An Intelligent Framework for Text Mining and Analysis Using Deep Learning

Zeyad Farooq Lutfi[1] , Muna Abdul Hussain Radhi[2] ,Esraa Jaffar Baker[3] , Samira Abdul-Kader Hussain[4]

[1,2,3.4]***Computer Science Department, Collage of Science, Mustansiriyah University, Baghdad-Iraq***

[1]zeyadfa6@uomustansiriyah.edu.iq

[2]muna.ali@uomustansiriyah.edu.iq

[3]es-alshaibany@uomustansiriyah.edu.iq

[4]samiracs@uomustansiriyah.edu.iq

Abstract

The rapid digital news expansion compels the determination of powerful large-scale media analytics. Despite an abundance of news topic classification use cases, the frameworks for evaluating Bare Bones Traditional Machine Learning Protocols for Machine Learning model tradeoffs tend to emphasize predictability while ignoring tradeoffs for both efficiency and reproducibility of the framework.

This paper presents a relatively simple and lightweight news mining framework leveraging DistilBERT.

The DistilBERT framework evaluated using the AG News Dataset publicly available and has been experimental with a multi-seed approach (42, 7, 21) for the purpose of establishing more systematic rigor. It also provides a detailed account of metrics recorded and compared against standard classification techniques. Utilized were TF-IDF and both Logistic Regression and Linear Support Vector Machines (L.S.V.M.).

For the developed DistilBERT framework, a mean accuracy of 0.956 ± 0.005 and mean Macro-F1 of ± 0.004 were achieved. Perhaps of most importance is the DistilBERT framework presents a trade off for being lightweight. To study this trade off, we analyze the dual framework edges characterized by lesser classification and news category discovery.
The full confusion matrices and class-wise performance evaluations illustrate that this approach is both thorough and deeply robust for all news.

Keywords

News Mining, Topic Classification, DistilBERT, Knowledge Distillation, Media Analytics, Reproducible Research, Machine Learning Benchmarking.

1. Introduction

The ever-increasing content on the internet news demands rapid mining technologies that scale with the volume. Such frameworks allow the analysis of news items, media tracking, and analysis of elaborate data to inform actionable decisions. Mining or organizing massive data sets that are unstructured, as they grow uncontrollably, has come to rely on news related topic categorization. This impacts personalized news recommendation, trend prediction, suppression of news and data leaks, and media analysis that incorporate artificial intelligence or other intelligent technologies. Therefore, classifying news items is key for addressing the problem of information overload and drawing conclusions that make sense from the digital stream [1],[2].

Through the process of knowledge distilling [3], DistilBERT is a notoriously light framework that works by approximating larger models, compressing the model by limiting the number of parameters baked into the model, and then by doing model running by limiting the time spent in the inference. "Single-run accuracy" has seen popularity, though, and has therefore limited news categorization processes. Many neglect experimental reproducibility, the stability of the results, and elaborate benchmark evaluations against AI. Constructing models of this sort is highly untrustworthy and does not reflect the number of real-world applications of AI [4], [5].

1.1 Research Gap and Contributions

A comprehensive review of the available studies identifies numerous gaps in news categorization. Few frameworks with multi-seed critical evaluations that facilitate reproducibility have been established. Next, comparisons between lighter transformers and more traditional baselines (SVMs, Logistic Regression, etc.) lack cohesion, or, at the very least, have poor consistency in the experimental setups. The computational aspect (training/inference time, in conjunction with the prediction gain) has been largely ignored in scenarios where the available resources (e.g., media) have been severely limited. Finally, the literature has not identified specific processes of reuse that have been clearly established[5,6,7].

This paper focuses on the “intelligent and repeated” classification of news topics by using the fine-tuned DistilBERT. Here are the main contributions of this paper:

- **Methodological Rigor:** The avoidance of “lucky” initializations is resolved by the multi-seed evaluation.
- **Systematic Benchmarking:** A true "head-to-head" comparison between TF-IDF based Logistic Regression and Linear SVMs has been made.
- **Multi-Dimensional Evaluation:** A full graded assessment is provided by using the confusion matrices, class-wise performance assessment, and the visibility of the individual metrics (Accuracy, Macro-F1, and Weighted-F1).

- **Efficiency-Aware Analysis:** By the use of the granularity approach, the assessment of the operational feasibility of the model is made by the detailed training/inference time.

The study uses an efficient pipeline to combine cutting-edge deep learning with real-world news analytics.
The remainder of this paper is structured as follows: Section 2 explores related work in news classification and transformer architectures. Section 3 describes DistilBERT's preprocessing and training. Section 4 details the experimental and baseline comparisons. Section 5 follows with the analysis of the performance, stability, and efficiency of the system. Section 6 proposes future areas for research.

The outlined pipeline makes use of a streamlined pipeline that makes the incorporation of higher order deep learning models amenable to advanced practical news analysis.

The subsequent sections of the paper explain the related research in news classification and the transformer models. Section 2 deals with the related works, Section 3 deals with the pre-processing and training of DistilBERT. Section 4 deals with the experiments and baseline settings and Section 5 deals with conclusion and future research suggestions.

2. Related Work

Statistical linguistic heuristics and neural systems for news topic classification. Section contextualizes the framework in mining and computational linguistics.

2.1 Statistical Baselines and Early Neural Architectures

SVMs and Linear Classifiers utilized sparse vectors through news categorization techniques like the bag-of-words and TF-IDF. These techniques tend to be simple and efficient. However, news categorization experiences difficulties in terms of semantic and temporal aspects in the data. Hence, they tend to be simplistic in terms of continuous and temporally distant aspects.

To address the challenges posed by traditional methods, RCNN was suggested for sequence-based applications, while Kim [8] demonstrated CNNs for sentence classification, and the Hierarchical Architecture of Attention Networks [10] and the use of multi-level attention models to improve text interpretation.

Modern systems use more sophisticated con representation techniques.

2.2 The Transformer Revolution and Model Compression

The Natural Language Processing (NLP)paradigm shift occurred due to Transformer and self-attention [11]. Pioneers of pre-training and fine-tuning approaches, BERT [1], and RoBERTa [6] have achieved impressive results on many news corpora.

Knowledge distillation has shrunk models in size and speed. DistilBERT [12], which is 40% smaller and 60% faster than BERT, still preserves 97% of its performance capabilities. Despite the growing popularity of lightweight transformers, there is a gap in scientific rigor (see Table 1) that includes statistical stability and reasonable benchmarking against baselines.

Table 1. Comparative Analysis of Representative News Topic Classification Approaches and the Proposed Framework

Study	Task / Dataset	Model(s)	Classical Baselines	Reproducibility (Multi-seed/Stats)	Efficiency Metrics	Key Contribution
Kim (2014)	Sentiment Classification	CNN	Yes (N-grams)	Limited	Partial	Pioneered CNNs for NLP tasks.
Zhang et al. (2015)	AG News	Char-level CNN	Yes (BoW/TF-IDF)	Limited	Partial	Demonstrated scalability on large datasets.
Yang et al. (2016)	Document Classification	HAN	Yes	Limited	Partial	Introduced hierarchical attention.
Devlin et al. (2019)	General NLP	BERT	Not Emphasized	Not Emphasized	Not Emphasized	Established the pretraining paradigm.
Sanh et al. (2020)	Efficient NLP	DistilBERT	vs. BERT	Reported	Yes (Size/Speed)	Model compression via distillation.
Proposed Framework	AG News	DistilBERT (Fine-tuned)	Yes (LR, SVM)	Yes (Mean ± Std)	Yes (Train/Inference Time)	Efficiency-aware, reproducible pipeline.

2.3 Benchmarking and Reproducibility Gaps

In recent times, there has been a "reproducibility crisis" in applied NLP [13]. The stochastic initialization of weights could improve confidence because of "single-run" accuracy in most studies. In addition, the lack of timing during training/inference, as well as class-level error breakdowns, limits their application when resources are limited.

Table 1 provides a clear demonstration that the proposed framework applies multiple seed experiments and efficiency profiling.

3. Proposed DistilBERT-Based Framework

The news topic classification pipeline combines speedy transformers with statistical validation. Our system focuses on achieving prediction performance and resource allocation efficiency while ensuring predictability.

3.1 Framework Overview

The system design includes the critical features listed here that will assist in the conversion of raw data into an appropriate evaluation of:

o Preprocessing Agnosticism: News is normalized in all algorithms.

o Benchmarking Double Layer: Comparison of performance between the DistilBERT model and standard TF-IDF-based linear classifiers such as Logistic Regression and SVM.

o Stochastic Stability Element: The multiple seeding evaluation approach (S={42, 7, 21}) is incorporated to eliminate the effects of initialization bias.

o Performance Assessment: The focus here is to determine the effect of training and inference latencies on the practicality of the system.

Figure 1. Workflow Diagram of the Proposed System Design.

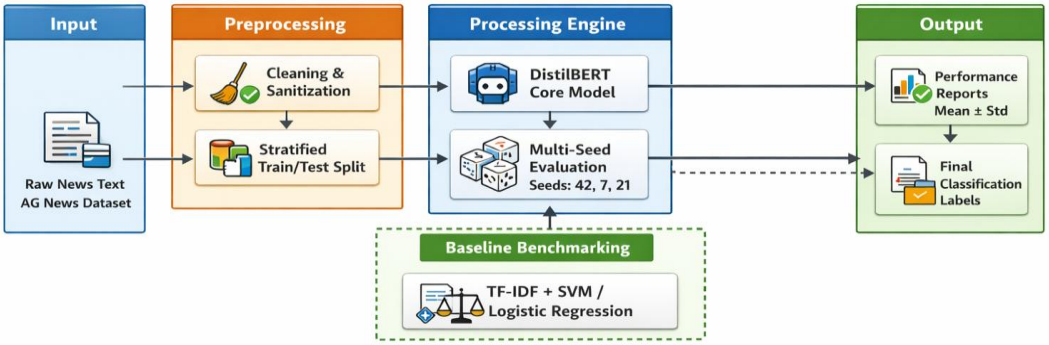

Figure 1: Proposed DistilBERT Based Architecture for News Classification Framework

Figure 2 illustrates a sequence diagram for evaluating a news classification system using DistilBERT based data retrieval.

First, the researcher obtains the AG News dataset where preprocessing techniques have already been applied to eliminate text and special characters, as well as perform stratified splitting. Afterward, TF-IDF + Logistic Regression / Linear SVM baseline models are trained alongside with fine-tuning the DistilBERT architecture under several random seeds to increase the reliability of findings. All four metrics, such as accuracy, Macro-F1, weighted-F1, and confusion matrices, are used to measure the prediction accuracy of all aforementioned models. Additionally, training and inference times are

evaluated under different random seeds, making this approach robust and repeatable.

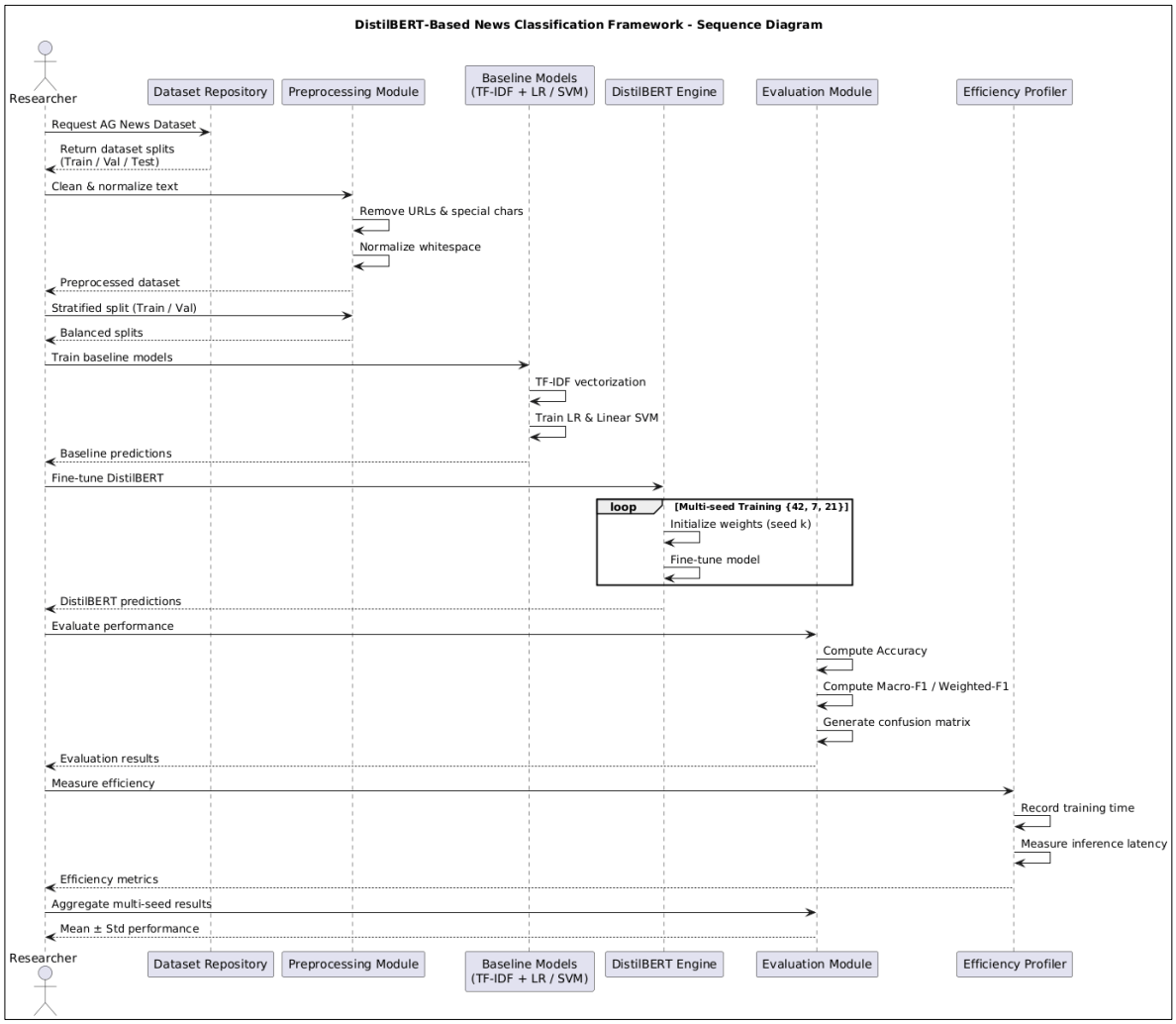

Figure 2: Sequence diagram of the proposed system.

3.2 Data Preprocessing and Dataset Characteristics

For data quality control, leading and trailing spaces, hyperlinks, special characters, and extra spaces were cleaned.

Experimented using AG News Dataset. The training set was further divided into training and validation sets to ensure experimental rigour; however, the test set remained the same as originally provided. To avoid model bias, Table 2 shows the uniform distribution of classes in the dataset, refer Figure 3.

Table 2. Label Distribution Across Dataset Splits

Subset	World	Sports	Business	Sci/Tech	Total
Training	26,991	26,966	27,100	26,943	108,000
Validation	3,009	3,034	2,900	3,057	12,000
Test	1,900	1,900	1,900	1,900	7,600

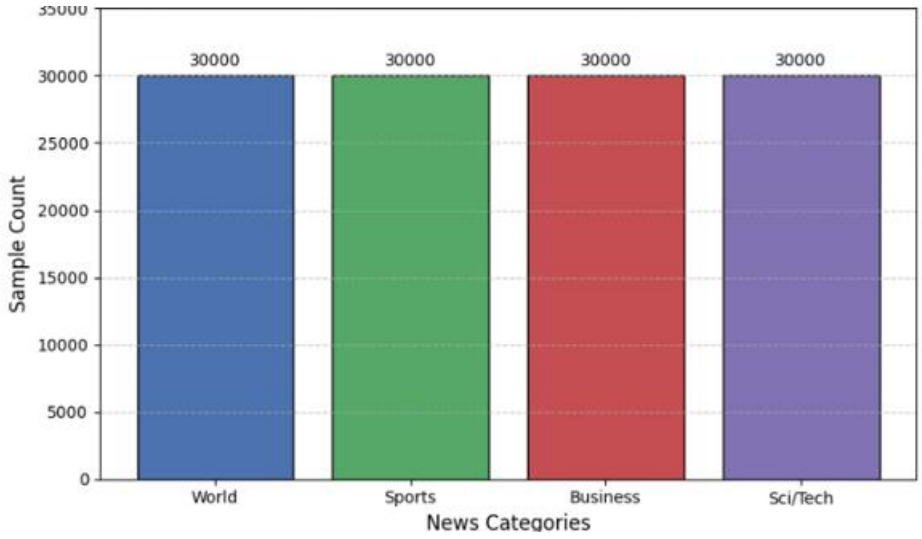

Figure 3 class distribution bar chart for AG News.

Figure 3 show Macro-F1 is fit for evaluation since each area (World, Sports, Business, Sci/Tech) has ~30,000 samples, ensuring balance and verifying its use.

3.3 Model Architecture and Mathematical Formulation

Framework uses DistilBERT-base-uncased architecture.The model processes input sequences to generate a conualized representation, where$h \in R^{dh}$ represents the pooled output of the **[CLS] token**.

The classification head transforms this representation into logits via:

$z = W h + b$ ……(1)

where W and b are the learnable weight matrix and bias vector. Class probabilities are derived using the Softmax function, and the predicted label $\hat{y}$ is obtained by:

$\hat{y}$ = arg $\max_k$, softmax (zk) ……(2)

$$\hat{y} = \arg\max_k \text{softmax}(z_k)$$

where k in {1….., 4}.

4. Experimental Setup and Baseline Models

4.1 Configuration and Baselines

Two conventional baselines—TF-IDF + Logistic Regression and Linear SVM—were used to compare rigorously. To guarantee comparability and fairness, these models were trained using DistilBERT data splits, see figure 4.

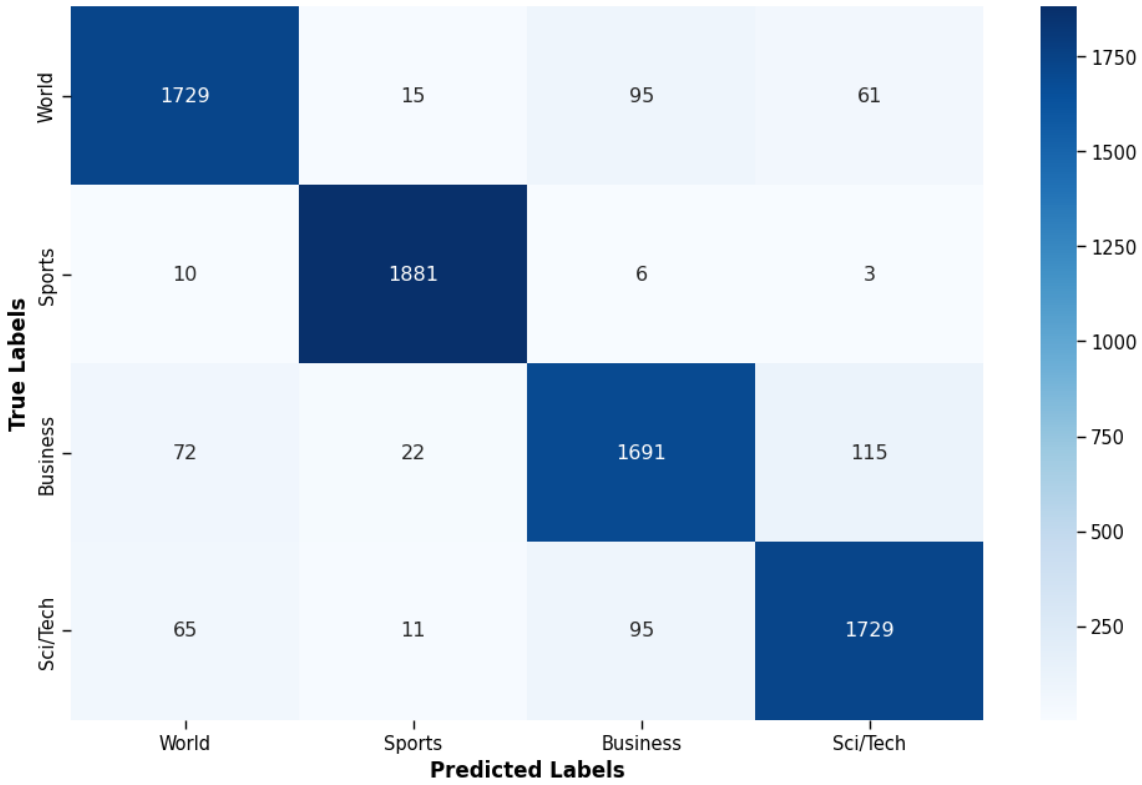

Figure4: Confusion Matrix of the Baseline Model (TF-IDF + Linear SVM)

4.2 Training Protocol

DistilBERT refined using Table 3 hyperparameters. Mixed-Precision Training (FP16) optimized GPU memory and accelerated convergence.

Table 3. Optimized Hyperparameters for Fine-Tuning DistilBERT

Parameter	**Configuration**
Optimizer	AdamW
Learning Rate	(2 × 10^ − 5
Batch Size	16 (Training) / 32 (Evaluation)
Maximum Sequence Length	128 tokens

Training Epochs	3
Evaluation Strategy	Multi-seed aggregation (n = 3)

5. Experimental Results and Comparative Analysis

5.1 Baseline Performance

The results obtained using the TF-IDF + Linear SVM are presented in table 4. In-depth analysis of the confusion matrix reveals that incorrect classifications mainly arise due to confusion between the "Business" and "Sci/Tech" classes. This demonstrates how difficult it is to classify articles with overlapping meanings of technology, especially business-oriented technologies, see figure 5.

Table 4. Class-wise Performance (TF-IDF + Linear SVM)

Class	Precision	Recall	F1-score
World	0.95	0.91	0.93
Sports	0.97	0.99	0.98
Business	0.90	0.89	0.89
Sci/Tech	0.93	0.91	0.92
Macro Average	**0.94**	**0.93**	**0.93**

Figure 5: Performance vs. Efficiency Trade-off

5.2 DistilBERT Statistical Stability

DistilBERT greater resilience in all stochastic runs. Table 5's results (mean ± std) indicate a stable model with a minimal standard deviation.

Table 5. Aggregated DistilBERT Results (Mean ± Std)

Metric	Mean ± Std
Accuracy	0.956 ± 0.005
Macro-F1	0.953 ± 0.004
Weighted-F1	0.955 ± 0.005

5.3 Comparative Evaluation and Efficiency Analysis

Final comparison in Table 6 shows performance-efficiency trade-off. While DistilBERT needs further training, (1,800s), it offers a significant 2.6% - 3.6% gain in Macro-F1 over classical baselines. Most crucially, a 25s inference delay for all test sets.(approx. 3.2ms per sample) makes it highly feasible for real-time media analytics systems.

Table 6. Final Comparative Matrix: Baselines vs. DistilBERT

Model	Accuracy	Macro-F1	Train Time (s)	Inference Time (s)
TF-IDF + Logistic Regression	0.92	0.91	120	1.2
TF-IDF + Linear SVM	0.93	0.93	350	1.8
DistilBERT	0.956	0.953	1,800	25.0

5.3 Discussion

The experimental evidence proves that the suggested DistilBERT model-based framework is more efficient in news topic classification than traditional TF-IDF-based models. The steady rise in all three measures of Accuracy, Macro-F1, and Weighted-F1, suggests that contextual representations are able to model the relationship between semantics that cannot be easily represented in linear models, especially between overlapping categories (like Business and Sci/Tech).

The main advantage of the suggested approach will be its statistical stability. The multi-seed assessment plan has low standard deviations on all metrics, which proves that the performance improvements obtained are not weak and occur because of positive random initialisation. This is in response to one of the limitations of previous research that uses single-run assessments. In terms of efficiency, though, despite needing more training, DistilBERT has lower inference latency, which makes it appropriate to real-time or near-real-time media analytics. This is a trade-off that is acceptable considering the high level of performance gains and reliability. Comprehensively, the paper points out the fact that lightweight transformers when considered in strict and reproducible experimental conditions bring a viable trade-off in terms of accuracy, efficiency and deployability. The evaluation should be continued in future on more datasets and domains to have further evaluation of the generalization capability.

6. Conclusion

The method utilizes DistilBERT1 for efficient classification of news article topics. After an exhaustive benchmark comparison and multi-seed statistical analysis of the architecture, the method proved to be both adaptable and efficient.

Media analytics can use lightweight transformer models to bridge the gaps between traditional ML and deep learning architecture analytics using confusion matrices, per-class statistics, and comparison of training and inference times.

The results show that the approach outperforms the classical baselines with an average accuracy of 0.956 ± 0.005 and a Macro-F1 score of $0.953 \pm 0.004$4. Because of the lower implementation costs of the approach, as evident from the results, it is most appropriate for low-resource constrained environments. The well-structured usable, reproducible reference baseline for this approach advances the field of intelligent media analytics by integrating both prediction accuracy and computational efficiency.

To validate the approach externally, future work will focus on multilingual and specialized domain news collections. Explainable AI is of practical value when it provides interpretability and transparency. In the analysis of live news events, lightweight transformers, and online/incremental learning, real-time and adaptive lightweight transformers could be of use.

References

[1] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).

[2] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Transformers: State-of-the-art natural language processing. InProceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations 2020 Oct (pp. 38-45).

[3] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2.

[4 Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning--based classification: a comprehensive review. ACM computing surveys (CSUR). 2021 Apr 17;54(3):1-40.

[5] Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A survey on classification: From shallow to deep learning. arXiv preprint arXiv:2008.00364. 2020 Aug 2.

[6] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.

[7] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for classification. Advances in neural information processing systems. 2015;28.

[8] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. 2014 Aug 25.

[9] Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for classification. InProceedings of the AAAI conference on artificial intelligence 2015 Feb 19 (Vol. 29, No. 1).

[10] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. InProceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies 2016 Jun (pp. 1480-1489).

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[12] Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D. Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984. 2020 Apr 6.

[13] Ethayarajh K. How conual are conualized word representations. Comparing the geometry of BERT, ELMo, and GPT-2 Embeddings. 2019 Sep;2.