iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33808

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: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Sentiment Analysis using Customer Reviews

Abhi Goyal[1], Ananya Garg[2], Dhara Dhakad[3], Riya Airen[4], Shefali Parmar[5]
[1, 2, 3, 4]*Department of Computer Engineering, MPSTME NMIMS*
[5]*Department of Computer Science Engineering, SVIIT SVVV*

Abstract: ***The customer review analysis is an important part of the decision-making of any organization. This data has been generated daily in huge amounts with the increasing digitalization. Not just organizations but also the customers need this analysis for their decision-making as well. For these purposes, the sentiment of the review needs to be calculated. This paper proposes the use of existing methods and adding some features that could increase the accuracy of the analysis. This feature extraction has been done using Parts of Speech (POS) Tagging. Also, several pre-existing models are compared on the same data available, and Naive Bayes has been proved as the best classification algorithm for sentiment analysis.***
Keywords: Sentiment Analysis, Parts of Speech Tagging, Naïve Bayes, Confusion Matrix, Stemming

I. INTRODUCTION

Sentiment Analysis is an application of Natural Language Processing (NLP). It is a text analysis method that detects the sentiment of the text (i.e., positive, negative, and neutral). In recent times, users tend to express their opinion more freely, understanding this user's sentiment is essential for any business. Lots of reviews are generated every day with the increasing trend of online shopping. This data helps businesses to expand themselves based on insights generated from the analysis. While purchasing online, users rely on other users' reviews, therefore the opinion matters. Enterprises when making crucial decisions for making a profit by marketing their products, predicting sales trends, analyses the huge amount of customer responses available as reviews. With the help of sentiment analysis, it is possible to analyse the huge amount of data available and extract the user's opinion, which may help both the organization as well as customers to achieve their motive.

A. Sentiment Analysis Can be Done on Three Levels

1) Document
2) Sentence
3) Aspect

At a document level, the entire document is analysed at one time. It gives the primary opinion of the context. It assumes that reviews belong to a single person. In Sentence Level, the document is broken down into sentences. The main concern is to find the target of the sentence that is used to define the sentiment. At the last aspect level, data is being categorized by aspect and identifies sentiment attributes to each one. It identifies fine-grained polarity [10].

II. LITERATURE SURVEY

An There are several approaches to perform sentiment analysis, these approaches can further be divided into two main sub-categories, one based on lexicons and the other is the machine learning approach. In machine learning, we have two ways of supervised and unsupervised learning. Many previous works are focused on these two sub-categories. There are two ways in machine learning: one is a probabilistic method and the other is Linear classification. Much research is being made in the Naive Bayes Method which is a probabilistic approach. Naive Bayes It calculates the probability of each aspect. For probability calculation, the word must be converted into text and that is done by taking the frequency count of the word in the whole sentence. Naive Bayes is being described as a highly scalable and simple classifier technique that requires a smaller number of parameters [1].
"MALLET" is another package that is used for NLP.[2] This compares two devices based on various parameters and puts an application of opinion mining for the customer's benefits while choosing products. [3] uses the sentiment polarity categorization along with the POS Tagging. POS Tagging stands for parts of speech tagging where each word has been assigned its parts of speech. Some semi-supervised learning methods use WordNet as a lexicon-based data dictionary to convert features into target words. After this conversion, a probabilistic method can be used to get the sentiment scores [4]. There are algorithms such as Naive Bayes [1][5][7][9], Max Entropy, Boosted Tree, Decision Tree [6], SVM [5], Random Forest, KNN, Logistic Regression [7] which are used for sentiment analysis. In all algorithms, Naive Bayes is better at simplicity level, Memory Requirement and Time Complexity Factor with a good Accuracy and Performance.[4].

Dependency Parsing[8] has been a new concept introduced in research. It's a kind of syntax parsing process that can generate the dependency grammar for a sentence. In some parts, it shows more f1 scores as compared to sentiment analysis.

III.METHODOLOGY

A. Data Collection

The data is taken from the updated version of the Amazon review Dataset released in 2014 [11]. The dataset belongs to the cell phone and Accessories category and has about 1,128,437 reviews. Each review includes the following information: 1) reviewer Id; 2) asin; 3) reviewer Time; 4) vote; 5) style; 6) review text; 7) overall; 8) summary; 9) review time .The initial dataset was available in JSON format and it has been converted into CSV format for future pre-processing.

B. Data Pre-Processing

Title Pre-processing is a data mining technique that is used to transform and convert the raw data into data that is useful and efficient for model selection. Several methods can be used for pre-processing. In this process, we eliminate the irrelevant data in the data set like stop words and punctuations. Here lower Casing, removing special and unwanted characters, Tokenization, Stop Word Removal, and Stemming has been applied. Lower Casing converts every letter of the sentence from uppercase to lowercase. Removing special and unwanted characters is used to remove punctuations or any special characters like emojis. Tokenization refers to splitting a sentence into words, this is useful for the vectorization process that will be used later. Stop word refers to the words that do not contribute to the meaning of the sentence, for example, words like the, he, have, they, etc. these words have no sentiment therefore it needs to be removed. NLKT Corpus has a set of stopwords that are considered to have no contribution to the meaning of the sentence. Stemming is a technique for minimizing word inflection to its root forms, such as mapping a group of words to the same stem, even though the stem isn't a true word in the Language

C. Model Selection

No Model selection is the process of selecting one machine learning model from the collection of other models tested considered for training and testing the dataset. The final model is selected based on certain evaluation parameters such as accuracy, f1 score, precision, recall, and time required to execute the process. The algorithms that are compared are Naive Bayes, SVM, KNN, Logistic Regression, Decision Tree, and Random Forest.

For predictive modelling, Naive Bayes [1][5] is a simple but surprisingly effective algorithm. It's a classification method based on Bayes' Theorem and the presumption of predictor independence. A Naive Bayes classifier, in simple terms, assumes that the existence of one function in a class is unrelated to the presence of any other feature. It's a classification approach based on Bayes' Theorem and the predictor independence assumption. In simple terms, a Naive Bayes classifier assumes that the presence of one function in a class has no bearing on the presence of any other feature.

The Support Vector Machine (SVM) is one of those supervised machine learning algorithms that can be used to solve both solving classification and regression problems. It is, however, mostly used to solve classification problems. In this, each data observation is plotted as a point in n-dimensional space, with the value of each feature as the value of a particular coordinate. After that, classification is done by finding the n. Following that, we classify the data by finding the hyper-plane that separates the two classes.

The KNN algorithm is a fundamental supervised machine learning algorithm for solving classification and regression problems. It's simple to set up and understand, but it has the downside of being noticeably slower as the amount of data in use increases works by calculating the distances between a query and all the examples in the data, selecting the K closest examples to the query, and for classification, it then votes for the most frequent label and in case of regression it averages the label.

The classification algorithm logistic regression is used to assign observations to a distinct set of groups. Email spam or not spam, online transactions fraud or not fraud, and tumour malignant or benign are some examples of classification issues. The logistic sigmoid function converts the contribution of logistic regression into a probability value. The relationship between one dichotomous dependent variable and one (categorical or continuous) independent variable is investigated using logistic regression analysis. This differs from linear regression analysis, which uses a continuous variable as the dependent variable.

The Decision Tree algorithm [6] is part of the supervised learning algorithms family. Unlike other supervised learning algorithms, the decision tree algorithm can be used to solve regression and classification problems. By learning basic decision rules inferred from prior data, a Decision Tree can be used to construct a training model that can be used to predict the class or value of the target variable. We start from the root of the tree while using Decision Trees to predict a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and jumps to the next node based on the comparison.

Random forest is a versatile, easy-to-use machine learning algorithm that, in most cases, produces excellent results even without hyper-parameter tuning. Because of its simplicity and versatility, it is also one of the most widely used algorithms. It creates a "forest" out of an ensemble of decision trees, which are normally trained using the "bagging" process. The bagging method's basic premise is that combining different learning models improves the overall outcome. Instead of looking for more relevant features, the random forest adds more randomness to the tree as it grows.

D. POS Tagging

Parts of Speech Tagging refers to assigning part of speech to individual words in a sentence. It marks up the words in text format for a particular part of a speech based on its definition and context. It oversees reading text in a language and assigning a token (Parts of Speech) to each word. Grammatical marking is another name for it. Nouns, pronouns, adjectives, adverbs, verbs, and other words fall under this category. There are pre-defined abbreviations for each part of speech. Machine Learning has predefined methods for POS Tagging that are useful for this process. Based on Parts of Speech certain features are being chosen for building the model. These 5 different permutations and combinations of adjectives, verbs, and adverbs are being compared for best features. These parts of speech are taken into consideration and passed for further model building.

E. Training & Testing

From the 11 lakhs of data available 5 lakh of data is being used for the training and testing purpose. The dataset is divided into 70-30 ratio for testing and training purposes. Python library Sklearn has a method that is used for this purpose. Initially, the 70% dataset is used for training the model and after this, the rest of the 30% data is used for the testing of the model.

F. Result

In this step, the reviews are being classified as positive, negative, and neutral. The accuracy is calculated using the confusion matrix, which is used to calculate the accuracy, precision, recall, F1 score, and Mean Square error is being calculated for the same. Confusion matrix is the method used to summarize the classification algorithm. Many times, simple accuracy comparison can be misleading when we have an unequal number of observations in each class or have more than two classes therefore it is essential to check the other parameters. The confusion matrix gives the total count of correct and incorrect predictions broken down by each class. In the Confusion matrix, we have 4 main terminologies, which are: “true positive” for correctly predicted event values; “true negative” for correctly predicted non-event values; “false positive” for incorrectly predicted event values; “false negative” for incorrectly predicted non-event values.

IV. CONCLUSIONS

Our experimental results show that Parts of Speech (POS) Tagging is effective while performing sentiment analysis, there are several POS tags for it, but adjectives, adverbs, and verbs contribute more towards the sentiment of any sentence as shown in Fig. 1. We carried out different combinations of these three POS and among them, adjectives, and verbs, provide the upper hand in terms of the number of words required for training the model hence it has less time complexity. Adjectives and Verbs just require 107508 words to train the model whereas without POS tagging it takes 462517 words to train upon as shown in Fig. 2.

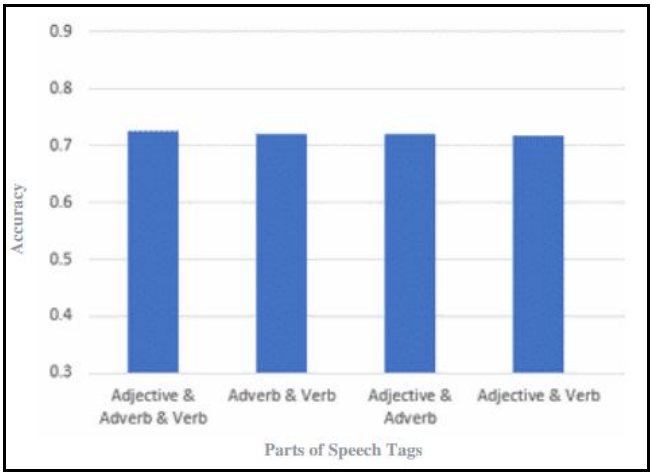


Fig. 1. Accuracy of all Parts of Speech Tagging Combination

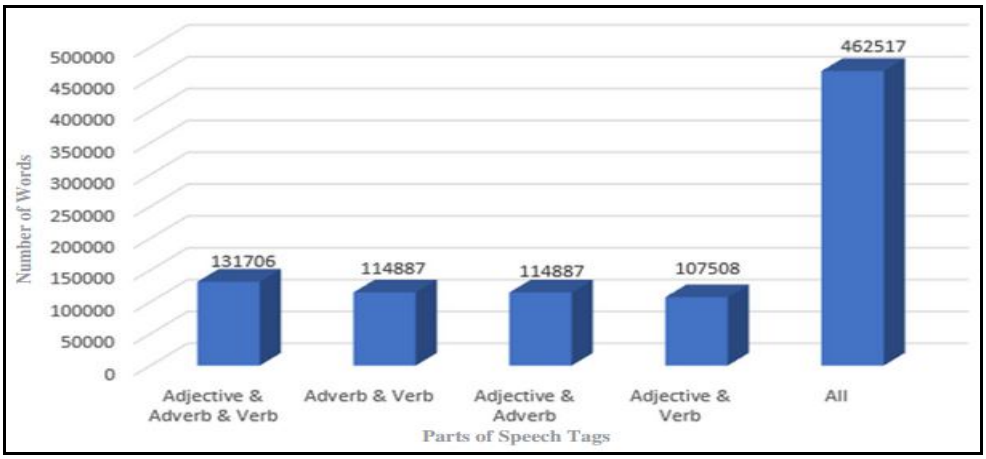


Fig. 2. Comparison of Words Based on Parts of Speech Tagging Combination

Our model comparison of 6 different algorithms shows that Naive Bayes is a better fit on our data as shown in Table I. Naive Bayes has certain advantages over other algorithms, it uses probability, easy to implement, and simple. Also, it compares both continuous and discrete variables. The main advantage of Naive Bayes is that it is fast and can be used in real-time predictions as well. It is based on the conditional independence assumption; this is an important factor of using this algorithm for real-time predictions. After the final implementation of the outcomes, we were able to make an accurate prediction with an accuracy of about 81.33%.

TABLE I
Performance Matrix

	Accuracy	Precision	Recall	F1-score	Root Mean Square Error	Time Required (in min)
KNN	71.9	75.2	95.6	0.83	0.782	0.062
Logistic Regression	68.2	83.2	86.3	0.84	0.718	0.167
Naive Bayes	**76.7**	79.6	**94.3**	**0.86**	**0.584**	**0.002**
SVM	73.8	**84.6**	86.1	0.84	0.589	0.088
Decision Tree	68.5	80.8	82.1	0.80	0.773	0.125
Random Forest	72.3	72.3	84.5	0.77	0.783	0.185

V. RESULTS

According to our experiment, the Naive Bayes classification proved to be the best fit among all the 6 algorithms used. It provides you certain benefits in terms of accuracy, simplicity, and time complexity. Further, we experimented with some combinations of POS tags, which were adjectives, verbs, and adverbs, among those combinations of adjectives and verbs, proved to have better speed with a smaller number of words to train on and providing us almost the same results. There may be a slight difference in the accuracy level when using this Parts of Speech (POS) Combination, but the difference is almost negligible. Furthermore, we can find other POS tags that affect the accuracy and make the model accurate with more level of accuracy. This experiment can be the extent to analyse the emojis as well because emoji holds a lot more emotions than a sentence, and in this social world use of emojis has increased tremendously.

REFERENCES

[1] P. P. Surya and B. Subbulakshmi, "Sentimental Analysis using Naive Bayes Classifier," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2019, pp. 1-5. IEEE Issue: 14 November 2019

[2] P. V. Rajeev and V. S. Rekha, "Recommending products to customers using opinion mining of online product reviews and features," 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], Nagercoil, 2015, pp. 1-5, doi: 10.1109/ICCPCT.2015.7159433

[3] Pankaj, P. Pandey, Muskan and N. Soni, "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 320-322, doi: 10.1109/COMITCon.2019.8862258

[4] Mali, Digvijay & Abhyankar, M & Bhavarthi, Paras & Gaidhar, K & Bangare, Manoj. (2016). SENTIMENT ANALYSIS OF PRODUCT REVIEWS FOR E-COMMERCE RECOMMENDATION.

[5] S. Pradha, M. N. Halgamuge and N. Tran Quoc Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 2019, pp. 1-8, doi: 10.1109/KSE.2019.8919368.

[6] W. Songpan, "The analysis and prediction of customer review rating using opinion mining," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, 2017, pp. 71-77, doi: 10.1109/SERA.2017.7965709.

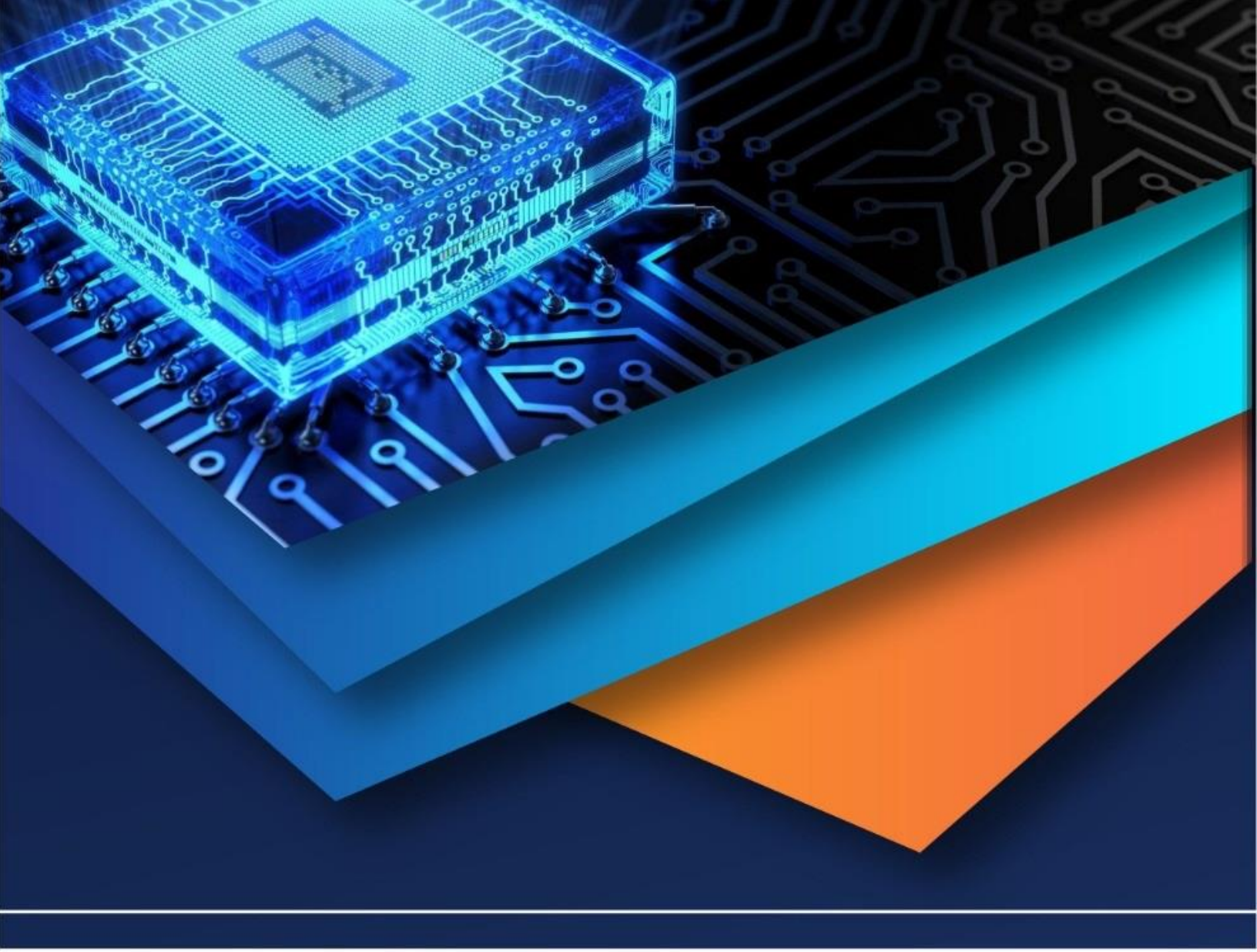
[7] K. L. S. Kumar, J. Desai and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-4, doi: 10.1109/ICCIC.2016.7919584.

[8] L. Li, "Identification of informative reviews enhanced by dependency parsing and sentiment analysis," 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), Wuhan, China, 2016, pp. 476-479, doi: 10.1109/CCI.2016.7778968.

[9] S. Vanaja and M. Belwal, "Aspect-Level Sentiment Analysis on E-Commerce Data," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, 2018, pp. 1275-1279, doi: 10.1109/ICIRCA.2018.8597286

[10] C. Rangu, S. Chatterjee and S. R. Valluru, "Text Mining Approach for Product Quality Enhancement: (Improving Product Quality through Machine Learning)," 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 2017, pp. 456-460, doi: 10.1109/IACC.2017.0100.

[11] k Jianmo Ni, Jiacheng Li, Julian McAuley, Empirical Methods in Natural Language Processing (EMNLP), 2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY