iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81404

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 14 Issue IV Apr 2026- Available at www.ijraset.com

Detection of Toxic Comments Using Deep Learning with XLM-R

G. Sivaramakrishna[1], K. S. Faizz Ahmad[2], V. Chennakesava Rao[3], G. Phani Gopi Chand [4], T. Venkatesh[5]
Department of Computer Science and Engineering, University College of Engineering and Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, India

ABSTRACT: ***Socialmedia represent one of the main ways for communicate with each other across the world. However, the openness of these platforms also increases the spread of offensive, abusive and harmful comments. Such toxic content can negatively affect user experience and create unsafe digital environment. Detecting such harmful comments becomes more difficult. This study proposes a tool for detecting toxic comments through implementing a deep learning mechanism, which is supported by use of the XLM-R (Cross-Lingual Language Model RoBERTa) transformer model. To build the model, a large dataset including both toxic and non-toxic comments. The trained model was integrated with a web application that detects toxicity in real time social media platforms. This system can help improve online safety and increase the quality of communication.***

INDEX TERMS: ***Toxic comment detection, NLP (Natural Language Processing), XLM-R, deep learning, online safety.***

I. INTRODUCTION

The rapid expansion of social media has significantly changed the way people communicate and share information. Millions of users post comments, opinions, and messages every day. The comments contain harmful and toxic content such as abusive language. Such content can negatively affect individuals and create an unsafe online environment.

This research proposes a toxic comment detection system that can identify toxic and non-toxic comments. If the comment contains any toxic words, it is unable to post the comment but successfully posts the comment when it does not contain any toxic words. The main goal of this project is to improve online safety by developing a system that works effectively and improves the safety of online communication environment.

A. Research Contributions

The key contributions of this research are:

- Development of a toxic comment detection system designed for English language comments.
- Utilization of a dataset containing 159,571labeled comments categorized into six toxicity types.
- Implementation and fine-tuning of the XLM-R transformer model for classification.
- Integration of the trained model into a web application capable of real-time toxicity detection.
- Experimental evaluation demonstrating strong performance compared with traditional methods.

B. Background

Digital communication platforms have become an important part of everyday life. Social media allows individuals to communicate easily. Despite this advantages, online platforms frequently contain abusive messages. Exposure to toxic comments can discourage users from participating in discussions and create negative online experiences.

C. Problem Definition

The continuous increase usage of social media platforms has led to a significant rise in toxic language and harassment. Traditional methods take more time to detect with less accuracy. Therefore, there is a need to design a system that can automatically identify the harmful text. This research resolves these challenges by applying deep learning models like XLM-R to detect comments in English for safer online communication.

©IJRASET: All Rights are Reserved

D. Problem Purpose

The key purpose of this project is to develop an automated, real-time toxic comment detection system using advanced NLP and deep learning models. The system aims to protect users from harmful content, improve the quality of online platforms, and support safer communication.

E. Scope of the Problem

This project focuses on a design system using the XLM-R transformer model for detecting comments in English. It can identify the harmful content in the comments. If the comment contains any toxic words, it is unable to post the comments, and successfully posts the comment when it does not contain any toxic words. It involves using a realistic dataset, training a transformer-based model for accurate classification, and integrating it into a web application that performs real-time toxicity detection. This system aims to support safer online communication.

F. Problem Features

The main problem is the difficulty in detecting toxic comments in online social media platforms. Users often write comments in informal ways, including spelling mistakes, abbreviations,slang, and inconsistent grammar which makes automatic detection more challenging. In addition, toxic comments can appear in different forms such as insults, threats, obscene language, and hate speech. Because of this variation in writing styles and expressions, identifying harmful content becomes a complex task. Therefore, there is a need for a smart and flexible system that can understand informal text and accurately detect toxic content. Such a system should be capable of analyzing user comments effectively and identifying different types of toxicity in real time.

II. LITERATURE SURVEY

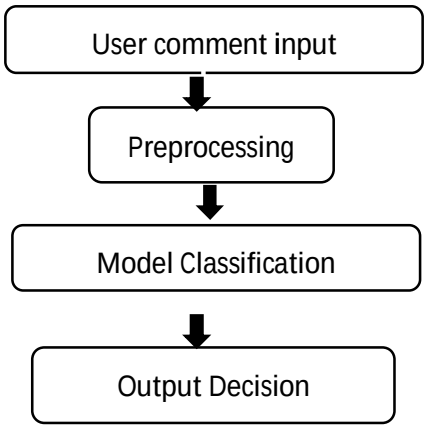
The problem of detecting toxic comments and abusive language in online text has been studied by many researchers in recent years. Early research focused mainly on English language content, using traditional machine learning methods such as support vector machines (SVM) and Naive Bayes classifiers. These models were trained using manually extracted features from text and worked well for simple text but struggled with complex languages, slang, and informal expressions commonly used in social media.

With the advancement of deep learning techniques, particularly through the use of neural networks for toxicity detection. Models based on Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) were introduced and showed improved performance over traditional methods. These models improved performance by automatically learning meaningful representations from data.

More recent studies have applied transformer-based models, such as BERT and its multiple versions, which are better at capturing the meaning and context of words in sentences. This model works well for complex and mixed language inputs. Despite these advantages, most existing systems still focus on single language with low accuracy.

III. EXISTING SYSTEM

Various techniques have been proposed for detecting toxic comments in online platforms.



1) Keyword-based or rule-based systems

These are the earliest approaches to detecting toxic comments. They rely on a predefined list of toxic words. Each comment is scanned, and if a match is found with any word in the toxic keyword list, the comment is flagged as toxic.

2) Traditional Machine LearningBased System

Machine learning algorithms can learn patterns from labeled datasets. Common techniques include:

- Naive Bayes
- Support Vector Machine
- Logistic Regression
- Random Forest Classification

These models learn patterns from labeled datasets instead of matching keywords. Although these methods improve accuracy but still struggling to capture complex relationships.

3) Deep Learning-Based System

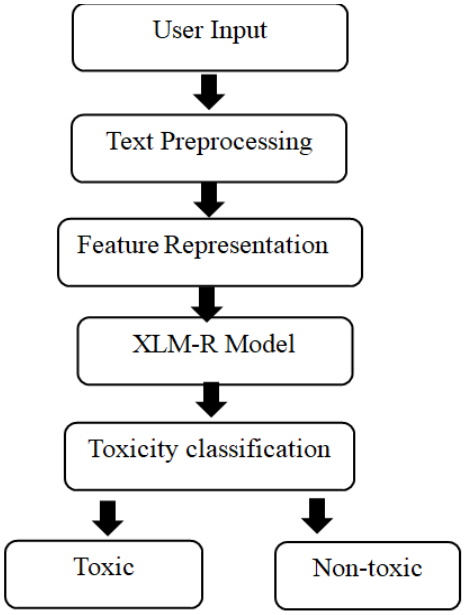
Deep learning models, such as LSTM (Long Short-Term Memory), BiLSTM, and CNN (Convolutional Neural Network), can detect the toxic comments. These automatically extracts features from data and can capture sequential relationships between words.

4) Disadvantages of Existing Systems

- Limited language supports
- Poor handling of large data
- Unable to find toxic percentage

IV. PROPOSED SYSTEM

The proposed system is includesthe following steps:



The proposed system is designed to automatically detect toxic comments in English with type. It can handle comments in informal ways, including spelling mistakes, abbreviations, slang, and inconsistent grammar, which is commonly used on social media. If the comment contains any toxic words, it makes it impossible to post the comment, and it successfully posts the comment when it does not contain any toxic words. The main goal of the system is to increase the quality of online communication.

Dataset

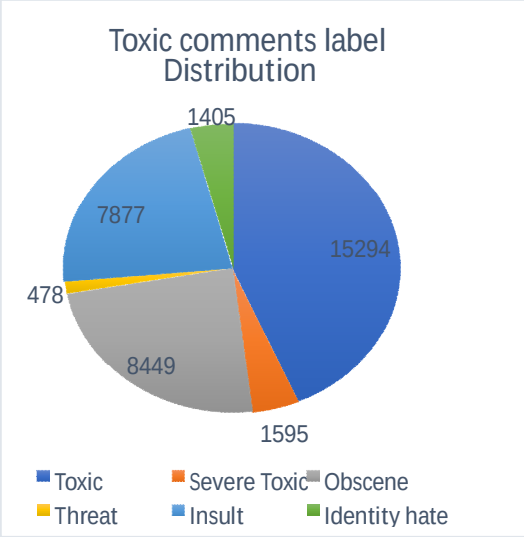
The system can be implemented by using the dataset named as Jigsaw toxic comment classification challenge.

It contains of 159,571 comments and categories included:

- Toxic
- Severe-Toxic

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

- Obscene
- Threat
- Insult
- Identity Hate



V. METHODOLOGY

1) Dataset Preparation

Before training the model, the dataset is divided into training and testing with the ratio of 80:20.

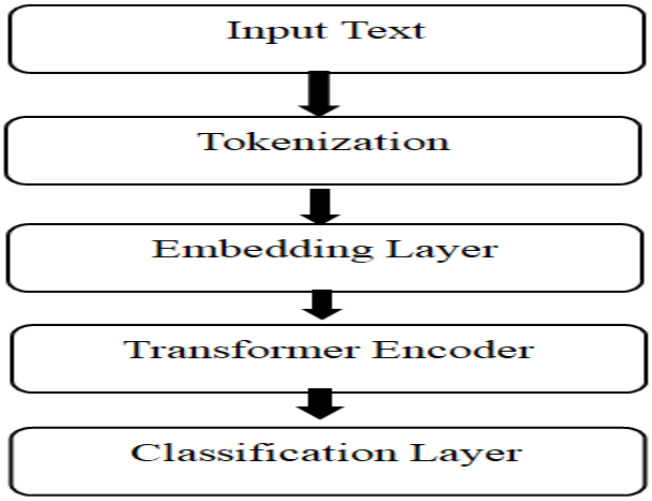
2) Data Preprocessing

Before using the models, the text must be preprocessed in the following ways:

- Cleaning:Remove unnecessary characters and noise.
- Tokenization:Break the words into smaller units known as tokens.
- Stop Word Removal: Removing the stop words like “the,” “is,” “at” in the sentence.
- Lowercasing: Converting all the text to lower case.
- Lemmatization / Stemming: Reduces the word to the root word.
- Normalization: Converting slang and abbreviations commonly used on social media into their standard representation.

3) Model Architecture

The XLM-R is a multilingual transformer model, that supports multiple languages. It gives better accuracy than the other models in detecting the toxic comments.



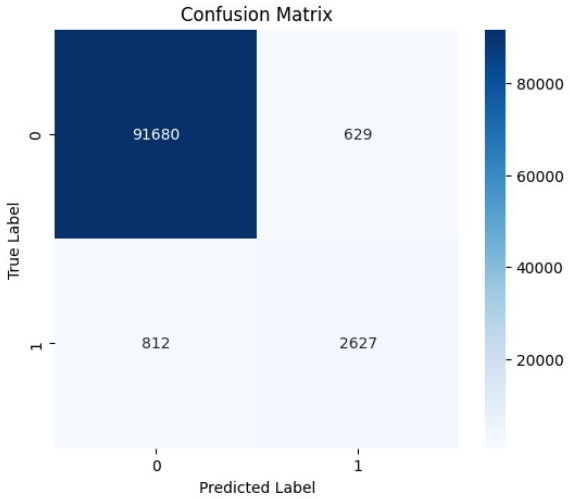
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) Model Training

The training process involves fine-tuning pre-trained XLM-R model using the labeled dataset. The model parameters are uses AdamW optimizer and cross-entropy loss. Training was performed on a GPU environment.

5) Model Evaluation

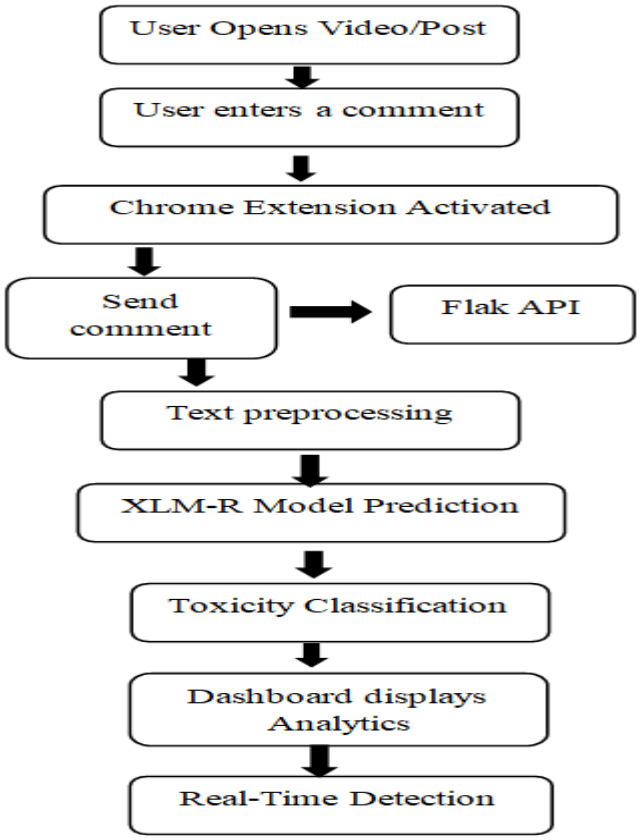
Confusion matrices is used to evaluate the performance of the model.



6) Model Deployment

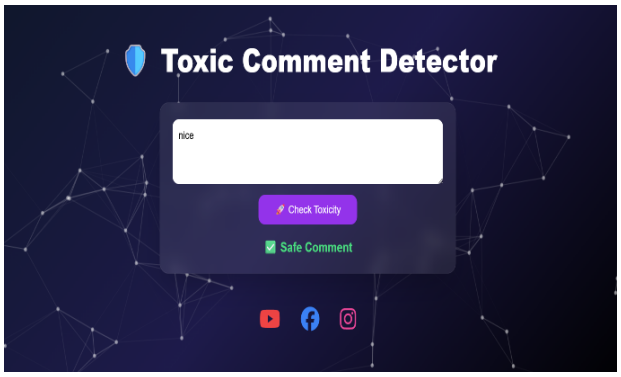
After training the model, it was integrated with the web application. The system supports the real-time detection of toxic comments in social media it supports platforms like You Tube, Instagram, Facebook.

VI. SYSTEM WORFLOW

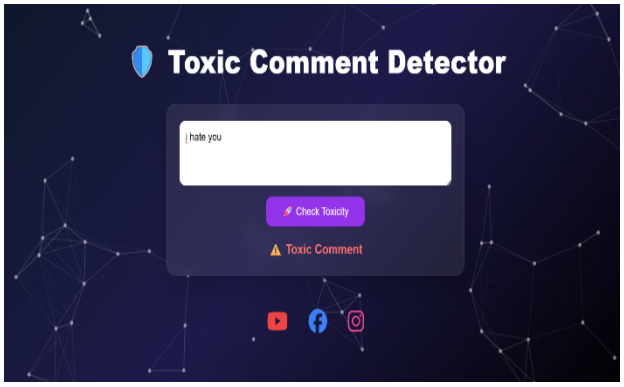


VII. RESULTS

The implemented system successfully detects the comments.



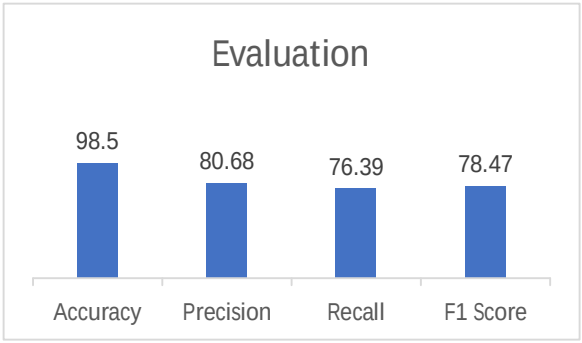
It shows as safe when it does not have any toxic word and allows to post comment in social media.



If the comment contains any toxic words, it shows the probability score of toxicity types and make not able to post comment in social media.

Accuracy, Precision, Recall and F1 Score are calculated based on confusion matrix results.

The results of the proposed model for the experiment are as follows:



VIII. CONCLUSION

The Toxic Comment Detection System aims to automatically detect the toxic comments in English, and classify the comments into nontoxic, toxic, threat, abusive, insult, and severe toxic by using the advanced Transformer model XLM-R. It prevents users from posting comments that contain toxic words and successfully posts the comment when it does not contain any toxic words in it, and also displays the toxicity probability.

Overall,the proposed system demonstrates how deep learning techniques can be effectively applied to online platforms for safer communication.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IX. FUTURE SCOPE

The current system effectively detects toxic commentsby using the XLM-R model. However, there are several ways to improve and extend the system in the future.

1) Support for more languages

There is a need to extend the system to support multiple languages for better online communication.

2) Domain Adaptation

The model can also be integrated with the different social media platforms for safer communication.

3) Real-Time Deployment

This model can also be fine-tuned for real-time detection in live chat systems.

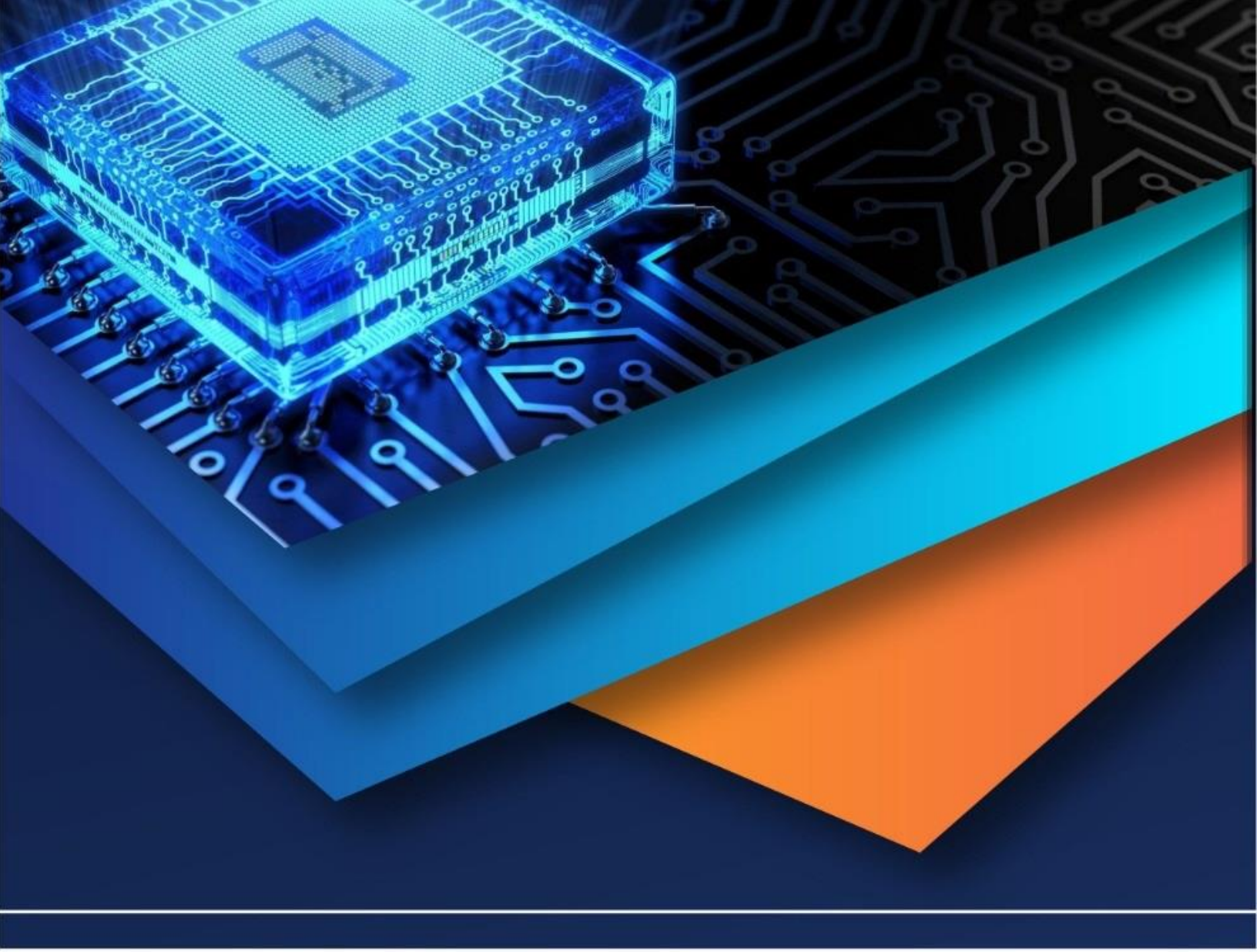
4) Integration with Modern Tools

The detection model can be integrated with tools like warning messages and blocking for better communication.

REFERENCES

[1] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532– 1543, 2014.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, pp. 1–17, 2018.
[4] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Unsupervised Cross-lingual Representation Learning at Scale," Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451, 2020.
[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace Transformers: State-of-the-art Natural Language Processing," arXiv preprint arXiv:1910.03771, pp. 1–15, 2020.
[6] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," arXiv preprint arXiv:1801.06146, pp. 1–12, 2018.
[7] T. K. Ho, "Random Decision Forests," Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR), pp. 278–282, 1995.
[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, pp. 1–12, 2013.
[9] T. G. Dietterich, "Ensemble Methods in Machine Learning," Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, pp. 1–15, 2000.
[10] Google Jigsaw, "Toxic Comment Classification Challenge," Kaggle Dataset, 2018. Available: https://www.kaggle.com/c/jigsaw-toxiccomment-classification-challenge
[11] P. Ruder, "Neural Transfer Learning for Natural Language Processing," PhD Thesis, National University of Ireland, Galway, 2019.
[12] A. Mozafari, F. Farahani, M. Farahani, M. Farahani, "A Survey on Multilingual Text Classification: Methods, Datasets, and Challenges," ACM Computing Surveys, vol. 54, no. 7, 2021.
[13] S. Ruder, P. Ghaffari, J. Breslin, "Character-level and Multi-lingual Approaches for Toxic Comment Detection," arXiv preprint arXiv:2005.10242, 2020.
[14] X. Yang, Y. Wu, C. Li, "Cross-lingual Text Classification with Pre-trained Language Models," Proceedings of the 2021 Conference on Empirical Methods in NLP, 2021.
[15] [V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," NeurIPS Workshop on Energy Efficient Machine Learning, 2019.
[16] D. Pires, E. Schlinger, D. Garrette, "How multilingual is Multilingual BERT?" ACL 2019, pp. 4996–5001.
[17] S. Garg, S. Jain, A. Kumar, "Detecting Toxic Comments in Indian Languages Using XLM-R," Proceedings of the 2022 International Conference on NLP Advances, 2022.
[18] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," ICLR 2018.
[19] A. Sun, C. Zhang, "A Comprehensive Survey of Text Classification Algorithms," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020.
[20] A. AI4Bharat, "IndicNLP Dataset for Multilingual NLP in Indian Languages," Available at: https://ai4bharat.org/indicnlp, 2021.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)