iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40567

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 10 Issue III Mar 2022- Available at www.ijraset.com

Performance of Text Classification Methods in Detection of Hate Speech in Media

Srinedhi Thanvanthri[1], Shivani Ramakrishnan[2]
[1]*Student, Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai, India*
[2]*Student, Department of Electronics and Communication Engineering, College of Engineering, Anna University, Guindy, Chennai, India*

Abstract: ***With the increased popularity of social media sites like Twitter and Instagram over the years, it has become easier for users of the sites to remain anonymous while taking part in hate speech against various peoples and communities. As a result, in an effort to curb such hate speech online, detection of the same has gained a lot more attention of late. Since curbing the growing amount of hate speech online by manual methods is not feasible, detection and control via Natural Language Processing and Deep Learning methods has gained popularity. In this paper, we evaluate the performance of a sequential model with the Universal Sentence Encoder against the RoBERTa method on different datasets for hate speech detection. The result of this study has shown a greater performance overall from using a Sequential model with*** **a multilingual USE layer.**
Keywords: Hate Speech Detection, RoBERTa, Universal Sentence Encoder, Sequential model.

I. INTRODUCTION

This decade has seen a rapid rise in the use of social media and networking sites. Sites like Twitter, Facebook, and Instagram have become some of the most frequently used sites for the general public to share their thoughts and opinions on various matters. However, these sites have also seen a rapid increase in the amount of hate speech and offensive language being published and shared by the users.

With today's climate, the detection of hate speech and offensive language online has become a necessity and the companies owning these sites have had to take on the responsibility of censoring and keeping in check the hateful content. This need for censorship has led to the automation of detection of hate in text as manually checking the content shared by users has become near impossible due to the sheer amount of content that exists on these platforms. The need for automation of hate detection has in turn led to the research and development of several machine learning and natural language processing methods for the same. More recently, deep learning and neural networks have gained attention over classical machine learning algorithms in such tasks.

In this study, we conducted a set of experiments mainly focussing on the use of Neural Networks with techniques like RoBERTa and Universal Sentence Encoder in analyzing and accurately detecting whether a certain text can be considered hate speech or not. We approached the task as a binary classification problem and compared the performance of the two models on the given data based on evaluation metrics used.

II. LITERATURE SURVEY

Recently, researchers have shown an increased interest in automatic hate speech detection in social media. Most of the current studies in the literature have primarily modeled the problem as a supervised classification task, whether using classical machine learning approaches or deep neural network approaches, we conducted a set of experiments mainly focussing on the use of Neural Networks with techniques like RoBERTa and Universal Sentence Encoder . Kwok and Y. Wang [16] proposed a supervised machine learning model to detect racist hate speech against black people in the Twittersphere and suggested improving the model by considering other features such as sentiment features and bigrams. Daniel Cera et al [13] present TensorFlow Hub sentence embedding models having good task transfer performance.

They show how model complexity, resources, and transfer performance are related. Comparisons are done between baselines with and without transfer learning, as well as baselines with word-level transfer.

Aljero et al [5] proposed a genetic programming (GP) model for identifying hate speech in which each chromosome represents a classifier with a universal sentence encoder as a feature. For the four publicly accessible hate speech datasets, the suggested GP model showed better results compared to other algorithms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

III. DATASETS

In this section, we describe the datasets used in our study of Hate Speech Detection methods. We used three datasets namely, the ETHOS binary dataset, the HatEval dataset, and a dataset consisting of text and labels sourced from sites like Kaggle. For final comparison of performance, we combined a few datasets for the model to be able to detect hate speech in general and not specific cases of the same.

IV. EXPERIMENTS

In this section, we elaborate on the experiments conducted in our study of Hate Speech Detection. We tested the performance of two model architectures on different hate speech datasets. These two architectures include - A Universal Sentence Encoder based Neural Network approach and a RoBERTa based approach.

A. Preprocessing

We applied several preprocessing techniques to the data before feeding the same input to our models. Since the data consists of tweets and text from social media sites, the text contains emojis, emoticons, hashtags and other symbols, as well as numbers. In preprocessing, we remove all these unnecessary symbols and characters, remove stopwords, and clean the text further by removing punctuation.

We must arrange the model input sequence in a certain way to meet pre-training. To accomplish so, you must first tokenize and then numericalize the sentences appropriately in the Roberta-based technique. The problem is that each pre-trained model that we will fine-tune requires the exact same pre-process - tokenization & numericalization - than the pre-process used during the pre-train part.

B. Universal Sentence Encoder Based Approach

The Universal Sentence Encoder encodes text into high dimensional vectors. It condenses a sentence as a 512-dimensional sentence embedding, uses this same embedding to complete many tasks, and modifies the sentence embedding based on the errors it makes. Because the same embedding must perform several generic tasks, it will only capture the most useful information while ignoring noise.

This then produces a general embedding that can then be used for tasks such as text classification. The Universal Sentence Encoder comes with two different architectures, namely, the Transformer architecture and the Deep Averaging Network architecture. In this study, we make use of the Multilingual Universal Sentence Encoder variant with the Transformer architecture.

In this study, we develop and build a sequential model with the Multilingual Universal Sentence Encoder being added as a layer. The sequential model is built with several layers with an input layer, the USE layer is then added, a dense layer, a dropout layer with a rate of 0.3, and finally the output layer and uses the ‘adam’ optimizer. Since the task at hand is essentially a binary classification task, binary cross-entropy was used as our loss function with accuracy, precision, recall etc as our evaluation metrics. The model was tested with different parameters and the most optimal configuration was used.

C. *RoBERTa Based Approach*

RoBERTa: A Robustly Optimized BERT Pretraining Approach by Yinhan Liu and et al, [15] first introduced the RoBERTa model and is an improvement on BERT, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rate. Each model architecture is related with three different sorts of classes:

A model class for loading and storing a specific pre-train model.

A tokenizer class is used to pre-process data in order to make it compliant with a specific model.

A configuration class that allows you to load and save your settings.

RoBERTa shares the same architecture as BERT, but it employs a byte-level BPE as a tokenizer (as does GPT-2) and a different pre-training technique.

Self-training methods with transformer models have achieved state-of-the-art performance on most NLP tasks.RoBERTa shares the same architecture as BERT, but it employs a byte-level BPE as a tokenizer (as does GPT-2) and a different pre-training approach. RoBERTa is trained on BookCorpus, roberta-base” was trained on 1024 V100 GPUs for 500K steps, and transfer learning is incorporated. Cross entropy loss was used as the loss function and Adam optimiser was used as the optimiser. For the dataset under study different hyper parameters were modified and finetuning of the model was done using based on the validation loss.

V. EVALUATION OF EXPERIMENTAL RESULTS

A. Universal Sentence Encoder Based Model

From our experiments, the Universal Sentence Encoder based model works well even with a smaller amount of data. The performance of this model has been measured with the help of some evaluation metrics including AUC, FPR, Precision, Recall, etc. The performance of the model is as shown.

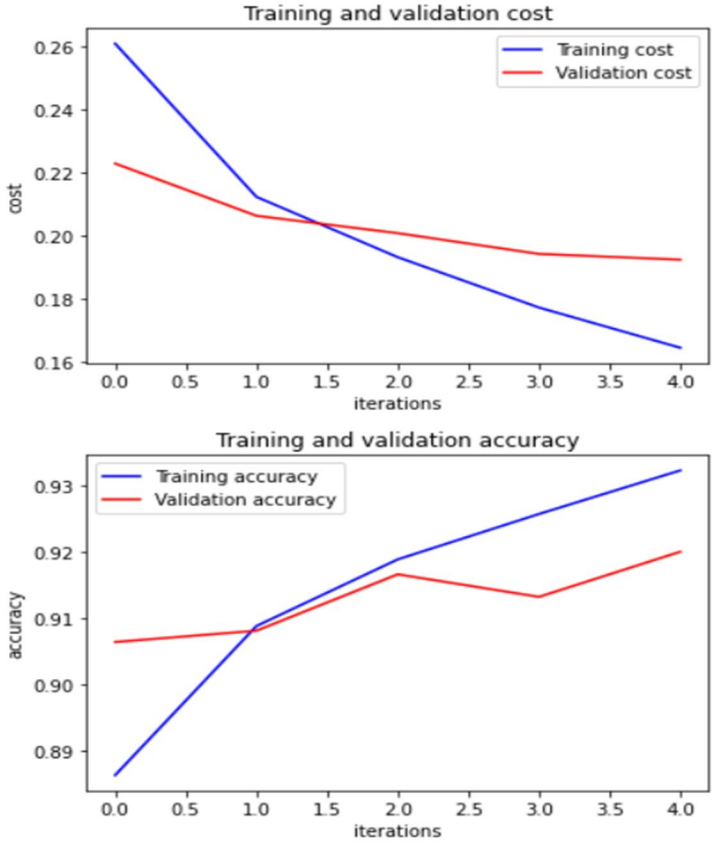


Fig. 1 Measure of Accuracy and Cost for USE based model

```
AUC:  0.9544843133466375
      threshold       fpr     1-fpr       tpr      diff
1077    0.19003  0.114275  0.885725  0.885983  0.000258
Accuracy score: 0.8856847791547505
```

Fig. 2 Evaluation Measure Results for USE based model

```
Classification Report:
              precision    recall  f1-score   support

    negative       0.98      0.89      0.93     10676
    positive       0.58      0.89      0.70      1912

    accuracy                           0.89     12588
   macro avg       0.78      0.89      0.82     12588
weighted avg       0.92      0.89      0.89     12588
```

Fig. 3 Classification Report for the USE based model

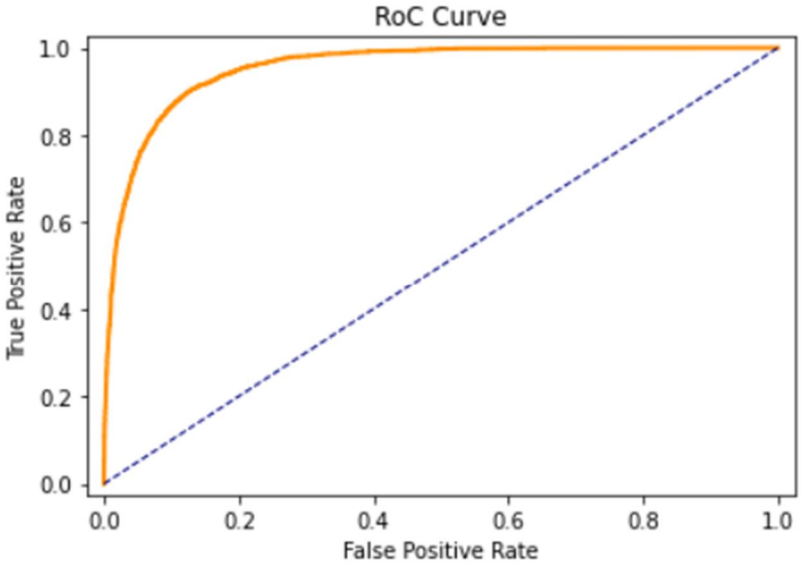


Fig. 4 ROC Curve for USE based model

B. Roberta Based Model

For the datasets under study, Roberta based model has been evaluated with metrics including Accuracy, Precision, Recall.It exhibited good validation accuracy of 0.87 and validation loss of 0.43,however ,as it is a classification task with imbalance in the dataset, the model's precision and recall was also studied which showed average results.

```
Validation Loss : 0.43666316450169657
Validation Accuracy : 87.29096989966555
```

Fig. 5 Evaluation results for Roberta model

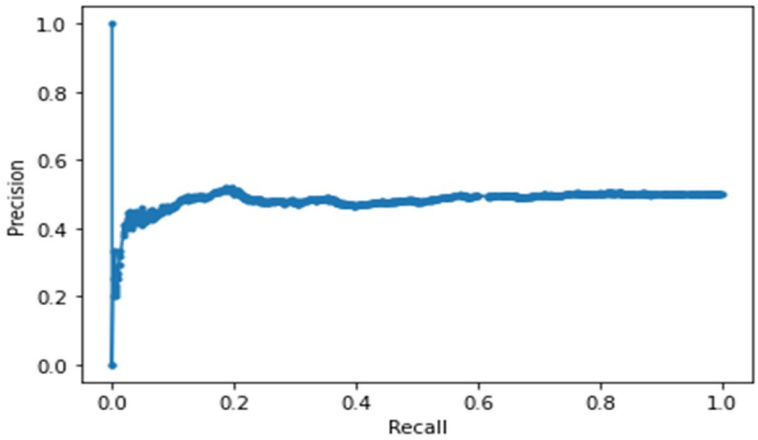


Fig 6 Precision recall curve for Roberta based model

VI. RESULTS

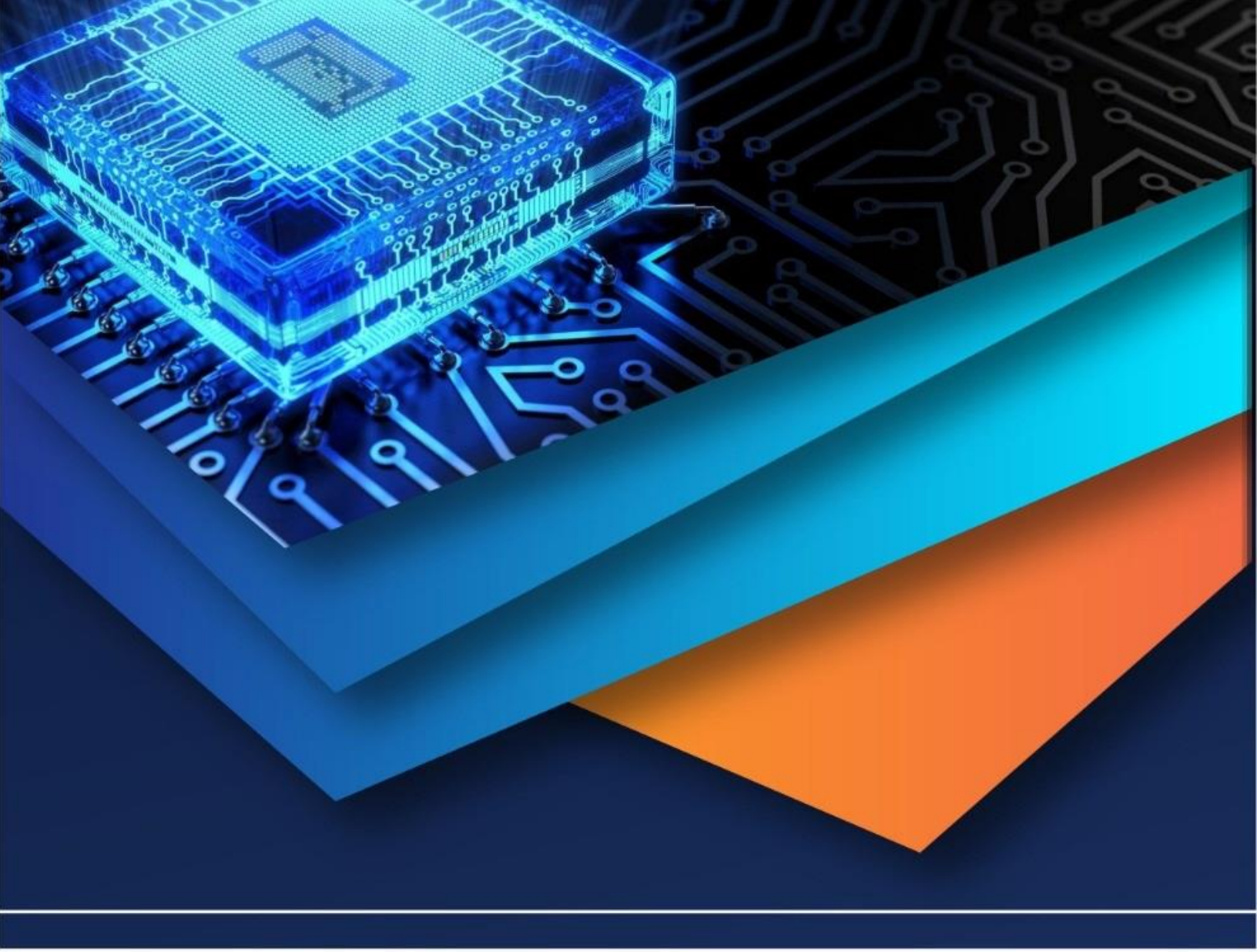
The section discusses the experimental results of the proposed model on four different datasets and provides a comparison between our results and other methods on the same datasets.The experiments were implemented using Python 3.7. Using the Roberta approach on the data sets under study, a validation accuracy of 87.51% was achieved after fine tuning the model.The Universal sentence encoder showed a better performance with 91.96 % validation accuracy and predicted instances of hate speech more accurately compared to the Roberta approach for the same conditions and dataset.

VII. CONCLUSIONS

In this study, we analyse two common and popular text classification methods and compare their performance in Hate Speech Detection. We conducted several experiments with our models and tested them on multiple datasets. From these experiments, we found that the model using Universal sentence encoder performs better than the Roberta model and also provides more reliable and accurate predictions overall. These models can also be further extended to perform multiclass classification with a more detailed set of classes like offensive or aggressive language and even against a certain community. Overall, it is clear that natural language processing methods like those used in this study are crucial to the detection and control of hate against persons and communities online.

REFERENCES

[1] Alshalan, Raghad & Al-Khalifa, Hend. (2020). A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. Applied Sciences. 10. 8614. 10.3390/app10238614.
[2] Ioannis Mollas, , Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. "ETHOS: an Online Hate Speech Detection Dataset." (2020).
[3] Schmidt, Anna & Wiegand, Michael. (2017). A Survey on Hate Speech Detection using Natural Language Processing. 1-10. 10.18653/v1/W17-1101.
[4] B. Pariyani, K. Shah, M. Shah, T. Vyas and S. Degadwala, "Hate Speech Detection in Twitter using Natural Language Processing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1146-1152, doi: 10.1109/ICICV50876.2021.9388496.
[5] Aljero, Mona & Dimililer, Nazife. (2021). Genetic Programming Approach to Detect Hate Speech in Social Media (July 2021). IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3104535.
[6] Zampieri, Nicolas & Illina, I. & Fohr, Dominique. (2021). Improving Automatic Hate Speech Detection with Multiword Expression Features.
[7] Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion (2017)
[8] Indurthi, Vijayasaradhi & Syed, Bakhtiyar & Shrivastava, Manish & Chakravartula, Nikhil & Gupta, Manish & Varma, Vasudeva. (2019). FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. 70-74. 10.18653/v1/S19-2009.
[9] Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., GuajardoCespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. ACL, Brussels, Belgium (Nov 2018).
[10] Basile, Manuela. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter." . In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 54–63). Association for Computational Linguistics, 2019.
[11] Gambäck, B.; Sikdar, U.K. Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 85–90.
[12] Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell. 2018, 48, 4730–4742.
[13] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder. arXiv:1803.11175, 2018.
[14] Kwok, I.; Wang, Y. Locate the hate: Detecting tweets against blacks. In Proceedings of the Twenty-seventh AAAI Conference on Artificial Intelligence, Washington, DC, USA, 14–18 July 2013.
[15] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (cite arxiv:1907.11692)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY