SOHAG GAIN

Academic CV | Machine Learning & Multilingual NLP

Dhaka, Bangladesh • sohag@sohaggain.com • sohaggain.com/academic.php • github.com/sohaggain

ACADEMIC PROFILE

Computer Science & Engineering graduate with research experience in multilingual text classification, comparing classical machine learning, neural architectures, and transformers under a unified evaluation framework. Research interests include natural language processing, reliable AI systems, retrieval-augmented generation, model evaluation, and practical deployment trade-offs.

EDUCATION

BSc in Computer Science & Engineering

Jamalpur Science & Technology University · February 2025 · CGPA: 3.55 / 4.00

RESEARCH THESIS

Benchmarking Classical Machine Learning and Transformer Architectures for Bangla–English News Classification: A Large-Scale Unified Evaluation with Morphological Complexity Analysis

Research question. How do classical machine-learning models, neural networks, and transformer architectures compare on large-scale Bangla and English news classification when they are evaluated under one preprocessing and evaluation framework?

Data. 434,563 Bangla news articles from Prothom Alo across 15 categories and 127,600 English AG News articles across four categories: 562,163 articles in total.

Methods. Logistic Regression, Naive Bayes, SVM, Random Forest, XGBoost, BiLSTM, BERT, and LSTM+BERT. Classical models used Bag-of-Words, one-hot, and TF-IDF representations; experiments used a stratified 80/10/10 train/validation/test split.

Reported findings. BERT achieved the strongest test accuracy: 92.95% for Bangla and 94.50% for English. The strongest classical TF-IDF + SVM baseline reached 90.95% and 91.03%, respectively, supporting a practical lower-resource deployment option.

Additional analysis. Quantitative inflectional-density analysis classified 98.8% of Bangla articles within the medium-to-high complexity ranges, adding a measured view of morphology relevant to model performance.

Full paper: sohaggain.com/assets/thesis/Sohag_Gain_BSc_Thesis.pdf

RESEARCH AND TECHNICAL SKILLS

Methods Supervised classification, model comparison, experimental design, ablation-minded analysis, performance evaluation	**NLP** Text preprocessing, TF-IDF, embeddings, transformers, Bangla and English classification, retrieval systems
Tools Python, scikit-learn, deep learning workflows, BERT-based models, data analysis	**Systems** FastAPI, RAG, vector databases, API integrations, Docker, PostgreSQL, Git/GitHub

SELECTED APPLIED SYSTEMS

- **RAG knowledge-base automation** — document ingestion, vector retrieval, and grounded AI answers; demonstrates an applied extension of NLP and information-retrieval work.
- **Multi-agent business OS** — agent orchestration with explicit responsibilities and human review; demonstrates system design for reliable AI-assisted operations.
- **Custom MCP server** — standardized authenticated tool access for AI agents; demonstrates API, integration, and deployment design.

RESEARCH SCOPE AND LIMITATIONS

- The Bangla corpus comes from a single news source and has substantial class imbalance; the English AG News corpus has four broad categories, limiting direct cross-language difficulty comparisons.
- The inflectional-density measure uses a curated set of 17 suffixes and does not cover all derivation, sandhi, or compounding. A fuller morphological analyzer is a clear next step.
- Reported model results come from one seeded run without variance estimates or significance testing. Further work can evaluate category-level performance, additional corpora, multilingual models, and alternative hybrid architectures.

SUPPORTING SCHOLARSHIP DOCUMENTS

Academic records: University, HSC, and SSC transcripts and testimonials.
Language and references: Medium of Instruction certificate and three recommendation letters.
Additional activity evidence: Programming contest participation, volunteer activity, and Robotics & IoT workshop certificates. Documents are shared individually with application committees; identity documents are kept private.

SCHOLARSHIP APPLICATION MATERIALS

For each application, align the statement of purpose with the target programme and connect the proposed study to the thesis in Bangla–English text classification, multilingual NLP, and reliable machine-learning evaluation. Provide official records and referee letters directly to the application committee as required.