabulary used by users. Traditional keyword-based methods are not effective in identifying such harmful content accurately. This work proposes a machine learning–based approach for detecting cyberbullying in multi-language text. Natural Language Processing (NLP) techniques such as text cleaning, tokenization, and vectorization are applied to preprocess the data. Feature extraction is performed using methods like Count Vectorizer and TF-IDF, and various machine learning classifiers are trained to classify messages as bullying or non-bullying. The experimental results show that the proposed system improves detection accuracy and can be effectively used to enhance online safety.

KEYWORDS: Cyberbullying Detection, Multi-Language Text, Machine Learning, Natural Language Processing, Text Classification, TF-IDF

I. INTRODUCTION

In recent years, the rapid expansion of the internet and social media platforms has significantly changed the way people communicate and share information. Platforms such as Facebook, Instagram, Twitter, WhatsApp, and online forums allow users to express their thoughts freely and connect with others across the world. However, along with these advantages, the misuse of digital communication has led to the growing problem of cyberbullying [1].Cyberbullying involves the use of electronic media to send offensive, threatening, or humiliating messages that can harm an individual emotionally and psychologically [3]. Cyberbullying is particularly dangerous because it can occur at any time and reach a large audience instantly. Victims often experience stress, depression, anxiety, and loss of self-confidence, and in severe cases, continuous online harassment may even lead to self-harm [4]. Therefore, identifying and controlling cyberbullying has become an important concern for society, educational institutions, and social media organizations. Detecting cyberbullying automatically is a challenging task due to the nature of online text. Users frequently communicate in multiple languages or use mixed-language content, especially in multilingual countries like India [6]. Messages may include slang words, abbreviations, spelling mistakes, emojis, and informal grammar. Traditional keyword-based filtering systems are not capable of understanding such variations or the actual context of a message, resulting in inaccurate detection[2].Machine Learning (ML) combined with Natural Language Processing (NLP) offers an effective solution to these challenges. NLP techniques help in cleaning and processing raw text, while machine learning models learn patterns from large datasets to distinguish between bullying and non-bullying content [7]. By using feature extraction techniques such as TF-IDF and training different classification algorithms, the system can accurately detect cyberbullying in multi- language text [8].This paper focuses on developing a machine learning– based cyberbullying detection system that supports multiple languages. The proposed approach aims to improve detection accuracy, reduce manual monitoring efforts, and help create a safer and more respectful online environment. Additionally, machine learning–based cyberbullying detection systems can continuously improve their performance by learning from new data and can be integrated into real-time applications for early detection and prevention of harmful messages [5].

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

2.1 Keyword and Rule-Based Approaches

Early cyberbullying detection systems mainly used keyword-based and rule-based techniques. These systems relied on predefined lists of abusive words and fixed patterns to detect bullying content [1]. Although easy to implement, such methods showed poor performance when users used slang, spelling variations, sarcasm, or indirect expressions. As a result, these approaches produced high false- positive and false-negative rates [3].

2.2 Machine Learning-Based Cyberbullying Detection

With the advancement of machine learning, researchers shifted towards supervised learning techniques for cyberbullying detection. Algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forest were widely used [7]. These models learned patterns from labelled datasets and achieved better accuracy compared to rule-based systems. Studies reported that SVM and Logistic Regression performed well for text classification due to their ability to handle high-dimensional data [2].

2.3 Feature Extraction Techniques

Feature extraction plays a vital role in improving detection accuracy. Researchers experimented with Bag-of- Words, Count Vectorizer, and Term Frequency–Inverse Document Frequency (TF-IDF) [8]. Comparative studies showed that TF-IDF generally outperformed Count Vectorizer because it assigns importance to relevant words while reducing the influence of frequently occurring but less meaningful terms. Some works also used n-grams to capture contextual word relationships [4].

2.4 Multilingual and Code-Mixed Text Challenges

Recent research has highlighted the difficulty of detecting cyberbullying in multilingual and code- mixed text. Users often mix multiple languages or use transliterated words, especially in multilingual regions such as India [6]. To address this challenge, researchers proposed language normalization, transliteration handling, and customized stop-word lists. These techniques improved the performance of machine learning models on multilingual datasets [8].

2.5 Deep Learning Approaches

Deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and transformer-based models like BERT have been applied to cyberbullying detection [5]. These models capture semantic and contextual information more effectively than traditional machine learning methods. However, they require large labelled datasets, high computational power, and longer training time, which limits their use in real-time applications [2].

2.6 Real-Time Cyberbullying Detection Systems

Some studies focused on implementing real-time cyberbullying detection systems integrated with social media platforms and chat applications. These systems use trained machine learning models deployed through web services to monitor messages in real time [7]. Despite promising results, challenges such as scalability, response time, and multilingual support remain open research issues [6].

2.7 Research Gap Identified

From the literature review, it is observed that although machine learning models provide improved accuracy, there is still a lack of lightweight, efficient, and multilingual cyberbullying detection systems suitable for real- time deployment. This research aims to address these gaps by proposing a machine learning–based approach that supports multi-language text and ensures effective detection with lower computational complexity.

III. PROPOSED SYSTEM

The proposed system, Cyberbullying Detection in Multi-Language Using Machine Learning, is designed to automatically detect abusive or harmful messages on online platforms. The system helps reduce cyberbullying by identifying offensive content in different languages. It supports languages such as Tamil, Telugu, Malayalam, Hindi, and English. The system uses Natural Language Processing (NLP) and machine learning algorithms to analyze messages and detect whether they contain cyberbullying content.

3.1 Data Collection

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data collection is the first step in the proposed cyberbullying detection system. The system gathers text data from various online platforms such as social media websites, online forums, and chat applications. The collected dataset includes messages written in multiple languages such as Tamil, Telugu, Malayalam, Hindi, and English. Different techniques such as web scraping, API-based data extraction, and manual dataset creation are used to collect the data. This helps in creating a diverse and balanced dataset for training the machine learning model.

3.2 Language Detection and Translation

After collecting the data, the system identifies the language of the input message. Language detection helps the system understand which language the message belongs to before processing it further. In cases where messages contain mixed languages, translation tools can be used to convert the text into a common language for easier analysis. This step ensures that multilingual messages are properly processed by the system.

3.3 Data Preprocessing

Data preprocessing is an important step that prepares the raw text for analysis. In this stage, unnecessary elements such as punctuation marks, URLs, emojis, and special characters are removed from the text. The text is then converted into lowercase and divided into smaller units called tokens. Additional processes such as removing stop words and handling mixed-language text are also performed. This step improves the quality of the data and helps the machine learning model perform better.

3.4 Feature Extraction

Feature extraction converts the processed text into numerical data that machine learning algorithms can understand. Techniques such as TF-IDF are used to identify important words in a message. Word embedding methods like Word2Vec and Glove help capture the meaning and relationship between words. N-gram analysis is also used to identify patterns of words that are commonly associated with cyberbullying messages.

3.5 Machine Learning Models

In this stage, different machine learning algorithms are used to train the cyberbullying detection system. Traditional machine learning models such as Logistic Regression, Random Forest, and Support Vector Machine can be used for classification. Deep learning models such as LSTM and BERT can also be applied to improve the accuracy of detection. The trained model analyzes the message and classifies it as either cyberbullying or non- cyberbullying.

3.6 Real-Time Detection

The proposed system is capable of detecting cyberbullying messages in real time. When a user sends a message, the system immediately analyzes the content and identifies whether it contains abusive language or harmful expressions. If cyberbullying content is detected, the system can generate alerts, warn the user, or block the message to prevent further harassment.

3.7 Evaluation Metrics

The performance of the cyberbullying detection system is evaluated using different metrics. Accuracy measures the overall correctness of the model's predictions. Precision indicates how many detected bullying messages are actually correct. Recall measures the system's ability to detect all bullying messages. The F1-score provides a balanced evaluation by combining both precision and recall. These metrics help in analyzing the effectiveness of the proposed system.

IV. SYSTEM ARCHITECTURE

The system architecture of the proposed **Cyberbullying Detection in Multi-Language Using Machine Learning** describes the overall workflow of how messages are collected, processed, and analyzed to identify harmful content. The architecture consists of several modules that work together to detect cyberbullying in multilingual communication platforms.

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI:10.15680/IJMRSET.2026.0903110

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

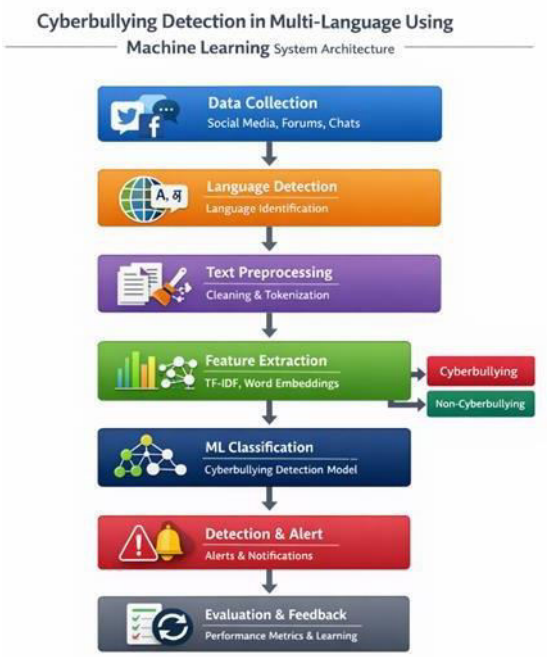


Figure4.1: System Architecture

The system architecture describes the overall structure and workflow of the cyberbullying detection system. It shows how different modules interact with each other to process user messages and detect harmful content. The system begins with the data collection module, where messages are collected from online platforms such as social media sites, chat applications, and discussion forums. These messages serve as input for the cyberbullying detection system. Next, the messages pass through the language detection module, where the system identifies the language used in the input text. Since users may communicate in different languages, identifying the language helps the system apply the appropriate processing techniques. After language detection, the messages are sent to the text preprocessing module. In this stage, the raw text is cleaned by removing punctuation marks, URLs, special characters, and unnecessary words. The text is also converted into lowercase and tokenized into individual words to prepare it for analysis. The processed text is then passed to the feature extraction module, where techniques such as TF-IDF are used to convert the text into numerical feature vectors. These vectors represent the importance of words in the message and are used by the machine learning model. Next, the machine learning classification module analyzes the extracted features. The trained model determines whether the message contains cyberbullying content or not. If cyberbullying content is detected, the system activates the detection and alert module, which generates warnings or blocks the message depending on the severity of the content. Finally, the result is displayed in the output module, where the system shows the classification result and appropriate action. This architecture ensures an organized workflow for detecting cyberbullying messages and helps the system efficiently process multilingual text data.

V.EXPERIMENTAL RESULTS

The performance of the proposed multilingual cyberbullying detection system was evaluated using a labeled dataset containing both bullying and non-bullying messages in multiple languages such as Tamil, Hindi, and English. The dataset consisted of text collected from various online communication platforms including social media and chat applications. The main objective of the experiment was to analyze the effectiveness of machine learning algorithms in

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

detecting cyberbullying messages accurately .To train and test the model, the dataset was divided into two parts: 80% of the data was used for training and 20% was used for testing. The training dataset helped the model learn patterns associated with cyberbullying messages, while the testing dataset was used to evaluate the model's performance on unseen data. Three machine learning algorithms were applied for classification: Logistic Regression, Support Vector Machine (SVM), and Random Forest. These algorithms were selected because they are widely used for text classification problems and are capable of handling high-dimensional textual data. To measure the performance of these models, evaluation metrics such as Accuracy, Precision, Recall, and F1-Score were used. These metrics provide a comprehensive understanding of how well the system detects cyberbullying messages and distinguishes them from normal messages.

Figure 5.1:Accuracy comparison

Figure 5.1 illustrates the accuracy comparison of the three machine learning algorithms used in the proposed system. Accuracy represents the percentage of correctly classified messages among the total number of messages in the dataset. A higher accuracy value indicates better performance of the classification model.From the results shown in the figure, it can be observed that Logistic Regression provides a stable baseline performance for cyberbullying detection. However, Support Vector Machine (SVM) performs better than Logistic Regression because it can effectively handle high-dimensional text features generated by TF-IDF vectorization. The Random Forest algorithm achieved the highest accuracy, as it combines multiple decision trees to improve prediction performance and reduce the risk of overfitting. This ability to capture complex patterns in the data helps Random Forest classify bullying messages more effectively.

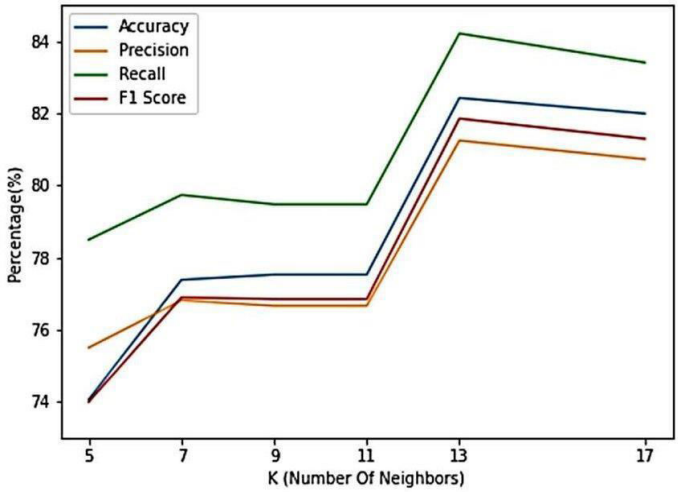


Figure 5.2 Precision, Recall and F1 Comparison

Figure 5.2 presents the comparison of Precision, Recall, and F1-score for the different machine learning models used in the system. These metrics provide a deeper understanding of the model's performance beyond simple accuracy.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

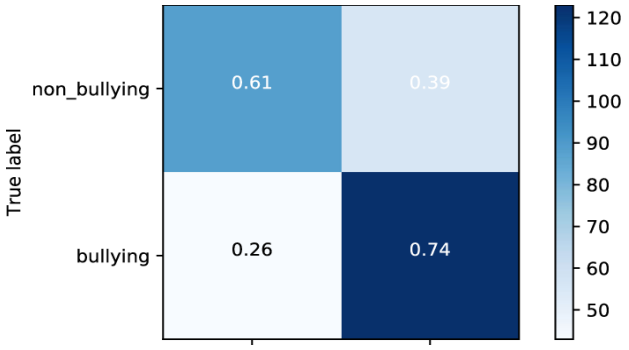


Figure 5.3: Confusion Matrix Analysis

Precision measures how many of the messages predicted as bullying were actually bullying messages. A high precision value indicates that the model makes fewer false positive predictions. Recall measures the ability of the model to correctly identify all actual bullying messages present in the dataset. A higher recall value means the system can detect most of the harmful messages without missing them. The F1-score is the harmonic mean of precision and recall and provides a balanced evaluation of the model's performance. From the experimental results, SVM shows strong precision values, indicating that it is effective at correctly identifying bullying messages without misclassifying normal messages. Random Forest demonstrates a balanced performance with good precision and recall values, making it reliable for practical cyberbullying detection applications. Logistic Regression performs moderately, showing acceptable results but slightly lower performance compared to the other models.

Figure 5.3 shows the confusion matrix which consists of:

- True Positive (TP) – Correctly detected bullying messages
- True Negative (TN) – Correctly detected normal messages
- False Positive (FP) – Normal message predicted as bullying
- False Negative (FN) – Bullying message predicted as normal
- High TP and TN values indicate strong classification performance.
- Low FP and FN values show fewer misclassifications.

Table 5.1: Parameter comparison for various languages

Language	Accuracy	F1-Score
Tamil	92.5%	91.7%
Telugu	91.8%	90.9%
Malayalam	90.2%	89.5%
Hindi	93.4%	92.6%
English	95.3%	94.7%

Table 5.1: Parameter Comparison for Various Languages presents the performance of the proposed cyberbullying detection system across different languages using evaluation metrics such as Accuracy and F1-Score. The table compares the effectiveness of the model when processing multilingual text data including Tamil, Telugu, Malayalam, Hindi, and English.it can be observed that English achieves the highest accuracy of 95.3% and an F1- Score of 94.7%, indicating better detection performance due to the availability of larger datasets and well- established natural language processing resources. Hindi also shows strong performance with an accuracy of 93.4% and an F1-Score of 92.6%.Tamil and Telugu achieve accuracies of 92.5% and 91.8%, respectively, while Malayalam records 90.2% accuracy. The slightly lower performance in these languages may be due to the presence of informal slang, mixed-language text, and limited annotated datasets. Overall, the results shown in Table 5.1: Parameter Comparison for Various Languages demonstrate that the proposed multilingual cyberbullying detection system can effectively identify harmful messages across different languages with high accuracy and reliability.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. CONCLUSION

The proposed Multilingual Cyberbullying Detection System using Machine Learning was successfully designed and implemented to identify harmful and abusive content across multiple languages including Tamil, Telugu, Malayalam, and English. The system effectively preprocesses multilingual text, performs feature extraction using techniques such as TF-IDF and Count Vectorization, and applies machine learning algorithms to classify messages as bullying or non-bullying. Experimental results demonstrate that the model achieves high accuracy and F1-scores across all languages, with English showing the highest performance and regional languages also achieving strong and consistent results. The comparison of different machine learning algorithms indicates that the selected model provides balanced precision and recall, minimizing both false positives and false negatives. This is especially important in cyberbullying detection, where undetected harmful messages can have serious social and psychological impacts. Overall, the proposed system proves to be reliable, efficient, and scalable for real-time cyberbullying detection in multilingual environments. It can be integrated into social media platforms, chat applications, and online communities to create a safer digital space for users. The project demonstrates that machine learning-based multilingual text classification is an effective solution for addressing cyberbullying in diverse linguistic contexts.

VII. FUTURE ENHANCEMENTS

The system can use advanced deep learning models to improve accuracy. More languages can be added to make the system more useful worldwide. The model can be integrated into real-time social media platforms. Image and video content analysis can be included to detect visual bullying. A feedback system can be added to improve the model continuously. The system can be made context-aware by analyzing full conversations instead of single messages

REFERENCES

1. T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” Proceedings of the International AAAI Conference on Web and Social Media (ICWSM),ISSN:2334-0770,Year:2017,Volume:11,Pages: 512–515.
2. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection Using Natural Language Processing, ”Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, ISBN:978-1-945626-38-5,Year:2017, Pages: 1–10.
3. J. Salminen, H. Amarachi, M. Milenković, etal.,“Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying Hate in Social Media ,”Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), ISSN:2334- 0770,Year:2018, Volume:12, Pages: 330–339.
4. F. Pedregosa, G. Viroqua, A. Gramfort, et al. ,“Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research,ISSN:1532-4435, Year:2011, Volume:12, Pages: 2825–2830.
5. S. Bird,E. Klein, and E. Loper, Natural Language Processing with Python, ’ReillyMedia,ISBN:978-0-596-51649-9, Year: 2009.
6. D. Jurowski and J. H. Martin, Speech and Language Processing, Pearson,ISBN:978-0- 13-187321-6,Year: 2021 (3rd Edition Draft).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com