PORT CITY INTERNATIONAL UNIVERSITY

Detecting and Explaining Bangla Idiom Misuse: A Comparative Study of Neural Approaches with Reason Generation

Supervised by:

Mrs. Farzina Akther

Assistant Professor

Department of CSE

Presented by:

Md. Ashiful Hoque Chowdhury

CSE 02707396

Batch 27A1

CONTENTS

INTRODUCTION

This study addresses contextual Bangla idiom misuse by introducing a manually annotated dataset with correctness labels and human-written explanations. It benchmarks ML, DL and transformer models and integrates an explainable framework using BanglaT5 to generate human-readable justifications. The work contributes new dataset, empirical baselines and interpretable models for advancing Bangla NLP and educational applications.

MOTIVATION

1. No automated idiom support for Bangla learners, especially in underserved contexts.
2. Weak idiom handling in existing Bangla NLP systems, harming downstream tasks.
3. Severe resource scarcity, with no large annotated idiom datasets or explanation-rich benchmarks.
4. Lack of explainability in black-box models, limiting educational trust and usability.

OBJECTIVES

1. Build a Bangla idiom dataset with correct/incorrect labels and short human explanations.
2. Construct idiom dictionary with meanings and idiom_id.
3. Generate explainable outputs in Bangla for better human understanding.
4. Evaluate performance against ML, DL baselines and transformer-based models.

LITERATURE REVIEW

TITLE	AUTHOR	YEAR AND PUBLISHER	FINDINGS
1. Implicit Knowledge-Augmented Prompting for Commonsense Explanation Generation PDF	Y. Ge, H. T. Yu, C. Lei, et al.	2025, Springer	**Aim:** To help AI models better explain why certain statements don't make sense. **Dataset Size:** ComVE + Commonsense Explanation dataset with ~12,000 sentences. **Model Performance:** Two-stage Identification and Prompting increases performance over LLMs baselines upto BLEU: ~631%, METEOR: ~192%, ROUGE: ~204%. **Limitation:** Focused only on explanation generation not classification and also doesn't handle idioms or non-literal language. **Language:** English **Align:** Explanation generation phase after classification.
2. A Hybrid Approach for Bengali Sentence Validation PDF	Juel Sikder, Prosenjit Chakraborty, Utpol Kanti Das, Krity Dhar	2024, Springer	**Aim:** Automated grammatical correctness checker for Bengali sentences **Dataset Size:** Bengali Sentence Validation dataset with ~5,000 sentences. **Model Performance:** Hybrid CNN + BiLSTM with POS tagging and linguistic rules. Accuracy: ~98%, F1-Score: ~0.98, AUC: ~0.99 **Limitation:** Struggles with idioms and nonstandard sentences. **Language:** Bangla. **Align:** POS and Classification.
3. BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla PDF	Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin	2022, NAACL	**Aim:** BERT-based Natural Language Understanding model pretrained in Bangla. Benchmarked NLU tasks: Text classification, sequence labeling and span prediction. **Dataset Size:** ~27.5 GB Bangla text from 110 websites. **Model Performance:** BanglaBERT - BERT model pretrained on Bangla text. **Limitation:** Limited to Bangla may not generalize well to other languages. **Language:** Bangla **Align:** POS tagging, idiom detection and general Bangla embeddings.

LITERATURE REVIEW

TITLE	AUTHOR	YEAR AND PUBLISHER	FINDINGS
4. SemEval-2020 Task 4: Commonsense Validation and Explanation (ComVE) PDF	C. Wang et al. (organizers)	2020, ACL	**Aim:** To test if AI can tell what makes sense in language and explain why it doesn't and mimicking human commonsense reasoning. **Dataset Size**: ComVE sentence pairs with ~12,000 instances. **Model Performance**: Subtask A: CN-HIT-IT.NLP – 97.0%, Subtask B: ECNU-SenseMaker – 95.0%, Subtask C: ANA – Human score 2.10/3, BUT-FIT – BLEU 22.4 **Limitation**: Explanation generation still hard. **Language**: English. **Align:** Commonsense validation and explanation.
5. An Empirical Framework of Idioms Translator From Bengali to English: Rule Based Approach PDF	Khatun / Hussain — metadata varies	2019, Conference 2020, ResearchGate	**Aim:** To develop a system for translating Bengali idiomic sentences into English using context-sensitive grammar and a top-down parsing algorithm. **Dataset Size**: Idiom-rich corpora with ~15,000 sentences. **Model Performance**: Rule-based idiom translation ≈ 85.33% accuracy. **Limitation**: Limited coverage and weak generalization. **Language**: Bangla → English. **Align:** Idiom detection and replace the sentence with literal meaning.
6. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa PDF	J. Briskilal, C. N. Subalalitha	2022, Elsevier	**Aim:** To classify idiomatic and literal sentences using an ensemble of pre-trained transformer models. **Dataset Size:** TroFi dataset with 3,737 sentences and 1,470 expert-annotated sentences. **Models Performance:** BERT achieved 85% accuracy, RoBERTa achieved 88% accuracy and the ensemble model achieved 90% accuracy. **Limitation:** Focuses only on binary idiom–literal classification **Language:** English **Align:** Idiom vs. literal sentence classification using transformer-based learning.

LITERATURE REVIEW

TITLE	AUTHOR	YEAR AND PUBLISHER	FINDINGS
7. Parts of Speech Tagging in Bengali for MWEs Detection PDF	Primary authors: Abedin, Purkayastha et al.	~2015 (IJCA Volume 99), International Journal of Computer Applications	**Aim:** Develop POS tagset and tagging approach for Bengali to detect Multiword. **Dataset Size:** Not explicitly large-scale (focus on tagset design + examples). **Model Performance:** Rule layered tagging approach effective for MWEs (Acc: 85%). **Limitation:** Rule-based and limited scalability. **Language:** Bangla. **Align:** Bangla-specific linguistic processing for MWEs/idioms.
8. TituLLMs: A Family of Bangla LLMs with Comprehensive Benchmarking PDF	Shahriar Kabir Nahin et al.	2025 (arXiv 2502.11187; ACL Findings 2025)	**Aim:** Introduce first large pretrained Bangla LLMs from Llama-3.2 base. **Dataset Size:** 5 new benchmarking datasets created. **Model Performance:** TituLLMs outperform base multilingual versions (Acc: 90%). **Limitation:** Not always superior to multilingual baselines on all tasks. **Language:** Bangla. **Align:** Bangla LLM pretraining + benchmarking.
9. CROW: Benchmarking Commonsense Reasoning in Real-World Tasks PDF	Mete Ismayilzada et al.	2023, EMNLP (arXiv 2310.15239)	**Aim:** Create multi-task benchmark evaluating commonsense. **Dataset Size:** Manually curated multi-task set (details in paper: several subtasks). **Model Performance:** State-of-the-art models tested and highlights gaps (Acc: 85%). **Limitation:** Focus on evaluation benchmark, not new model training. **Language:** English **Align:** Commonsense reasoning benchmarking in applied tasks.
10. "I'm Not Mad": Commonsense Implications of Negation and Contradiction PDF	Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, Yejin Choi	2021, NAACL	**Aim:** Introduce ANION dataset for commonsense implications under negation. **Dataset Size:** Curated examples of negated/contradictory statements. **Model Performance:** Models struggle with nuanced implications (Acc: 90%). **Limitation:** English-focused; negation handling remains hard for NLI models. **Language:** English. **Align:** Commonsense inference in contradictory/negated language.

METHODOLOGY

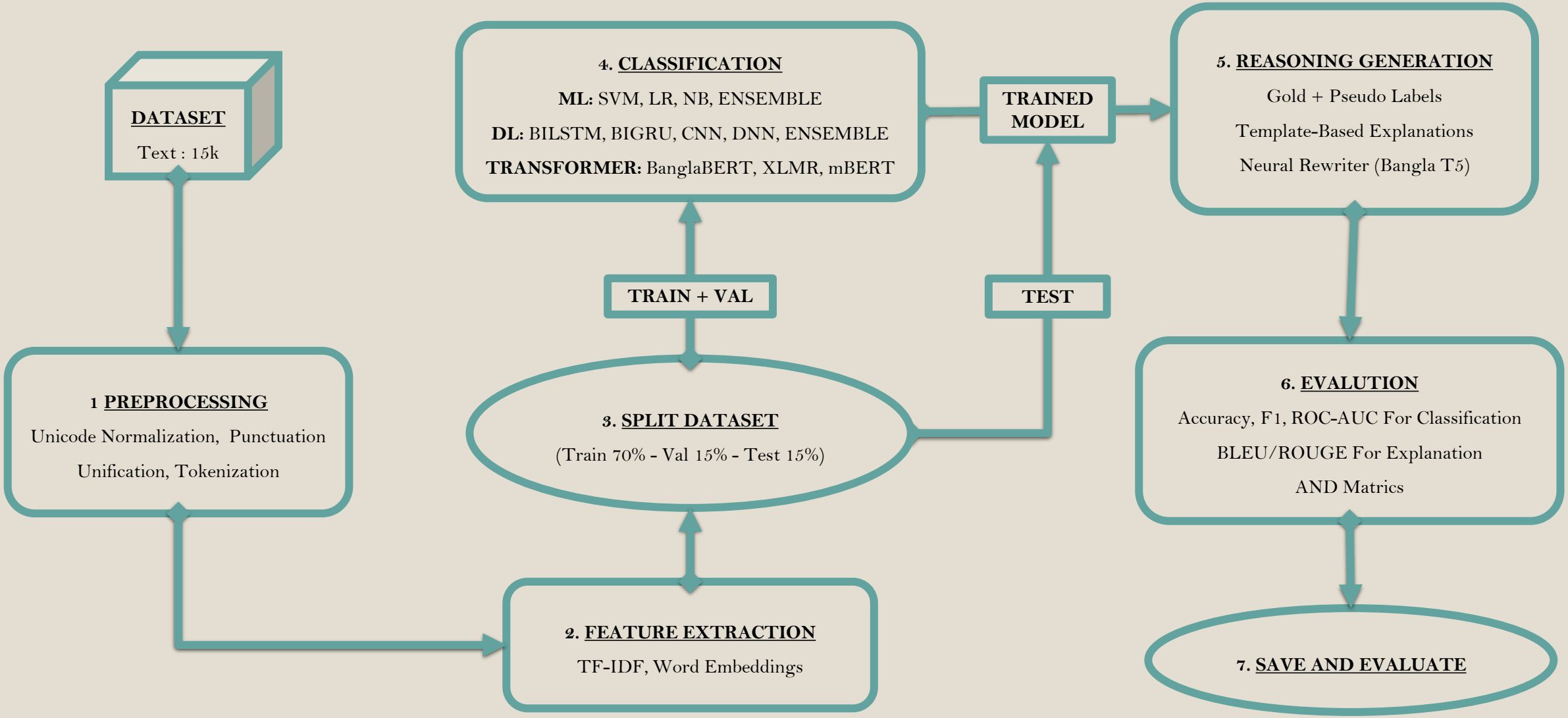

Figure 1: Methodology

DATASET

Table 1: Dataset Analysis Overview

Category	Metric	Value
Overall Statistics	**Total Instances**	**15,670**
	Total Idioms	**1,316**
	Avg. Sentences per Idiom	11.91
	Correct Usage (s_label=1)	7,836 (50.01%)
	Incorrect Usage (s_label=0)	7,834 (49.99%)
	Avg. Sentence Length	10.27 words
	Avg. Reason Length	18.5 words
Train Split	Instances	10,969 (70%)
	Correct Usage	~5,484
	Incorrect Usage	~5,485
Validation Split	Instances	2,350 (15%)
	Correct Usage	~1,175
	Incorrect Usage	~1,175
Test Split	Instances	2,351 (15%)
	Correct Usage	~1,177
	Incorrect Usage	~1,174
	Annotation Hours	~600 person-hours

DATASET

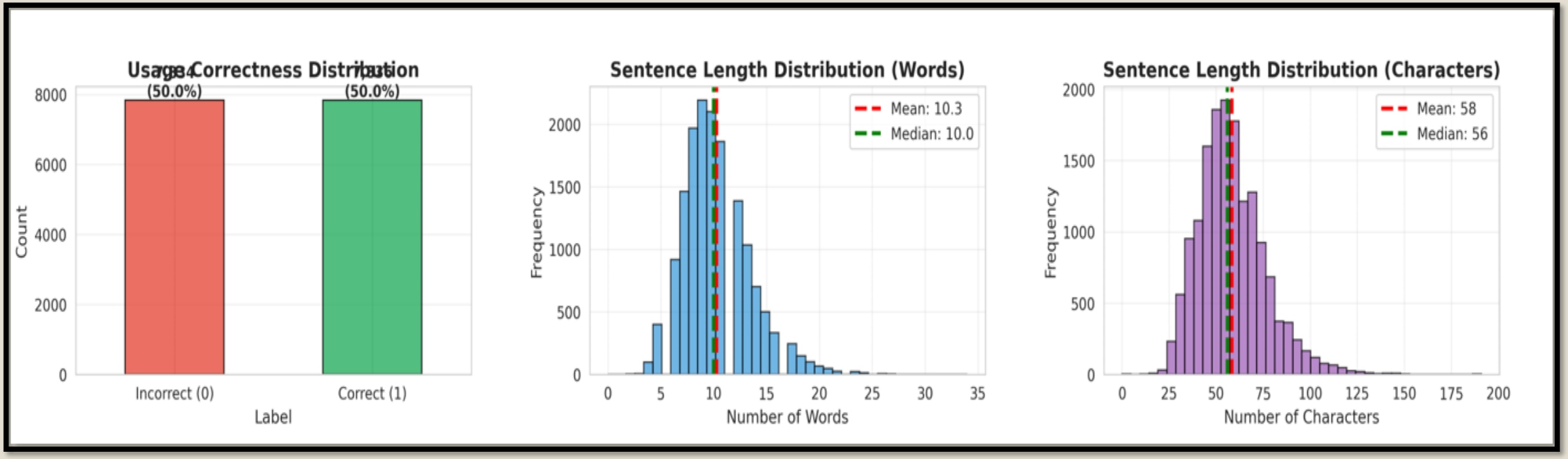

Figure 2: Dataset Analysis Plots

DATASET

Dictionary Structure:

1. {"idiom": "কথার ফুলঝুরি", "meaning": "অতিরিক্ত কথা বলা"}
2. {"idiom": "চোখের মণি", "meaning": "খুব প্রিয়"}

Dataset Structure:

1. {"idiom_word": "গভীর জলের মাছ", "idiom_meaning": "চালাক ব্যক্তি", "sentence": "গভীর জলের মাছ ধরার জন্য আমি বড়শি কিনেছি", "i_label": 1, "s_label": 0, "reason": "উক্ত 'গভীর জলের মাছ' বাগধারাটির অর্থ হলো 'চালাক ব্যক্তি', কিন্তু এটি 'আসল মাছ ধরা' বোঝাতে ব্যবহার করা হয়েছে, যা ভুল"}
2. {"idiom_word": "অগ্নিপরীক্ষা", "idiom_meaning": "কঠিন পরীক্ষা", "sentence": "দীর্ঘদিন রোগভোগের পর সুস্থ হয়ে ওঠাটা ছিল তার জন্য এক অগ্নিপরীক্ষা", "i_label": 1, "s_label": 1, "reason": "উক্ত 'অগ্নিপরীক্ষা' বাগধারাটির অর্থ হলো 'কঠিন পরীক্ষা' এবং এটি সঠিকভাবে ব্যবহৃত হয়েছে।"}

USED TOOLS & PLATFORMS

Development & Training

- Google Colab – Cloud GPU & CPU resources
- Python 3.10 – Data processing & modeling

Libraries & Frameworks

- PyTorch / TensorFlow – Deep Learning
- Scikit-learn – ML baselines
- BNLTK / Indic NLP – Bangla NLP preprocessing
- NumPy / Pandas – Data handling & computation

Data Sources & Support

- Bangla Academy Dictionaries & Online Resources

Collaboration & Storage

- Google Drive / GitHub – Dataset & model storage
- Overleaf / MS Word / LaTeX – Paper writing

PREPROCESSING

- Unicode NFC normalization.
- Removal of zero-width characters and whitespace.
- Removed duplicates and empty fields.

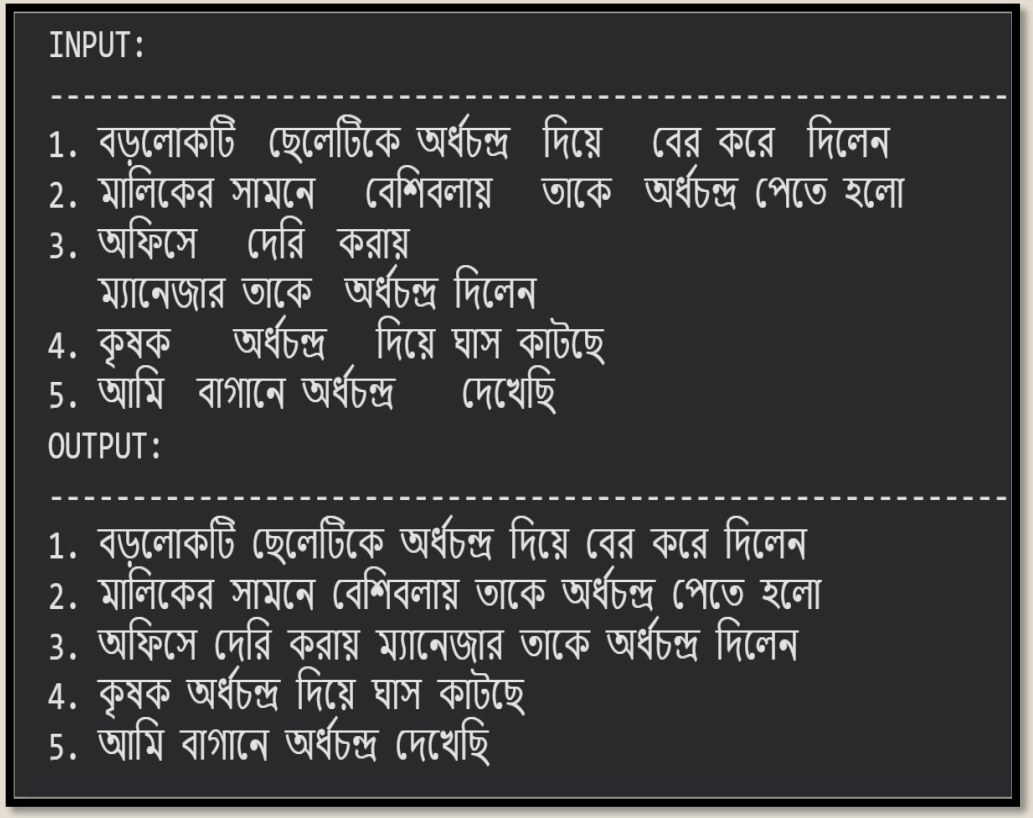

Figure 3: Preprocessing Result

FEATURE EXTRACTION

- ML uses TF-IDF to create sparse 5,800-dim vectors from character and word n-grams.
- DL models map tokenized sequences, max 180 tokens to dense 128-dim embedding.
- Transformers models generate contextual 768-dim embedding using.

```
Feature Matrix Shapes:
   Train: (10969, 5800) (1,068,101 non-zero values)
   Val:   (2350, 5800) (225,273 non-zero values)
   Test:  (2351, 5800) (226,741 non-zero values)

Sparsity:
   Train: 98.32%
   Val:   98.35%
   Test:  98.34%

Vectorizer Configurations:

   Sentence Vectorizer:
      - Analyzer: char_wb
      - N-gram range: (3, 6)
      - Max features: 5000
      - Vocabulary size: 5,000

   Idiom Meaning Vectorizer:
      - Analyzer: word
      - N-gram range: (1, 2)
      - Max features: 500
      - Vocabulary size: 500

   Idiom Word Vectorizer:
      - Analyzer: char_wb
      - N-gram range: (2, 4)
      - Max features: 300
      - Vocabulary size: 300
```

Figure 4: Feature Extraction Result

RESULT AND DISCUSSION

1. LR

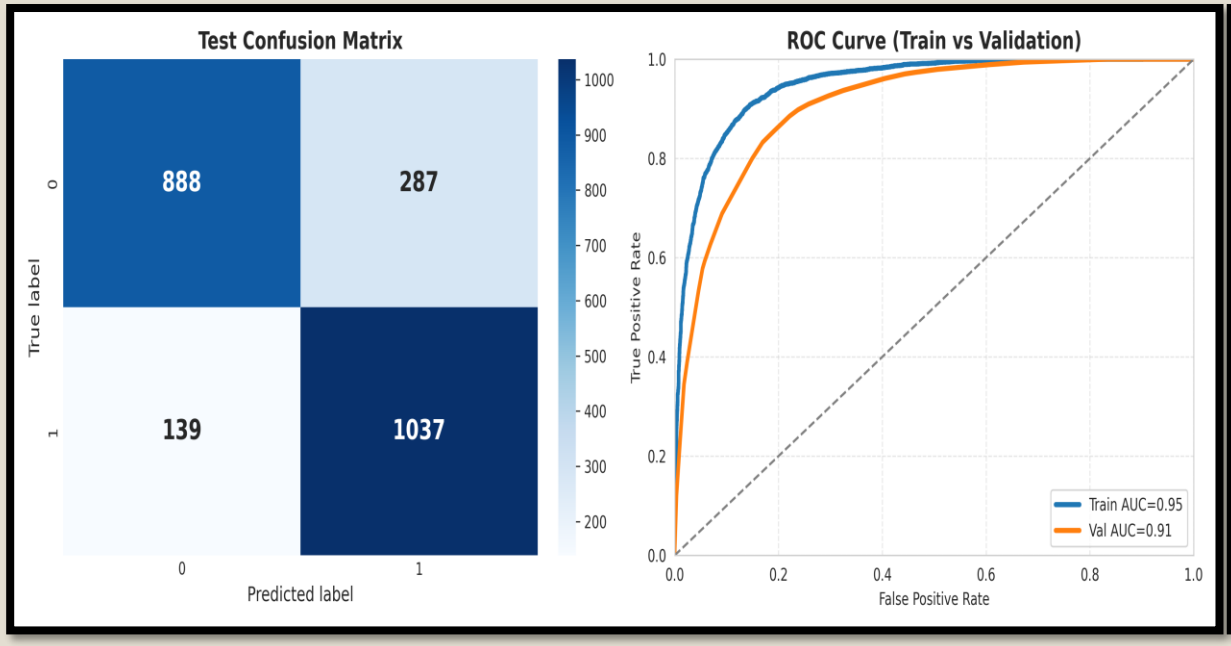

Figure 5: Confusion Matrices and ROC Curve of LR

2. SVM

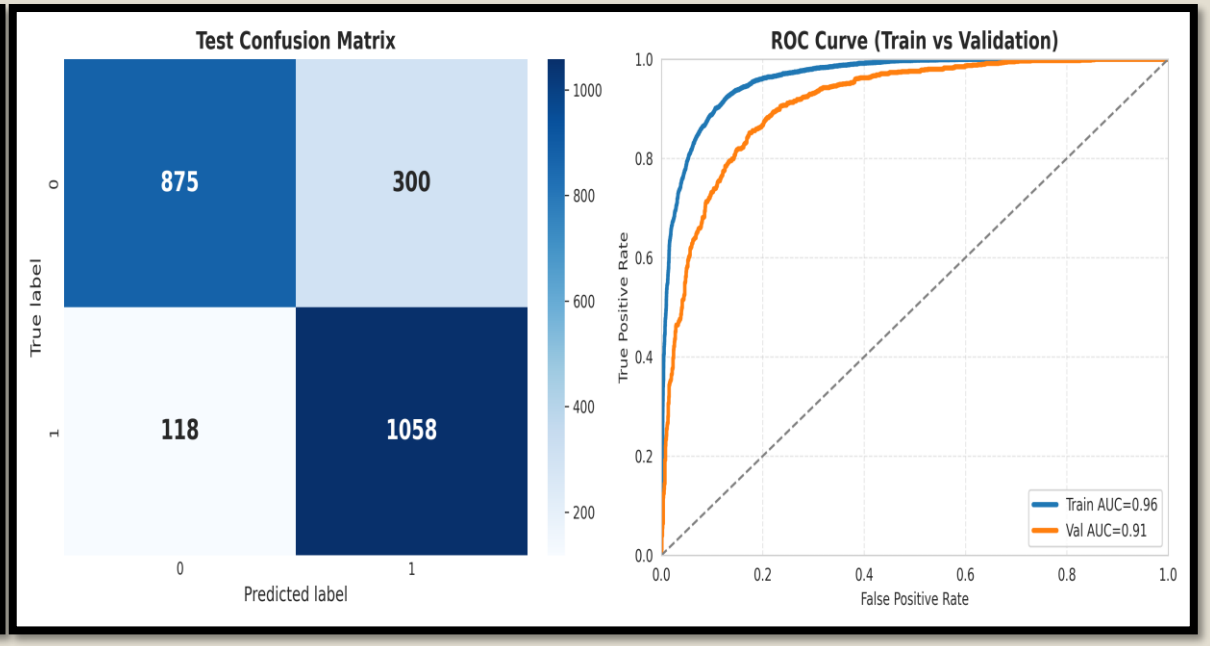

Figure 6: Confusion Matrices and ROC Curve of SVM

RESULT & DISCUSSION

3. Naive Bayes

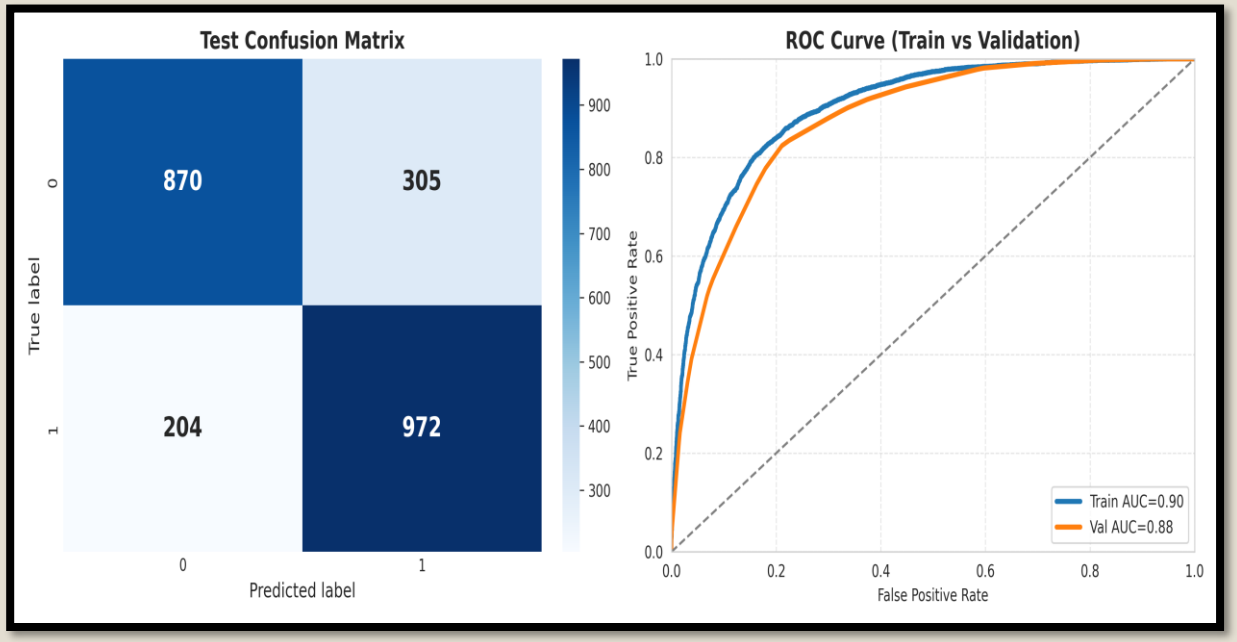

Figure 7: Confusion Matrices and ROC Curve of NB

4. ML ENSEMBLE

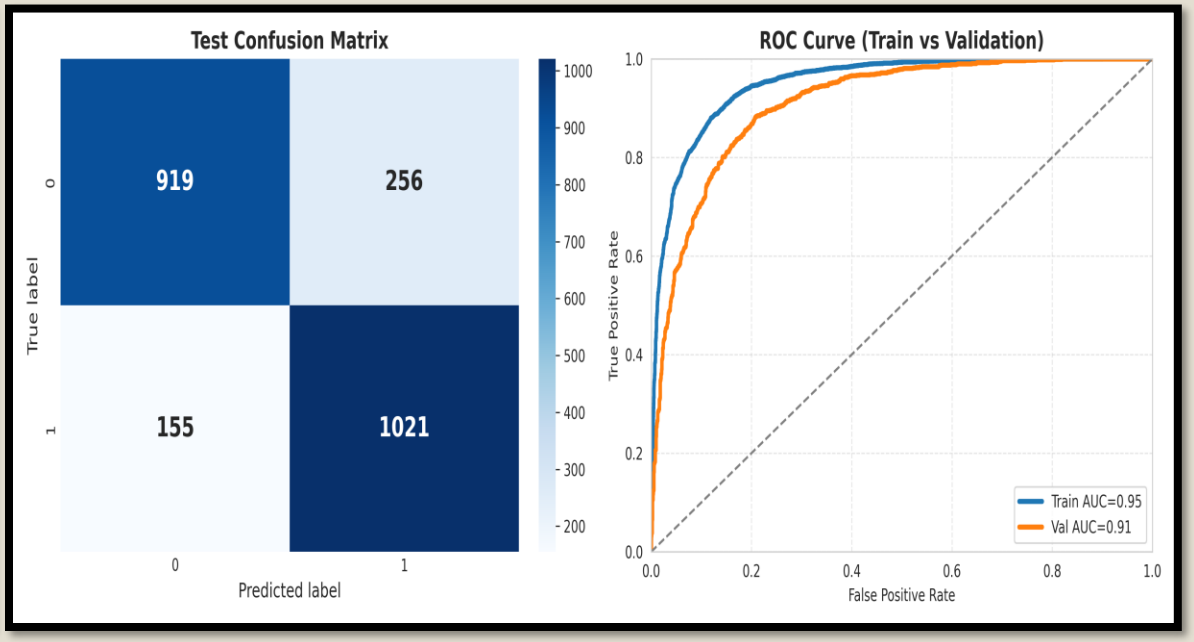

Figure 8: Confusion Matrices and ROC Curve of ML Ensemble

RESULT & DISCUSSION

5. BILSTM

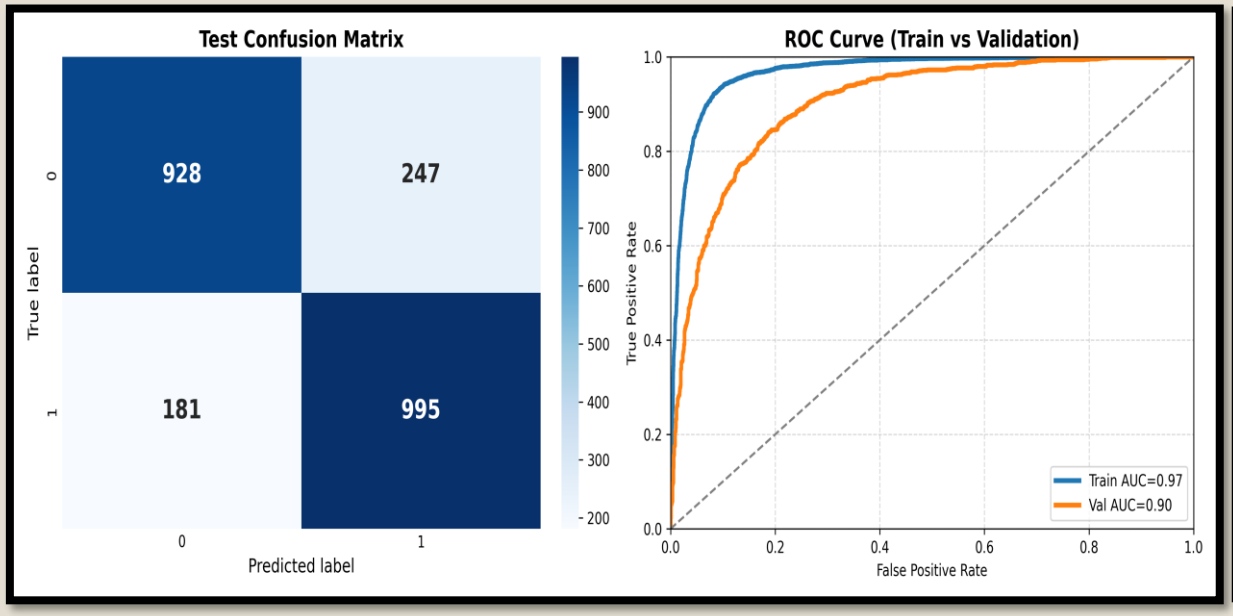

Figure 9: Confusion Matrices and ROC Curve of BILSTM

6. BIGRU

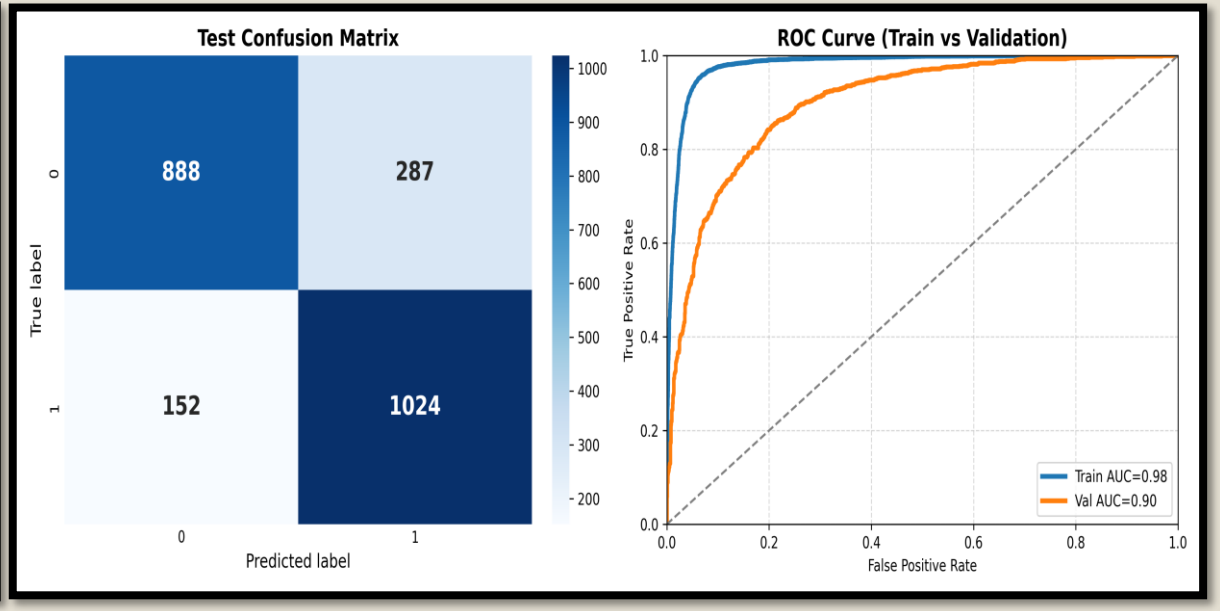

Figure 10: Confusion Matrices and ROC Curve of BIGRU

RESULT & DISCUSSION

7. CNN

8. DNN

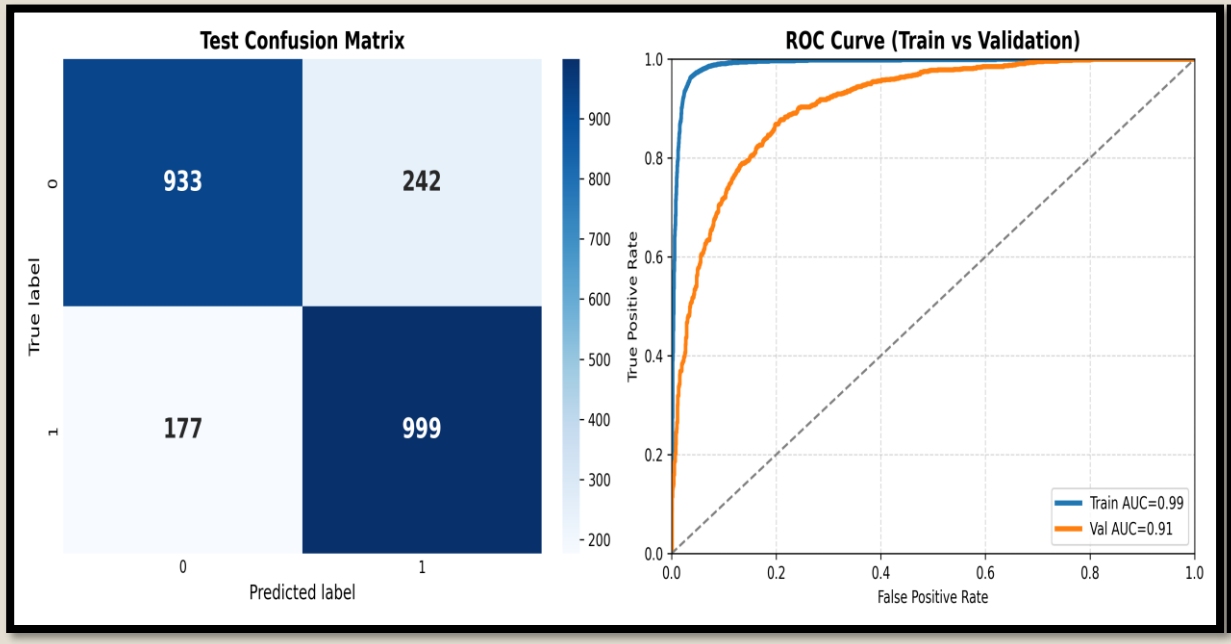

Figure 11: Confusion Matrices and ROC Curve of CNN

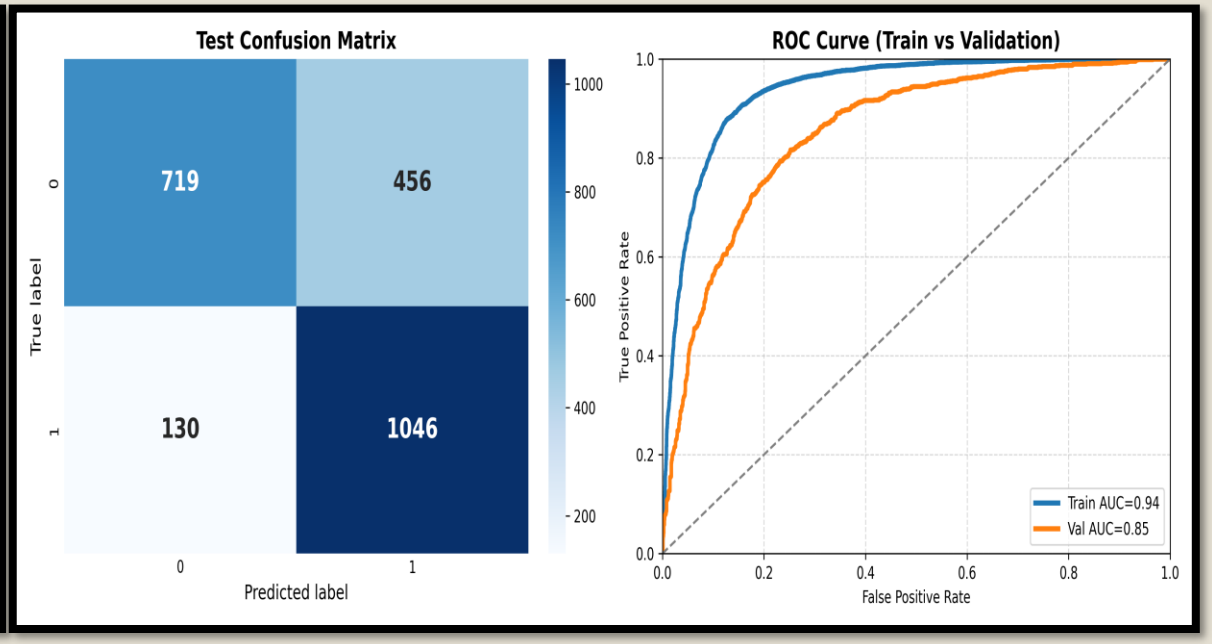

Figure 12: Confusion Matrices and ROC Curve of DNN

RESULT & DISCUSSION

9. DL ENSEMBLE

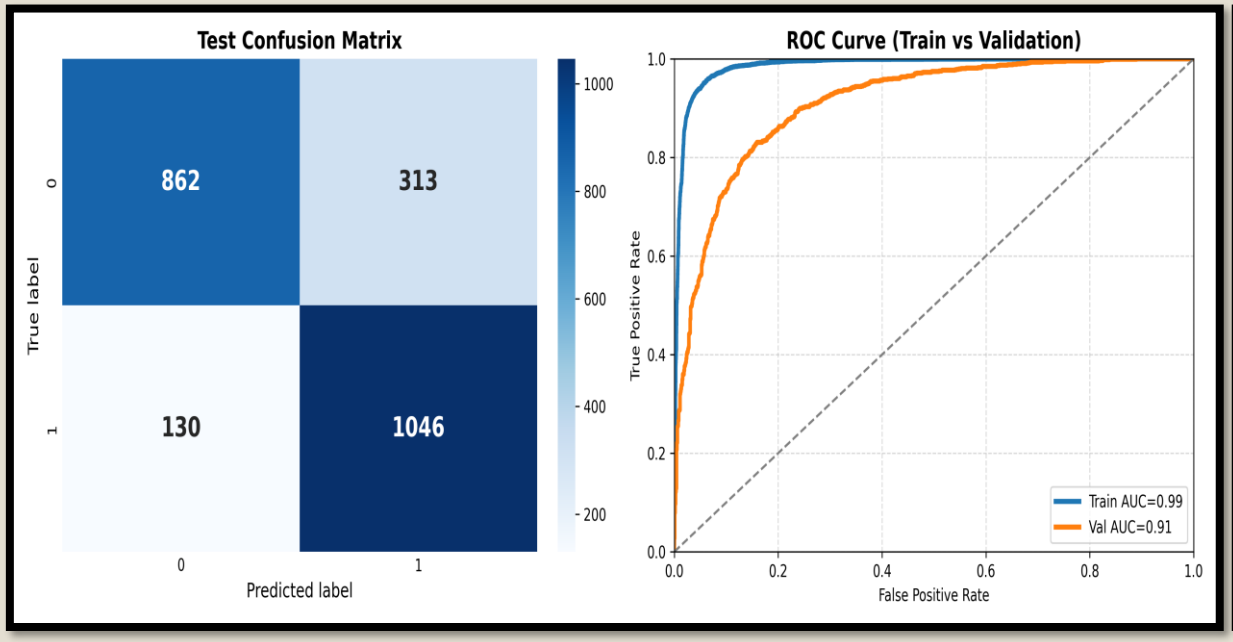

Figure 13: Confusion Matrices and ROC Curve of DL Ensemble

10. M-BERT

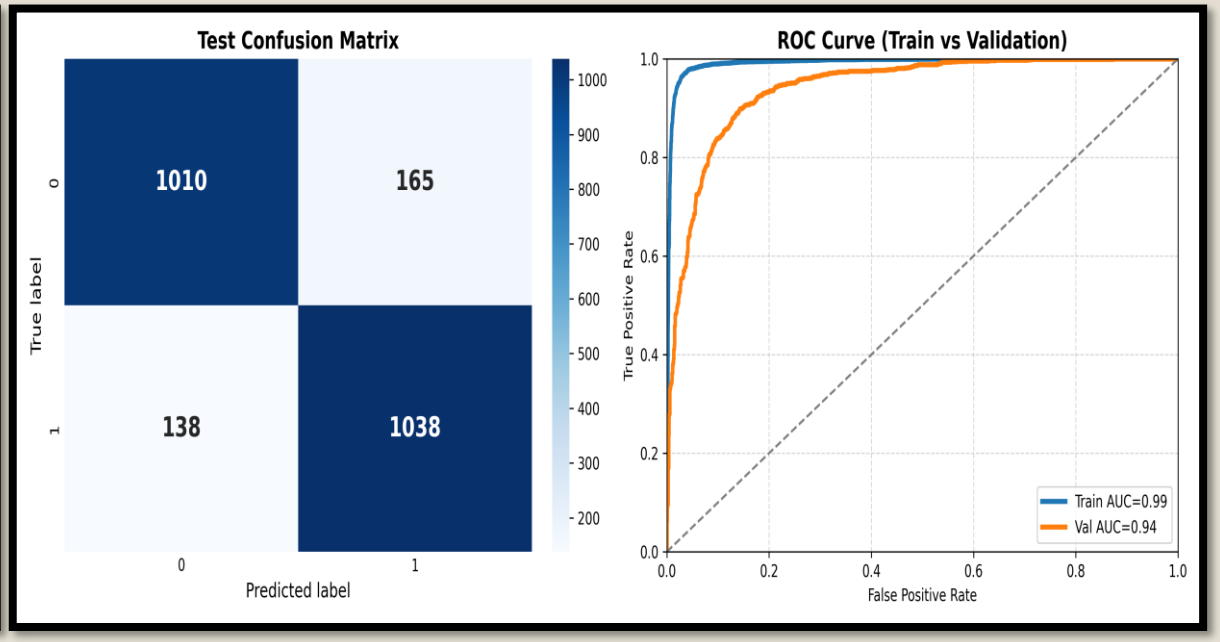

Figure 14: Confusion Matrices and ROC Curve of mBERT

RESULT & DISCUSSION

11. BanglaBERT

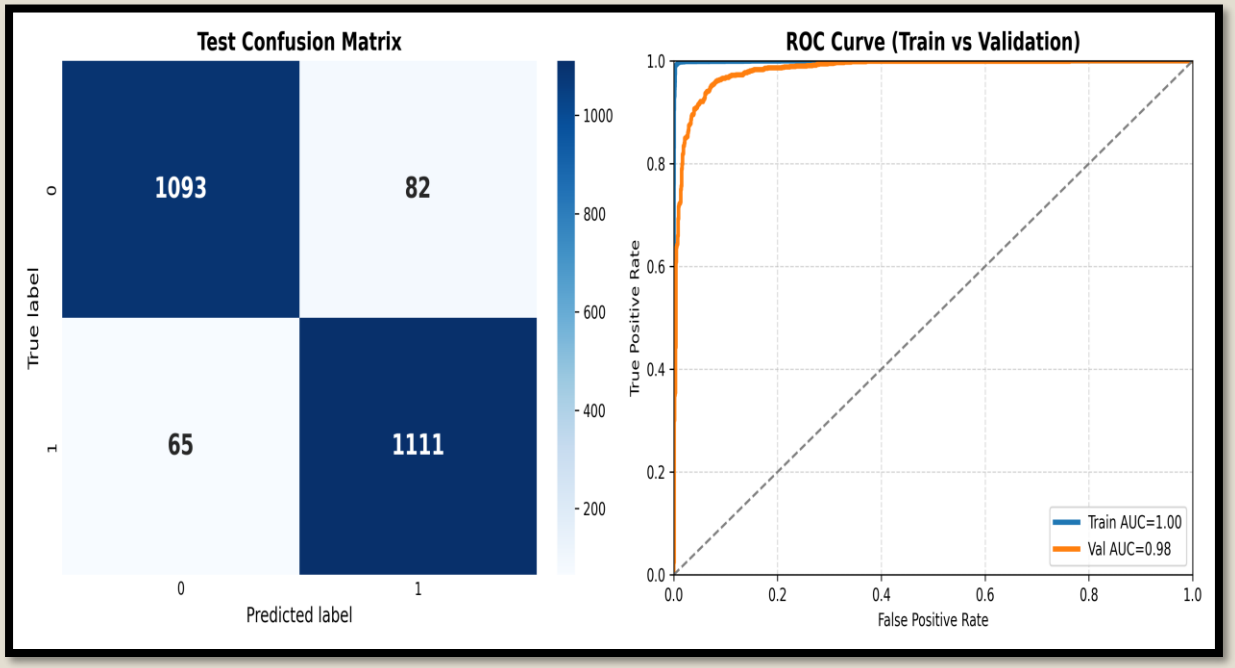

Figure 15: Confusion Matrices and ROC Curve of BanglaBERT

12. XLM-R

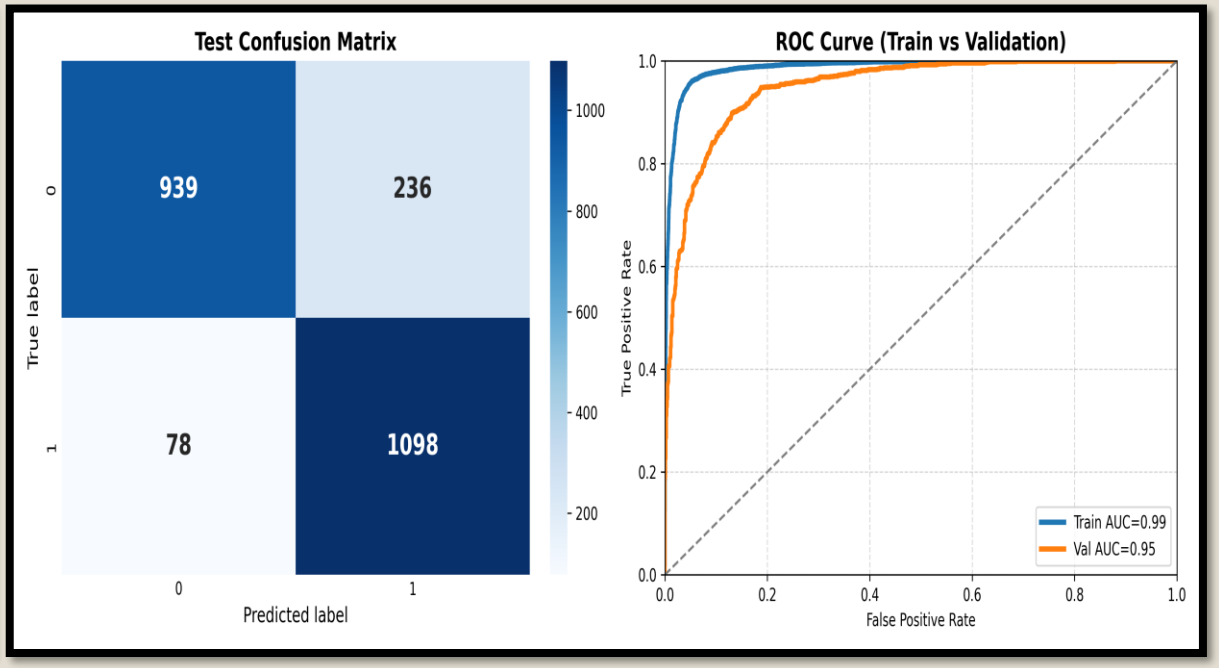

Figure 16: Confusion Matrices and ROC Curve of XLM-R

RESULT & DISCUSSION

13. BanglaT5

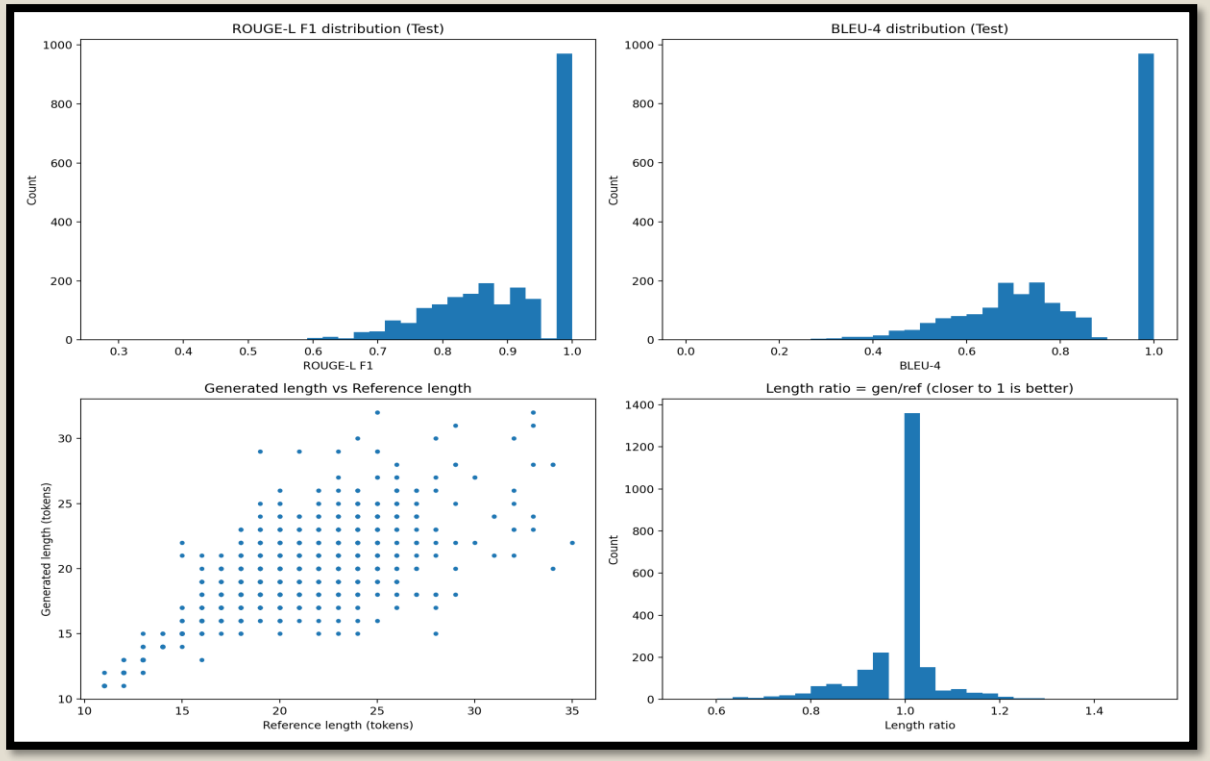

Figure 17: Reason Generation Metrics

```
METRICS SUMMARY:
num_samples: 2351
rougeL_f1_mean: 0.9036605583171314
rougeL_f1_median: 0.9230769230769231
bleu4_mean: 0.8196802616346881
bleu4_median: 0.803154665668484
gen_len_mean_tokens: 16.86176095278605
ref_len_mean_tokens: 17.317311782220333
```

Figure 18: Classification Report of BanglaT5

RESULT & DISCUSSION

Tabel 2: Performance of Models

Model Type	Model Name	Accuracy	F1-score	ROC-AUC
TRANSFORMER	**BanglaBERT**	**94%**	**94%**	**98%**
	XLM-R	88%	88%	95%
	mBERT	87%	86%	94%
ML	SVM	83%	83%	91%
	ML Ensemble	82%	82%	91%
	LR	83%	83%	90%
	Naive Bayes	79%	79%	87%
DL	DL Ensemble	83%	83%	90%
	BiLSTM	82%	82%	90%
	BiGRU	81%	81%	90%
	DNN	77%	78%	85%
	CNN	83%	83%	91%

RESULT & DISCUSSION

Tabel 3: Comparative Analysis

Aspect	Briskilal & Subalalitha (2022)	My Work (2026)
Task	**Idiom vs. Literal Classification**	**Idiom Misuse Detection + Explanation**
Language	English (high-resource)	Bangla (low-resource, 230M speakers)
Dataset Size	TroFi Dataset ~5,207 sentences	Custom Bangla Dataset ~15,670 sentences
Idioms Covered	50 English verbs	1,316 distinct Bangla idioms Manually constructed
Models Evaluated	BERT-base, RoBERTa-base, Ensemble	3 ML models, 4 DL models, 3 Transformers, 2 Ensembles, BanglaT5
Best Model	BERT + RoBERTa Ensemble	BanglaBERT
Preprocessing	Removed punctuation, Stopwords, Lowercase, Removed punctuation only	Unicode NFC normalization, Zero-width character removal, Whitespace normalization, Punctuation unification, No stop-word removal
Feature Engineering	Pre-trained embeddings only	TF-IDF (5,800 features), Learned embeddings (128-dim), Pre-trained contextual (768-dim)
Accuracy	**90% (Ensemble)**	**94% (BanglaBERT)**
F1-Score	**89% (Ensemble)**	**94% (BanglaBERT)**
ROC-AUC	Not reported	0.99 Validation and 1.00 Training
Explainability	**None (Black-box classification)**	**BanglaT5 achieves ROUGE-L = 0.90 and BLEU-4 = 0.82 on 2,351 samples**
Novel Contribution	First BERT + RoBERTa ensemble for idiom classification	First Bangla idiom misuse dataset, First explainable idiom processing, Comprehensive comparative study, Two-stage explanation generation

RESULT & DISCUSSION

CLASSIFICATION EXAMPLE:

```
Example 1
Sentence    : রাজনীতিতে জয়ী হতে হলে গভীর জলের মাছ হতে হয়
Prediction : 1 | Prob(1) = 0.997
GT Label    : 1 | CORRECT ✅

Example 2
Sentence    : নতুন ব্যবসায়ী সত্যিই গভীর জলের মাছ
Prediction : 1 | Prob(1) = 0.998
GT Label    : 1 | CORRECT ✅
```

REASONING EXAMPLE:

```
Example 15
Sentence    : বাড়ি তৈরির জন্য টাকা দেওয়ার কথা বলে এখন সে হাত গুটিয়ে নিয়েছে, যেন গাছে তুলে মই কাড়া
Label       : 1
Generated Reason:
উক্ত 'গাছে তুলে মই কাড়া' বাগধারাটির অর্থ হলো 'সাহায্যের আশা দিয়ে সাহায্য না করা' এবং এটি সঠিকভাবে ব্যবহৃত হয়েছে।
--------------------------------------------------------------------------------

Example 16
Sentence    : তিনি শুধু সাহায্য করেই ক্ষান্ত হননি, আমাকে গাছে তুলে মই কাড়ার পর আরও উপরে ওঠার সিঁড়িও দিলেন
Label       : 0
Generated Reason:
উক্ত 'গাছে তুলে মই কাড়া' বাগধারাটির অর্থ হলো 'সাহায্যের আশা দিয়ে সাহায্য না করা', কিন্তু এটি 'আক্ষরিক অর্থে গাছে ওঠা' বোঝাতে ব্যবহার করা হয়েছে, যা ভুল
--------------------------------------------------------------------------------
```

CHALLENGES

1. Context Dependence: Idiom meaning changes with context.
2. Low-Resource Language: Limited annotated datasets and pre-trained Bangla models.
3. Data Sparsity: Many idioms appear infrequently, limiting model learning.
4. Complex Linguistics: Morphological richness and compound word forms in Bangla increase preprocessing and feature extraction complexity.
5. Explainability: Generating natural, human-aligned reasoning for model predictions is hard.
6. Generalization: Ensuring models perform well on unseen idioms and diverse text sources.
7. Computational Constraints: Transformer fine-tuning requires significant GPU resources, which may limit scalability.

CONCLUSION

This research develops a framework for detecting and explaining Bangla idiom misuse using a new dataset of 15,670 sentences. **BanglaBERT** achieved 94% accuracy and a two-stage BanglaT5 model generated explanations aligned with human reasoning (ROUGE-L mean 0.90). Results show transformer embeddings excel in context understanding, supporting explainable AI for education. The dataset and models could be publicly released for future low-resource language research.

FUTURE WORK

Future work includes expanding the dataset with more idioms, spoken language and fine-grained labels. And also exploring advanced neural models, multimodal, neuro-symbolic approaches. Improving explanation generation with contrastive methods and applying the framework to other low-resource languages and extending techniques to related NLP tasks like metaphor, sarcasm and error analysis.

REFERENCES

[1]Y. Ge et al., “Implicit knowledge-augmented prompting for commonsense explanation generation,” Knowledge and Information Systems, vol. 67, Jan. 2025, doi: https://doi.org/10.1007/s10115-024-02326-w.

[2]J. Sikder, P. Chakraborty, U. K. Das, and K. Dhar, “A hybrid approach for Bengali sentence validation,” Artificial Intelligence Review, vol. 57, no. 11, Oct. 2024, doi: https://doi.org/10.1007/s10462-024-10795-2.

[3]A. Bhattacharjee et al., “BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla,” arXiv (Cornell University), vol. 4, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2101.00204.

[4]C. Wang, S. Liang, Y. Jin, Y. Wang, X. Zhu, and Y. Zhang, “SemEval-2020 Task 4: Commonsense Validation and Explanation,” arXiv (Cornell University), vol. 2, Jan. 2020, doi: https://doi.org/10.48550/arxiv.2007.00236.

[6] J. Briskilal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” Information Processing & Management, vol. 59, no. 1, 2022, Art. no. 102756, doi: https://doi.org/10.1016/j.measen.2022.100434.

REFERENCES

[7] A. Abedin and B. S. Purkayastha, “Parts of Speech Tagging in Bengali for MWEs Detection,” International Journal of Computer Applications, vol. 99, no. 17, pp. 1–6, Aug. 2014, doi: 10.5120/17485-8182.

[8] S. K. Nahin et al., “TituLLMs: A Family of Bangla LLMs with Comprehensive Benchmarking,” arXiv preprint arXiv:2502.11187, 2025, doi: https://arxiv.org/abs/2502.11187.

[9] M. Ismayilzada et al., “CROW: Benchmarking Commonsense Reasoning in Real-World Tasks,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, Dec. 2023, doi: https://doi.org/10.48550/arXiv.2310.15239.

[10] L. Jiang, A. Bosselut, C. Bhagavatula, and Y. Choi, “‘I’m Not Mad’: Commonsense Implications of Negation and Contradiction,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Online, Jun. 2021, pp. 180–191. doi: https://doi.org/10.48550/arXiv.2104.06511.

THANK YOU