iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74148

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 13 Issue IX Sep 2025- Available at www.ijraset.com

Analyzing Sentiments in Amazon Product Reviews through NLP Techniques

Rekha[1], Shyamol Banerjee[2]
[1]*Research Scholar,* [2]*Asst. Prof, Dept. of Computer Science and Engg., Shriram College of Engineering & Management, Gwalior*

Abstract: ***This research presents a thorough and systematic approach to classifying the sentiment of Amazon product evaluations using a deep learning framework based on a bidirectional reinforcement learning network. Utilising a large-scale dataset that has been structured-preprocessed into sentiment labels, the study effectively incorporates numerical ratings and review language. Semantic embedding using previously trained GloVe vectors to a text normalisation, and meticulous data cleaning are some of the methods used in the study to guarantee that the model can grasp all the nuances and context-dependent information in language. Because it efficiently regulates the bidirectional flow of information, BiLSTM is necessary for the model to learn via past and future word correlations inside a review. To prevent overfitting and enhance generalisation, the design is fine-tuned with the use of nationwide pooling, batch normalization, including dropout layers. Stratified sampling is a must for data division in order to ensure adequate representation across sentiment classes, especially considering the inherent bias in players material. Performance evaluation using metrics like F1-score, recall, accuracy, and precision shows that the suggested model is effective. The accuracy indicator (0.9124) and the F1-score (0.9123) stand out among the others, both of which have very high values. The methodology has demonstrated its value by reliably labelling assessments as positive, neutral, or negative. Full exploratory data analysis also shows that class balance and preprocessing have a major effect on the model's performance. Finally, this study lays out a scalable and extensible deep learning-based method that works well in sentiment analysis for e-commerce and customer feedback systems, among other real-world uses.***

Keywords**: *Sentiment Analysis, Amazon Reviews, NLP Techniques, Deep Learning, BiLSTM.***

I. INTRODUCTION

Sentiment analysis, often known as opinion mining, is one of the several uses of natural language processing (NLP) that has lately grown in importance, especially for assessing customer reviews posted by people who have shopped online. Online reviews are more important than ever before when consumers are deciding what to buy. Businesses seek customer input in order to enhance their offerings. One of the biggest online retailers in the world, Amazon, has a plethora of product reviews written by actual customers. The disorganised structure, colloquial tone, and recurring ambiguity make reviews like this both easy and difficult to read. Marking reviews as "good," "negative," or "neutral" is a primary goal of sentiment analysis. With this data, automated systems may learn more about client happiness, spot problems, and change marketing tactics as necessary. Here, natural language processing, also known as NLP, plays a pivotal role in converting spoken words into structured formats that machine learning systems can comprehend. Using traditional classifiers such as Naïve Bayes or supporting vector machines for sentiment analysis was once the norm, following feature selection by hand using methods like bag-of-words or TF-IDF.[1]–[3]. The field of deep learning has made great strides recently. Models such as attention-based architectures, recurrent neural network systems (RNNs), and LSTM networks, which stand for Long Short-Term Memory, are becoming more complex while making it easier to understand context and categorise emotions.

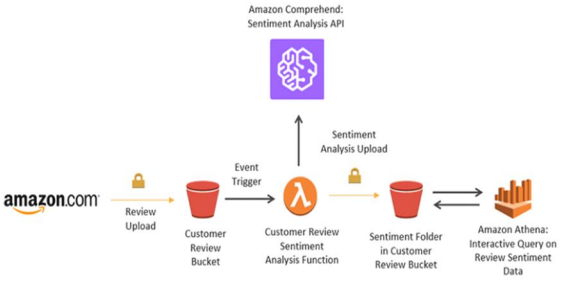


Fig. 1 Amazon Product Reviews [4]

We can learn more about people's emotions thanks to these models' ability to identify reviewers' sequential relationships and semantic linkages. More than just finding out what consumers think about Amazon merchandise is within the capabilities of NLP approaches. Emotion recognition, aspect-based sentiment analysis, and fine-grained sentiment analysis are all features that aid in understanding client sentiment. Case in point: a review may praise the smartphone's battery life yet be critical of its picture quality.[5]–[8]. With aspect-based sentiment analysis, the system can independently examine these emotions, providing businesses with more actionable insights. The availability of labelled datasets and open-source language-processing tools like NLTK, SpaCy, and Hugging Face Transformers has also made sentiment analysis models accessible to everybody. By equipping models with deep contextualised word embeddings, pre-trained language models like BERT (a bidirectional encoder representation from Transformers) have transformed natural language processing. These models now have a better grasp of sarcasm, negation, and polysemy. One example of how sentiment detection is in a constant state of flux is the shift from statistical and rule-based models to ones based on deep learning. Tokenisation, stop-word removal, stemming, lemmatisation, and vectorisation are still crucial pretreatment procedures for ensuring acceptable data and effective model performance when it comes to Amazon product reviews. The task becomes significantly more challenging when dealing with issues such as imbalanced datasets, fraudulent reviews, and code-mixed language. For this reason, it is common practice to employ hybrid approaches that combine rule-based techniques with AI or deep learning techniques in order to fortify and clarify the system. There are many practical applications of sentiment analysis in the corporate sector, demonstrating its significance outside the realm of academia. The marketing team can use sentiment data to make sure their promotions are in line with consumer wants, the customer service team can use it to prevent problems indicated in negative reviews, and the product development team can use it to prioritise which features to improve. Additionally, automated sentiment analysis systems can comprehend and process massive review volumes instantly. Because of this, online retailers like Amazon are able to maintain a dynamic and adaptable approach to consumer contact. Ethical considerations regarding the use of sentiment analysis algorithms should centre on data privacy, algorithmic bias, and openness.[9], [10]. Since models are trained using publicly available evaluations, it is crucial to protect user data and ensure that the analysis does not perpetuate preexisting biases or discriminating against specific user groups. The importance of easily understandable model predictions is growing in domains where AI-generated insights are utilised to make critical business decisions. By analysing Amazon product reviews using natural language processing techniques, this study aims to build a robust classification model that can properly determine customers' emotions. To categorise sentiments, we investigate the efficacy of various preprocessing techniques, feature extraction algorithms, and classification systems. While some of these techniques use more modern deep learning architectures, others rely on more traditional machine learning models. Through a comprehensive comparison, the study aims to demonstrate the advantages and disadvantages of each method. Both theoretical and practical applications in the business sector will benefit from this. Businesses can gain valuable insights into client sentiment and how to respond promptly by utilising this study's findings on emotion analysis and its contribution to the academic literature. It accomplishes this by utilising state-of-the-art approaches and models in natural language processing (NLP). Written evaluations of hotels, films, and social media posts are just a few examples of how you might put this study's findings to use. In today's data-rich and interconnected society, this demonstrates how versatile and useful NLP-driven sentiment analysis is.[11].

II. LITERATURE REVIEW

Wang 2024 et al. understand internet shopping reviews, you need to apply sentiment analysis. Looks into Amazon reviews using Word2Vec and Support Vector Machine (SVM). Word2Vec displays how words are related to each other in terms of meaning after preprocessing and feature extraction. SVM is used to train and improve sentiment categorization. The combined method captures complicated text patterns, which makes it more accurate and faster. Experimental results suggest that this method works better than standard ones, allowing for better sentiment recognition. The results give us a solid base for studying how people feel and what they think, which helps e-commerce sites make smart choices and encourages more research into advanced methods of sentiment analysis[12].

Shaik 2024 et al. build a good product, you need to know how customers feel. This project builds a prediction pipeline that uses pre-trained BERT and T5 models to find aspects and analyses sentiment in review data. The models use certain features to sort reviews into three groups: favorable, negative, or neutral. They were trained on both synthetic and labelled datasets. Both models were improved, with BERT getting 92% accuracy and beating T5 in precision, recall, F1-score, and efficiency when it came to eco-friendly products. BERT is chosen for the final pipeline because it is very good at looking at user reviews. These insights back up product creation that is based on data and meets the needs and wants of customers[13].

Latha 2024 et al. Customers are more satisfied when online retailers use sentiment analysis on customer evaluations. When building an automated recommendation system, machine learning and deep learning models are utilized. We remove stop words, stem, lemmatize, and tokenise the text data from 60,000 Amazon reviews that are favorable, negative, or neutral. This makes it easier to classify the data and cuts down on training time. The TF-IDF vectorization technique converts legible text into fast-processing numerical vectors. We put these vectors into a modified convolutional neural network model to figure out how people feel. The suggested model gets an average accuracy of 97.40% on the Amazon Product Reviews dataset, which means it can do good analysis to provide better product recommendations[14].

Chenglerayen 2024 et al. examines the efficacy of BERT in predicting customer sentiment towards Amazon products using review data. Our objective is to compare BERT to popular models in order to determine its superiority in collecting user opinions. These models include Logistic Regression, TF-IDF, Random Forest, BERT, Naive Bayes, and SVM. Using its context-aware capabilities, BERT tackles problems with traditional methods. In order to measure a model's performance, we use the F1-score, recall, precision and precision. The anticipated outcome is that organisations will have more data at their fingertips for sentiment analysis, allowing them to refine their goods in response to consumer feedback. The findings also hope to persuade other sectors to use sentiment analysis in additional contexts.[15].

Boril 2024 et al. uses web scraping, text cleaning with a custom lexicon, and Bag-of-Words (BoW) feature extraction to develop a framework for sentiment analysis of online product evaluations. Cosine similarity and neural network models are used to classify sentiment. The method comprises evaluating performance and analyzing language. Using polarity salience and exponential-based weighting on unigram and bigram BoW vectors in a case study of Amazon reviews leads to fewer errors and more accurate results. Using salience to reduce dimensionality makes models work better. Word clouds show the salience matrix, which is also used to rank important terms and bigrams. Anyone can use the dataset and framework again[16].

TABLE: 1 LITERATURE SUMMARY

Authors/year	methodology	Research gap	Findings
Kumar/2024 [17]	Convolutional Neural Networks (CNN).	Lack of comparative analysis across traditional, deep, and transformer models.	BERT outperformed all models in accuracy, precision, recall, and AUC.
Hashmi/2024 [18]	SVM	Limited integration of contextual product information in sentiment analysis models.	FastText with boosting models achieved highest accuracy across all datasets.
Arwa/2021 [19]	Logistic Regression, Random Forest	Lack of comparative evaluation across vectorization techniques and classifiers.	BERT outperformed all models in multiclass and binary classifications.
Alharbi/2021 [20]	GLRNN with FastText achieved highest accuracy.	Limited studies compare recurrent models with multiple embedding techniques comprehensively.	GLRNN with FastText outperformed others on unbalanced dataset accuracy.
Rao/2021 [21]	Sarcasm-aware sentiment analysis using classifiers.	Lack of sarcasm detection in traditional sentiment classification approaches.	SVM achieved highest accuracy in sarcasm-aware sentiment classification tasks.

III. RESEARCH METHODOLOGY

This research lays forth a systematic approach to building a deep learning sentiment classifier for use with Amazon product reviews. Accumulating a large quantity of data, including ratings and reviews, is the initial stage. As part of data preparation, you can clean up language, fill in missing values, and convert ratings into sentiment labels. If you want to tokenise and embed cleaned text, you can utilise GloVe visuals with 200 dimensions. By combining Spatial Drop out, GlobalMaxPooling1D, or thick layers with Batch Normalisation and Dropout, we are able to construct a Bidirectional LSTM model. In the end, a softmax layer determines whether the outcomes are good, neutral, or bad. Guaranteeing a uniform distribution of data is the goal of stratified sampling. One way to evaluate a model's performance is by looking at its accuracy, recall, accuracy, or F1-score.

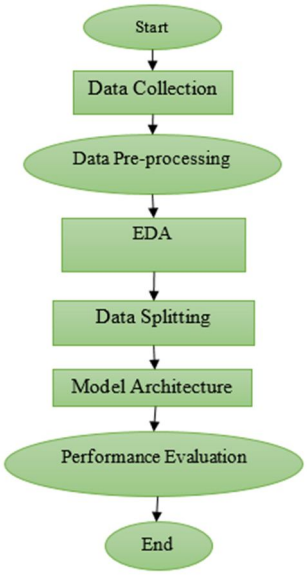


Fig. 2 Proposed Flow Chart

A. Data Collection

The information utilised for this research comes from the Amazon Product Reviews database, which may be accessed at this URL: https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews. There is a wealth of information about product reviews in this database covering a variety of topics, including books, clothing, and technology. Typical review metadata includes product ID, user name, review score (rating), time, summary, and full text in addition to helpfulness score and review score. When it comes to sentiment analysis, we solely utilise the Text and Score columns. The whole text of the review is in the Text column, while the review's numerical rating is in the Score" column. This dataset is ideal for developing and testing NLP models due to its abundance of real-world intricacy, noise, and diverse language usage. Training AI models, especially those with complex architectures that rely on large amounts of data, becomes much easier due to the abundance of reviews. We parse the data after downloading it in JSON format so that process it.

B. Data preprocessing

Making sure the input text is clean, organised, and ready to train machine learning models is a critical step in data preparation. The initial stage in maintaining accurate data is to remove rows with missing values in the "Text" or "Score" columns. When creating a sentiment label, the "Score" column works well. "Positive" refers to reviews with a score of 4 or higher, "Neutral" to those with a score of 3 or lower, and "Negative" to those with a score of 3 or below. For multi-class categorisation, this converts the integers into categories. The next step is to complete the text preparation. After stripping the text of non-alphanumeric characters and making it entirely lowercase, NLTK divides it into tokens. By combining stemming with the Porter Stemmer, we may eliminate common stop words and other words that do not contribute to the meaning of the text. We discard tokens with less than three characters and reassemble cleaned ones. What follows is the tokenisation and sequencing of the cleaned text by Keras' Tokeniser. To do this, it uses an integer index mapping to the top 5,000 words. To maintain consistency in the input size, sequences are paddinged to 200 tokens. To facilitate their usage with categorical cross-entropy loss, emotion labels undergo one-hot encoding. This neural network uses a pre-trained GloVe embedding matrix, which consists of 200-dimensional vectors, to provide its embedding layer with rich semantic word representations.

C. Exploratory Data Analysis (EDA)

Data distribution, structure, and patterns can be better understood with the aid of exploratory data analysis (EDA). To make informed choices regarding preprocessing and modelling, this data is helpful. We begin by checking the dataset size and looking for missing values, particularly in the Score and Text columns. The majority of reviews are positive, while a small percentage are neutral or negative, as seen in a class distribution plot. This suggests an imbalance. Stratified sampling is clearly required in this case. Word clouds for different feeling classes show you the most frequently used words. To illustrate the difference, you will often see terms like "love," "wonderful," and "outstanding" in favourable evaluations and "poor," "waste," and "disappointed" in negative ones. Counting the number of words or bytes in each review allows us to determine the average length of reviews. This demonstrates the need to pad or shorten the model input. Distributions of scores and the correlation between inspection length and emotional tone are also considered. In particular, for tasks like sequence management and class balance, this analysis guides our decisions regarding data preprocessing and model construction.

1) Distributions

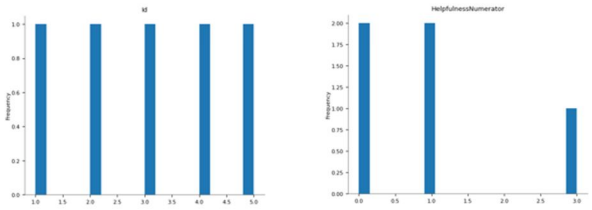


Fig. 3 Histograms of ID and Helpfulness Numerator

Fig. 3 Shows histograms for the ID and Helpfulness Numerator characteristics. The ID histogram shows that there are no duplicates by showing that each review has a unique identification. The "Helpfulness Numerator" histogram shows a right-skewed distribution, which means that most reviews got low helpfulness scores. This could mean that users didn't interact with or give comments on how valuable they were.

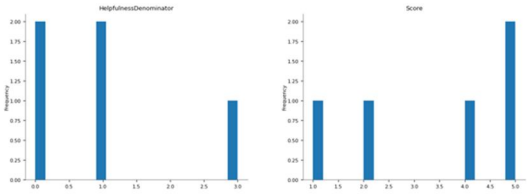


Fig. 4 Histograms of Helpfulness Denominator and Score

Fig. 4 Shows histograms of the Helpfulness Denominator and Score characteristics. The Helpfulness Denominator is substantially skewed to the right, which means that most evaluations didn't get many helpfulness votes. The "Score" histogram displays an uneven distribution, with more good ratings (mainly 4 and 5) than negative ones. This suggests that users tend to favour positive reviews.

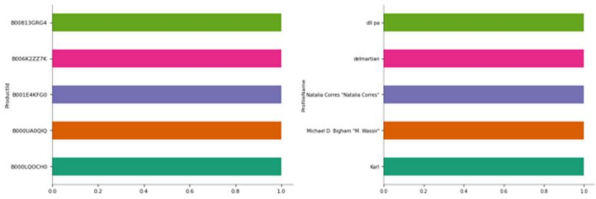


Fig. 5 Bar charts of Product ID, Profile Name

Fig. 5 shows bar charts that show how Product IDs and Profile Names are spread around. These graphs make it easier to see which goods get the most reviews and which reviewers are the most engaged. The graphic shows how often entries are made, which shows patterns in how popular products are and how engaged users are across the dataset.

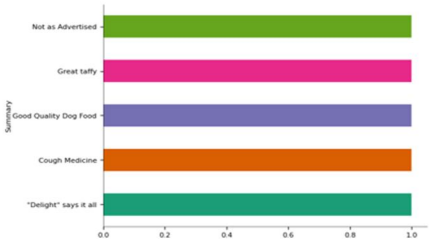


Fig. 6 Bar chart of review Summary

Fig. 6 shows a bar chart of the review summaries, with the most popular phrases used by reviewers highlighted. This visualization shows common summary terms, which can help you understand how customers feel and what themes keep coming up. It helps to know how people usually sum up their experiences with a product in a few words.

2) 2-d distributions

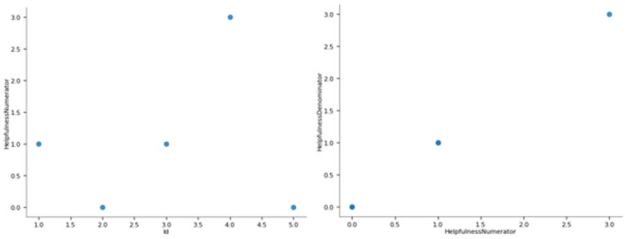


Fig. 7 Scatter plots of helpfulness and ID

Fig. 7 displays scatter plots of Helpfulness Numerator and Helpfulness Denominator vs D. The plots don't show a clear relationship, which means that helpfulness votes are spread out arbitrarily between review IDs. Most of the data points are grouped around lower values, which means that most reviews didn't get much user comment on how useful they were.

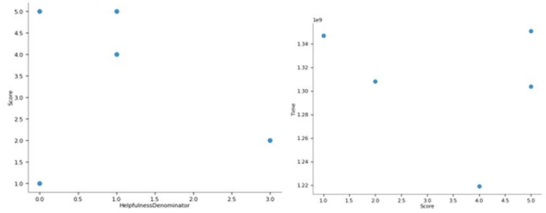


Fig. 8 Scatter plots of review metrics

Fig.8 shows scatter plots of different review measures, such as Helpfulness Numerator and Helpfulness Denominator. The graphs depict patterns that are sparse and spread out, with dense clustering at lower values. This means that even if most reviews don't help much, readers may sometimes leave additional comments or feedback on reviews that are ranked higher.

3) Values

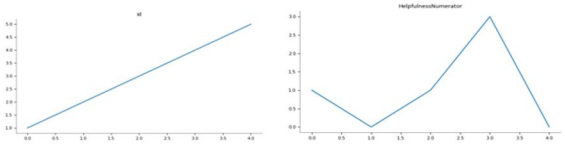


Fig. 9 Line plots of ID and Helpfulness Numerator

Fig 9. Shows line plots of ID against Helpfulness Numerator. The plot shows how helpfulness scores change among review IDs. Most of the time, the scores stay low, but there are some spikes that show a few highly scored reviews. This pattern implies that users are not consistently engaging, since only some reviews get a lot of helpfulness comments.

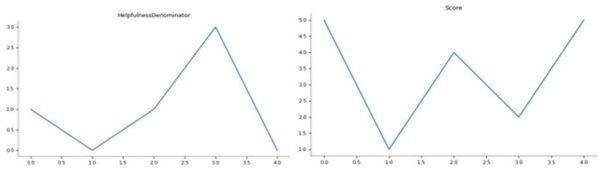


Fig. 10 Line plots of Helpfulness Denominator and Score

Fig 10. Shows line plots of the Helpfulness Denominator and Score. The plot illustrates that most reviews had low helpfulness denominator values, which means that not many users voted. The score values stay rather consistent, with a few surges here and there. This suggests that reviews that get a lot of stars don't usually get a lot of helpfulness evaluations.

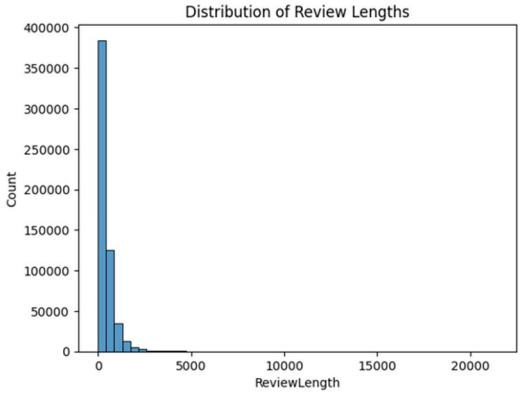


Fig. 11 Histogram of review lengths distribution

Fig 11. Shows a histogram of the distribution of review lengths, which is based on the number of words in each review. The distribution is tilted to the right, with most reviews being brief and the number of reviews going down slowly as the length of the review goes up. There are a few really long reviews, which shows that people write in different ways and with different levels of information.

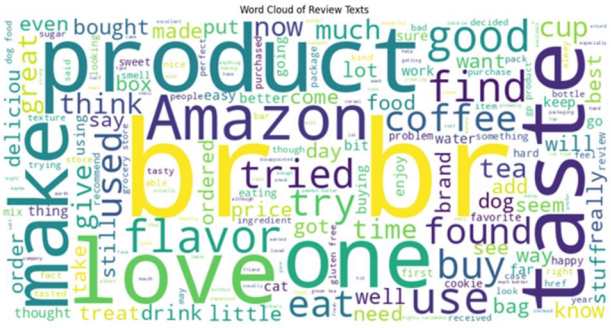


Fig. 12 Word cloud of review texts

Fig 12. Shows a word cloud made from the review texts that shows the words that show up the most. Words that are used a lot, like great, product, love, and use, are bigger, which shows that they are used a lot. The visualization gives a fast look at the main ideas and feelings that are shared in the reviews.

D. Data Splitting

To ensure a fair representation and evaluation of all sentiment classes, we employ a Stratified Train-Test separation to partition the dataset. Both the training and testing datasets maintain their original class distributions while using this strategy. In cases involving multiple classes, where the distribution of classes might not be uniform, this becomes very important. Using Scikit-learn's StratifiedShuffleSplit, we trained the model with 80% of the data and tested it with 20%. Sorting results in less or no bias in the sample by changing the ratios of the three sentiment categories (positive, neutral, and negative).

Before training starts, it is important to build a pre-trained embed matrix using GloVe embed (glove.6B.200d.txt). A 200-dimensional vector represents each word in the tokenizer's lexicon. A word's vector in GloVe takes up one row in the embedding matrix. If it isn't, then that row's value stays 0. The matrix describes the model's embedding layer. This layer is highly informative in terms of semantics, which speeds up training and improves overall performance.

E. Model Architecture

By utilising a Bidirectional Short-Term Long-Term Memory (BiLSTM) architecture, this sentiment categorisation model sorts product reviews into three distinct groups: good, bad, and negative. It excels at processing sequential material, such as natural language, because it can recall both the word order and their contextual meaning. To demonstrate the semantic relationships between words, the model makes use of pre-trained word embeddings. Additionally, it has BiLSTM layers that can process information in both directions, allowing it to grasp the full context. Dense layers combine features to make sure the classification is accurate. This combined method helps the model quickly handle complicated, noisy text input. This is helpful for activities that include more than one class of sentiment and helps us figure out how people feel about things in reviews.

1) Embedding Layer with Pre-trained GloVe Vectors: The model has a trainable embedding layer that starts with 200-dimensional GloVe vectors and turns words into dense semantic representations. Fine-tuning during training changes these embeddings so that they fit the different situations and emotions in the dataset. This helps the machine better understand and sift through different product reviews.

2) Spatial Dropout for Regularization: A SpatialDropout1D layer with a 0.3 rate loses full word embedding during training to avoid overfitting while making the model more general. This stops the model from relying too much on single words, which lets it learn more generic patterns that work better on text data that is sparse or noisy.

3) Bidirectional LSTM Layer: A BiLSTM layer with 128 units handles input in both directions, taking into account the words around it. When `return sequences=True`, it gives hidden states for all time steps, which lets downstream layers learn from every token. This improves the quality of the representation and the model's ability to classify sentiment.

4) Global Max Pooling Layer: GlobalMaxPooling1D comes after the BiLSTM layer and turns outputs of different lengths into fixed-length vectors by picking the highest value from each feature map. This shows the most essential patterns, cuts down on the number of dimensions, and keeps important information so that the model may focus on the most critical aspects for successful sentiment categorization.

5) Dense Block with Batch Normalization and Dropout: A dense layer with 64 units with ReLU activation improves characteristics that come from pooling. Batch Normalization makes training more stable and faster, while a 0.5 Dropout layer stops overfitting by randomly turning off neurons. These layers work together to help the model learn strong, non-linear patterns that will help it classify sentiment correctly.

6) Output Layer for Multi-class Classification: The last dense layer has three units: one for positive, one for neutral, and one for negative feelings. It uses softmax activation to make a probability distribution. This lets the model guess one type of feeling for each review. Softmax makes things easier to understand and works with categorical cross-entropy, which makes it perfect for jobs that involve classifying sentiment into more than one class.

TABLE: 2 HYPERPARAMETER TABLE OF THE MODEL

Component	Hyperparameter	Value / Setting
Embedding Layer	Embedding Dimension	200
	Pre-trained Embeddings	GloVe (glove.6B.200d.txt)
	Trainable	Yes
	Input Sequence Length	200 tokens
	Vocabulary Size	5000 (top frequent words)
Dropout Layer	Spatial Dropout Rate	0.3
LSTM Layer	Units	128
	Bidirectional	Yes
	Return Sequences	True
Pooling Layer	Type	Global Max Pooling 1D

Dense Layer	Hidden Units	64
	Activation	ReLU
	Batch Normalization	Yes
	Dropout Rate	0.5
Output Layer	Output Units	3 (Positive, Neutral, Negative)
	Activation	Softmax
Optimizer	Type	Adam
Loss Function		Categorical Crossentropy
Metrics		Accuracy, Precision, Recall, F1 Score
Training Params	Epochs	100 (with early stopping)
	Batch Size	128
Callbacks	EarlyStopping Patience	3 epochs
	ReduceLROnPlateau Factor	0.5
	ReduceLROnPlateau Patience	2 epochs
	Minimum Learning Rate	1e-6

To keep everything under control, the model uses a 128-unit BiLSTM layer, 200-dimensional trainable GloVe embeddings, and SpatialDropout1D, Batch Normalisation, and Dropout. We used categorical cross entropy and the Adam optimiser to train it. With early pausing and reduced learning rate, it runs for 100 epochs. Four metrics—recall, accuracy, precision, and F1-score—evaluate the outcomes.

IV. RESULTS & DISCUSSION

When it comes to recall, precision, accuracy, and the F1-score, the Bidirectional short- and long-term memory (BiLSTM) model truly shines. The low loss and high accuracy suggest that the training process or accurate projections are nearly error-free. The accuracy and recall ratings reveal how well the model controls for false positives and negatives. The F1-score indicates the classification performance of a balanced model. If the values of the training, validation, and testing metrics are similar, it means that there is no overfitting and that the model is generalisable. The fact that the model maintains its stability in the face of massive volumes of noisy Online review data is a testament to its strength. Based on the outcome, the BiLSTM model is capable of identifying many sentiment categories in practical e-commerce settings.

A. Accuracy

The number of right guesses divided by the total number of predictions is what accuracy means in sentiment categorisation. It can be useful, but it can also be misleading when the datasets aren't balanced and the classes that are most common make the results look better. So, to make sure that all sentiment categories do equally well in multi-class situations, accuracy should be looked at together with precision, recall, and F1-score.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

B. Loss

Loss tells you how far anticipated outputs are from real labels, which helps you improve your model. When doing multi-class sentiment analysis, categorical cross-entropy compares the projected probability to the true labels. Better performance means less loss. Loss shows how sure you are about your predictions, unlike accuracy. This helps you find overfitting or under fitting during training and validation.

$$Loss = -\frac{1}{m}\sum_{i=1}^{m} yi.\log(yi) \quad (2)$$

C. *Precision*

Precision is a measure of how well a model finds true positives among all the anticipated positives. When doing multi-class sentiment analysis, you take the average of the scores for each class. High accuracy lowers the number of false positives, which is useful in datasets that aren't balanced because it stops negative reviews from being wrongly classified as positive, which can lead to wrong product evaluations.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

D. *Recall*

Recall shows how well a model can find all the real positives, which shows how well it catches true sentiment instances. In sentiment analysis, high recall means that most of the relevant reviews are found accurately. It's important to catch important feelings, like negative feedback, because missing them could mean missing out on unhappy customers or making bad choices.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

E. *F1-score*

The F1-score combines precision and recall into one balanced metric, reflecting a model's overall effectiveness. It's especially useful in multi-class sentiment analysis with imbalanced data, offering a realistic view of performance. A high F1-score indicates accurate and comprehensive predictions, making it a reliable measure for evaluating classification models.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

F. *Confusion Matrix*

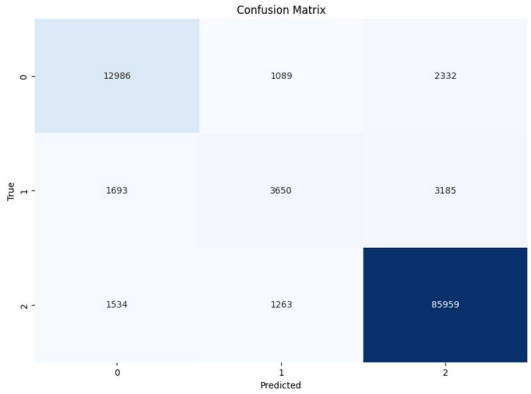


Fig. 13 Confusion Matrix of Sentiment Classification

Fig.13 shows the confusion matrix for sentiment classification, which shows how well the model did on both expected and real classes. It showsnally shows the right and wrong predictions for each sentiment category, giving information about the model's accuracy, trends of misclassification, and overall efficacy.

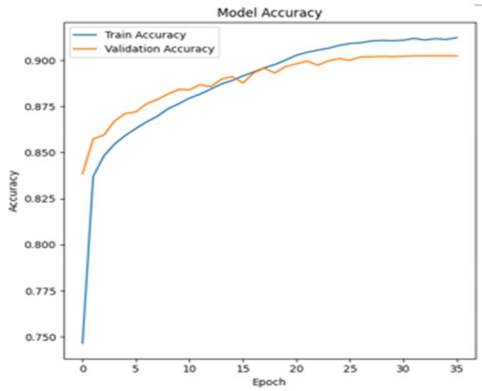


Fig. 14 Model accuracy: train and validation

Fig. 14 gives an indication of the model's accuracy across several epochs of training and validation. Learning is having the desired effect, as the graph shows that accuracy is increasing over time. Due to the close proximity of both the training and validation curves, overfitting is minimal. This proves that the Bidirectional LSTM model is good at adapting to novel datasets.

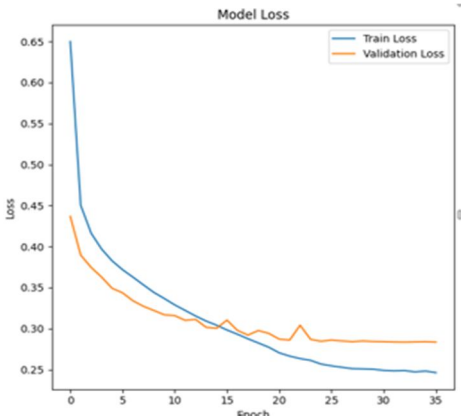


Fig. 15 Model loss: training and validation

Fig 15. Shows the training and validation loss curves over time, which show how the model is learning. A consistent drop in both losses shows that convergence is working. The small difference between the curves implies that the Bidirectional LSTM model was able to generalise well and didn't Overfit too much during the training procedure for sentiment classification.

TABLE: 3 Performance Evaluation During Training Of Proposed Bilstm Model

Metrics	Accuracy	Loss	Precision	Recall	F1-score
Value	0.9124	0.2463	0.9290	0.8962	0.9123

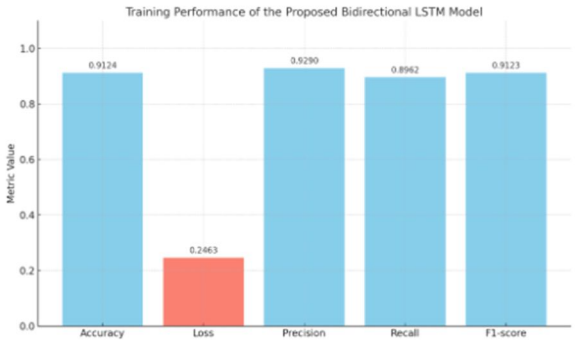


Fig. 16 Proposed Bi-LSTM Model Training Performance

The Bidirectional LSTM model shows strong performance with 0.9124 accuracy, 0.2463 loss, 0.9290 precision, 0.8962 recall, and a balanced 0.9123 F1-score. These metrics confirm its reliability in multi-class sentiment classification, demonstrating accurate, robust predictions for positive, neutral, and negative reviews in real-world textual datasets.

TABLE: 4 Comparative Analysis Between Existing Models And Proposed

Models	Accuracy	References
Random Forest	89%	[5]
CNN	78.062%	[22]
Proposed Bidirectional LSTM Model	91.24%	----------

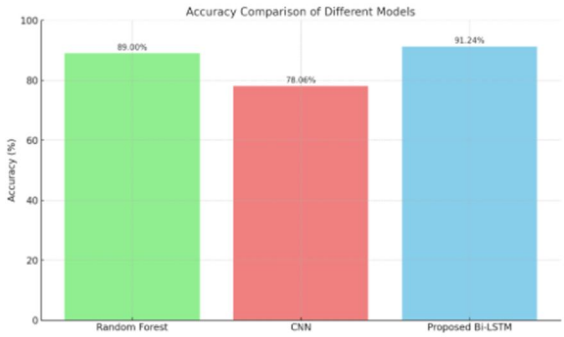


Fig. 17 Comparative Analysis

According to a comparative analysis, BiLSTM is the best model, with an accuracy rate of 91.24%. It does a great job of capturing context in sequential text. Random Forest gets an 89% score, which is average, but CNN gets a lower score of 78.06% because it can't handle sequences as well. BiLSTM is the best option for sentiment classification jobs since it understands context effectively.

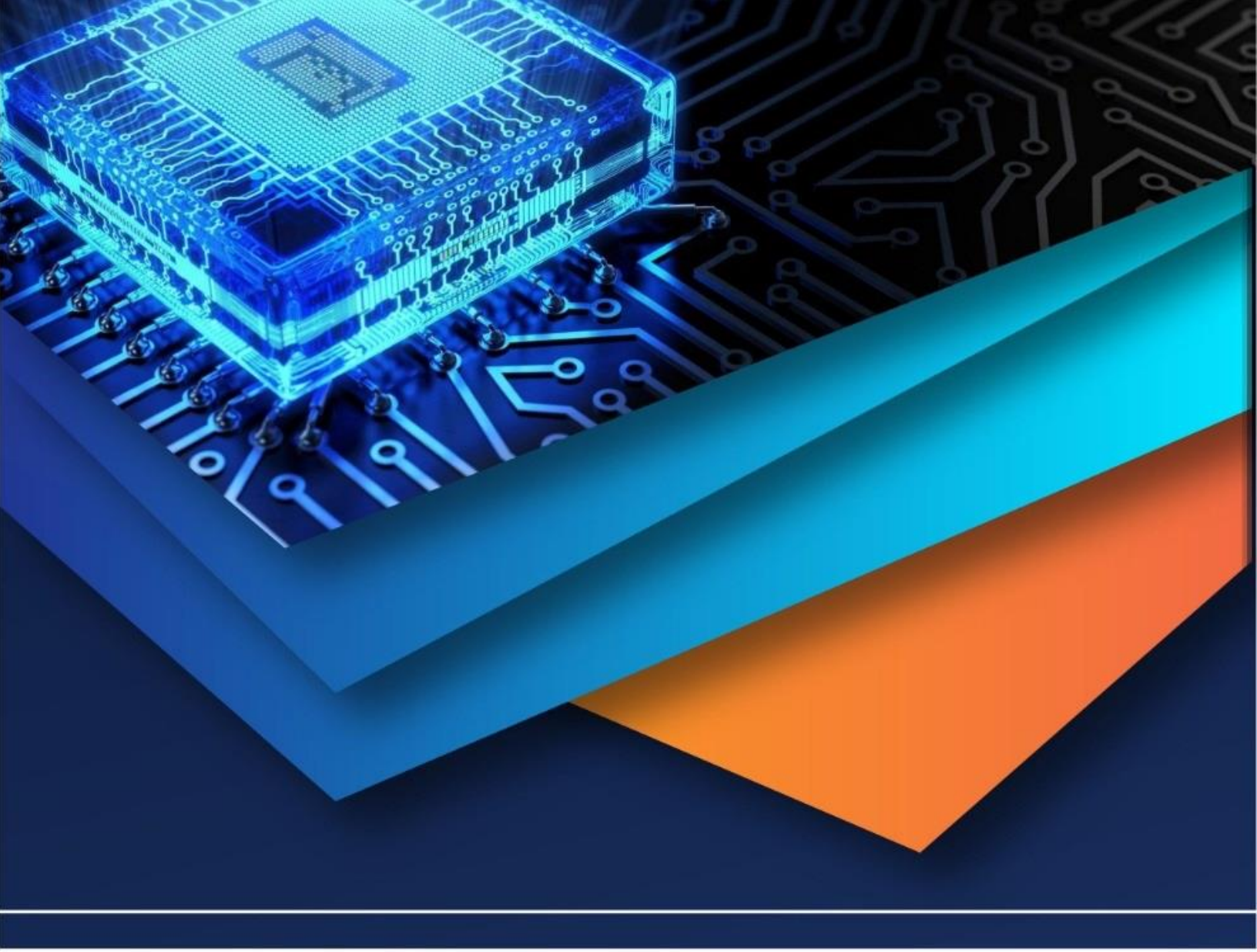
V. CONCLUSION

Finally, the results demonstrate that a deep learning-based Bidirectional short- and long-term memory (BiLSTM) model can reliably gauge customers' sentiments regarding Amazon purchases. The study lays a good groundwork for model training by carefully working with a large dataset. This involves cleaning up the text, adding sentiment labels, and using GloVe word embedding. The model can figure out complex contextual dependencies in review text since it has spatial dropout, BiLSTM layers, global max pooling, and dense layers with batch normalization and dropout. The BiLSTM model did quite well on the tests, getting 91.24% of the answers right and doing well on the other criteria as well. This means that it can arrange things quite well and is pretty reliable. The BiLSTM proved far better at handling the sequential and complex character of natural language than older models like Random Forest and CNN. This shows how powerful recurrent architectures are, especially when it comes to identifying emotions across several classes. The results not only show that the model can handle real-world, noisy text data, but they also imply that it could help businesses get useful information from customer evaluations. Overall, this work introduces a sentiment analysis pipeline that is both scalable and accurate and can be utilized on any e-commerce site. This will lead to a better customer experience and product development plans based on data.

REFERENCES

[1] M. Belal, J. She, and S. Wong, “Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis,” 2023, [Online]. Available: http://arxiv.org/abs/2306.17177
[2] H. Computational, M. Methods, M. Volume, and A. Id, “Retracted: Development of NLP-Integrated Intelligent Web System for E-Mental Health,” Comput. Math. Methods Med., vol. 2023, pp. 1–1, 2023, doi: 10.1155/2023/9780851.
[3] C. A. License, “Universities : A Research Based on Artificial Intelligence,” vol. 2022, 2023.
[4] “Amazon Product Reviews.” https://aws.amazon.com/blogs/machine-learning/detect-sentiment-from-customer-reviews-using-amazon-comprehend/ (accessed Jun. 28, 2025).
[5] D. C. Feng, W. J. Wang, S. Mangalathu, and Z. Sun, “Condition Assessment of Highway Bridges Using Textual Data and Natural Language Processing- (NLP-) Based Machine Learning Models,” Struct. Control Heal. Monit., vol. 2023, 2023, doi: 10.1155/2023/9761154.
[6] “Retracted: Sentiment Analysis of Statements on Social Media and Electronic Media Using Machine and Deep Learning Classifiers,” Comput. Intell. Neurosci., vol. 2023, pp. 1–1, 2023, doi: 10.1155/2023/9846879.
[7] D. Mahto, S. C. Yadav, and G. S. Lalotra, “Sentiment Prediction of Textual Data Using Hybrid ConvBidirectional-LSTM Model,” Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/1068554.
[8] C. Jin, Z. Song, J. Xu, and H. Gao, “Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics,” Wirel. Commun. Mob. Comput., vol. 2022, 2022, doi: 10.1155/2022/1167462
[9] Y. Wen, W. Li, Q. Zeng, H. Duan, F. Zhang, and S. Kang, “Syntactic Knowledge Embedding Network for Aspect-Based Sentiment Classification,” Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/1352028.
[10] V. Dogra et al., “A Complete Process of Text Classification System Using State-of-the-Art NLP Models,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/1883698.
[11] A. Vitetta, “Sentiment Analysis Models with Bayesian Approach: A Bike Preference Application in Metropolitan Cities,” J. Adv. Transp., vol. 2022, 2022, doi: 10.1155/2022/2499282.
[12] H. Wang, “Word2Vec and SVM Fusion for Advanced Sentiment Analysis on Amazon Reviews,” Highlights in Science, Engineering and Technology, vol. 85. pp. 743–749, 2024. doi: 10.54097/sw4pft19.

[13] M. K. Shaik Vadla, M. A. Suresh, and V. K. Viswanathan, "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT," Algorithms, vol. 17, no. 2, 2024, doi: 10.3390/a17020059.
[14] Y. M. Latha and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," ETRI J., vol. 46, no. 4, pp. 633–647, 2024, doi: 10.4218/etrij.2023-0162.
[15] K. Chenglerayen, "From Reviews to Results : Leveraging Amazon Feedback for Product Evolution From Reviews to Results : Leveraging Amazon Feedback for Product Evolution," no. December, 2024, doi: 10.32628/IJSRSET2411458.
[16] H. Boril, K. King, G. Strenski, A. Olson, and W. Husen, "Building Sentiment Analysis Pipeline: A Case Study On Amazon Reviews," 27th IEEE/ACIS Int. Summer Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2024 - Proc., pp. 189–194, 2024, doi: 10.1109/SNPD61259.2024.10673954.
[17] M. S. Kumar, S. Bhagath, N. Tharun, K. Kishorekumar, and M. Narendar, "Transformer-Based Sentiment Analysis : A Comparative Study of Machine Learning , Deep Learning , and BERT Models on Amazon Product Reviews," vol. 2, no. 1, pp. 74–81, 2024.
[18] E. Hashmi and S. Y. Yayilgan, A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews, no. 0123456789. Springer US, 2024. doi: 10.1007/s10660-024-09896-5.
[19] A. S. M. AlQahtani, "Product Sentiment Analysis for Amazon Reviews," Int. J. Comput. Sci. Inf. Technol., vol. 13, no. 3, pp. 15–30, 2021, doi: 10.5121/ijcsit.2021.13302.
[20] N. M. Alharbi, N. S. Alghamdi, E. H. Alkhammash, and J. F. Al Amri, "Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews," Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/5536560.
[21] M. V. Rao and C. Sindhu, "Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analysis," 2021 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2021, no. October, pp. 196–199, 2021, doi: 10.1109/WiSPNET51692.2021.9419432.
[22] I. E. Fattoh, F. Kamal Alsheref, W. M. Ead, and A. M. Youssef, "Semantic Sentiment Classification for COVID-19 Tweets Using Universal Sentence Encoder," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/6354543.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY