BERT-Powered AI News Classifier: An End-to-End Transformer Pipeline for Large-Scale Multi-Class News Topic Classification

Chinthala Sai Charan
Email: chinthala2004@gmail.com

***Abstract*—The exponential growth of digital journalism and online content distribution has increased the demand for automated systems capable of organizing and understanding news at scale. News topic classification is a foundational task for content recommendation, information retrieval, news summarization, and media analytics. In this work, we present a practical and deployment-oriented Transformer-based news classification system that categorizes news headlines into major topic classes using a fine-tuned BERT-family model. The proposed pipeline addresses key challenges commonly observed in real-world datasets, including inconsistent category labels, long-tail class distributions, and noise introduced by highly specific categories. To mitigate these issues, the system applies category normalization by retaining the top K most frequent categories and mapping remaining labels to an *Other* class, ensuring stable training and improved generalization. The model is trained using the Hugging Face Transformers framework with GPU acceleration and mixed precision. Additionally, confidence-aware predictions are computed using softmax probability scores to support interpretability in real-time inference. The final solution demonstrates a complete end-to-end workflow, including dataset preprocessing, label encoding, stratified splitting, tokenization, training with checkpointing, and prediction on unseen headlines. The project is positioned as an industry-relevant system suitable for internship evaluation, showcasing modern NLP engineering practices and reproducible experimentation.**

***Index Terms*—News Classification, Transformer Models, BERT, DistilBERT, Text Classification, NLP, Topic Classification, Deep Learning, Hugging Face**

I. Introduction

News consumption has evolved from traditional print media to a fast-paced digital ecosystem where headlines are generated continuously by newspapers, online media portals, and social platforms. The overwhelming volume of daily news makes it difficult for readers and platforms to organize, search, and consume content efficiently. Automated news topic classification helps address this challenge by assigning each headline or article to a topic category such as Politics, Sports, Business, Crime, Health, or Education.

News topic classification has wide practical relevance. It can be used for:

- Personalized news recommendation and feed ranking.
- Content tagging and cataloging for digital news platforms.
- Supporting summarization pipelines by grouping news into coherent topics.
- Media monitoring systems and analytics dashboards.
- Moderation and filtering pipelines to control exposure to sensitive categories.

Traditional machine learning approaches to text classification often relied on feature extraction techniques such as bag-of-words, TF-IDF, and n-gram representations, followed by classifiers such as Naïve Bayes, Support Vector Machines (SVM), or Logistic Regression. While these methods can be effective in small-scale settings, they struggle to

capture contextual semantics and long-range dependencies present in natural language. The emergence of Transformer architectures has significantly improved performance in NLP tasks by learning deep contextual representations.

Bidirectional Encoder Representations from Transformers (BERT) is one of the most influential Transformer-based models, enabling state-of-the-art performance in tasks such as sentiment analysis, question answering, and text classification. DistilBERT, a smaller and faster distilled version of BERT, provides comparable performance with reduced computational cost. This project leverages a BERT-family model for large-scale multi-class classification of news headlines, emphasizing reproducibility, robustness, and deployment readiness.

A. Motivation

This project was motivated by an internship assignment requiring the use of a news summarization dataset to build an LLM-based solution. While summarization is a common task, we aimed to build a system that remains strongly aligned with real-world industry needs. Topic classification is a core capability that enhances summarization pipelines and enables multi-stage news intelligence systems. By focusing on a Transformer-based classifier with practical preprocessing and checkpointing strategies, this work demonstrates a complete end-to-end NLP engineering workflow.

B. Key Contributions

The main contributions of this work are:

- An end-to-end Transformer pipeline for multi-class news topic classification.
- Category normalization and long-tail mitigation using top-K selection and an *Other* class.
- A scalable training workflow using GPU acceleration and mixed precision.
- Confidence-aware inference for interpretability and reliability.
- Reproducible workflow compatible with Jupyter notebooks and modern development environments.

II. Problem Statement

The objective is to design a system that takes a news headline as input and outputs a topic category. The dataset contains a large number of category values, many of which are rare, overly specific, or inconsistent. Training a classifier directly on such labels leads to unstable training and poor generalization. Therefore, the classification problem must be formulated carefully.

A. Task Definition

Given a news headline x, predict its category y from a finite set of topic labels:

$$f(x) \rightarrow y$$

where $y \in \{1, 2, \ldots, C\}$ and C is the number of categories after preprocessing.

B. Challenges

- **Noisy Labels:** Real-world category values may include long descriptions or merged tags.
- **Long-Tail Distribution:** Many categories have extremely low frequency.
- **Large-Scale Training:** The dataset is large, requiring efficient training strategies.
- **Reproducibility:** Training must be reproducible and robust to interruptions.

III. Related Work

Text classification has been widely studied in NLP. Early approaches relied on statistical models and manual feature engineering. Naïve Bayes and SVM-based classifiers were common for news classification tasks. However, such methods often struggle with polysemy, context-dependent meaning, and phrase-level semantics.

Deep learning models such as CNNs and LSTMs improved performance by learning distributed word embeddings and sequential patterns. Yet, recurrent architectures can be limited by vanishing gradients and inefficiency for long sequences.

Transformer-based models, introduced by Vaswani et al., revolutionized NLP by using self-attention mechanisms to model dependencies in parallel. BERT introduced bidirectional pretraining using masked language modeling, enabling strong

transfer learning for downstream tasks. DistilBERT reduced the computational cost through knowledge distillation, making it suitable for real-world applications where efficiency matters.

News classification using BERT has been explored in academic and industry settings. However, real-world datasets often contain noisy category labels and long-tail distributions. This work addresses these practical challenges using category normalization and robust preprocessing.

IV. Dataset Description

The dataset used in this project is provided in Excel format and contains news-related information including headlines, newspaper sources, published dates, and category labels. The dataset was originally curated for summarization tasks but is repurposed in this work for classification.

A. Dataset Fields

The dataset includes fields such as:

- **newspaper_name:** Source of the news headline.
- **published_date:** Publication date.
- **text / headline:** News headline text.
- **article_human_summary:** Human-written summary (not directly used in this model).
- **category / news_category / clean_category:** Category label.

B. Why Headline-Based Classification?

Headlines are concise and often contain high-signal information about the topic. Headline classification is valuable because:

- It is computationally cheaper than full-article classification.
- Headlines are widely available even when full content is behind paywalls.
- It enables fast real-time categorization.

V. Preprocessing and Label Normalization

Preprocessing is a critical step in transforming raw datasets into a stable learning problem. Real-world news datasets frequently contain inconsistent categories and rare labels that can significantly degrade training.

A. Data Cleaning

The following cleaning steps were applied:

- Removal of missing values in headline and category columns.
- Conversion of headline and category fields to string format.
- Standardization of column names.

B. Long-Tail Category Reduction

A major challenge was the presence of thousands of unique category values. Directly training a multi-class classifier with thousands of classes is computationally expensive and often results in poor generalization due to extreme class imbalance.

To address this, the top K most frequent categories were retained, and all remaining categories were mapped to *Other*:

$$category' = \begin{cases} category & \text{if } category \in TopK \\ Other & \text{otherwise} \end{cases}$$

This step reduced the classification space to $K+1$ classes, making training feasible and meaningful.

C. Final Category Distribution

After preprocessing, the dataset contains 11 categories (Top 10 + Other). Table I illustrates a representative distribution.

TABLE I: Final Category Distribution (Top 10 + Other)

Category	Count
Other	79046
Politics	48796
Business and Finance	33855
National News	33012
Local News	29853
International News	26443
Crime and Justice	24418
Sports	23290
Entertainment	19898
Health and Wellness	16434
Education	13710

VI. Model Architecture

A. Transformer Encoder for Classification

Transformer models represent text as contextual embeddings. For sequence classification, a special

token representation (e.g., [CLS] for BERT) is used as a pooled summary of the input.

The classifier head maps the pooled representation to logits:

$$z = Wh + b$$

where h is the pooled hidden state, and z is the logits vector.

A softmax function converts logits into probabilities:

$$p(y = i|x) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

B. Choice of DistilBERT

DistilBERT is used due to:

- Faster training compared to full BERT.
- Lower memory usage, suitable for mid-range GPUs.
- Competitive performance on classification tasks.

VII. Training Pipeline

The system was implemented using Hugging Face Transformers and PyTorch. The complete workflow is summarized in Algorithm 1.

Algorithm 1 End-to-End News Classification Pipeline

1: Load dataset from Excel file
2: Clean missing values and normalize column names
3: Select top K categories and map others to *Other*
4: Encode categories using LabelEncoder
5: Perform stratified train-test split
6: Tokenize text using Transformer tokenizer
7: Create PyTorch datasets for training and testing
8: Initialize Transformer classification model
9: Fine-tune model with checkpointing enabled
10: Predict categories for unseen headlines with confidence scores

A. Train-Test Split

A stratified split was used to preserve the class distribution:

- Training set: 90%
- Test set: 10%

B. Tokenization

Tokenization converts raw text into input IDs and attention masks. A maximum length of 128 tokens was chosen to balance speed and performance.

C. Training Configuration

The training configuration included:

- Batch size: 16
- Epochs: 2
- Learning rate: 2×10^{-5}
- Weight decay: 0.01
- Mixed precision (FP16) enabled on GPU

D. Checkpointing and Recovery

Checkpointing was enabled to support recovery from interruptions. This feature was critical in ensuring reproducibility and stability during long training runs.

VIII. Results and Observations

The model was successfully trained for 2 epochs. Training logs indicated stable convergence with final training loss around 1.0. The training process was accelerated using an NVIDIA GPU and mixed precision training.

A. Training Output Summary

The training completed with:

- Global steps: 39,236
- Training loss: approximately 1.05
- High throughput due to GPU acceleration

B. Qualitative Prediction Examples

Example predictions produced by the system:

- Headline: “Government announces new policy for digital education” $\rightarrow$ Education
- Headline: “Stock markets fall amid global uncertainty” $\rightarrow$ Business and Finance
- Headline: “India wins thrilling cricket match” $\rightarrow$ Sports

C. Confidence Scoring

Confidence scoring is computed using the softmax probability of the predicted class. This provides interpretability, especially when headlines contain ambiguous terms.

IX. Discussion

The results demonstrate that Transformer-based models are highly effective for topic classification of news headlines. The category normalization strategy was essential for converting a noisy real-world dataset into a stable classification problem. The inclusion of checkpointing ensured resilience to system interruptions and improved the practical usability of the training pipeline.

A. Why Category Normalization Matters

Without label normalization, the dataset contained thousands of unique category values. Training such a classifier would be inefficient and produce poor generalization. Mapping rare categories to *Other* is a standard industry practice to manage long-tail distributions.

B. Deployment Considerations

For deployment, the trained model can be used as an inference service to classify incoming headlines in real time. Since DistilBERT is relatively lightweight, it can run efficiently on CPU-based servers or GPU-enabled systems depending on throughput requirements.

X. Limitations

- The model uses only headlines and does not incorporate full article context.
- The system does not perform sentiment analysis or misinformation detection in the current version.
- The dataset categories are derived from real-world sources and may contain subjective labeling.

XI. Future Work

Potential future enhancements include:

- Adding sentiment analysis as a parallel pipeline using pretrained sentiment models.
- Extending the classifier to full-article classification for richer context.
- Implementing explainability methods such as attention visualization.
- Integrating the classifier into a web-based demo using Streamlit or FastAPI.

XII. Conclusion

This paper presented a practical Transformer-based news classification system that categorizes news headlines into major topics using a fine-tuned BERT-family model. The project addressed key real-world dataset challenges through category normalization and long-tail mitigation using an *Other* class. Training was performed using Hugging Face Transformers with GPU acceleration, mixed precision, and checkpointing for reliability. The resulting model provides accurate topic predictions and confidence scores, making it suitable for real-world integration and internship-level evaluation. The end-to-end nature of the project highlights the value of Transformer architectures in scalable NLP applications.

References

[1] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT*, 2019.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[4] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP: System Demonstrations*, 2020.