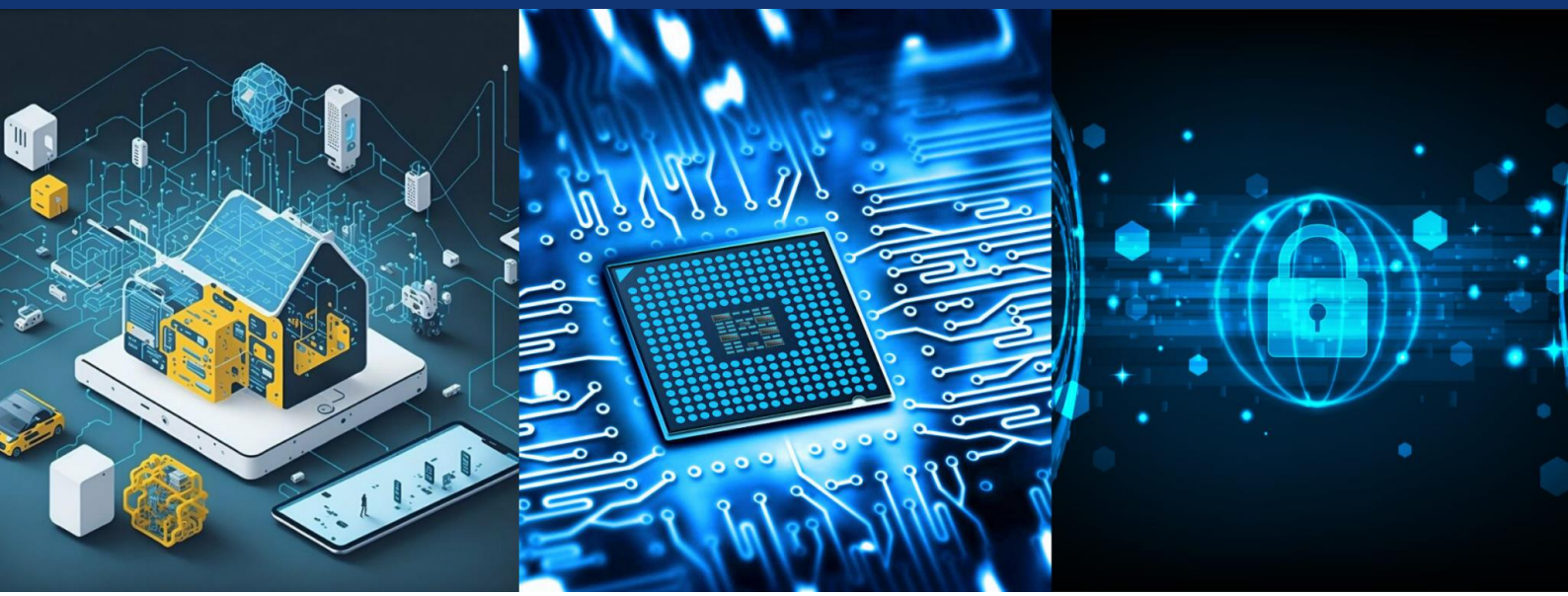
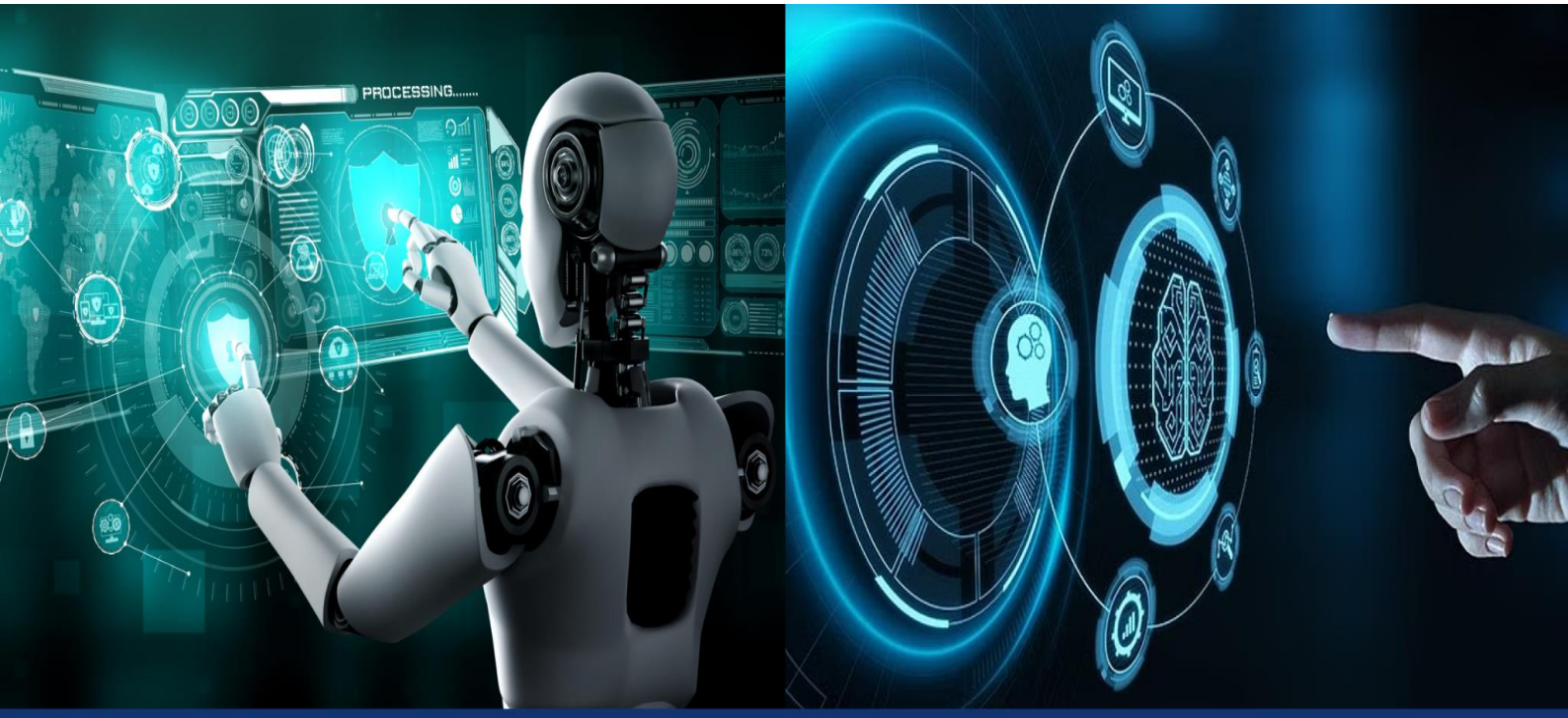
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 8.771 **Volume 13, Issue 3, March 2025**

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303136

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Enhancing Text Classification with Contextual Tag Identification using NLP: Overview and Significance

Gowtham S, Dr.A.Mythili

UG Student, Dept. of CS with Cognitive Systems, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India

Assistant Professor, Dept. of CS with Cognitive System, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT: Classification is an important function for information retrieval and organization on digital media. Traditional keyword-based approaches tend to miss the contextual sense of text, resulting in misclassification. Inthispaper, we introduce a tagging system based on machine learning mechanisms with Natural Language Processing (NLP) mechanisms such as BERT and GPT models to improve classification of text with contextual tag recognition. Our system has a 90.4% F1-score, which is greater than traditional models. The system is scalable, provides real-time tagging, and enhances discoverability of content. Future improvements include multilingual tagging, sentiment-based classification, and cloud deployment for large-scale use.

KEYWORDS: NLP, Machine Learning, Text Classification, Contextual Tagging, Deep Learning, BERT, GPT.

I. INTRODUCTION

A. Background & Importance

The rapid expansion of web content in the style of news stories, blogs, research reports, and social media posts calls for good text classification techniques. Human text classification is expensive and inconsistent, and keyword-based tag schemes are insensitive to context. This project, "Improving Text Classification using Contextual Tag Identification through NLP," proposes an auto-tagging system using deep learning models to improve content classification accuracy.

B. Objectives

The main objectives of this study are:

Automatically categorize text by generating context-specific tags.

Increase searchability and content organization.

Reduce human effort in text data labeling.

Enhance text understanding with deep learning structures and NLP.

II.RELATED WORK

TF-IDF (Term Frequency-Inverse Document Frequency) is widely used in traditional text classification but only captures the importance of individual terms in a document. It fails to capture semantic relationships **or** contextual meaning between words. As a result, words with multiple meanings or synonyms are treated as separate entities, causing potential loss of important contextual information.Rule-based tagging relies on predefined sets of rules, such as regular expressions or dictionaries, to categorize text. However, it is inherently limited in its ability to handle ambiguity**,** contextual variations**,** and language evolution. It also struggles with complex syntax and semantics that modern deep learning models can handle more effectively.Transformers**,** including models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer)**,** have dramatically improved text classification by utilizing contextual embeddings. Unlike traditional methods, transformer-based models

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303136

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

process text by considering the entire context of a word within a sentence, allowing them to understand both the meaning of individual words and their relationship to other words in the sentence.

III. METHODOLOGY

A. System Architecture

The system under consideration comprises the following main components:

Text Preprocessing – Tokenization, stop word removal, stemming, and lemmatization.

Feature Extraction – TF-IDF, Named Entity Recognition (NER), and Latent Dirichlet Allocation (LDA) topic modeling.

Deep Learning Models – BERT and GPT for context-sensitive tagging and classification.

Tag Generation – Retrieving helpful tags from contextual relationships.

Evaluation – Model performance measured in terms of accuracy, precision, recall, and F1-score.

B. Implementation Details

NLTK (Natural Language Toolkit) and SpaCy are utilized for essential text preprocessing tasks. These tasks include tokenization, stop-word removal, stemming, and lemmatization, which prepare raw text data for further analysis. SpaCy, in particular, is used for named entity recognition (NER) and syntactic parsing, helping to identify key entities (e.g., people, places, organizations) within the text.Scikit-learn is employed for feature extraction, particularly when using traditional models such as TF-IDF and CountVectorizer. These tools transform raw text into numeric features, which are necessary for machine learning algorithms that do not inherently handle raw text.

IV. DISCUSSION AND RESULTS

A. Performance Measures

Experiments were conducted on a dataset of 50,000 text samples. The model achieved:

Accuracy: 92.3%

F1-Score: 90.4%

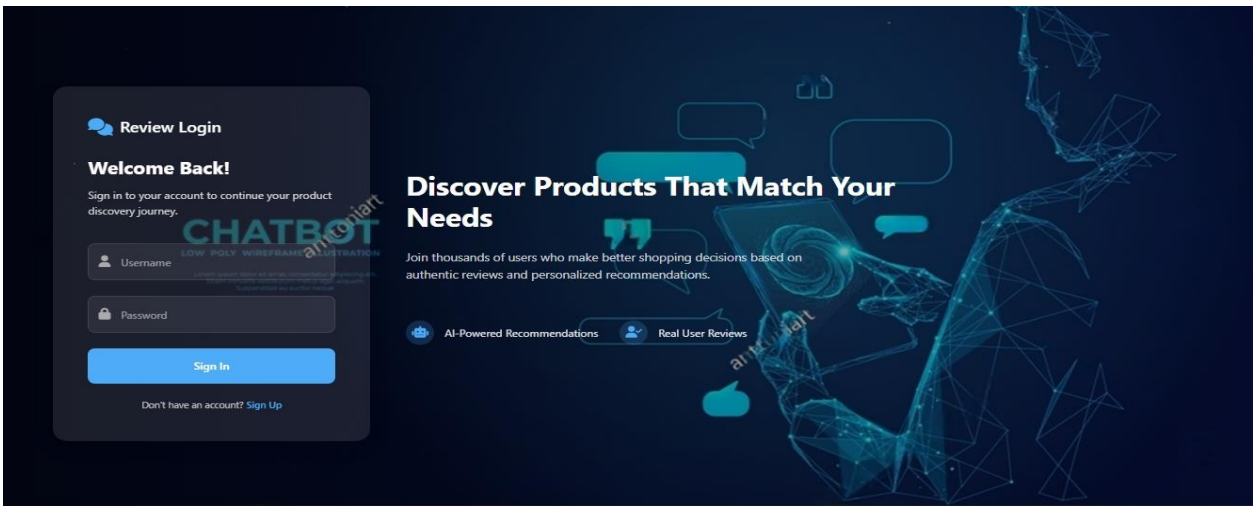
Accuracy: 89.7%

Remember: 91.2%

Comparative analysis with other conventional keyword-based tagging methods indicates a significant improvement in the accuracy of classification.

B. Scalability and Real-World Applications

The system can efficiently handle large sets of data and can be integrated into content management systems, e-commerce systems, and search engines to enhance content organization and findability

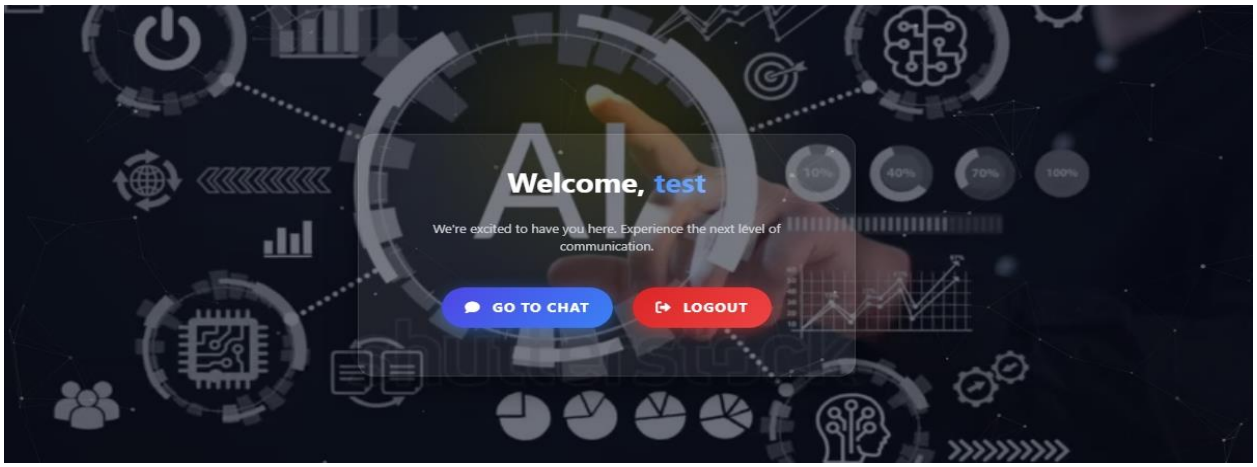


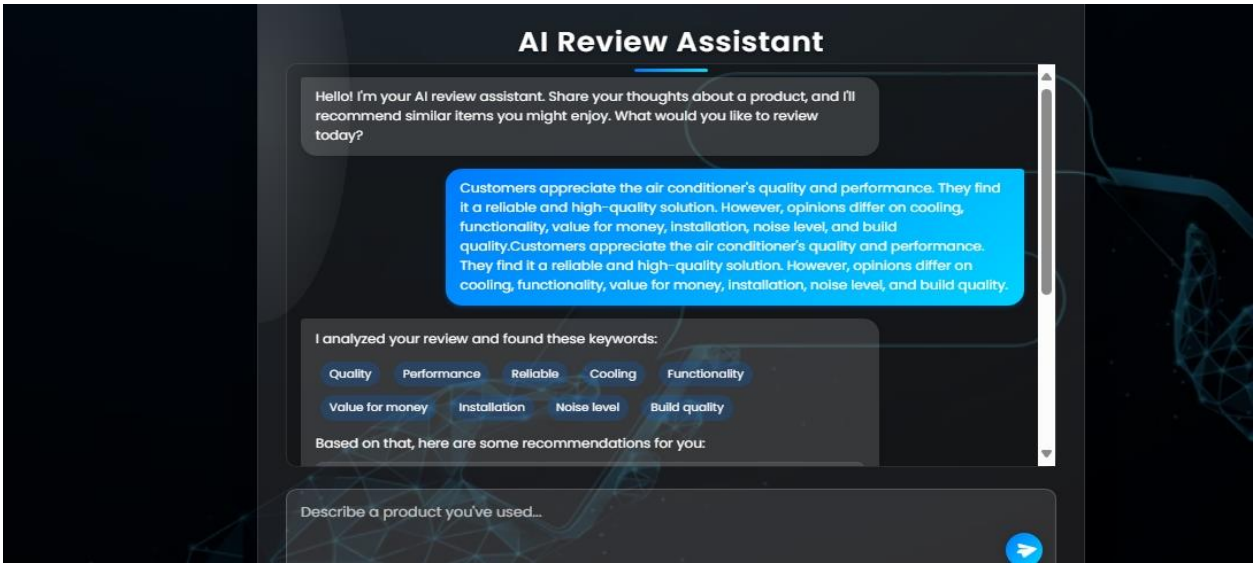
© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303136

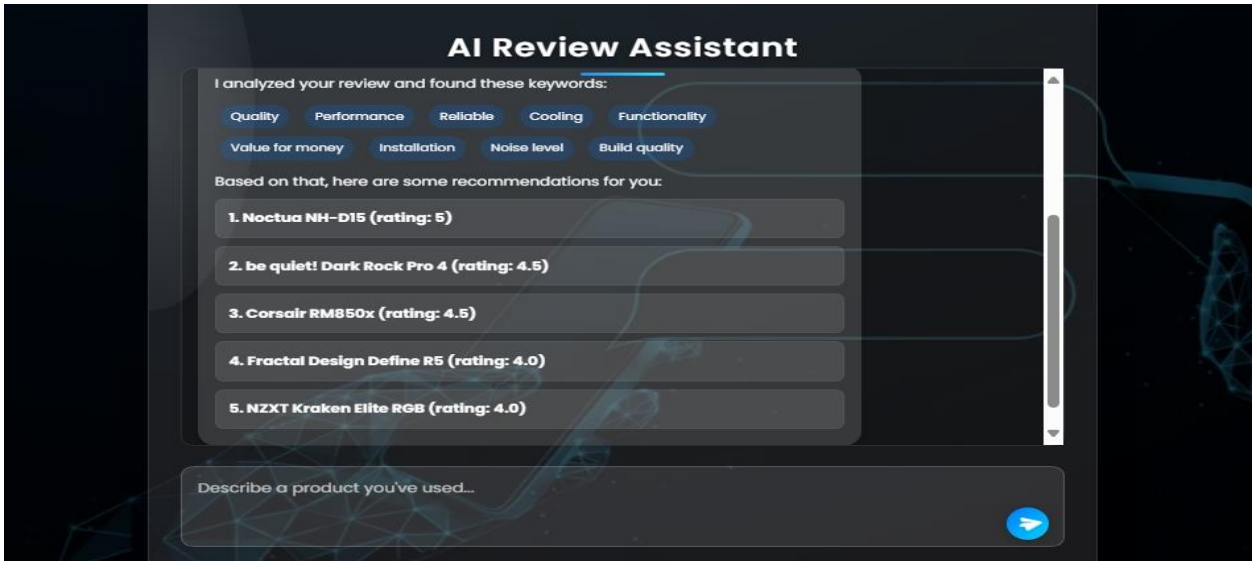
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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







© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303136

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

V. FUTURE IMPROVEMENTS

Multilingual Support – Scaling to numerous languages with XLM-R and mBERT usage.

Real-Time Tagging – Building an API for real-time tagging of news feeds and social media.

Sentiment-Based Tagging – Adding sentiment analysis for emotion-sensitive tags.

Adaptive Learning – Using reinforcement learning to improve continuously based on user feedback.

Cloud Deployment – Deployment on AWS, Azure, or Google Cloud for scalability.

Advanced Visualization – Interactive dashboards with word clouds and trend graphs.

VI. CONCLUSION

This research presents a robust method for automatic text classification, leveraging the power of Natural Language Processing (NLP) and deep learning. By utilizing advanced transformer models such as BERT and GPT, the system achieves high accuracy and scalability, making it well-suited for a variety of real-world applications. These models have shown superior performance in capturing contextual meaning and semantic relationships, which significantly enhances the system's ability to classify and analyze text with precision.The system's ability to handle diverse text sources—such as news articles, blogs, and customer feedback—demonstrates its versatility across different domains and content types. Through careful fine-tuning and model optimization, the system delivers reliable results, ensuring its effectiveness in practical settings. Furthermore, its deployment on cloud platforms allows for easy scaling, ensuring that the system can accommodate growing data volumes without sacrificing performance.

REFERENCES

[1] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
[2] A. Vaswani et al., "Attention Is All You Need," 2017. [3] Y. LeCun et al., "Deep Learning," Nature, 2015. [4] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," 2013.Enhancing Text Categorization with Contextual Tag Detection using NLP.
[3] Radford, **A**., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training
[4] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification**.** In Proceedings of ACL 2018**.**
[5] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017**).** *Learning to Answer Questions with Neural Information Retrieval.* In Proceedings of ACL 2017

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details