CS 404 NLP Final Project Proposal

Title: News Headline Topic Classifier Using Traditional ML and Deep Learning

---

Content

Application Description:

This project involves building an NLP-based text classification system that predicts the topic category of a news headline. For example, let's say there is a short headline, "NASA launches new satellite", the model will classify it into one of four categories: **World, Sports, Business, or Science/Technology**. The model will be trained on the AG News dataset containing thousands of labeled news headlines.

Significance:

In an era of information overload, automating the categorization of news content is crucial for improving user experience on news platforms, apps, or aggregators. This project demonstrates how NLP and supervised learning techniques can be used to build practical tools with real-world impact.

Project Goal(s):

- Build a baseline model using traditional machine learning (TF-IDF + Logistic Regression/SVM).
- Train a deep learning model using dense neural networks on headline vectors.
- Fine-tune a transformer-based model (e.g., DistilBERT) for headline classification.
- Evaluate and compare the performance of all three models using standard metrics.
- Visualize results and provide an interface to test new headlines.

Core NLP Techniques That Will Be Employed:

- Text preprocessing: tokenization, stopword removal, lowercasing
- TF-IDF vectorization for classical ML models

- Word embeddings and deep neural networks
- Transformer-based fine-tuning using pre-trained BERT/DistilBERT
- Multiclass classification (4 categories)
- Performance evaluation: Accuracy, F1-score, Confusion Matrix

---

The Data We Intend to Use:

- **AG News Dataset**
    - ~120,000 news headlines labeled into four categories
    - Source: https://www.kaggle.com/datasets/amananandrai/ag-news-classification
    - Includes balanced data across categories

Tools and Libraries:

- Python
- scikit-learn (Logistic Regression, TF-IDF)
- TensorFlow or PyTorch (for DNN)
- HuggingFace Transformers (for fine-tuning DistilBERT)
- pandas, numpy (data wrangling)
- matplotlib, seaborn (visualization)
- Streamlit (optional – for demo interface)

Justification About the Feasibility:

- The dataset is public, clean, and well-labeled, requiring minimal preprocessing.
- The baseline model can be implemented in 1 day using scikit-learn.

Potential Challenges:

- Ensuring class balance and avoiding overfitting in the DNN model
- Handling short-text data effectively (few words per headline)

- Transformer fine-tuning may require GPU support
- Ensuring fair comparison across different models (standard metrics, same splits)

Team Members and Workload Distribution (if applicable):

Name	**Role**	**Responsibilities**
Bibek Sharma	NLP Developer & Researcher	Dataset handling, TF-IDF modeling, BERT fine-tuning
Ankit Paudel	DL Engineer	DNN architecture, training, evaluation
Sohan Lama	Frontend/Visualization	Charts, model output display, Streamlit demo

This is our tentative project model. We welcome your feedback and are fully prepared to refine or modify the project based on your suggestions or recommendations.