International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

An Optimized Transformer Ensemble Model for Context-Aware Speech Act Classification

M. Ganesh[1], Mogulagani Ankitha[2], Komakula Sathwika[2], Kaniganti Chandu[2], Pogula Nagaraju[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

ABSTRACT

The rapid growth of digital communication has increased the need for understanding user intent in textual data, leading to the development of speech act classification systems. Historically, traditional approaches relied on rule-based methods and basic machine learning techniques, which struggled to capture the complexity of natural language. The main problem lies in accurately classifying unstructured and context-dependent text into meaningful categories such as assertive, directive, expressive, and question. Conventional systems often depend on manual feature extraction techniques like bag-of-words and TF-IDF, which fail to capture semantic relationships and contextual dependencies. These limitations result in lower accuracy, poor generalization, and difficulty in handling imbalanced datasets. To address these challenges, there is a need for a more robust and intelligent system that can effectively process textual data and improve classification performance. This research proposes a transformer-driven approach that utilizes eXtreme Language Network (XLNet) for feature extraction, combined with multiple machine learning models including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and a Boosting Fusion Model (BFM) integrating Light Gradient Boosting Machine (LGBM) and Categorical Boosting (CB). The system incorporates Natural Language Preprocessing (NLP) preprocessing techniques and Synthetic Minority Over-sampling Technique (SMOTE) based data balancing to enhance learning efficiency. The proposed system demonstrates significant improvements in accuracy and reliability, with the BFM achieving an accuracy of 96.33%, outperforming LR, RF, and SVM.

Key words: SMOTE (Synthetic Minority Over-sampling Technique), Light Gradient Boosting Machine (LGBM), Categorical Boosting (CatBoost), Support Vector Machine (SVM), Boosting Fusion Model (BFM), Natural Language Processing (NLP).

1. INTRODUCTION

The increasing reliance on digital communication platforms has created a growing need for intelligent systems capable of understanding human intent from textual data. Speech act classification plays a vital role in this context by identifying the functional purpose of a sentence, such as whether it expresses a statement, question, directive, or emotion [1]. Accurate classification of speech acts is essential for applications such as conversational systems, virtual assistants, and automated customer support, where understanding user intent directly impacts system effectiveness [2]. However, the complexity of natural language, including ambiguity, contextual dependency, and variability in expression, makes this task challenging [3].

Traditional approaches to speech act classification have largely relied on manual feature extraction techniques and basic statistical representations, which often fail to capture deeper semantic relationships within text [4]. These methods struggle to handle unstructured and context-rich data, resulting in limited performance and reduced generalization capabilities. With the advancement of deep contextual representations, as shown in figure 1. transformer-based models have demonstrated significant

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

improvements in capturing semantic and syntactic information from text [5]. These models provide richer feature representations that enhance the performance of downstream classification tasks.

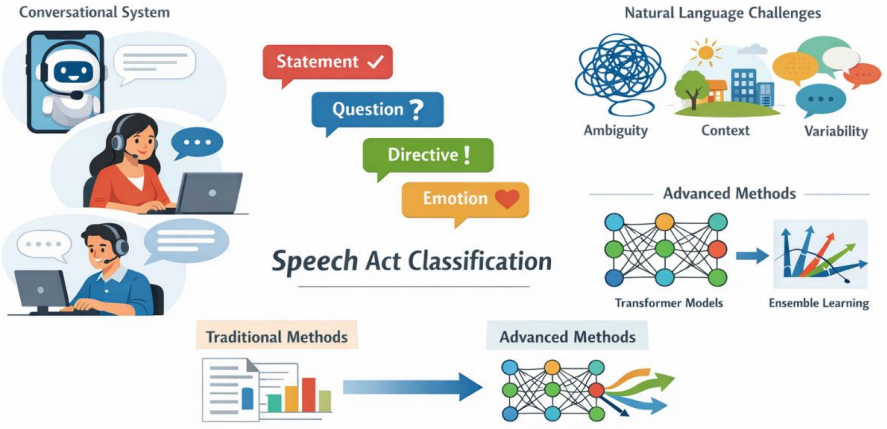


Figure 1: Overview of speech act.

In addition, combining multiple learning models through ensemble techniques has emerged as an effective strategy for improving classification accuracy and robustness [6]. Ensemble models leverage the strengths of different algorithms to reduce individual model limitations and enhance overall predictive performance. By integrating transformer-driven feature extraction with ensemble learning strategies, it is possible to achieve more reliable and scalable speech act classification [7]. This research focuses on developing an efficient framework that combines advanced feature representation with ensemble-based learning to improve classification outcomes and address the limitations of conventional methods.

2. LITERATURE SURVEY

Wieczorkowska et al. [8] proposed Speech-based communication between users and machines is a very lively branch of research that covers speech recognition, synthesis, and, generally, natural language processing. Speech corpora are needed for training algorithms for human–machine communication, especially for automatic speech recognition and for speech synthesis. Generative artificial intelligence models also need corpora for training for every language implemented. Therefore, speech corpora are constantly being created. In this paper, we discuss how to create high-quality corpora. The technical parameters of the recordings and audio files are addressed, and a methodology is proposed for planning speech corpus creation with an emphasis on usability.

Park et al. [9] focused on the individual modules of such a system, and there is an evident lack of research on a dialog framework that can integrate and manage the entire dialog system. Therefore, in this study, we propose a framework that enables the user to effectively develop an intelligent dialog system. The proposed framework ontologically expresses the knowledge required for the task-oriented dialog system's process and can build a dialog system by editing the dialog knowledge. In addition, the framework provides a module router that can indirectly run externally developed modules. Further, it enables a more intelligent conversation by providing a hierarchical argument structure (HAS) to manage the various argument representations included in natural language sentences. Lieskovská et al. [10] focused Emotions are an integral part of human interactions and are significant factors in determining user satisfaction or customer opinion. speech emotion recognition (SER) modules also play an important role in the development of human–computer interaction (HCI) applications. A tremendous

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

number of SER systems have been developed over the last decades. Attention-based deep neural networks (DNNs) have been shown as suitable tools for mining information that is unevenly time distributed in multimedia content. The attention mechanism has been recently incorporated in DNN architectures to emphasise also emotional salient information.

Pallewela et al. [11] proposed speech-based applications, it has become essential to improve audio-based emotion expression. However, there is a lack of specificity and agreement in current emotion annotation practice, as evidenced by conflicting labels in many human-annotated emotional datasets for the same speech segments. Previous studies have had to filter out these conflicts and, therefore, a large portion of the collected data has been considered unusable. In this study, we aimed to improve the accuracy of computational prediction of uncertain emotion labels by utilizing high-confidence emotion labelled speech segments from the IEMOCAP emotion dataset. We implemented an audio-based emotion recognition model using bag of audio word encoding (BoAW) to obtain a representation of audio aspects of emotion in speech with state-of-the-art recurrent neural network models. Our approach improved the state-of-the-art audio-based emotion recognition with a 61.09% accuracy rate, an improvement of 1.02% over the BiDialogueRNN model and 1.72% over the EmoCaps multi-modal emotion recognition models. Augustyniak et al. [12] proposed method, called frequentiment, is based on calculating the frequency of features (words) in the document and averaging their impact on the sentiment score as opposed to documents that do not contain these features. Afterwards, we use ensemble classification to improve the overall accuracy of the method. What is important is that the frequentiment-based lexicons with sentiment threshold selection outperform other popular lexicons and some supervised learners, while being 3–5 times faster than the supervised approach.

3. PROPOSED METHODOLOGY

The proposed methodology follows a structured and systematic pipeline for analysing and classifying speech acts from textual data in an efficient and scalable manner. It begins with data ingestion, preprocessing, and contextual feature extraction, followed by multi-model learning and predictive analysis. Transformer-based feature extraction is used to capture semantic and contextual relationships within the text, which are then utilized by multiple machine learning models for classification, as shown in figure 2. The integration of baseline models with an advanced ensemble approach improves classification accuracy, robustness, and adaptability. The overall framework ensures efficient handling of large-scale textual data while maintaining consistency in performance.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

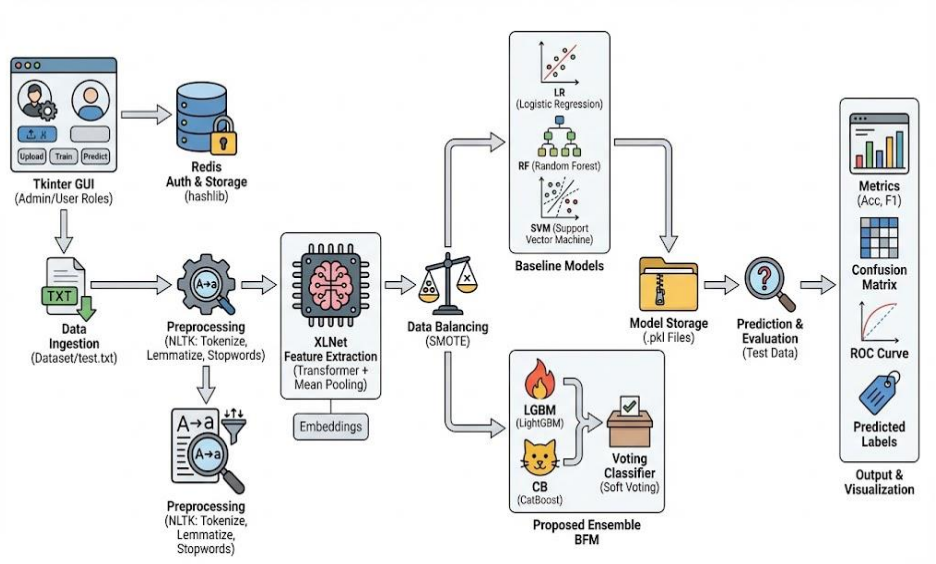


Figure 2: System architecture

User Interface (Tkinter GUI)

- The user and administrator interact with the system through a desktop-based graphical interface developed using Tkinter.
- Admin users can access high-level functions including dataset uploading, XLNet feature extraction, SMOTE balancing, and model training.
- Standard users are restricted to a simplified dashboard for performing predictions on test data and viewing classification results.

Security & Authentication (Redis & Hashlib)

- This study utilizes Redis as a lightweight, high-speed key-value store to manage user credentials and roles.
- To ensure data integrity and security, passwords are never stored in plain text; instead, they are processed through SHA-256 hashing before being committed to the database.
- The system validates roles (Admin vs. User) during login to determine which dashboard functionalities are accessible.

Raw Data Ingestion (Dataset Input)

- The input consists of structured text files containing raw dialogue or sentences labeled with their corresponding speech acts.
- This data serves as the foundation for the entire training and testing pipeline, flowing directly into the preprocessing module.

NLP Preprocessing & Cleaning (NLTK)

- Textual data undergoes a multi-stage cleaning process including tokenization, stop-word removal, and lemmatization.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- The research utilizes NLTK to reduce words to their base forms (e.g., "running" to "run"), which reduces noise and improves the consistency of the feature set.

Transformer Feature Extraction (XLNet)

- This is the primary intelligence engine of the study. The cleaned text is passed through a pre-trained XLNet model.
- The system performs "Mean Pooling" on the hidden states to generate a fixed-length numerical vector (embedding) for every sentence.
- These embeddings capture the bidirectional context and semantic depth required for accurate intent classification.

Data Balancing (SMOTE)

- Because speech act datasets are often imbalanced, this research applies SMOTE.
- It generates synthetic examples for minority classes (like "Question" or "Directive") to ensure the classifiers are trained on a perfectly balanced distribution of labels.

Existing Baseline Models (LR, RF, SVM)

- The balanced **XLNet** features are first fed into three established baseline classifiers:
  - **LR:** A linear model used to establish a predictive baseline.
  - **RF:** A tree-based ensemble used to capture non-linear relationships.
  - **SVM:** A model that optimizes class separation in high-dimensional space.
- These models provide the necessary benchmarks to evaluate the improvements offered by the proposed approach.

Proposed Boosting Fusion Model (BFM)

- The BFM represents the core innovation of this research, acting as a "Deep-Decision Fusion" layer.
- It combines the strengths of two gradient boosting frameworks:
  1. **LGBM:** Optimized for high-speed training and efficiency with large feature sets.
  2. **CB:** Specialized in handling categorical patterns and reducing model overfitting.
- The BFM uses a "Soft Voting" mechanism to aggregate probabilities from both LGBM and CB, resulting in a more robust and accurate classification than any single model.

Prediction Results & Evaluation

- The final output is the classified Speech Act label for the input text.
- The system generates a comprehensive suite of metrics, including Accuracy, Precision, Recall, and F1-Score, along with visual aids like Confusion Matrices and ROC Curves to validate the performance of the BFM against the baselines.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

4. RESULTS AND DISCUSSION

The results of this research demonstrate the effectiveness of combining transformer-based feature extraction with multiple machine learning models for speech act classification. The system evaluates the performance of LR, RF, SVM, and the proposed BFM integrated with LGBM and CB using metrics such as accuracy, precision, recall, and F1-score. Experimental outcomes show that transformer-driven features significantly improve classification performance compared to traditional text representations. The comparative analysis highlights the strengths and limitations of each model under the same feature conditions. The proposed BFM achieves improved prediction consistency due to its ensemble learning capability. Visualization techniques such as confusion matrices and ROC curves further support the evaluation process.

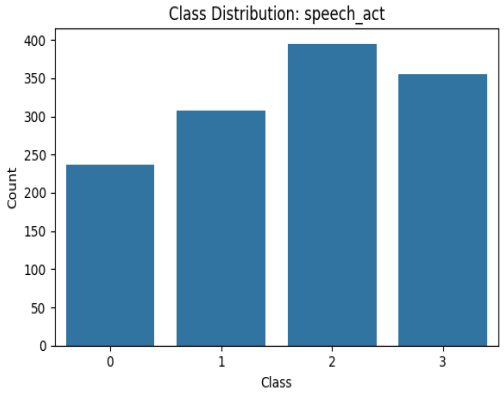


Figure 4: Class Distribution of Speech Acts in the Dataset.

Figure 4 illustrates the class distribution of the speech act dataset used in this research. It depicts the frequency of each class label, providing insight into how the data is distributed across different speech act categories. The figure highlights the presence of class imbalance, where certain classes have higher sample counts compared to others. This imbalance can influence model learning and prediction performance if not properly addressed. The visualization supports the need for techniques such as SMOTE to ensure balanced training data.

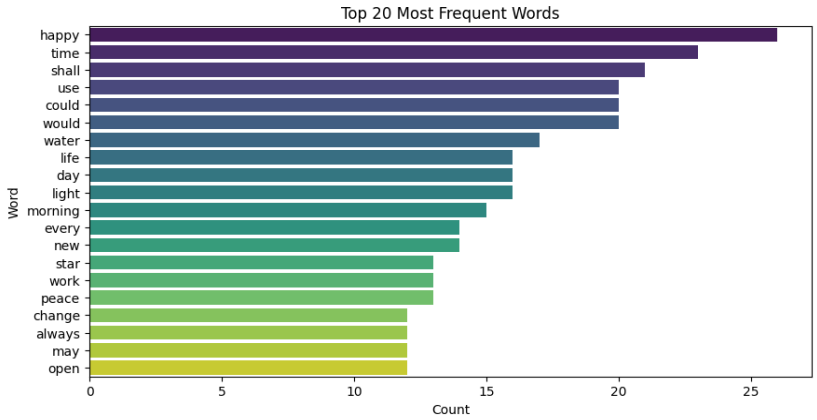


Figure 5: Bar Chart of the Top 20 Most Frequent Words in the pre-processed Dataset.

Figure 5 illustrates the bar chart representation of the top 20 most frequent words in the speech act dataset. It depicts the frequency distribution of commonly occurring terms, highlighting their relative

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

importance within the text corpus. The figure provides a clear comparison of word usage based on count, enabling better understanding of dominant linguistic features. These frequent words reflect recurring patterns and contribute significantly to feature representation during model training. The visualization supports the identification of meaningful textual attributes that influence classification performance.

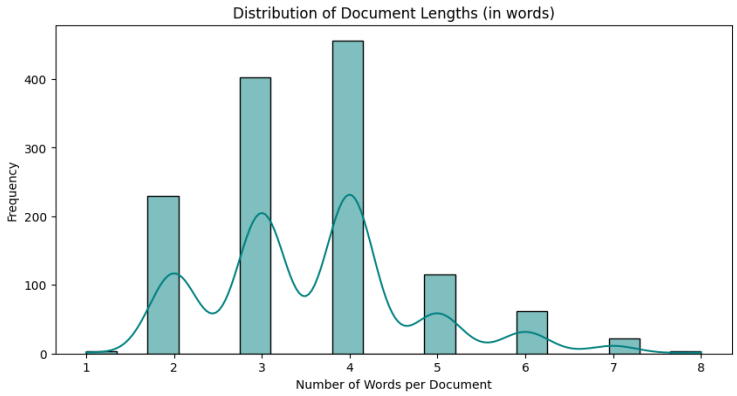


Figure 6: Distribution of Document Lengths (in Words) in the dataset.

Figure 6 illustrates the distribution of document lengths in the speech act dataset, measured in terms of the number of words per text instance. It depicts how the dataset is composed of varying text lengths, with most samples concentrated within a specific range. The figure provides insight into the frequency of short and long textual inputs, highlighting the overall structure of the dataset. This analysis helps in understanding the nature of input data and its impact on feature extraction and model performance. It also supports decisions related to sequence handling in transformer-based models like XLNet.

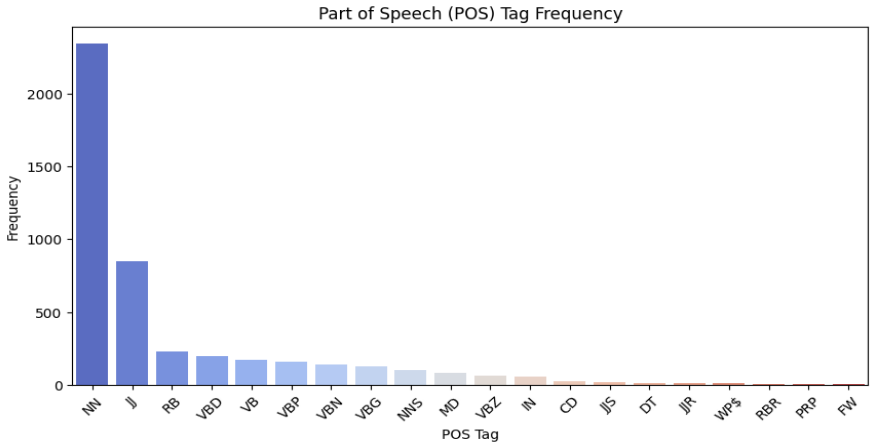


Figure 7: Frequency Distribution of Part-of-Speech (POS) Tags in the dataset.

Figure 7 illustrates the frequency distribution of Part-of-Speech (POS) tags in the speech act dataset. It depicts the occurrence of different grammatical categories such as nouns, verbs, adjectives, and others within the textual data. The figure highlights that certain POS tags appear more frequently, indicating their dominance in sentence construction. This analysis provides insight into the linguistic structure and syntactic patterns present in the dataset. It helps in understanding how different word types contribute to speech act classification.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

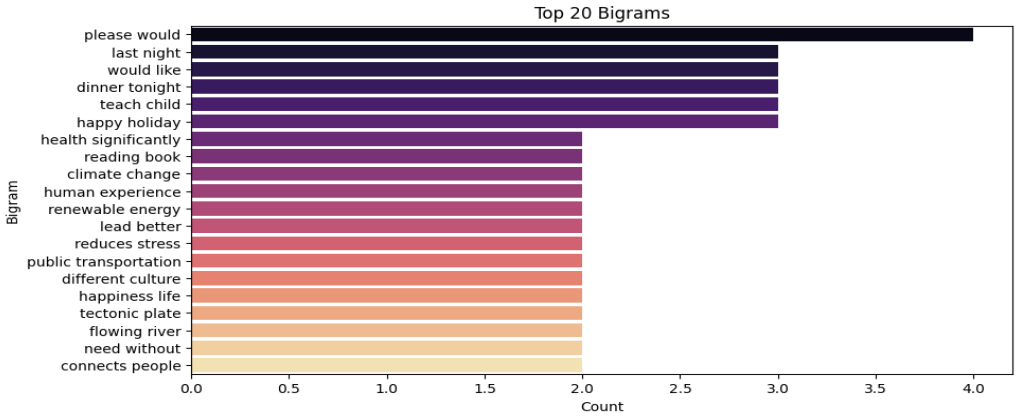


Figure 8: Bar Chart of the Top 20 Most Frequent Bigrams in the Pre-processed dataset

Figure 8 illustrates the bar chart representation of the top 20 most frequent bigrams in the speech act dataset. It depicts commonly occurring word pairs that capture contextual relationships between consecutive terms. The figure highlights meaningful phrase-level patterns that go beyond individual word analysis. These bigrams provide deeper insights into sentence structure and semantic associations within the dataset. Such patterns play an important role in improving feature representation for classification models.

Figure 9 illustrates the class distribution of the speech act dataset after applying SMOTE-based balancing. It depicts how all classes are uniformly represented with nearly equal sample counts, eliminating the imbalance observed in the original dataset. The figure highlights the effectiveness of SMOTE in generating synthetic samples for minority classes. This balanced distribution improves the learning capability of models such as LR, RF, SVM, and BFM. It ensures that the models do not become biased toward any class.

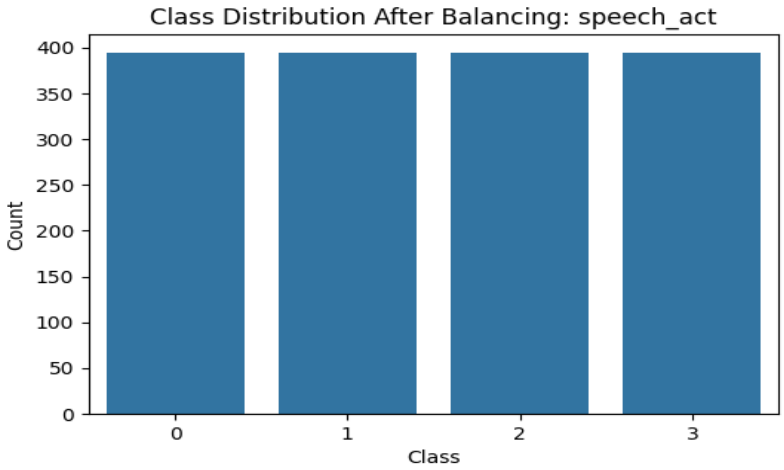


Figure 9: Class distribution after balancing for speech acts.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

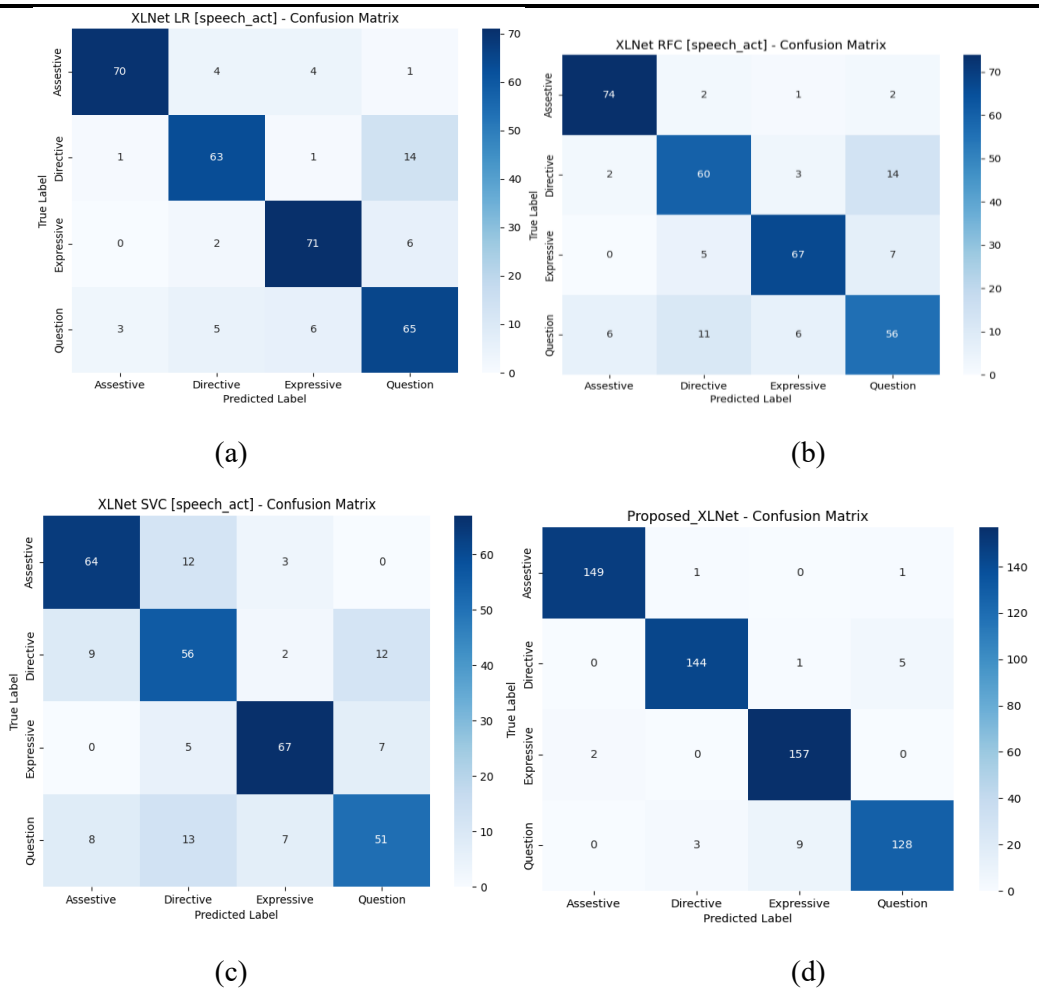


Figure 10: Confusion matrix obtained using XLNet (a) LR model. (b) RFC model. (c) SVM model. (d)BFM.

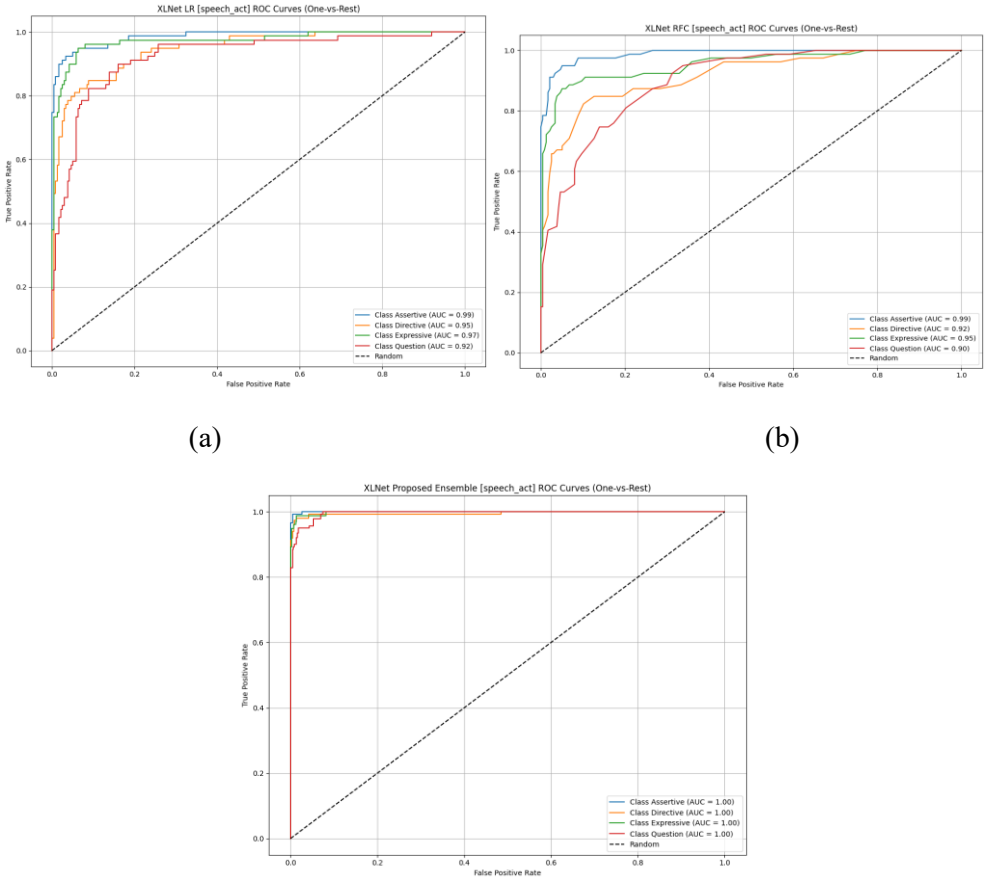
Figure 10 (a) illustrates the confusion matrix of the LR model using XLNet features for speech act classification. It depicts the relationship between actual and predicted class labels across all categories. The figure highlights that most predictions are correctly classified along the diagonal, indicating good model performance. Some misclassifications are observed between closely related classes, reflecting overlapping linguistic patterns.

Figure 10 (b) illustrates the confusion matrix of the RF model based on XLNet feature extraction. It depicts improved classification performance with a higher number of correct predictions compared to baseline models. The figure shows that RF effectively captures complex patterns within the dataset. However, minor misclassifications still occur in certain classes due to feature similarities. This representation highlights the robustness of RF in speech act classification.

Figure 10 (c) illustrates the confusion matrix of the SVM model using XLNet embeddings. It depicts the classification performance across different speech act categories with moderate accuracy. The figure shows that SVM performs well for some classes but struggles with others, resulting in noticeable misclassifications. These errors are mainly due to overlapping feature distributions in high-dimensional space. The visualization provides insight into the limitations of SVM for this task.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Figure 10 (d) illustrates the confusion matrix of the proposed BFM model combining LGBM and CB with XLNet features. It depicts a significantly higher number of correct predictions across all classes, as seen along the diagonal. The figure highlights minimal misclassification compared to other models, demonstrating superior performance. The ensemble approach effectively captures both linear and complex patterns in the data.



(a) (b)

(c)

Figure 11: ROC Curve obtained using XLNet (a) LR model. (b) RFC model. (c) BFM.

Figure 11 (a) illustrates the ROC curves of the LR model using XLNet features for speech act classification. It depicts the trade-off between true positive rate and false positive rate for each class in a one-vs-rest setting. The curves show strong performance with high AUC values across most classes, indicating effective discrimination capability. Slight variations among classes reflect differences in classification difficulty. Overall, the figure demonstrates that LR performs well when combined with transformer-based features.

Figure 11 (b) illustrates the ROC curves of the RF model based on XLNet feature extraction. It depicts improved classification capability with consistently high true positive rates across classes. The curves indicate strong model performance with competitive AUC values, particularly for dominant classes. Some variations in curve shapes suggest differences in class separability. The visualization confirms that RF effectively captures complex patterns in the dataset.

Figure 11 (c) illustrates the ROC curves of the proposed BFM model integrating LGBM and CB with XLNet features. It depicts near-perfect classification performance with AUC values approaching 1.0

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

for all classes. The curves are closely aligned with the top-left corner, indicating excellent true positive rates and minimal false positives. This highlights the superior discriminative power of the ensemble model.

The comparative analysis evaluates the performance of different machine learning models used in this research, including LR, RF, SVM, and the proposed BFM integrating LGBM and CB with XLNet-based feature extraction. It focuses on comparing models using standard evaluation metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in speech act classification. The analysis highlights how each model performs under the same feature conditions, ensuring a fair comparison. It also examines strengths and limitations of individual models in handling complex and imbalanced data. Visualization techniques such as confusion matrices and ROC curves further support performance comparison. The results demonstrate that ensemble-based approaches provide improved classification consistency and accuracy.

Table 1: Performance evaluation obtained using XLNet with (a) SVM model. (b) RFC model. (c) LR model. (d) BFM.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	75.31	75.44	75.31	75.28
RF	81.32	81.26	81.32	81.28
LR	85.12	85.47	85.12	85.21
BFM	96.33	96.37	96.21	96.26

Table 1 has contained different models clearly shows variations in classification effectiveness across LR, RF, SVM, and BFM. The SVM model achieved an accuracy of 75.31%, indicating moderate performance in handling speech act classification. The RF model improved the results with an accuracy of 81.32%, demonstrating better capability in capturing complex patterns. The LR model further enhanced performance, achieving an accuracy of 85.12%, showing its effectiveness with XLNet features. The proposed BFM model significantly outperformed all other models with an accuracy of 96.33%, highlighting the strength of combining LGBM and CB. Similar trends are observed in precision, recall, and F1-score, confirming consistent performance improvements.

	text	Predicted_Output
0	The sky is blue and the sun is shining brightl...	Assertive
1	Scientists have confirmed that water boils at ...	Assertive
2	I believe that honesty is the best policy in a...	Assertive
3	The capital city of France is Paris, located i...	Assertive
4	Exercise improves both physical and mental hea...	Assertive
5	Reading books expands your knowledge and vocab...	Assertive
6	Climate change is affecting weather patterns a...	Assertive
7	Why did the chicken cross the road safely?	Question
8	Who won the championship game last night?	Question
9	Which book should I read first in series?	Question
10	When will the train arrive at station?	Question
11	Whose turn is it to take out trash?	Question
12	Can you pass the salt please quickly?	Question
13	Could you explain the instructions once more?	Question
14	Would you like some tea with sugar?	Question
15	Go to bed at nine sharp.	Question
16	Wake up early for school tomorrow.	Directive

Figure 12: Sample predictions on new test data.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Figure 12 illustrates the prediction output of the system on test data using the trained classification model. It depicts how input textual sentences are processed and assigned corresponding speech act labels such as Assertive, Question, and Directive. The figure highlights the effectiveness of the trained models, particularly the proposed BFM, in accurately classifying different types of text. It represents the final stage of the pipeline where preprocessing, XLNet feature extraction, and model inference are combined to generate results. The output demonstrates the system's capability to handle real-time input and produce meaningful predictions.

5. CONCLUSION

The research presents an effective approach for speech act classification by integrating XLNet-based feature extraction with multiple machine learning models such as LR, RF, SVM, and the proposed BFM integrating LGBM and CB. The system successfully processes unstructured textual data through NLP preprocessing and transforms it into meaningful feature representations. Experimental results demonstrate that transformer-based embeddings significantly enhance classification performance compared to traditional methods. Among all models, the proposed BFM achieved the highest accuracy of 96.33%, showing superior capability in capturing complex linguistic patterns. The use of SMOTE further improved model generalization by addressing class imbalance issues. Visualization techniques such as confusion matrices and ROC curves validated the robustness of the models. The system also ensures secure user interaction through Redis-based authentication and provides a user-friendly GUI for seamless operation. Performance improvements are evident in terms of higher accuracy, better precision-recall balance, and reduced misclassification. The modular design of the system enhances scalability and maintainability for future extensions. The research demonstrates a reliable and efficient solution for automated speech act classification in real-world scenarios.

REFERENCES

[1] J. R. Searle, "Speech Acts: An Essay in the Philosophy of Language," Cambridge University Press, 1969.
[2] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 3rd ed., Pearson, 2020.
[3] C. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.
[4] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, 1988.
[5] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.
[6] T. G. Dietterich, "Ensemble Methods in Machine Learning," Multiple Classifier Systems, 2000.
[7] Z. Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding," NeurIPS, 2019.
[8] Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora. Applied Sciences. 2025; 15(4):1848. https://doi.org/10.3390/app15041848
[9] Park Y, Kang S, Seo J. An Efficient Framework for Development of Task-Oriented Dialog Systems in a Smart Home Environment. Sensors. 2018; 18(5):1581. https://doi.org/10.3390/s18051581
[10] Lieskovská E, Jakubec M, Jarina R, Chmulík M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics*. 2021; 10(10):1163. https://doi.org/10.3390/electronics10101163

[11] Pallewela N, Alahakoon D, Adikari A, Pierce JE, Rose ML. Optimizing Speech Emotion Recognition with Machine Learning Based Advanced Audio Cue Analysis. *Technologies*. 2024; 12(7):111. https://doi.org/10.3390/technologies12070111

[12] Augustyniak Ł, Szymański P, Kajdanowicz T, Tuligłowicz W. Comprehensive Study on Lexicon-based Ensemble Classification Sentiment Analysis. *Entropy*. 2016; 18(1):4. https://doi.org/10.3390/e18010004