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 exponential growth of digital news material has increased the demand for strong, efficient, and scalable mining methods for large-scale media analytics and news intelligence systems. Deep learning-based language models have performed well in news topic classification, but current literature prioritizes prediction accuracy over experimental reproducibility, computational efficiency, and tough benchmarking vs. conventional machine learning benchmarks.

The present research proposes a carefully and repeatably designed news mining framework based on the lightweight DistilBERT transformer architecture.

The proposed methodology verified using the famous AG News dataset. The systems use multi-seed. Experimental evaluation to improve methodological soundness. (Seeds: 42, 7, 21), detailed metric reporting, and a fair comparative analysis against traditional classifications, specifically TF-IDF combined with Logistic Regression and Linear Support Vector Machines (SVM).

DistilBERT-based framework outperforms baselines in experiments, attaining a mean accuracy of 0.956 ± 0.005 and a Macro-F1 score of 0.953 ± 0.004. Notably, this framework is suitable for resource-constrained applications since it has less processing overhead than larger transformer structures. Complete confusion matrices and class-wise performance assessments demonstrate the approach's statistical stability and resilience across all news categories.

Keywords

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

1. Introduction

Fast-growing online news content requires efficient and scalable mining technologies. Frameworks enable automatic news analysis, media monitoring, and data-driven decision-making. News topic categorization is now the major method mining organizes massive unstructured data. Personalized news suggestion, trend forecasting, misinformation mitigation, and intelligent media analytics are affected. Thus, classifying news articles is essential for managing information overload and drawing useful conclusions from digital streams[1],[2].

DistilBERT is a common lightweight transformer architecture for simulating big models. DistilBERT reduces parameter count and inference latency while retaining most of its performance via knowledge distillation [3]. Though popular, "single-run" accuracy restricts news classification research. Many disregard experimental reproducibility, statistical stability, and extensive benchmarking against machine learning baselines. Modelling without multi-seed assessment, class-wise error analysis, and systematic efficiency evaluations lacks scientific credibility and real-world applicability[4], [5].

1.1 Research Gap and Contributions

An analysis of existing literature shows many news classification research gaps. Replicable experimental frameworks with multi-seed evaluations for statistical reliability are few. Second, lightweight transformer-traditional baseline comparisons (e.g., SVM, Logistic Regression) are typically fragmented or undertaken under inconsistent experimental settings. The computational dimension—training/inference time vs prediction gain—is commonly disregarded in resource-constrained media situations. Lastly, research lacks clear reuse processes[5,6,7].

DistilBERT fine-tuning is used to classify news topics intelligently and repeatedly in this article. Paper's main contributions:

- **Methodological Rigor:** Utilizing multi-seed evaluation to avoid "lucky" initializations.
- **Systematic Benchmarking:** Conducting a "head-to-head" comparison with TF-IDF-based Logistic Regression and Linear SVM under identical environmental settings.
- **Multi-Dimensional Evaluation:** Providing a granular analysis through Accuracy, Macro-F1, and Weighted-F1 metrics, supplemented by confusion matrix visualizations and class-specific performance breakdowns.

- **Efficiency-Aware Analysis:** Quantifying the operational feasibility of the model through detailed training and inference time assessments.

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 includes experimental and baseline settings, whereas Section 5 offers performance, stability, and efficiency data. Section 6 concludes with research suggestions.

2. Related Work

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

2.1 From Statistical Baselines to Early Neural Architectures

Sparse vectors were utilized for traditional news classification. as Bag-of-Words (BoW) and TF-IDF, together with linear classifiers like Logistic Regression and Support Vector Machines. Although computationally efficient and interpretable, these methods cannot capture conual intricacies and long-term semantic linkages in diverse news corpora.

To circumvent these constraints, early neural methods included hierarchical feature learning. Recurrent CNNs (RCNNs) were proposed for sequential [8, 9], while Kim [8] showed that CNNs can classify sentences. HAN hierarchy [10]. Improve text interpretation using multi-level attention models. Modern systems employ deeper con representations than these models.

2.2 The Transformer Revolution and Model Compression

Natural Language Processing was transformed by Transformer and self-attention [11]. Pretraining-and-fine-tuning pioneers BERT [1] and RoBERTa [6] performed well on numerous news datasets.

Knowledge distillation made models small and efficient. While 40% smaller and 60% quicker than BERT, DistilBERT [12] retains 97% of its performance. Though lightweight transformers are becoming increasingly popular, a critical literature review (Table 1) shows a lack of scientific rigor, notably in statistical stability and fair benchmarking with 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 NL tasks.
Zhang et al. (2015)	AG News	Char-level CNN	Yes (BoW/TF-IDF)	Limited	Partial	Demonstrat scalability o 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 compressio via distillation.
Proposed Framework	AG News	DistilBERT (Fine-tuned)	Yes (LR, SVM)	Yes (Mean ± Std)	Yes (Train/Inference Time)	Efficiency-aware, reproducibl pipeline.

2.3 Benchmarking and Reproducibility Gaps

Recently, applied NLP has had a "reproducibility crisis" [13]. Stochastic weight initialization may boost confidence due to "single-run" accuracy in many investigations. The absence of training/inference time measurements and class-wise error analysis limits these models' resource-constrained applicability.

Table 1 illustrates that the suggested framework uses multi-seed experimental approaches and systematic efficiency profiling to bridge theoretical and real deploy ability.

3. Proposed DistilBERT-Based Framework

The end-to-end news topic classification pipeline employs a fast transformer and thorough statistical validation. Our system balances prediction performance and resource usage while protecting repeatability.

3.1 Framework Overview

The design includes the following fundamental components to move raw data to robust evaluation:

- Agnostic Preprocessing: Standardizing news across all models.
- Dual-Tier Benchmarking: Comparing DistilBERT against traditional TF-IDF-based linear classifiers (Logistic Regression and SVM).
- Stochastic Stability Module: Implementing a multi-seed evaluation strategy (S={42, 7, 21}) to mitigate initialization bias.
- Efficiency Profiling: Measuring training and inference latency to assess real-world deployability.

Figure 1. Procedural Flow of the Proposed Framework. The architecture exhibits raw news input, agnostic preprocessing, stratified splitting, and DistilBERT engine multi-seed fine-tuning. The framework averages stochastic initialization data for repeatable performance measures.

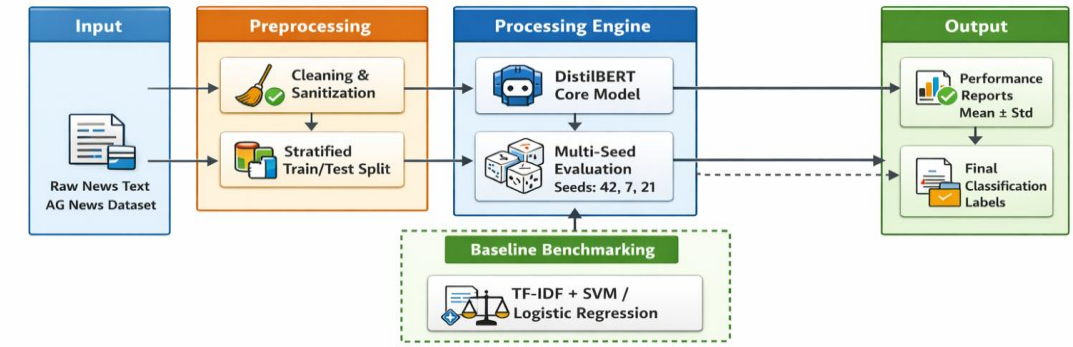

Figure 1: Proposed DistilBERT-Based News Classification Framework Architecture

Figure 2 illustrates the sequence diagram displays the process of the news classification system based on DistilBERT data retrieval to performance analysis. The researcher gets the AG News dataset which has been preprocessed to clean text, special characters, and stratified splitting. TF-IDF + Logistic Regression / Linear SVM are trained using the baseline models and DistilBERT is fine-tuned using various random seeds to ensure statistical stability. Accuracy, Macro-F1, weighted-F1 and confusion matrices are used to evaluate predictions of all models. Training and inference times are profiled, with multi-seed results combined to give strong and repeatable performance measures. It is a sequence that provides

a framework of news classification that is systematic, reliable, and deployable.

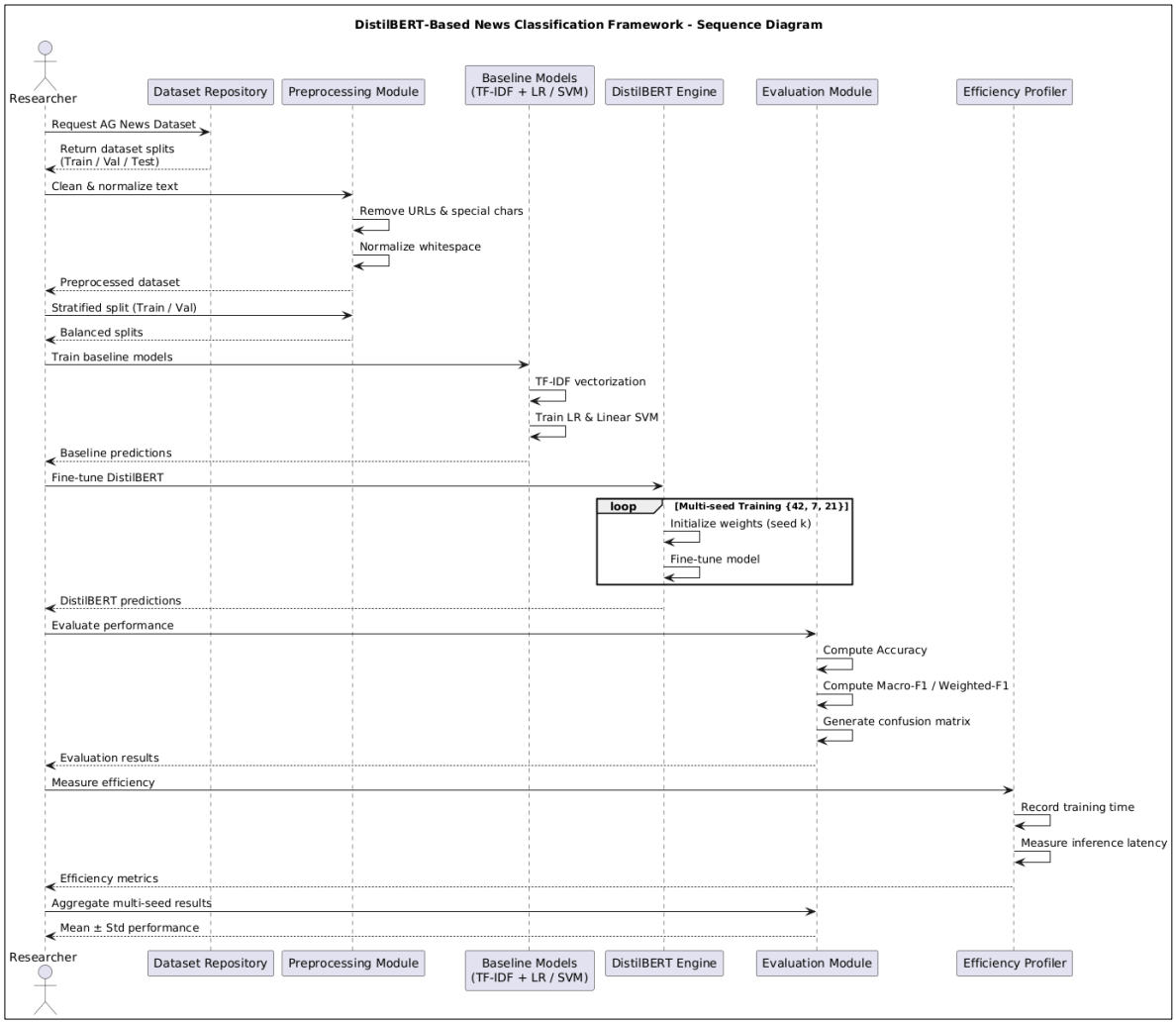

Figure 2: Sequence diagram of the proposed system.

3.2 Data Preprocessing and Dataset Characteristics

To assure data quality, leading/trailing whitespace, URLs, special characters, and redundant whitespace were removed and normalized. Using AG News experimentally. The training fraction was stratified into training and validation sets to guarantee experimental integrity, but the original test split was kept for final evaluation. Avoiding model bias, Table 2 demonstrates the dataset's near-uniform class distribution, see 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

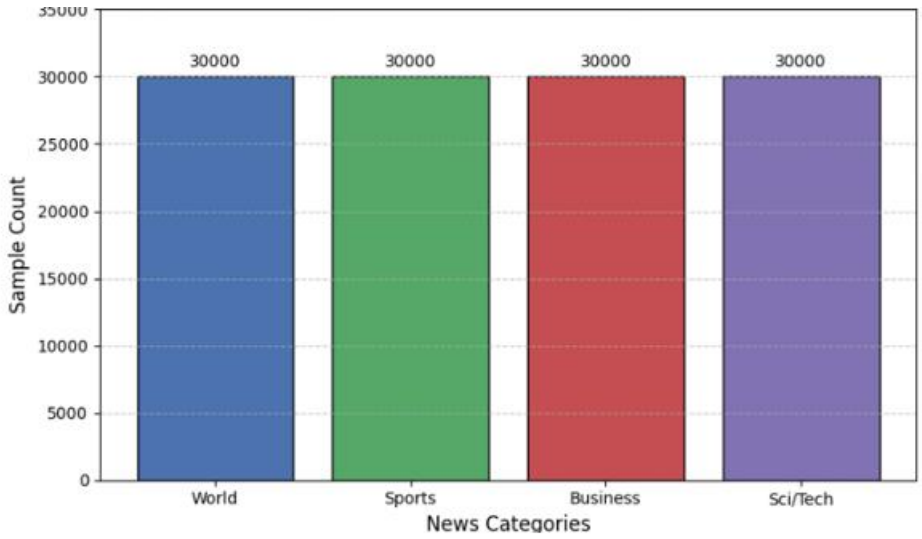

Figure 3 class distribution bar chart for AG News.

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.

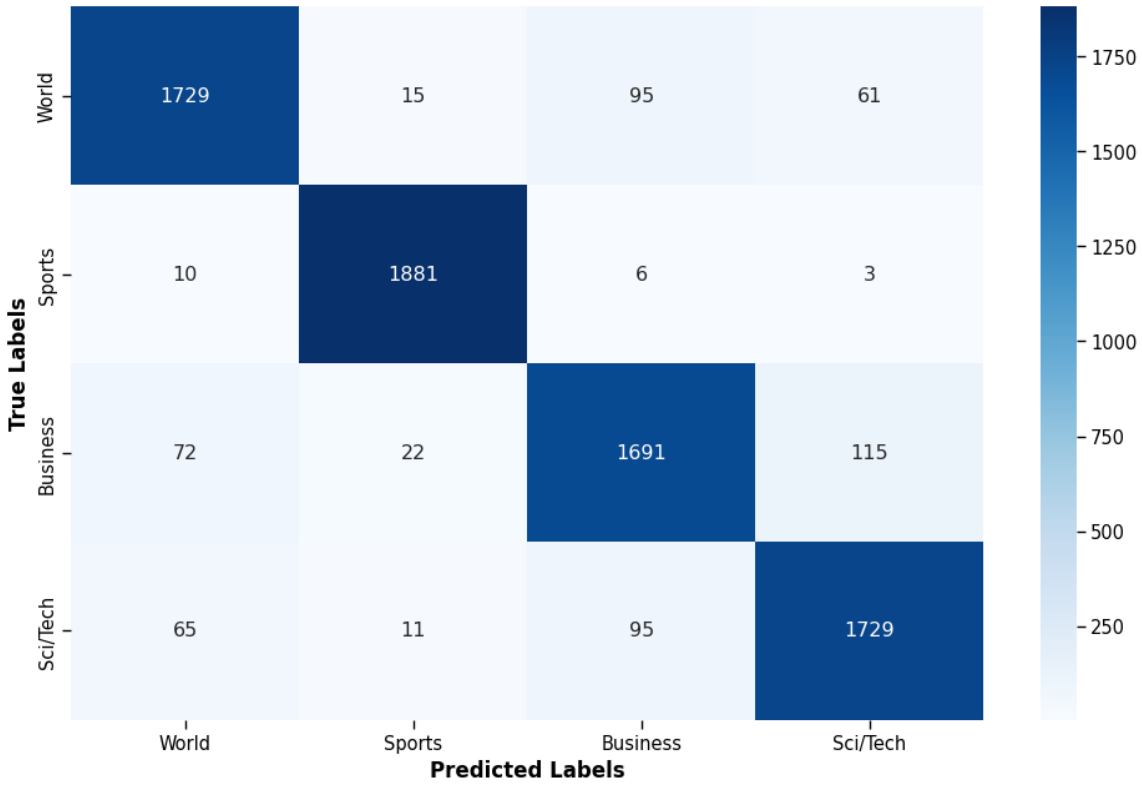

Figure4: Confusion Matrix of the Baseline Model (TF-IDF + Linear SVM)

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 TF-IDF + Linear SVM initial performance is shown in Table 4. A detailed investigation of the Confusion Matrix shows that misclassifications mostly occur between "Business" and "Sci/Tech" categories, despite the model's 0.93 accuracy. The semantic overlap in technological business news challenges conventional frameworks, 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

This research used DistilBERT1 to effectively categorize news topics. Due to rigorous testing against classical baselines and multi-seed statistical assessment, the framework is predictive and computationally efficient.

In media analytics, lightweight transformer models may bridge the gap between standard machine learning and large-scale deep learning models, according to confusion matrices, class-wise metrics, and training/inference time evaluations.

Results show that DistilBERT-based framework beats baselines with a mean accuracy of 0.956 ± 0.005 and a Macro-F1 score of $0.953 \pm 0.004$4. It's appropriate for resource-constrained applications since it has less computational overhead than larger transformer topologies. The well-documented, repeatable reference baseline in this study establishes intelligent media analytics research by combining predicted effectiveness with computational cost.

For generalizability, future research will include multilingual and domain-specific news corpora. Explainable AI meets real-world needs by increasing transparency and interpretability. For adaptive and real-time analytics in dynamic news instances, lightweight transformer architectures and online or incremental learning may increase performance.

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.