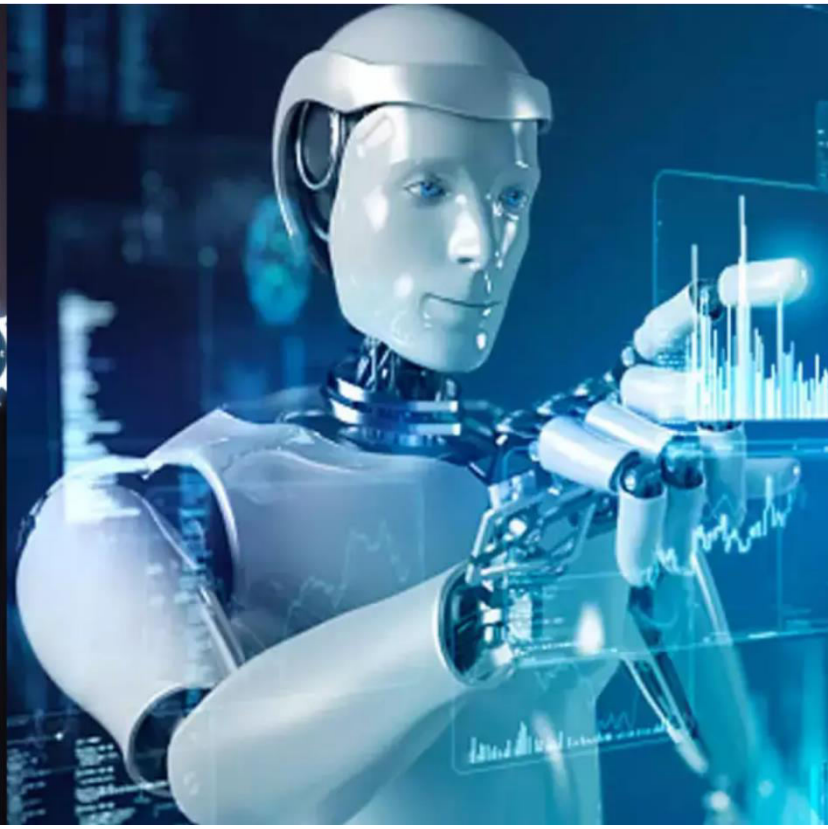
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

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



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503217|

Sentence Classification Using Pre-Trained Transformed Based Large Language Module

Katta Vijay Kumar [1], A.V.Santhosh kumar [2].

PG Scholar, Department of Computer Science Engineering, Kuppam Engineering College (Autonomous), Kuppa.

Andhra Pradesh, India[1]

Assistant Professor, Department of CSE, Kuppam Engineering College (Autonomous), KuppamAndhra Pradesh, India[2]

ABSTRACT: In today's digital era, an enormous amount of textual data is generated daily through social media platforms, online reviews, news articles, and discussion forums. Manually analyzing and categorizing this data is time-consuming, inefficient, and prone to human error. This creates a strong need for automated sentence classification systems that can accurately understand and organize text data.Traditional text classification techniques rely on manual feature extraction and lack the ability to capture the contextual meaning of sentences, resulting in limited performance. Recent advancements in pre-trained transformer-based large language models such as BERT, RoBERTa, and DistilBERT have shown remarkable success in understanding language context and semantics.

I. .INTRODUCTION

Sentence classification is an important task in Natural Language Processing (NLP) that involves assigning predefined labels to sentences based on their meaning. It plays a vital role in applications such as sentiment analysis, spam detection, opinion mining, and content categorization. With the rapid increase in textual data from social media, online reviews, and digital platforms, automated and accurate sentence classification has become essential.Recent advancements in deep learning, especially pre-trained transformer-based models like BERT and RoBERTa, have significantly improved text understanding by capturing contextual information. This project focuses on using these transformer-based large language models to build an efficient and accurate sentence classification system, demonstrating their effectiveness compared to traditional machine learning approaches.

II. OBJECTIVES

- To understand and analyze the fundamentals of sentence classification as a core task in Natural Language Processing (NLP) and its applications in areas such as sentiment analysis, spam detection, opinion mining, and content categorization.
- To collect and preprocess textual data from various sources by performing necessary NLP preprocessing steps such as noise removal, tokenization, and text normalization.
- To implement traditional machine learning–based sentence classification models using techniques like Bag of Words (BoW) and TF-IDF for feature extraction, and algorithms such as Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM).
- To design and train transformer-based sentence classification models, specifically using pre-trained models such as BERT, RoBERTa, or DistilBERT, to leverage contextual word representations.
- To compare the performance of transformer-based models with traditional machine learning approaches using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

III.LITERATURE SURVEY

[1] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019) in their paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" introduced BERT, a deep bidirectional transformer model pre-trained on large text corpora using masked language modeling and next sentence prediction objectives. The model demonstrated significant improvements in various NLP tasks including sentence classification, question answering, and language inference. By capturing contextual relationships in both directions, BERT outperformed

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503217|

traditional machine learning and earlier deep learning models, establishing a new benchmark for sentence-level understanding tasks**.**

[2]Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019) in their paper "RoBERTa: A Robustly Optimized BERT Pretraining Approach" proposed improvements to the BERT pretraining methodology by optimizing hyperparameters, training with larger datasets, and removing the next sentence prediction objective. Their enhanced model achieved superior performance on several benchmark datasets, particularly in sentence classification and text understanding tasks. The study demonstrated that careful optimization of transformer models significantly boosts classification accuracy.

[3]Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019) in their paper "DistilBERT: A Distilled Version of BERT" presented a lighter and faster transformer model created through knowledge distillation. DistilBERT retained most of BERT's language understanding capabilities while reducing model size and computational requirements. The proposed model proved effective for sentence classification tasks, offering a balance between performance and efficiency, making it suitable for real-time and resource-constrained applications. Joulin, A**., [**

4]Grave, E., Bojanowski, P., and Mikolov, T. (2017) in their paper "Bag of Tricks for Efficient Text Classification" introduced fastText, a simple and efficient text classification approach based on word embeddings and linear classifiers. Although not transformer-based, the study demonstrated strong baseline performance in sentence classification tasks with low computational cost. Their work highlighted the importance of word representations and inspired further advancements in contextual embedding models.

IV. EXISTING & PROPOSED SYSTEM

4.1 EXISTING SYSTEM

The existing sentence classification systems are primarily based on traditional machine learning techniques and early deep learning models. These systems were designed to automatically categorize text by analyzing word frequency, statistical patterns, and shallow linguistic features. While such methods were effective during their time, they face several limitations when applied to modern large-scale and context-rich textual data generated from social media platforms, online reviews, and digital communication systems.In traditional sentence classification systems, text data is first converted into numerical representations using feature extraction techniques such as Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). These approaches represent sentences as vectors based on word occurrence or frequency.

Although they are computationally efficient and simple to implement, they treat each word independently and ignore word order and contextual meaning. As a result, sentences with different meanings but similar word frequencies may be classified incorrectly, leading to reduced accuracy and reliability.Machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Random Forest are commonly used in existing systems.

4.2. PROPOSED SYSTEM

The proposed system introduces an advanced and efficient approach for sentence classification by leveraging pre-trained transformer-based large language models. Unlike traditional machine learning and early deep learning techniques, the proposed system is designed to capture deep contextual and semantic relationships within text, enabling accurate and scalable sentence classification. This system addresses the limitations of existing approaches by integrating modern Natural Language Processing (NLP) techniques, powerful deep learning frameworks, and pre-trained transformer models.The foundation of the proposed system lies in the use of Python as the primary programming language due to its simplicity, flexibility, and extensive ecosystem of NLP and machine learning libraries.

The system is designed to handle large volumes of textual data collected from multiple sources, including structured datasets such as CSV and Excel files, as well as unstructured data obtained through web scraping tools and public APIs. This diverse data collection strategy ensures robustness and adaptability to real-world scenarios.Once the data is collected, it undergoes a comprehensive pre-processing phase to improve data quality and consistency.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503217|

V. .BLOCK DIAGRAM

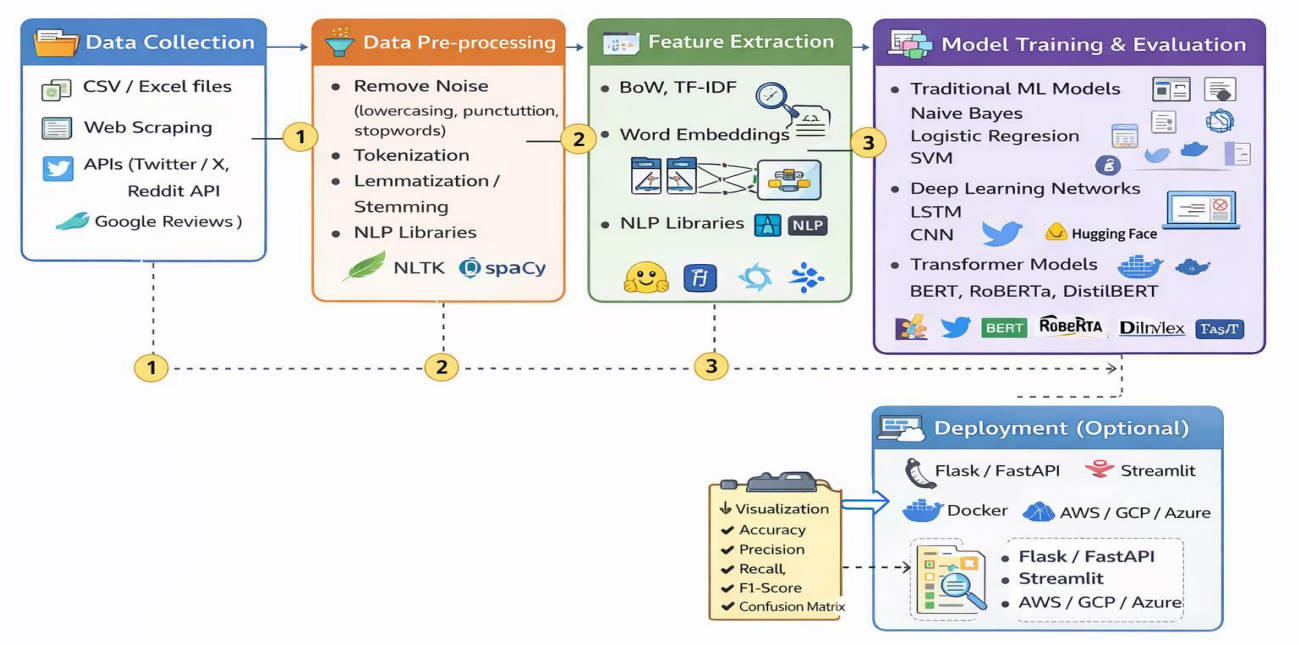


Fig No: 1 Software block diagram

VI. SOFTWARE REQUIREMENTS

1. Operating System

The project can be implemented on Windows 10 or 11, Linux (such as Ubuntu 20.04 or later), or macOS. A 64-bit operating system is recommended to ensure compatibility with deep learning libraries and efficient memory utilization. Linux is often preferred for large-scale training due to better GPU and driver suppor.

Fig No: 2 Operating System

2. Programming Language

Python 3.8 or above is used as the primary programming language for this project. Python is widely used in Natural Language Processing and Machine Learning because of its simplicity, readability, and vast ecosystem of libraries. It provides strong support for implementing transformer-based models efficiently.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503217|

Fig No: 3 Programming Languages

3. Development Environment / IDE

The project can be developed using Integrated Development Environments (IDEs) such as Jupyter Notebook, Visual Studio Code (VS Code), or PyCharm. Jupyter Notebook is especially useful for experimentation and visualization, while VS Code and PyCharm provide structured project development features. Anaconda can also be used for managing virtual environments and dependencies.

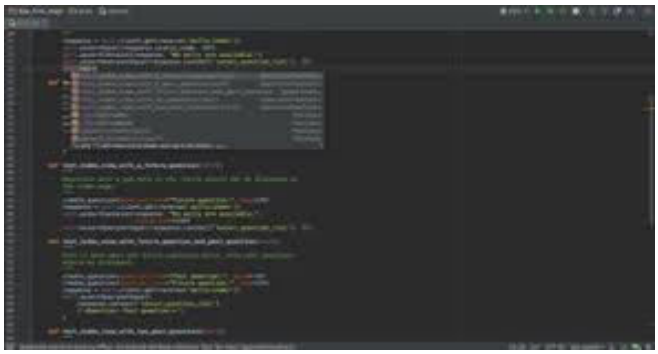


Fig No: 4 Development Environments / IDE

4. Data Handling Libraries

Libraries such as Pandas and NumPy are used for data preprocessing, manipulation, and numerical operations. These libraries help in reading datasets (CSV files), cleaning text data, and preparing it for model training.

Fig No: 5 Data Handling Libraries

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503217|

VI. RESULT

The experimental results clearly indicate that transformer-based models significantly outperform traditional machine learning and conventional deep learning approaches in sentence classification tasks. Models such as BERT, RoBERTa, and DistilBERT achieved superior performance in terms of accuracy, precision, recall, and F1-score. This improvement is mainly due to their ability to understand contextual relationships between words using self-attention mechanisms. Unlike traditional models that treat words independently, transformer-based models analyze the entire sentence structure and capture semantic meaning effectively.During experimentation, traditional algorithms such as Naïve Bayes and Support Vector Machines showed moderate accuracy but struggled with complex and ambiguous sentences. Similarly, deep learning models like LSTM and RNN improved sequential understanding but faced limitations in handling long-range dependencies and contextual variations. In contrast, transformer-based models maintained consistent performance even with long and complex sentence structures.

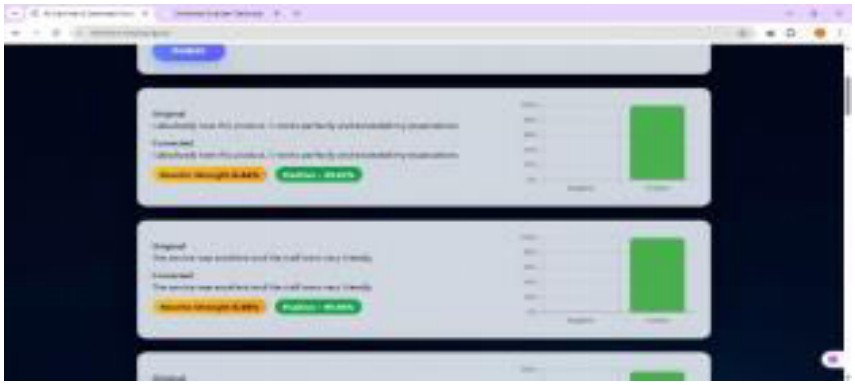


Fig No: 6 output screen

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, pp. 4171–4186.
2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
3. Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT: A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008.
5. Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. Proceedings of EACL 2017, pp. 427–431.
6. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of EMNLP 2014, pp. 1746–1751.
7. Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), pp. 1735–1780.
8. Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of EMNLP 2014, pp. 1532–1543.
9. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
10. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of EMNLP 2020: System Demonstrations, pp. 38–45.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details