INTRO TO NLP - CS7.401

NATURAL LANGUAGE INFERENCE

Team: Two Six

Submitted By :
Harshita Harshita (2023201002)
Shreyansh Shrivastava (2023201006)

Course Instructor :
Prof. Manish Shrivastava
Prof. Rahul Mishra

ABSTRACT

This presentation outlines our project on Natural Language Inference (NLI), focusing on the crucial task of textual entailment recognition. NLI plays a pivotal role in understanding and processing human language, enabling machines to determine if the meaning of one text fragment can be inferred from another. Our project aims to explore innovative approaches to improve the accuracy and efficiency of textual entailment recognition, leveraging state-of-the-art machine learning techniques and comprehensive datasets such as MultiNLI, SICK, and SNLI.

DATASET OVERVIEW

1) **SNLI (Stanford Natural Language Inference):**
   Size: 570,152 sentence pairs.
   Usage: Benchmarking NLP models in textual entailment and inference tasks.
2) **MNLI (Multi-Genre Natural Language Inference):**
   Size: 393,695 sentence pairs.
   About: Same as SNLI but covers diverse text genres.
   Usage: Evaluating models' ability to generalize across different textual domains.
3) **SICK (Sentences Involving Compositional Knowledge):**
   Size: 9,927 sentence pairs.
   Task: Assess relatedness and entailment, focusing on compositional meaning.
   Usage: Evaluating model’s understanding of sentence semantics and compositional meaning.

APPROACHES

- Logistic Regression
- BiLSTM’s
- Transformers - BERT , RoBERTa, ALBERT
- Max Epoch Selection
- Ensemble with Max Voting
- Epoch Max and Ensemble Aggregation
- RoBERTa with PEFT (LoRA)

LOGISTIC REGRESSION

Overview: Logistic regression is a popular statistical method used for binary and categorical outcome prediction. In the field of Natural Language Processing (NLP), it's utilized to handle Natural Language Inference (NLI) tasks, which determine if a premise text logically follows, contradicts, or is neutral to a hypothesis text.

Feature Representation: We benchmark our logistic regression model against traditional NLP baseline TF-IDF and word embeddings. These features are fundamental in transforming textual data into a numerical format that logistic regression can process, emphasizing the importance of word frequency and distribution across documents.

Model Training and Evaluation

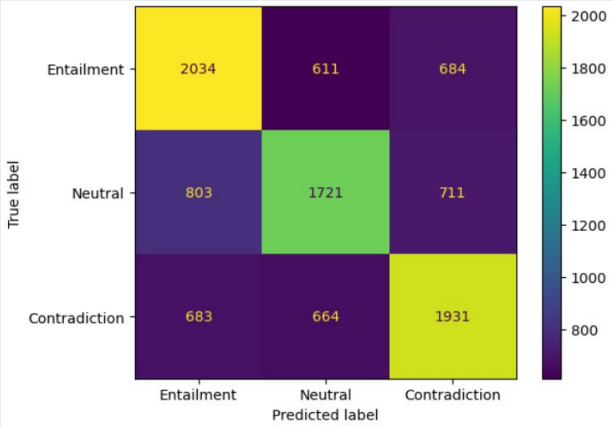
- **Data Preprocessing**: The initial phase involves meticulous cleaning of the dataset, removal of unlabeled entries, and transformation of text into vectors using TF-IDF, which emphasizes the significance of words based on their frequency and distribution across documents.

- **Model Training Strategy**: Logistic regression is applied post-vectorization to predict the probability of textual relationships. We set a high iteration count (max_iter=10,000) to ensure convergence and minimize prediction errors during model training.

Model Training and Evaluation

Model Performance: Stanford NLI (SNLI) Dataset

- Overall Accuracy: 57.77%
- Entailment: Precision 0.58, Recall 0.61, F1-Score 0.59
- Neutral: Precision 0.57, Recall 0.53, F1-Score 0.55
- Contradiction: Precision 0.58, Recall 0.59, F1-Score 0.58



Confusion Matrix Analysis:

- True Positives for each class: Entailment 2034, Neutral 1721, Contradiction 1931
- Notable Misclassifications highlight significant overlap in textual cues across categories.
- **Insights**: The results demonstrate moderate effectiveness with a balanced performance across classes, though there is a noticeable challenge in distinguishing overlapping linguistic features.

Model Training and Evaluation

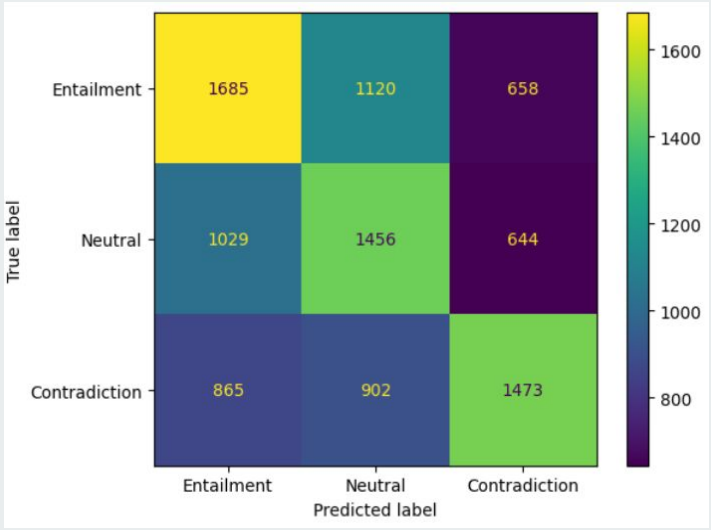
Model Performance: Multi-Genre Natural Language Inference (MultiNLI) Dataset

- **Performance:** Accuracy at 46.93%, with lower precision and recall suggesting difficulty in consistent label identification.
- **Misclassification Details:** High confusion particularly between "Entailment" and "Neutral" categories.

Model Performance: The Sentences Involving Compositional Knowledge (SICK)Dataset

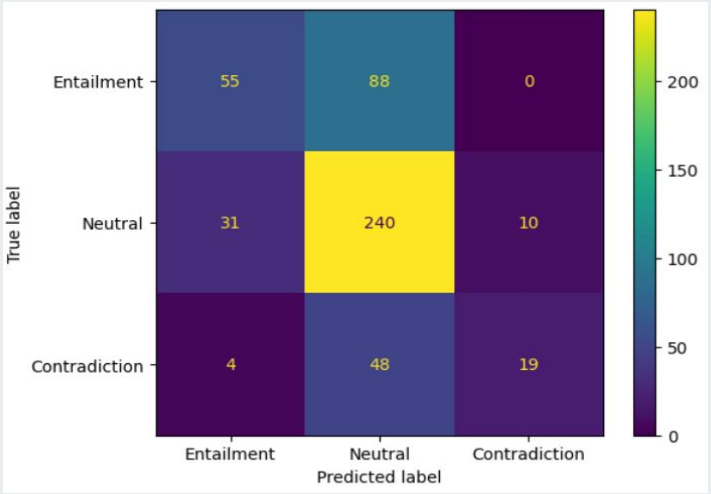
- **Performance:** Higher accuracy at 63.43%, but precision, recall, and F1-scores show significant issues in distinguishing "Entailment" and "Contradiction."
- **Misclassifications:** Particularly high misclassification rates from "Entailment" to "Neutral", and vice versa.

Model Training and Evaluation



MNLI Logistic Regression

SICK Logistic Regression



Challenges and Recommendations

Challenges:

- The model struggles with significant overlap in linguistic cues across different labels, leading to high rates of misclassification.
- Inconsistent performance across different datasets underscores the need for enhanced feature engineering.

Recommendations:

- **Feature Engineering:** Incorporate more sophisticated methods such as deep learning-based word embeddings to better capture contextual nuances.
- **Model Complexity:** More advanced machine learning algorithms or ensemble methods that might provide a more robust framework for handling complex NLI tasks.

Conclusion: Emphasize the potential of logistic regression as a baseline and its enhancement with advanced techniques to improve its efficacy and reliability in real-world NLP applications.

Bi LSTMs

The BiLSTM Model uses:

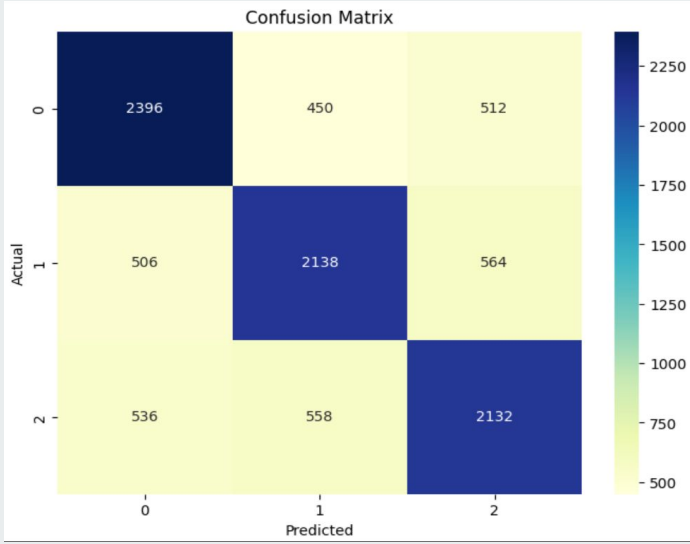
- Embedding Layer: Converts words into vectors.
- BiLSTM: Captures contextual information from both text directions.
- Concatenation: Merges hidden states for a full contextual view.
- Fully Connected Layer: Classifies the relationship between text pairs.
- Optimizer: Adam()
- Loss Function: CrossEntropy()

```
BiLSTMModel(
  (embedding): Embedding(2170, 100)
  (lstm): LSTM(100, 128, batch_first=True, bidirectional=True)
  (fc): Linear(in_features=512, out_features=3, bias=True)
)
```

Model Training and Evaluation

Overview:(SNLI)

- Accuracy: 68.08%
- Precision: 68%
- Recall: 68.08%
- F1 Score: 68.06%



Issues:

Misclassifications: Neutral vs. Contradiction

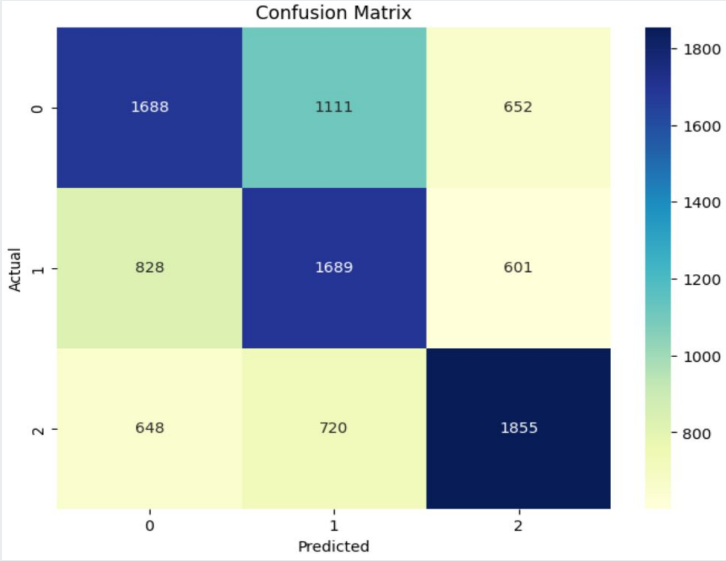
Recommendations:

Enhancements: Attention mechanisms, Data augmentation

Conclusion:

Potential Improvement: Enhanced linguistic distinction capabilities

Model Training and Evaluation



Overview:(MNLI)

- Accuracy: 53.43%
- Precision: 54%
- Recall: 54%
- F1 Score: 54%

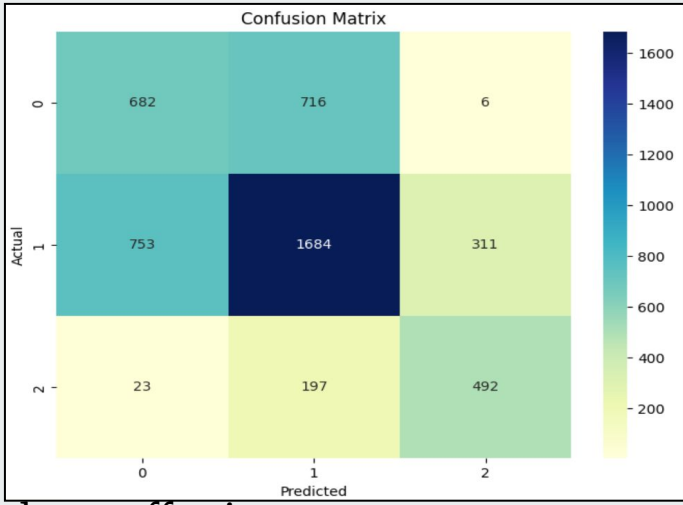
Performance:

- Entailment: Lower effectiveness with 49% recall.
- Neutral: Moderate recall at 54%, lowest precision.
- Contradiction: Best performance with 58% recall and 60% precision.

Conclusion:

- Improvement Needed: Enhance precision for neutrals and recall for entailments. Consider advanced modeling techniques to boost overall performance.

Model Training and Evaluation



Overview:(SICK)

- Accuracy: 58.76%
- Precision: 57%
- Recall: 60%
- F1 Score: 58.83%

Performance:

- Entailment: Least effective, with only 49% recall.
- Neutral: Highest precision and good recall, indicating moderate effectiveness.
- Contradiction: Best recall at 69%, showing strongest model response.

Conclusion:

- Focus Areas: Enhance entailment detection and refine overall model features to improve precision and recall.

Transformers

- BERT, RoBERTa, and ALBERTa models are trained and evaluated individually.
- After individual evaluation, predictions from all models are combined using a majority voting ensemble method.
- **Epoch Max:** Combine predictions by selecting the most commonly predicted class across different epochs
- **Majority Voting Ensemble:** Combine predictions from multiple models by selecting the most commonly predicted class.
- **Ensemble + Epoch Max:** Combine predictions by selecting the most commonly predicted class across different epochs across all models and selecting the class with the highest average probability.
- These methods leverage the strengths of individual models to improve overall prediction accuracy and robustness.

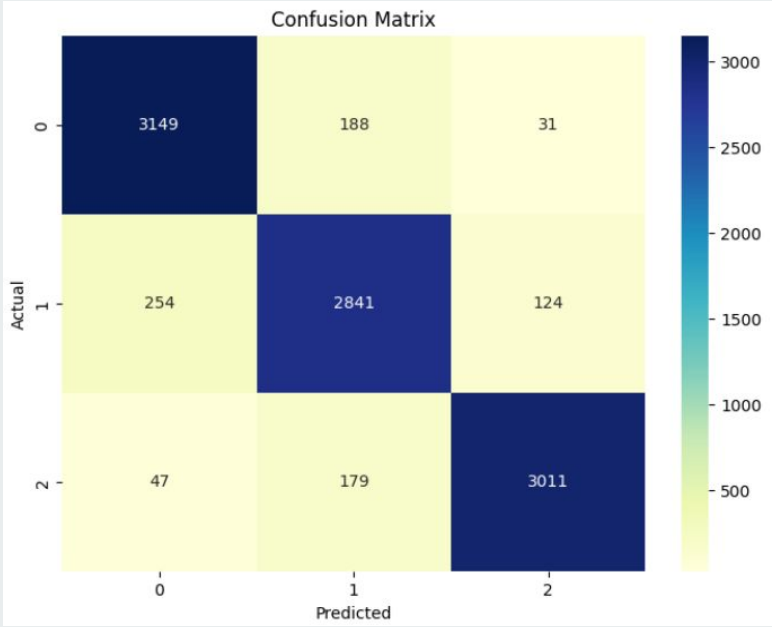
Model Training and Evaluation

Overview:(SNLI) - ALBERT + RoBERTa Ensemble

- Accuracy: 91.62%
- Recall: 91.62%
- F1 Score: 91.63%

Performance:

- Entailment: 3149 out of 3368 correct.
- Neutral: 2841 out of 3219 correct.
- Contradiction: 3011 out of 3237 correct.



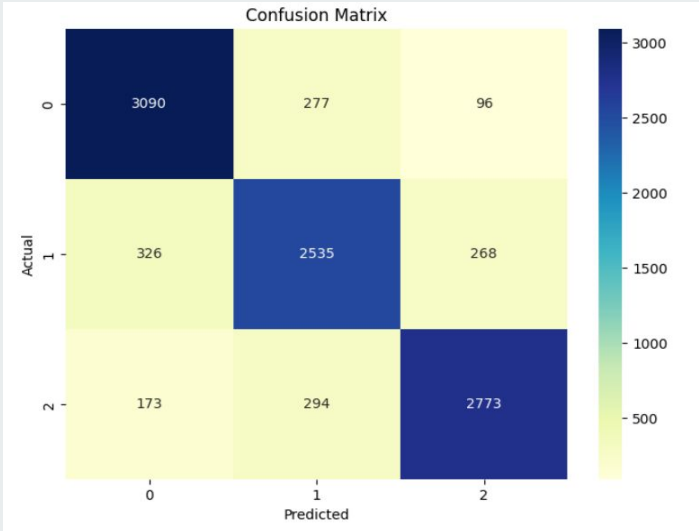
Conclusion:

- The BERT and ALBERT ensemble demonstrates high efficacy in NLI, excelling across all categories.

Model Training and Evaluation

Overview:(MNLI) - ALBERT + RoBERTa Ensemble

- Accuracy: 85.41%
- Recall: 85.41%
- F1 Score: 85.40%



Performance:

- Strong in identifying entailment and contradiction.
- High overall performance with balanced precision and recall across all classes.

Model Training and Evaluation

Overview:(SICK) - ALBERT + RoBERTa Ensemble

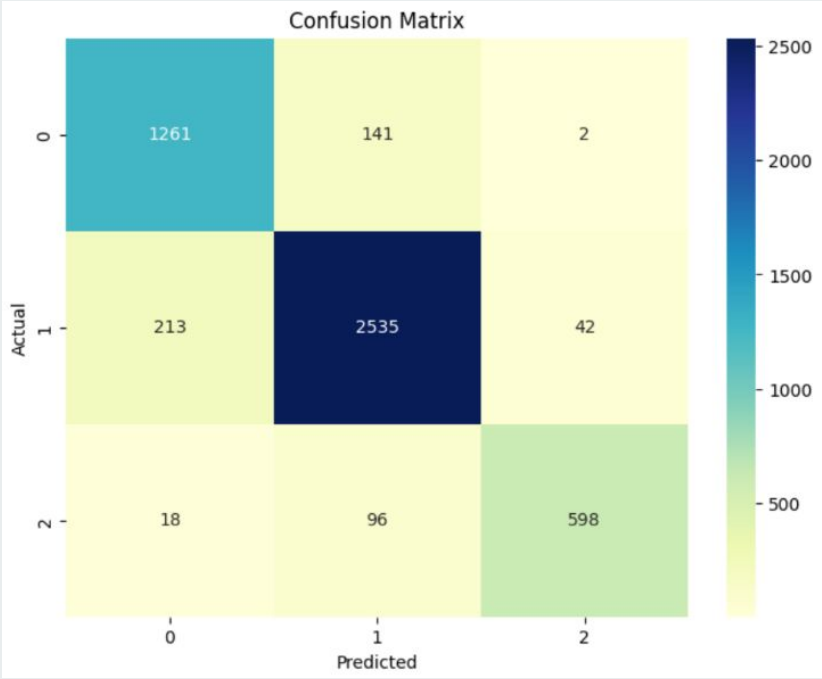
- Accuracy: 89.56%
- Recall: 89.56%
- F1 Score: 89.28%

Performance:

Entailment: Strong precision with minimal false positives.

Neutral: High accuracy, few misclassifications.

Contradiction: Excellent in distinguishing contradictions.



PEFT

Integrated Parameter Efficient Fine Tuning (PEFT) with LoRA (Low-Rank Adaptation) for the task of sequence classification within the domain of natural language inference (NLI).

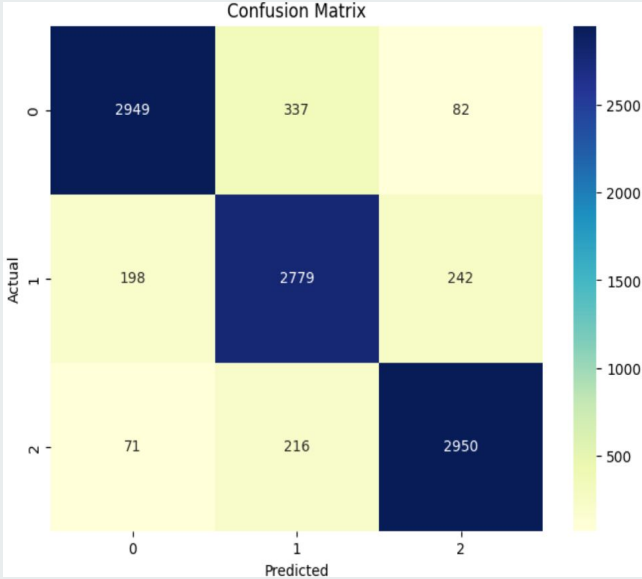
Configuration and Model Initialization:

- **Tokenization**: Utilizes the RobertaTokenizer to convert raw text data into a format that is suitable for the Roberta model, ensuring text strings are correctly split into tokens and encoded to numerical IDs.
- A Roberta For Sequence Classification is initialized for a three-class problem, representing typical NLI labels: entailment, contradiction, and neutral.
- **PEFT with LoRA Configuration**: Adapts the standard Roberta model by inserting trainable low-rank matrices into specific sub-modules (query, key, value) of the transformer layers. This approach aims to enhance the model's adaptability and performance on NLI tasks without drastically increasing the number of parameters.

Model Training and Evaluation

PEFT with LoRA Model on SNLI:

- Accuracy: 88.33%
- Precision: 88%
- Recall: 88%
- F1 Score: 88%



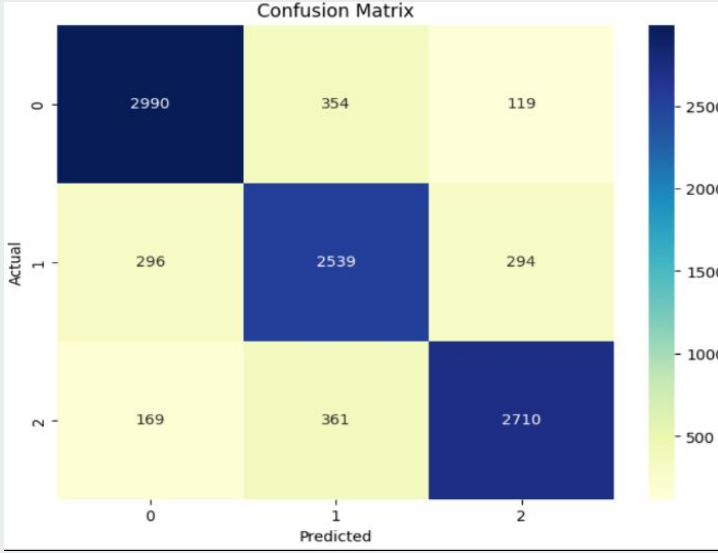
Conclusion:

- Strong performance in all categories, particularly in contradiction detection.
- Model demonstrates high precision in entailment and excellent recall in contradiction, underscoring its efficiency in discerning complex sentence relationships effectively.

Model Training and Evaluation

PEFT with LoRA Model on MNLI:

- Accuracy: 83.80%
- Recall: 83.80%
- F1 Score: 83.83%



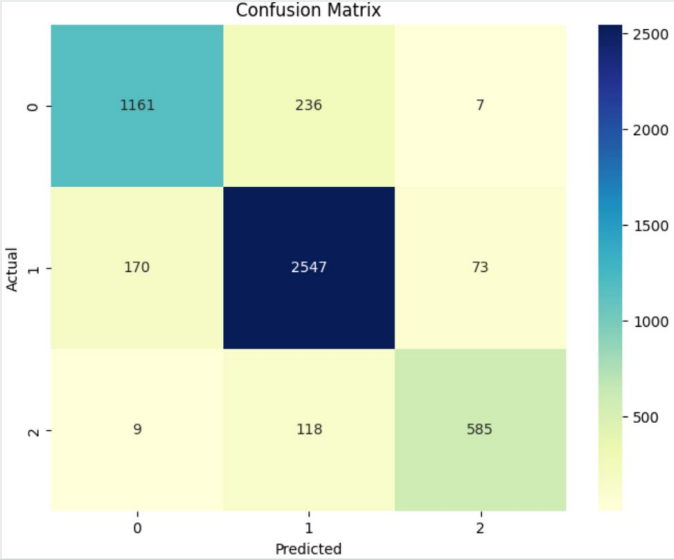
Highlights:

- Strong in contradiction and entailment detection.
- Consistently high performance across all categories, with slight room for improvement in neutral classification accuracy.

Model Training and Evaluation

PEFT with LoRA Model on SICK:

- Accuracy: 87.51%
- Precision: 87%
- Recall: 85%
- F1 Score: 86%



Conclusion:

- Good in identifying neutral statements with high recall.
- Strong overall classification capability with balanced precision and recall across categories.

Overall Results

MODELS	SNLI	MULTI - NLI	SICK
Logistic Regression	57.77%	46.93%	63.43%
BiLSTM's	68.08%	53.43%	58.76%
BERT	89.31%	78.80%	84.69%
RoBERTa	90.50%	84.29%	88.44%
ALBERT	89.64%	81.20%	86.91%
Max Epoch Selection - BERT	90.28%	80.64%	84.12%
Max Epoch Selection - RoBERTa	91.36%	85.25%	89.36%
Max Epoch Selection - ALBERT	90.36%	82.49%	88.57%
RoBERTa with PEFT (LoRA)	88.33%	83.80%	87.51%

Ensemble Results

MODELS	SNLI	MULTI - NLI	SICK
EWMV* - Bert + Roberta + Alberta	91.28%	84.87%	88.44%
EWMV* - Bert + Alberta	91.36%	82.82%	85.59%
EWMV* - Bert + Roberta	91.04%	84.07%	86.14%
EWMV* - Roberta + Alberta	90.97%	84.97%	89.30%
EMEA* - Bert + Roberta + Alberta	91.61%	85.20%	88.86%
EMEA*- Bert + Alberta	91.07%	83.50%	86.36%
EMEA* - Bert + Roberta	91.40%	85.02%	86.89%
EMEA* - Roberta + Alberta	**91.63%**	**85.41%**	**89.56%**

* EWMV - Ensemble with Max Voting
EMEA - Epoch Max and Ensemble Aggregation

Observation and Conclusion

Key Points:

Model Performance: Advanced transformer models and ensembles outperform traditional NLI approaches.

Best Models: RoBERTa and ALBERT consistently deliver superior results.

Optimal Strategies: Ensemble strategies like EWMV and EMEA optimize performance, with RoBERTa + ALBERT proving exceptionally effective.

Takeaway:

Leveraging combinations of advanced models through ensembles significantly boosts NLI performance, making them ideal for tackling complex linguistic tasks across diverse datasets

Thank You