EARIN Project
News Category Prediction Using Machine Learning and Transformer-Based Models

Aditya Kandla and Martyna Wielgopolan

Project Goal

The goal of this project is to build a predictive model that automatically classifies short news headlines into predefined categories such as *Politics*, *Science*, *Entertainment*, etc. The model should generalize well across topics and handle class imbalance, achieving high accuracy and balanced performance across all categories.

Introduction and Project Description

This project aims to design, implement, and evaluate multiple machine learning models capable of categorizing news headlines into topics using natural language processing (NLP) techniques. The classification problem is framed as a supervised learning task, where each headline is a short text input, and the category is the target label.

The project uses the **News Category Dataset** from HuffPost, which contains over 200,000 headlines published between 2012 and 2018. By comparing traditional machine learning methods with modern deep learning techniques such as **BERT**, we aim to highlight the trade-offs in complexity, interpretability, and performance.

Preliminary Assumptions of the Project

1. Algorithms Description with Examples

The following algorithms will be implemented and compared:

- **Logistic Regression + TF-IDF**: A strong and interpretable baseline for linear classification over sparse features derived from text.
- **Multinomial Naive Bayes (MNB)**: Effective for text classification when independence assumptions approximately hold.
- **Support Vector Machine (SVM)**: Suitable for high-dimensional sparse data.
- **BERT (Bidirectional Encoder Representations from Transformers)**: A transformer-based model fine-tuned for headline classification.

Examples:

- Headline: *"Apple introduces new iPhone model"* → Category: `TECH`
- Headline: *"Supreme Court rules on abortion law"* → Category: `POLITICS`

2. Dataset Description

We will use the **News Category Dataset** from Kaggle, which includes:

- Over 200,000 headlines
- Category labels (e.g., `BUSINESS`, `ENTERTAINMENT`, `WELLNESS`)
- Optional metadata: `short_description`, `authors`, `date`, `link`

The dataset contains over 40 unique categories. Class imbalance will be addressed using techniques such as stratified sampling or class weighting.

3. Plan of Tests/Experiments

1. Data preprocessing: lowercasing, punctuation removal, optional stopword filtering
2. Tokenization and vectorization:
   - TF-IDF for classical models
   - Token IDs and attention masks for BERT
3. Train-validation-test split: 80% / 10% / 10%, stratified
4. Model training and hyperparameter tuning
5. Model evaluation on test set
6. Comparison of model performance and error analysis

4. Methods of Result Visualization

- Confusion matrices
- Precision/Recall/F1 bar plots
- Training curves (loss/accuracy)
- t-SNE or PCA plots for embedding visualization

5. Quality Measures

- **Accuracy**: Overall percentage of correct predictions
- **Precision, Recall, F1-Score (per class)**: Class-specific performance
- **Macro F1-score**: Average performance across classes
- **Weighted F1-score**: Performance weighted by support
- **Training and inference time**: For efficiency comparison

Conclusion

This project outlines a comprehensive plan to develop an effective model for news headline classification using both classical and modern machine learning approaches. By evaluating models such as Logistic Regression, Naive Bayes, SVM, and BERT, we aim to identify the trade-offs between interpretability, computational efficiency, and accuracy. The use of proper preprocessing, visualization, and evaluation metrics will ensure a rigorous comparison. The final model can contribute to automated news categorization systems, improving content management and user experience in news platforms.