News Classification with Neural Networks: Embracing a Comprehensive Approach

by Classification Crusaders

Team Members:

1. Hrishikesh Rajesh Shenai (hms6207)
2. Vedant Sahai (vzs5356)
3. Sri Krishna Chaitanya Velamakanni (vzs5369)

Introduction

- **Project Context:** "This project advances into the realm of text classification, focusing on categorizing news articles using Neural Networks."
- **Techniques Used:** "We employ CNN, BiLSTM, and XLNet transformer models, chosen for their effectiveness in processing and understanding complex textual data."
- **Relevance and Application:** "Text classification, like POS tagging, is a cornerstone in NLP, crucial for tasks such as information retrieval, sentiment analysis, and automated content organization."
- **Objective:** "Our goal is to efficiently classify news based on brief summaries and descriptions, highlighting the capabilities of different neural network architectures in handling nuanced text."
- **Comparative Analysis:** "The project compares the performance of CNN, BiLSTM, and XLNet models, showcasing their strengths and limitations in text classification tasks."

Workflow Overflow

- Task and Motivation
- Exploratory Data Analysis
- Data Pre-processing
- Model Training
- Performance Evaluation and Hyperparameter Tuning
- Learnings and outcomes
- Results and comparative analysis amongst models

Analyzing Data

- Dataset Overview: Utilizes 210,294 news headlines from Huffington Post (2012-2022), each record featuring category, headline, and short description.
- Diversity in Categories: Encompasses 42 unique categories, providing a rich source for analysis. Top ten categories displayed for reference.
- Article Length Analysis: Articles vary significantly in length, averaging around 200 words, with some under 100 and others over 350 words, highlighting content diversity.
- Class Imbalance Insight: Notable class imbalance observed, with "Politics" having the highest count (approx. 31,000 articles) and "Home & Living" the lowest (around 4,000).
- Exploratory Data Analysis:
  - Analysis of average article lengths by category, guiding model feature and parameter selection.
  - Word cloud visualization to identify frequent themes, assisting in feature selection and addressing class imbalances.

Data Preprocessing

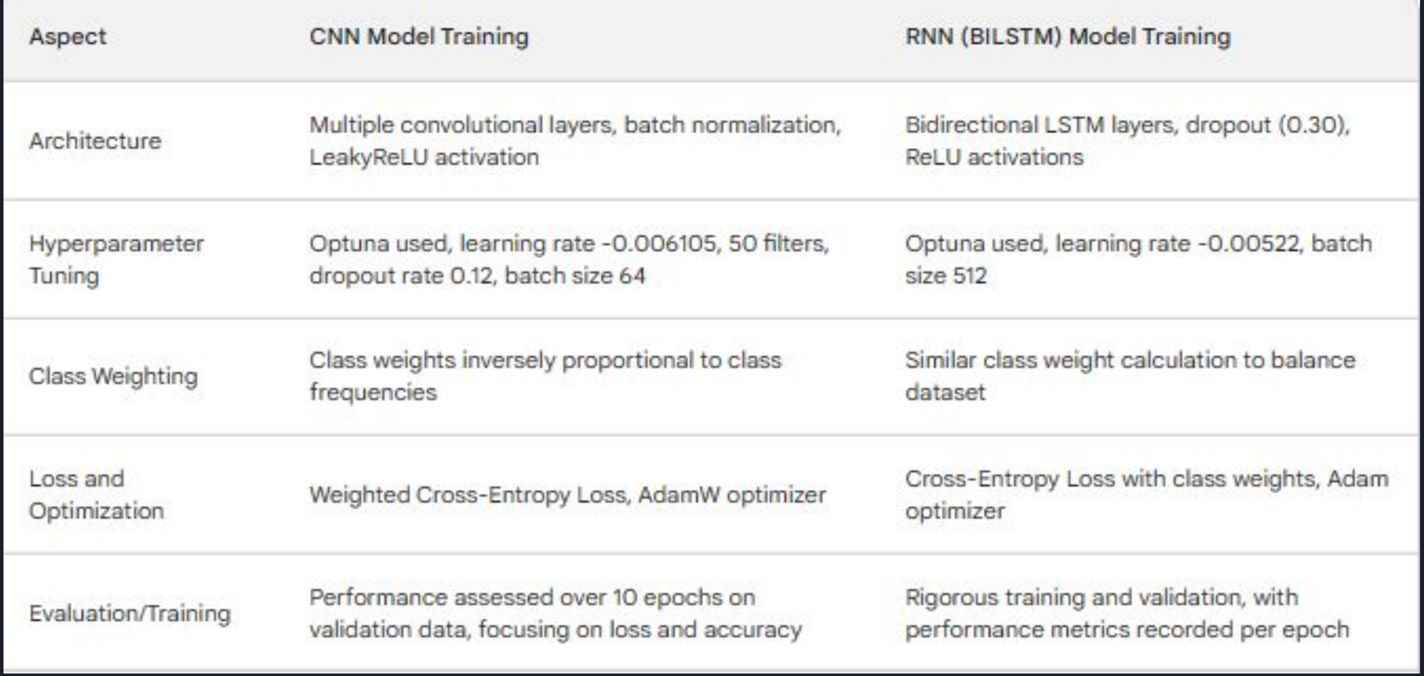
Steps involved in the data pre-processing:

- **Text Cleaning:** Removed stopwords, HTML tags, special characters, punctuation, and symbols for cleaner data.
- **Numbers Removal:** Replaced numbers with ‘#’ for uniform numerical representation.
- **Expanding Contractions:** Transformed shortened word forms to full versions to reduce ambiguity.
- **Word Lemmatization:** Simplified words to their base form, reducing vocabulary complexity.
- **Class Imbalance Mitigation:** Merged similar classes to enhance model generalization.
- **Outcome:** Streamlined and consistent dataset, improving model accuracy and efficiency in handling diverse news content.

Feature Engineering

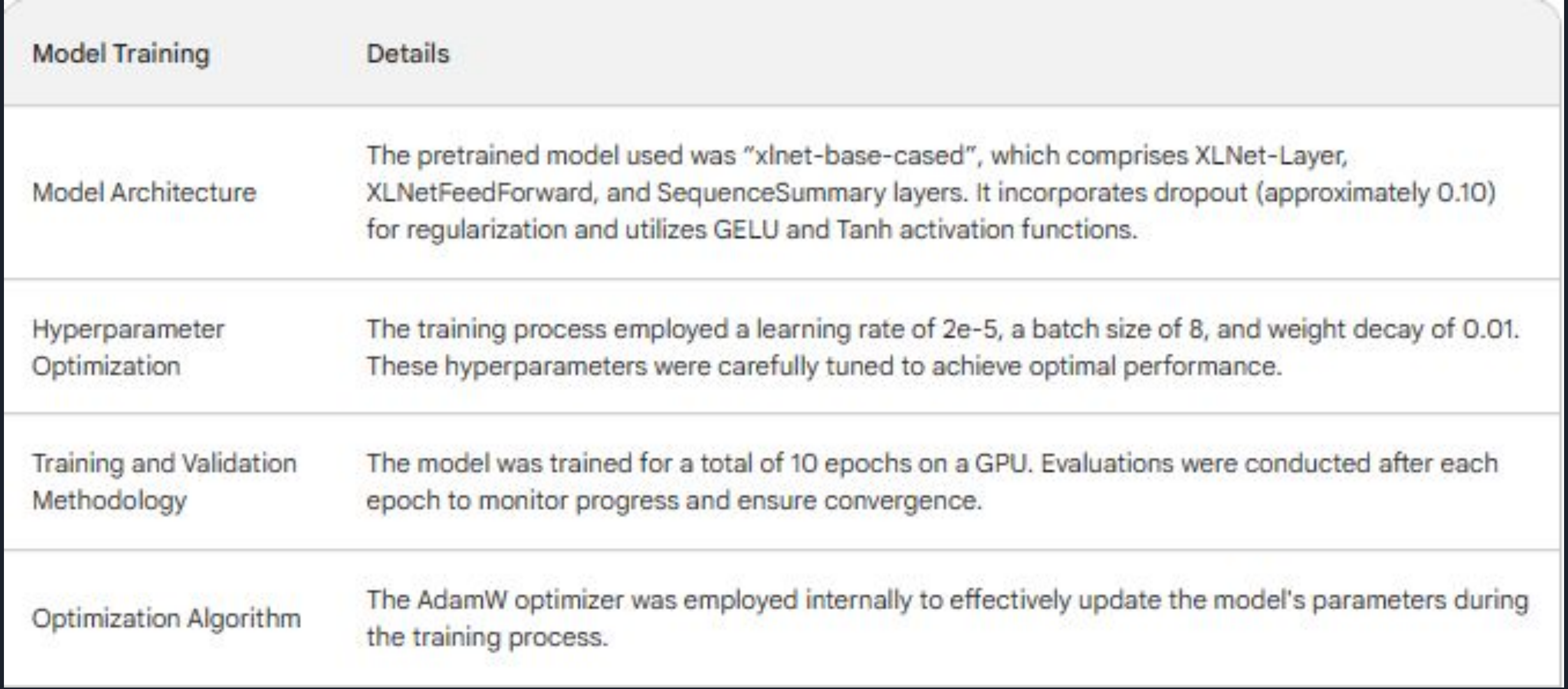
- **Data Preparation:** ’headline’ and ’short-description’ columns are combined into ’full-review’, removing any null values.
- **Category Filtering:** Targeted categories with more than 5000 reviews to ensure a balanced dataset.
- **Tokenization and Padding:** Text data are converted into padded integer sequences using Keras Tokenizer in the case of CNN and RNN and using XLNetTokenizer in the case of Transformer.
- **Word Embeddings:** Implemented 300-dimensional GloVe embeddings for enhanced word representation (for CNN and RNN).

Model Training (CNN vs RNN)



Aspect	CNN Model Training	RNN (BILSTM) Model Training
Architecture	Multiple convolutional layers, batch normalization, LeakyReLU activation	Bidirectional LSTM layers, dropout (0.30), ReLU activations
Hyperparameter Tuning	Optuna used, learning rate -0.006105, 50 filters, dropout rate 0.12, batch size 64	Optuna used, learning rate -0.00522, batch size 512
Class Weighting	Class weights inversely proportional to class frequencies	Similar class weight calculation to balance dataset
Loss and Optimization	Weighted Cross-Entropy Loss, AdamW optimizer	Cross-Entropy Loss with class weights, Adam optimizer
Evaluation/Training	Performance assessed over 10 epochs on validation data, focusing on loss and accuracy	Rigorous training and validation, with performance metrics recorded per epoch

Model Training (Transformers)



Model Training	Details
Model Architecture	The pretrained model used was "xlnet-base-cased", which comprises XLNet-Layer, XLNetFeedForward, and SequenceSummary layers. It incorporates dropout (approximately 0.10) for regularization and utilizes GELU and Tanh activation functions.
Hyperparameter Optimization	The training process employed a learning rate of 2e-5, a batch size of 8, and weight decay of 0.01. These hyperparameters were carefully tuned to achieve optimal performance.
Training and Validation Methodology	The model was trained for a total of 10 epochs on a GPU. Evaluations were conducted after each epoch to monitor progress and ensure convergence.
Optimization Algorithm	The AdamW optimizer was employed internally to effectively update the model's parameters during the training process.

Challenges and Obstacles in Handling Imbalanced Datasets

Overview:

- Imbalanced datasets pose a significant challenge in classification tasks due to the uneven distribution of classes.

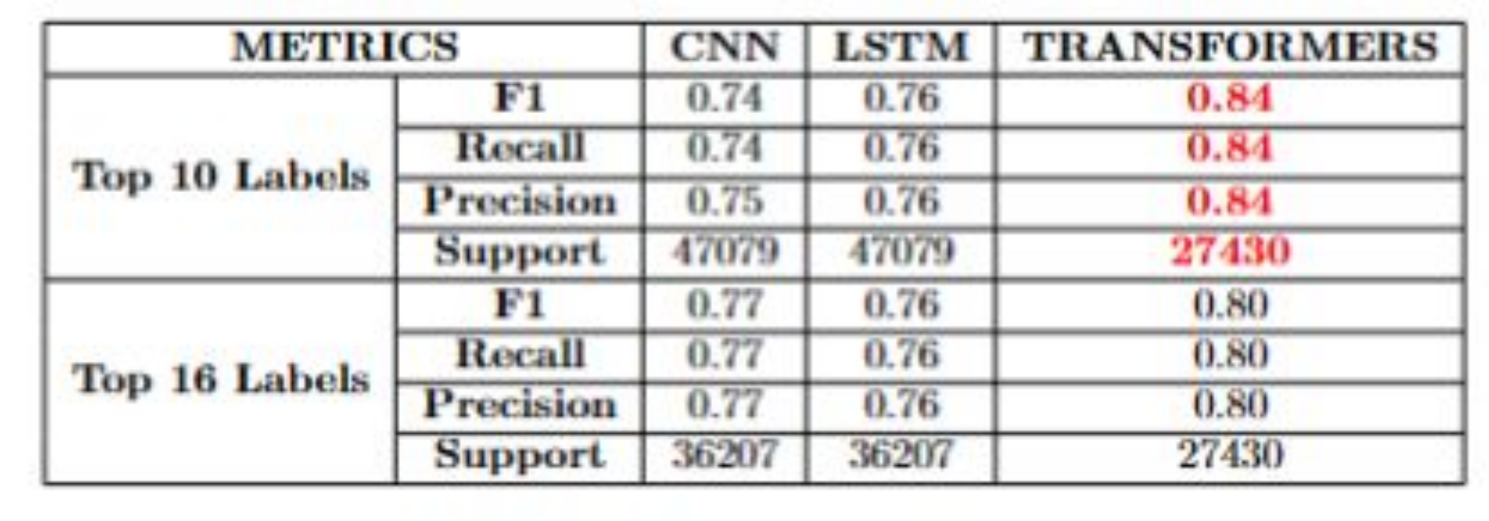
Issues:

- **Biased Model Learning:** Models tend to prioritize accuracy for the majority class, neglecting minority classes.
- **Poor Generalization:** Models fail to generalize well to unseen data due to underrepresentation of minority classes.
- **Difficulty in Capturing Minority Class Patterns:** Limited representation of minority classes hinders pattern recognition.
- **Data Collection Challenges:** Acquiring balanced data can be difficult or costly.
- **Model Sensitivity to Noise:** Outliers in unbalanced datasets disproportionately affect minority classes.
- **Impact on Decision Threshold:** Default thresholds may not align with unbalanced datasets, affecting precision and recall.
- **Difficulty in Model Interpretability:** Unbalanced datasets complicate decision boundary interpretation and feature importance assessment.

Approaches to overcome the challenges faced:

- **Undersampling:** Strategically reducing majority class samples, but may lead to information loss.
- **Hyperparameter Tuning:** Optimizing model parameters and using techniques like weighted loss functions.
- **Combining Minority Classes:** Consolidating similar classes to enhance generalization and interpretability.
- **Transfer Learning in Transformers:** Pre-trained XLNet Transformer effectively handles unbalanced datasets.

Performance Evaluation



METRICS		CNN	LSTM	TRANSFORMERS
Top 10 Labels	F1	0.74	0.76	0.84
	Recall	0.74	0.76	0.84
	Precision	0.75	0.76	0.84
	Support	47079	47079	27430
Top 16 Labels	F1	0.77	0.76	0.80
	Recall	0.77	0.76	0.80
	Precision	0.77	0.76	0.80
	Support	36207	36207	27430

Concluding Remarks

- **Innovative Approach:** Successfully integrated CNN and RNN (BiLSTM) models for advanced news article classification.
- **Enhanced Data Processing:** Implemented thorough preprocessing and exploratory data analysis, ensuring high-quality input for model training.
- **Model Performance:** Both models demonstrated robust accuracy and efficiency, with detailed comparative analysis highlighting unique strengths.
- **Practical Benefits:** Automated classification significantly improves news content management and user experience in digital news consumption.
- **Future Prospects:** Lays groundwork for broader NLP applications and potential adaptation to varied textual content.
- **Challenges and Learnings:** Overcoming project challenges enriched our understanding of NLP's dynamic nature and the importance of continual innovation.

Thank you !

Open for Q/A