SMOOTHED-MODERNBERT: CO-ATTENTIONAL SYNERGY OF PROBABILISTIC TOPIC MODELS AND MODERNBERT THROUGH DYNAMIC FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Document classification remains a critical challenge in natural language processing (NLP) as text volumes and thematic complexity escalate. Although transformer-based architectures like BERT excel at capturing contextual semantics, they often overlook the latent thematic structures inherent in document-level discourse. Conversely, probabilistic topic models effectively distill coarse-grained thematic patterns but struggle with nuanced contextual dependencies. To address these limitations, this study introduces a novel hybrid approach that synergizes the contextual depth of ModernBERT with the interpretable thematic representations of smoothed-Dirichlet-based topic models. Our model aligns token-level representations with document-level thematic distributions by optimizing contextual and topic objectives through a co-attention mechanism layer. By utilizing a dynamic fusion layer, where co-attention scores dynamically gate and blend BERT's embeddings with topic mixtures at each instance, the approach captures both fine-grained context and global theme interplay in a unified representation. Our method bridges a critical gap in the NLP methodology, paving the way for enhanced model generalizability in domains that require both thematic abstraction and contextual granularity. Empirical evaluations on benchmark corpora demonstrate consistent classification robustness over standalone approaches. To ensure the reproducibility of our experiments and encourage further research, we open-source our implementation code.

1 INTRODUCTION

Document classification is a fundamental task in natural language processing (NLP), which underpins applications such as news categorization, sentiment analysis, and information retrieval Devlin et al. (2019); Bao et al. (2019). Early methods relied on hand-crafted features and statistical models, but the exponential growth in text volume and complexity has driven a shift toward deep learning. Recurrent architectures such as LSTMs Clavié et al. (2021) and GRUs Ravanelli et al. (2018); Ahmed et al. (2023); Mortezapour Shiri et al. (2023) automated feature extraction and sequential patterns captured, although their inherently sequential nature limits parallelism and long-range dependency modeling Nam et al. (2017). Hybrid approaches that combine truncated attention with recurrent units or integrate self-attention into bidirectional GRUs have partially alleviated these issues Nam et al. (2017); Sun et al. (2019b); Jiang & Wang (2022).

The advent of transformers, particularly BERT with its multi-head self-attention and contextual embeddings Vaswani et al. (2017), has further transformed the field by allowing full parallel processing of entire sequences. Fine-tuning techniques have delivered state-of-the-art results across benchmarks Sun et al. (2019a); Wang et al. (2020a), and extensions combining BERT with capsule networks Wang et al. (2020b); Liu et al. (2012) or graph neural networks Li & Jia (2025); Qasim et al. (2022); Jamshidi et al. (2024); Davidson & Dym (2024) continue to push performance. Despite these advances, transformer models can overlook global thematic coherence, misread sarcasm or broader discourse, and offer limited interpretability. In contrast, probabilistic topic models Luo et al. (2022), such as smoothed Dirichlet distribution, identify coherent themes but struggle with contextual nuance Nallapati et al. (2007). Bridging this gap, hybrid frameworks like TopicBERT fuse Gaussian topic priors with BERT embeddings Chaudhary et al. (2020a), yet typically employ shallow concatenation that underutilizes the complementary strengths of each paradigm.

Despite these advances, reconciling token-level contextual precision with document-level thematic interpretability remains an open challenge. To this end, we propose Smoothed-ModernBER: co-attentional synergy of probabilistic topic models and ModernBERT through dynamic fusion (SD-MoBERT), a novel architecture that integrates ModernBERT with a dynamically smoothed Dirichlet topic model via a co-attentional synergy mechanism. Unlike prior shallow-fusion methods, our model jointly optimizes the dynamically fused contextual and thematic losses, fostering mutual reinforcement between granular semantics and global topics. We demonstrate that this integration yields better performance and interpretability across multiple benchmark corpora, bridging the methodological gap between contextual depth and thematic coherence. The main contributions of our studies are summarized as follows:

1. We propose a novel hybrid architecture that integrates ModernBERT's contextual semantics with smoothed-Dirichlet topic modeling, bridging neural and probabilistic paradigms to jointly optimize fine-grained context and interpretable thematic structures.
2. Dynamic co-attention fusion: We introduce a gated mechanism that dynamically blends token-level BERT embeddings with smoothed Dirichlet document-level topic mixtures, enabling adaptive weighting of local and global thematic information.
3. Empirical Validation and Reproducibility: We show that SD-MoBERT consistently outperforms baseline models and make our full implementation publicly available to facilitate future research and practical adoption `https://github.com/anonymousPapersSubmissions/Smoothed-ModernBERT`.

The remainder of this paper is organized as follows. Section 2 reviews related work on transformer encoders and topic modeling. Section 3 reviews the background studies, while Section 4 presents the proposed model. The experimental results and conclusion are presented in Section 5 and Section 6, respectively.

2 Related Work

Document classification has evolved through five key paradigms: traditional statistical methods, neural architectures, transformers, hybrid topic-neural frameworks, and co-attentional synergy. Early approaches relied on manually engineered features such as Bag-of-Words (BoW) Qader et al. (2019) and TF-IDF Christian et al. (2016), which quantified the importance of words, but ignored context. Bag-of-N-Grams Li et al. (2016) improved phrase representation, while bag-of-means models integrated word embeddings, although semantic nuances remained elusive.

Neural architectures addressed these limitations through character-level CNNs Zhang et al. (2015); Bielik et al. (2017), though fixed kernel sizes hindered long-range dependency modeling Yue et al. (2018). The compact CNN variants Talai & Kherici (2023) reduced parameters, but retained locality constraints. Sequential models such as LSTMs Clavié et al. (2021) and GRUs Michael et al. (2024) captured longer contexts but suffered from limited parallelism. Bidirectional GRUs with truncated drop loss Abbasi et al. (2024) mitigated class imbalance, yet sequential processing persisted as a bottleneck.

Transformers revolutionized the field via self-attention mechanisms Vaswani et al. (2017), with BERT Devlin et al. (2019) achieving state-of-the-art through bidirectional pre-training. ModernBERT Warner et al. (2024) scaled efficiency via flash attention but lacked document-level thematic coherence. Hybrid enhancements like capsule networks Wang et al. (2020b) and graph neural networks Li & Jia (2025) improved hierarchical features but struggled with global topic integration.

Hybrid topic-neural frameworks emerged to bridge thematic and contextual modeling. Probabilistic topic models (PTMs) Wang et al. (2022) like LDA Blei et al. (2003) abstracted themes but ignored word order. Neural topic models (NTMs) Wu et al. (2024); Ojo & Bouguila (2024) used variational autoencoders for continuous distributions, while Topic-BERT Chaudhary et al. (2020a) combined BERT with Gaussian topic vectors, a shallow fusion lacking synergy. Class-based TF-IDF clustering enhanced interpretability but limited classifier integration. Recent work includes SBERT-TM for short texts Cheng et al. (2023) and ensemble models like ENTM-TS Voskergian et al. (2024), though computational costs constrained scalability. Concurrently, co-attentional architectures Lee et al.

(2025) optimized feature fusion but overlooked probabilistic topic priors, leaving opportunities for deeper integration of thematic structure and contextual semantics.

2.1 Motivation: Toward Co-Attentional Synergy

Co-attention mechanisms have proven effective in multimodal reasoning by aligning heterogeneous representations Nam et al. (2017). However, their application to intra-textual fusion of context and themes remains underexplored. SD-MoBERT diverges from shallow fusion by employing a co-attention layer that dynamically aligns ModernBERT's token embeddings with smoothed Dirichlet topic distributions. This mutual reinforcement allows topic priors to guide attention heads toward thematically salient tokens, while contextual features refine topic coherence via variational inference. By unifying transformer efficiency, probabilistic topic modeling, and co-attentional interaction in a single, scalable architecture, SD-MoBERT transcends the limitations of each paradigm and offers a robust, informed solution for document classification in complex, heterogeneous corpora.

3 Background Studies

3.1 Smoothed Dirichlet Distribution (SD)

The smoothed Dirichlet distribution extends the conventional Dirichlet distribution by introducing regularization, making it a robust prior for categorical data in Bayesian frameworks. This adaptation is particularly advantageous for mitigating zero-probability issues in sparse categorical settings, such as emotion recognition in social media analytics Najar & Bouguila (2022), happiness modeling, and pain estimation Najar & Bouguila (2021). By redistributing probability mass across categories, smoothing enhances model stability Heckerman (1998) and generalizability Chen & Goodman (1999). Following Nallapati et al. (2007), a smoothed proportion $\mathbf{F}^u$ is derived from raw word counts using a tunable parameter λ:

$$\mathbf{F}^u \;=\; \frac{\mathbf{X}^s \;-\; (1-\lambda)\,\mathbf{X}^{GE}}{\lambda} \tag{1}$$

where $\mathbf{X}^s$ and $\mathbf{X}^{GE}$ denote the smoothed feature proportion and baseline word distribution (e.g., general English), respectively. The likelihood of observing $\mathbf{X}^s$ under the smoothed Dirichlet prior is:

$$p(\mathbf{X} \mid \boldsymbol{\alpha}, \varepsilon) = \frac{1}{B(\boldsymbol{\alpha}+\varepsilon)} \prod_{i=1}^{K} X_i^{(\alpha_i+\varepsilon)-1}, \quad \frac{1}{B(\boldsymbol{\alpha}+\varepsilon)} = \frac{\Gamma\big(\sum_i(\alpha_i+\varepsilon)\big)}{\prod_i \Gamma(\alpha_i+\varepsilon)}, \tag{2}$$

where K and $\varepsilon > 0$ denote the vocabulary size and smoothing parameter, respectively, X_i^s the smoothed feature, and α_i is the concentration parameters. The normalizer $B(\vec{\alpha})$ ensures a valid probability simplex.

In contrast to prior work that smooths raw inputs Nallapati et al. (2007), our method applies smoothing directly to the Dirichlet parameters and the latent representation. Preliminary experiments revealed that smoothing raw features induces covariate shifts in the feature representations, destabilizing training. Thus, our approach maintains feature consistency while enabling end-to-end optimization. This strategy aligns with the model's dynamic adaptation capabilities.

3.2 ModernBERT

ModernBERT builds upon BERT's bidirectional Transformer architecture to deliver powerful contextual embeddings while addressing the original's quadratic time and memory complexity in relation to sequence length Warner et al. (2024). By extending its maximum input length from 512 to 8,192 tokens, ModernBERT can capture long-range dependencies and global context in lengthy documents. A key innovation is FlashAttention, an optimized CUDA kernel that reorganizes attention computations to reduce memory accesses and fully exploit on-chip caches, yielding up to a two-fold speedup in self-attention layers Dao et al. (2022). Positional information is encoded using rotary positional embeddings (RoPE), which applies continuous rotation transformations to token representations and scales gracefully to very long sequences without the need for learned positional parameters Warner et al. (2024). To further mitigate computational costs, ModernBERT employs sequence packing and blockwise attention, splitting inputs into contiguous chunks and restricting attention to

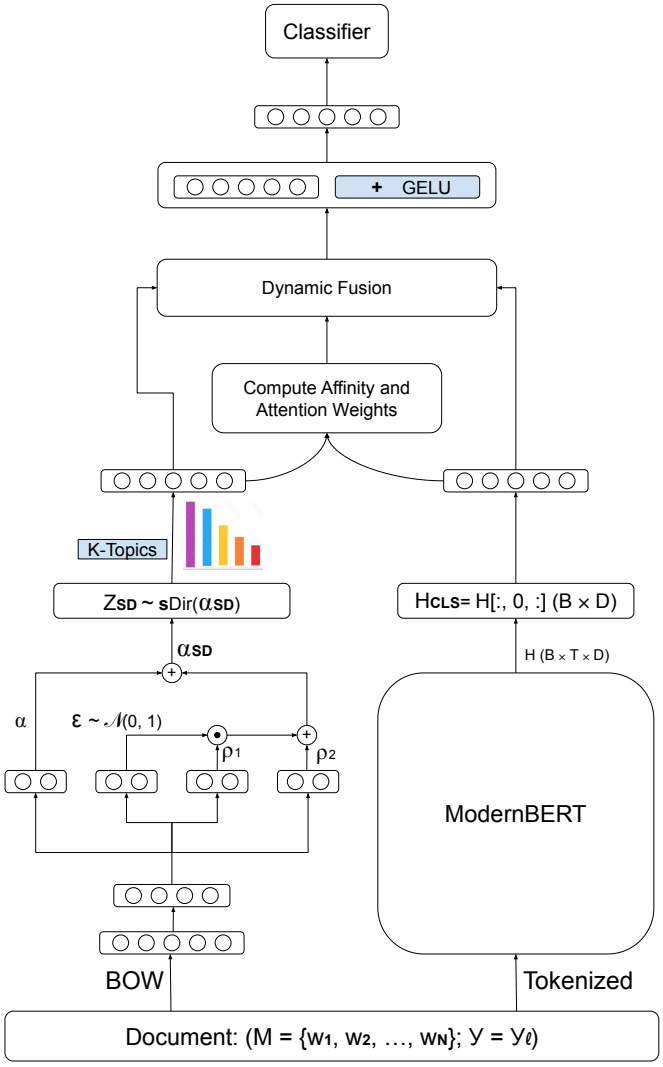

Figure 1: A schematic representation of the proposed SD-MoBERT model, leveraging smoothed Dirichlet neural topic model and ModernBERT.

intra-block and adjacent-block interactions; this achieves sub-quadratic complexity while preserving essential cross-chunk dependencies Warner et al. (2024). Finally, feed-forward sublayers incorporate low-rank matrix factorizations and sparse projection patterns that reduce parameter counts and confine expensive operations to the most informative tokens. These enhancements allow ModernBERT to handle long sequences of tokens efficiently, making it a scalable and context-rich foundation for hybrid models.

4 Proposed Model: Smoothed-ModernBERT (SD-MoBERT)

Figure 1 illustrates the architecture of SD-MoBERT, combining a neural topic model with an advanced transformer-based modernBERT. We employ ModernBERT because of its architectural innovations, such as support for up to 8,192 token contexts, FlashAttention, and rotary positional embeddings, which enable fast, memory-efficient processing of very long documents without sacrificing contextual depth Warner et al. (2024). Given a document $M = \{w_1, w_2, \ldots, w_N\}$ with label y, SD-MoBERT processes two parallel streams. Firstly, a normalized bag-of-words vector $\mathbf{X} \in \mathbb{R}^V$ (V = vocabulary size) for latent topic inference is generated. Secondly, a copy of the document is segmented into

subword tokens $\{t_n\}$ to generate a token sequence $\{t_1, \ldots, t_T\}$ ($T \leq 8192$) via ModernBERT's tokenizer, producing contextual embeddings $\mathbf{E} \in \mathbb{R}^{T \times D}$ (hidden size D), with [CLS] and [SEP] marking the start and the end.

In the generative process, we first draw from the neural topic model and infer a latent topic vector $\mathbf{Z} \in \mathbb{R}^K$ (K topics) under a smoothed Dirichlet prior:

$$\mathbf{Z}_{\text{SD}} \sim \text{sDir}(\boldsymbol{\alpha}_{\text{SD}}) \Longleftarrow \frac{\exp(\boldsymbol{\alpha}_{\text{SD}}^i)}{\sum_{i=1}^{K} \exp(\boldsymbol{\alpha}_{\text{SD}}^i)} \tag{3}$$

$$\boldsymbol{\alpha}_{\text{SD}} = \boldsymbol{\alpha} + \boldsymbol{\rho_2} + \boldsymbol{\epsilon} \odot \exp\big(\log \boldsymbol{\rho_1}\big) \in \mathbb{R}^K, \tag{4}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\rho_1}$, and $\boldsymbol{\rho_2}$ are the neural topic model's outputs, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Conversely, let the ModernBERT output be denoted by $\mathbf{h}_{\text{CLS}}$. We then project $\mathbf{Z}_{\text{SD}}$ and $\mathbf{h}_{\text{CLS}}$ each through a linear layer followed by a GELU activation to produce $\mathbf{Z}_{\text{SD}}^t$ and $\mathbf{Z}_{\text{CLS}}$, respectively. Next, we define the attention score $\mathbf{S}$ and the attention weight $\mathbf{Z}_{\text{att}}$ as:

$$\mathbf{S} = \langle \mathbf{Z}_{\text{SD}}^t, \mathbf{Z}_{\text{CLS}} \rangle_D + \mathbf{b}_0, \quad \mathbf{Z}_{\text{att}} = \sigma(\mathbf{S}) \tag{5}$$

where $\mathbf{b}_0$ and σ denote the attention bias weight and sigmoid function, respectively. Following this, we dynamically fuse representation as:

$$\mathbf{Z}_{\text{fused}} = \mathbf{Z}_{\text{att}} \mathbf{Z}_{\text{SD}}^t + (1 - \mathbf{Z}_{\text{att}}) \mathbf{Z}_{\text{CLS}}, \quad \mathbf{Z} = \tanh(\mathbf{Z}_{\text{fused}}) \in \mathbb{R}^{B \times D} \tag{6}$$

where $\mathbf{Z}$, B and D denote the latent representation, batch size, and sequence dimension, respectively. The latent representation is further projected through two linear layers and fed to the classifier, and we optimized with the joint loss:

$$\mathcal{L} = \underbrace{-\sum_i y_i \log \hat{y}_i}_{\mathcal{L}_{\text{CE}}} + \beta \underbrace{D_{\text{KL}}\big(q(\mathbf{Z} \mid \mathbf{X}) \| \text{sDir}(\boldsymbol{\alpha}_{\text{SD}})\big)}_{\mathcal{L}_{\text{KL}}}, \tag{7}$$

where β balances classification accuracy against topic coherence and $\mathcal{L}_{\text{CE}}$ denotes the classification loss. y_i and $\hat{y}$ represent the actual label and the prediction, respectively. $\mathcal{L}_{\text{KL}}$ Ojo et al. (2025) denotes the thematic loss that regularizes the latent space and penalizes the loss function to ensure that the model does not overfit. By aligning the thematic representations from the Dirichlet-based topic model with the contextual embeddings from ModernBERT through a co-attention mechanism, SD-MoBERT achieves a synergistic understanding of documents. This fusion enables the model to maintain interpretability through topic distributions while capturing nuanced contextual relationships, improving the performance of document classification tasks. See Section B for more details on the pseudocode for the generative process and Section B.1 for details on $\mathcal{L}_{\text{KL}}$.

5 Experimental Results

5.1 Experimental Settings

Please note that we conduct 30 separate experiments with different seeds using different validation sets at each experiment. Thus, we report the average value of our experiments over 30 runs. We explore the hyperparameter space using grid search to select the best combination of parameters for the experiment. We use a learning rate of $2e^{-5}$ with a warm-up of 10 and use AdamW optimizer, $\beta = 0.2$. We set the batch size and epoch to 8 and 20, respectively. We set the topic number of the smoothed Dirichlet component to 100. Section C presents the effect of hyperparameter tuning.

5.2 Datasets

We compare our proposed model with the baseline models on five widely used benchmark datasets, allowing insightful comparisons. The 20 Newsgroups (20NG) dataset Albishre et al. (2015) comprises 18,846 documents distributed across 20 categories, ranging from sports and politics to technology and religion. It contains 11,314 samples for training and 7,532 for testing. The Movie Review

(MR) dataset Haider Rizvi et al. (2025) contains 10,662 movie reviews balanced between 5,331 positives and 5,331 negatives for sentiment analysis. Ohsumed Haider Rizvi et al. (2025) consists of MEDLINE abstracts tagged in 23 categories of cardiovascular disease. It contains 7,400 documents, split into 3,357 for training and 4,043 for testing. Finally, we use the Reuters collection, drawn from the 1987 newswire, which is commonly evaluated via its R8 subset (8 classes, 5,485 training and 2,189 test documents) and R52 subset (52 classes, 6,532 training and 2,568 test documents) Moschitti & Basili (2004).

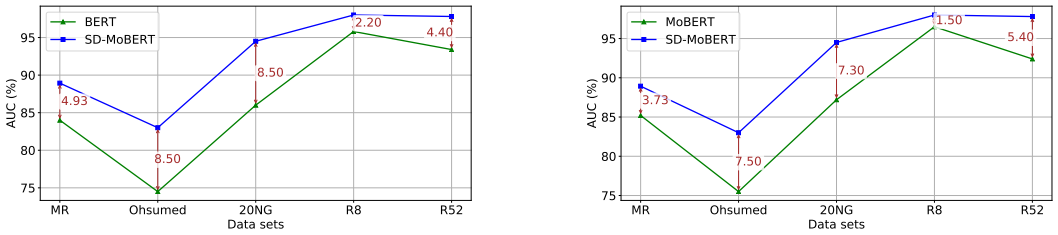

Figure 2: Analyses of the area under curve (AUC) of SD-MoBERT against BERT and MoBERT, $K = 100$, $\beta = 0.2$.

5.3 Baseline Models

To evaluate SD-MoBERT, we benchmark it against five close variants: BERT Devlin et al. (2019) and MoBERT Warner et al. (2024) without smoothed Dirichlet, SD-BERT (smoothed Dirichlet + BERT) Devlin et al. (2019), SD-RoBERTa (smoothed Dirichlet + RoBERTa-base) Masala et al. (2020), and SD-DistilBERT (smoothed Dirichlet + DistilBERT) Sanh et al. (2019), as well as a number of topic and graph-augmented models. These include TopicBERT-64/128 Chaudhary et al. (2020b), TextING Zhang et al. (2020), HyperGAT Ding et al. (2020), TextFCG Wang et al. (2023), TextSSL Piao et al. (2022), BertGCN Lin et al. (2021), GTC Liu et al. (2023), MHGAT Galke et al. (2022), and PaSIG-S Wang et al. (2025), providing a comprehensive backdrop for assessing the gains afforded by smoothed Dirichlet fusion in modern transformers.

5.4 Area Under Curve (AUC) Analysis of SD-MoBERT Against BERT and MoBERT

The AUC Çorbacıoğlu & Aksel (2023) plots the true positive rate (TPR) versus the false positive rate (FPR) to evaluate a model's ability to differentiate between classes across different thresholds. We use the trapezoidal rule to approximate the AUC, defined in Yeh et al. (2002)

$$AUC \approx \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} \tag{8}$$

Figure 2a illustrates the relative AUC gains of SD-MoBERT over the original BERT encoder across the five datasets. SD-MoBERT (blue squares) consistently outperforms BERT (green triangles), with improvements ranging from 2.2% to 8.5%. The largest gains occur on the 20 Newsgroups and Ohsumed corpora (both 8.5% gains), while even on Reuters R8, the margin remains substantial at 2.2%. Similarly, Figure 2b shows the AUC improvements of SD-MoBERT's relative to the MoBERT variant. Across all data sets, SD-MoBERT achieves gains between 1.5% on R8 and 7.5% on Ohsumed data sets. These results underscore the robustness of the smoothed Dirichlet in capturing the thematic structures inherent in document-level discourse and the dynamic fusion mechanism in enhancing discriminative power over the base ModernBERT architecture.

5.5 Performance Comparison of SD-MoBERT Against Baselines and Model Variants

Tables 1 and 2 present a detailed evaluation of SD-MoBERT relative to ten established baselines and five BERT-family variants across five benchmark datasets, reporting mean accuracy and F1

Table 1: Comparisons of the average test accuracy and F1 scores with their respective standard deviations. We evaluate SD-MoBERT alongside other baseline models across three datasets (MR, Ohsumed, and 20NG), $K = 100$, $\beta = 0.2$.

	Model	MR		Ohsumed		20NG	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
Baselines	TextING	79.75 ± 0.78	79.63 ± 0.85	73.51 ± 1.05	68.15 ± 0.77	85.13 ± 0.66	84.32 ± 0.12
	HyperGAT	76.64 ± 0.81	76.58 ± 0.92	66.55 ± 1.37	59.05 ± 1.84	83.29 ± 0.46	82.72 ± 0.24
	TextFCG	80.59 ± 0.29	80.56 ± 0.47	69.58 ± 0.39	56.16 ± 0.71	85.95 ± 0.33	84.91 ± 0.51
	TextSSL	75.74 ± 0.25	75.64 ± 0.38	62.01 ± 0.41	51.99 ± 0.78	79.55 ± 0.27	79.11 ± 0.65
	TopicBERT-64	85.21 ± 0.91	85.01 ± 0.76	72.31 ± 0.33	71.13 ± 0.48	83.86 ± 0.55	83.19 ± 0.82
	TopicBERT-128	86.89 ± 0.33	86.15 ± 0.64	74.10 ± 0.74	73.92 ± 0.22	82.60 ± 0.10	82.60 ± 0.41
	BertGCN	84.92 ± 0.84	84.05 ± 0.67	71.88 ± 0.52	62.72 ± 0.47	88.69 ± 0.45	88.02 ± 0.20
	GTC	77.22 ± 0.37	77.01 ± 0.24	69.72 ± 0.72	62.8 ± 0.11	87.03 ± 0.61	85.73 ± 0.40
	MHGAT	78.09 ± 0.73	77.24 ± 0.57	72.88 ± 0.84	65.04 ± 1.60	92.68 ± 0.30	91.94 ± 0.13
	PaSIG-S	87.05 ± 0.09	87.04 ± 0.09	81.18 ± 0.21	74.58 ± 0.42	93.21 ± 0.07	92.91 ± 0.08
Proposed Model Variants	BERT	85.72 ± 0.13	84.50 ± 0.41	76.94 ± 0.01	76.70 ± 0.00	85.33 ± 0.14	82.31 ± 0.01
	MoBERT	86.00 ± 0.05	84.9 ± 0.03	76.99 ± 0.02	76.51 ± 0.01	87.72 ± 0.33	85.24 ± 0.12
	SD-BERT	86.02 ± 0.02	85.39 ± 0.23	77.01 ± 0.02	76.90 ± 0.03	89.12 ± 0.11	87.03 ± 0.47
	SD-RoBERTa	88.10 ± 0.24	87.69 ± 0.39	79.82 ± 0.11	79.01 ± 0.13	92.55 ± 0.03	91.31 ± 0.05
	SD-DistilBERT	87.09 ± 1.31	84.59 ± 0.94	75.62 ± 0.01	75.11 ± 0.06	86.40 ± 0.16	81.60 ± 0.01
	SD-MoBERT	$\mathbf{88.97 \pm 0.02}$	$\mathbf{88.13 \pm 0.05}$	$\mathbf{83.49 \pm 0.04}$	$\mathbf{80.00 \pm 0.21}$	$\mathbf{95.27 \pm 0.05}$	$\mathbf{93.11 \pm 0.07}$

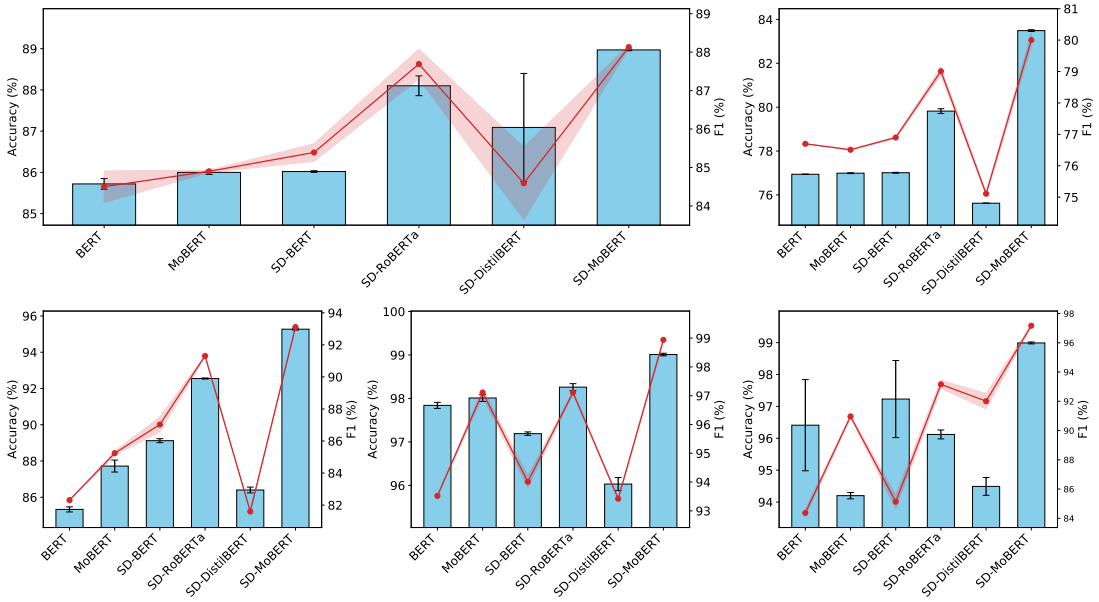

Figure 3: Comparison of the classification accuracy and F1 score in six transformer-based models on five text classification benchmarks. The bar plots (sky blue) depict mean test accuracy with the error bars, while the overlaid red lines trace mean F1 scores. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $K = 100$, $\beta = 0.2$.

Table 2: Comparisons of the average test accuracy and F1 scores with their respective standard deviations. We evaluate SD-MoBERT alongside other baseline models across two datasets (R8 and R52), $K = 100$, $\beta = 0.2$.

	Model	R8		R52	
		Accuracy	F1	Accuracy	F1
Baselines	TextING	97.45 ± 0.70	95.94 ± 0.63	94.95 ± 0.95	76.71 ± 0.87
	HyperGAT	96.43 ± 0.63	92.12 ± 1.51	94.24 ± 0.54	72.35 ± 1.83
	TextFCG	97.53 ± 0.34	92.44 ± 0.21	95.64 ± 0.15	69.13 ± 0.28
	TextSSL	97.31 ± 0.42	93.01 ± 0.33	93.97 ± 0.66	72.79 ± 1.41
	TopicBERT-64	93.01 ± 0.29	92.11 ± 0.63	72.89 ± 0.57	72.18 ± 0.98
	TopicBERT-128	93.94 ± 0.22	92.83 ± 0.51	73.42 ± 0.37	72.84 ± 0.29
	BertGCN	97.94 ± 0.73	94.60 ± 0.44	95.50 ± 0.44	52.30 ± 0.73
	GTC	97.21 ± 0.85	93.73 ± 0.64	94.51 ± 0.97	94.52 ± 0.77
	MHGAT	97.65 ± 0.47	93.09 ± 1.21	94.78 ± 0.37	76.74 ± 1.06
	PaSIG-S	$\mathbf{99.02 \pm 0.04}$	98.16 ± 0.12	98.34 ± 0.03	85.99 ± 1.52
Proposed Model Variants	BERT	97.84 ± 0.07	93.52 ± 0.01	96.41 ± 1.43	84.37 ± 0.25
	MoBERT	98.01 ± 0.08	97.11 ± 0.15	94.20 ± 0.10	90.97 ± 0.09
	SD-BERT	97.19 ± 0.04	94.01 ± 0.20	97.23 ± 1.21	85.14 ± 0.47
	SD-RoBERTa	98.26 ± 0.08	97.11 ± 0.05	96.12 ± 0.14	93.16 ± 0.28
	SD-DistilBERT	96.03 ± 0.15	93.42 ± 0.03	94.49 ± 0.28	92.01 ± 0.53
	SD-MoBERT	99.01 ± 0.03	$\mathbf{98.94 \pm 0.07}$	$\mathbf{98.99 \pm 0.03}$	$\mathbf{97.17 \pm 0.07}$

scores with their corresponding standard deviations. On the MR short-text sentiment classification task, PaSIG-S achieved an average accuracy of $87.05\% \pm 0.09$ and F1 score of $87.04\% \pm 0.09$. SD-MoBERT raises accuracy to $88.97\% \pm 0.02$, a $(88.97 - 87.05)/87.05 \times 100 \approx 2.3\%$ relative gain, and boosts F1 to $88.13\% \pm 0.05$, a $\approx 1.3\%$ improvement in F1. On the Ohsumed corpus, SD-MoBERT attains $83.49\% \pm 0.04$ accuracy, outperforming the best baseline (PaSIG-S: $81.18\% \pm 0.21$) by 2.85%, and achieves an F1 score of $80.00\% \pm 0.21$, a 7.26% over PaSIG-S's $74.58\% \pm 0.42$. These gains underscore SD-MoBERT's enhanced ability to disambiguate complex medical terminology where graph-based methods (e.g. HyperGAT) exhibit lower F1 score. For the 20 Newsgroups (20NG), SD-MoBERT reaches $95.27\% \pm 0.05$ accuracy, a 2.21% improvement over PaSIG-S's $93.21\% \pm 0.07$, and records an F1 of $93.11\% \pm 0.07$, surpassing the next-best model (MHGAT: $91.94\% \pm 0.13$) by 1.27%. SD-MoBERT shows its ability to distinguish semantically overlapping categories.

On R8, PaSIG-S achieves $99.02\% \pm 0.04$ accuracy and $98.16\% \pm 0.12$ F1, while SD-MoBERT records $99.01\% \pm 0.03$ (a negligible -0.01% change) and $98.94\% \pm 0.07$, corresponding to a $\approx 0.8\%$ F1 improvement. On Reuters R52, SD-MoBERT yields an F1 score of $97.17\% \pm 0.07$, representing a 12.98% increase over PaSIG-S's $85.99\% \pm 1.52$. Such a substantial margin highlights its robustness in hierarchical news classification, where error propagation across parent-child categories is a known challenge. When compared to other BERT variants, SD-MoBERT consistently delivers further gains. In the MR short-text sentiment benchmark, accuracy improves from BERT's $85.72\% \pm 0.13$ and MoBERT's $86.00\% \pm 0.05$ to $88.97\% \pm 0.02$, corresponding to relative increases of 3.25% and 2.97%, respectively. On Reuters R8, SD-MoBERT's F1 score of $98.94\% \pm 0.07$ exceeds SD-RoBERTa's $97.11\% \pm 0.05$ by 1.88%, demonstrating the efficacy of smoothed-Dirichlet regularization. Against SD-DistilBERT on Ohsumed, SD-MoBERT's F1 advantage of 6.51% (80.00% vs. 75.11%) further confirms that model compression without careful calibration can degrade performance on specialized domains. Across all five datasets, SD-MoBERT exhibits minimal performance variance (standard deviations between ± 0.02 and ± 0.07), in stark contrast to several baselines and variants (e.g. TextSSL on Ohsumed, GTC on R52, and SD-DistilBERT on MR), whose larger fluctuations signal instability. This consistency is attributable to the smoothed-Dirichlet fusion's ability to regularize confidence estimates and mitigate overfitting. As shown in Table 3, we evaluate whether the observed accuracy gains of SD-MoBERT over the best baseline (PaSIG-S) are statistically significant. As indicated in Table 3, all p-values ($\ll 0.05$), uniformly reject H_0, while the CIs remain vanishingly narrow. See more details in Section E.

$$H_0 \colon \mu_{\text{SD-MoBERT}} = \mu_{\text{PaSIG-S}} \quad \text{vs.} \quad H_1 \colon \mu_{\text{SD-MoBERT}} \neq \mu_{\text{PaSIG-S}} \tag{9}$$

5.6 Error-Bar Analysis of Accuracy and F1 Across Proposed Model Variants

Figure 3 presents the error bar across the five data sets. The accuracy bars exhibit consistently narrow error margins, typically under 0.5 %, indicating that each model's mean performance is highly stable over repeated runs. Notably, the unsmoothed BERT and MoBERT backbones show slightly wider accuracy-bar spreads (up to 1.3 % on SD-DistilBERT's R52 result), whereas the smoothed-Dirichlet variants (SD-BERT, SD-RoBERTa, SD-DistilBERT, SD-MoBERT) reduce that variability to under 0.3 %, reflecting more reliable convergence. In contrast, the F1 scores (red lines) display larger error bands, ranging from virtually zero for MoBERT on MR up to 0.94 % for SD-DistilBERT, highlighting that the precision-recall balance is intrinsically more sensitive in the architectures. Importantly, SD-MoBERT not only attains the highest mean accuracy and F1 in every data set but also maintains among the smallest F1-error spreads (≤ 0.07 %), underscoring its robustness in both overall correctness and class-balanced performance.

5.7 Limitations

SD-MoBERT requires manual selection of the topic count K, as too few topics yield overly broad themes and too many produce fragmented noise, necessitating costly hyperparameter searches. The added topic model and co-attention layer also incur extra parameters and runtime overhead, and learned topics may not transfer across domains without retraining. Future work will explore adaptive topic estimation and computation efficiency. See more discussion on the time complexity and runtime cost in Section F.

Table 3: Statistical analyses of SD-MoBERT over 30 runs using different validation sets and the best baseline model (PaSIG-S) accuracy. The bold values signify p-values that are below 0.05, CI and S denote the class interval, and standard deviation, respectively, $K = 100$, $\beta = 0.2$.

		MR	Ohsumed	20NG	R8	R52
SD-MoBERT	Mean (F1)	88.13	80.00	93.11	98.94	97.17
	Variance	$8.54e^{-4}$	$2.57e^{-2}$	$1.37e^{-3}$	$1.51e^{-3}$	$1.53e^{-3}$
	S	0.029	0.160	0.037	0.039	0.039
	CI	[88.120 – 88.140]	[79.943 – 80.057]	[93.097 – 93.123]	[98.926 – 98.954]	[97.156 – 97.184]
Best baseline (PaSIG-S)	F1	87.04	74.58	92.91	98.16	85.99
p-value		$\mathbf{2.378}e^{-47}$	$\mathbf{3.782}e^{-46}$	$\mathbf{3.087}e^{-23}$	$\mathbf{1.487}e^{-39}$	$\mathbf{5.488}e^{-73}$

6 Conclusion

This study addresses the critical challenge of document classification in NLP by harmonizing the complementary strengths of transformer architectures and probabilistic topic modeling. While ModernBERT captures nuanced contextual semantics and topic models distill interpretable thematic structures, their isolated applications leave a methodological gap in handling both granular context and global discourse. Our proposed framework bridges this divide through a novel co-attention mechanism that dynamically fuses token-level BERT embeddings with document-level smoothed-Dirichlet topic distributions, enabling joint optimization of contextual and thematic objectives. Empirical validation across benchmark corpora demonstrates that this synergistic approach achieves superior classification robustness, outperforming standalone models by effectively leveraging multi-granular semantic signals. The dynamic gating mechanism ensures adaptive weighting of contextual and thematic features, enhancing generalizability across domains requiring both precision and abstraction. By open-sourcing our implementation, we invite the community to build upon this work, advancing methodologies that unify local and global text representations. This contribution not only advances document classification but also establishes a blueprint for integrating neural and probabilistic paradigms in NLP, fostering models that balance interpretability with state-of-the-art performance. Future work will explore adaptive topic number estimation and multi-head co-attention to model richer interactions between topics and tokens.

References

Asima Akber Abbasi, Aneela Zameer, and Muhammad Asif Zahoor Raja. An enhanced strategy for minority class detection using bidirectional gru employing penalized cross-entropy and self-attention mechanisms for imbalance network traffic. *The European Physical Journal Plus*, 139(6): 530, 2024.

Shams Forruque Ahmed, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M Mofijur, ABM Shawkat Ali, and Amir H Gandomi. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617, 2023.

Khaled Albishre, Mubarak Albathan, and Yuefeng Li. Effective 20 newsgroups dataset cleaning. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pp. 98–101. IEEE, 2015.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot text classification with distributional signatures. *International Conference on Learning Representations paper*, 2019.

Pavol Bielik, Veselin Raychev, and Martin Vechev. Program synthesis for character level language modeling. In *International conference on learning representations*, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. TopicBERT for energy efficient document classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1682–1690, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.152. URL https://aclanthology.org/2020.findings-emnlp.152/.

Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. Topicbert for energy efficient document classification. *Association for Computational Linguistics*, 2020b.

Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

Huaqing Cheng, Shengquan Liu, Weiwei Sun, and Qi Sun. A neural topic modeling study integrating sbert and data augmentation. *Applied Sciences*, 13(7):4595, 2023.

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294, 2016.

Benjamin Clavié, Akshita Gheewala, Paul Briton, Marc Alphonsus, Rym Laabiyad, and Francesco Piccoli. Legalmfit: Efficient short legal text classification with lstm language model pre-training. *Association for Computational Linguistics*, 2021.

Şeref Kerem Çorbacıoğlu and Gökhan Aksel. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4):195, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Yair Davidson and Nadav Dym. On the holder stability of multiset and graph neural networks. *International Conference on Learning Representations paper*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. Be more with less: Hypergraph attention networks for inductive text classification. *Association for Computational Linguistics*, 2020.

Lukas Galke, Andor Diera, Bao Xin Lin, Bhakti Khera, Tim Meuser, Tushar Singhal, Fabian Karl, and Ansgar Scherp. Are we really making much progress in text classification? a comparative review. *Computational Linguistics*, 48(1):4038–4051, 2022.

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

Syed Mustafa Haider Rizvi, Ramsha Imran, and Arif Mahmood. Text classification using graph convolutional networks: A comprehensive survey. *ACM Computing Surveys*, 2025.

David Heckerman. A tutorial on learning with bayesian networks. *Learning in graphical models*, pp. 301–354, 1998.

Saman Jamshidi, Mahin Mohammadi, Saeed Bagheri, Hamid Esmaeili Najafabadi, Alireza Rezvanian, Mehdi Gheisari, Mustafa Ghaderzadeh, Amir Shahab Shahabi, and Zongda Wu. Effective text classification using bert, mtm lstm, and dt. *Data & Knowledge Engineering*, 151:102306, 2024.

TianTian Jiang and ZhanGuo Wang. Text classification using bigru with directional self-attention. In *2022 11th International Conference of Information and Communication Technology (ICTech))*, pp. 394–397. IEEE, 2022.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. *Hugging Face*, 2025.

Bofang Li, Zhe Zhao, Tao Liu, Puwei Wang, and Xiaoyong Du. Weighted neural bag-of-n-grams model: New baselines for text classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1591–1600, 2016.

Xi Li and Lili Jia. English text topic classification using bert-based model. *Journal of Computational Methods in Sciences and Engineering*, pp. 14727978251321982, 2025.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. Bertgcn: Transductive text classification by combining gcn and bert. *Association for Computational Linguistics*, 2021.

Boting Liu, Weili Guan, Changjin Yang, Zhijie Fang, and Zhiheng Lu. Transformer and graph convolutional network for text classification. *International Journal of Computational Intelligence Systems*, 16(1):161, 2023.

Qiwen Liu, Tianjian Chen, Jing Cai, and Dianhai Yu. Enlister: baidu's recommender system for the biggest chinese q&a website. In *Proceedings of the sixth ACM conference on Recommender systems*, pp. 285–288, 2012.

Zeping Luo, Shiyou Wu, Cindy Weng, Mo Zhou, and Rong Ge. Understanding the robustness of self-supervised learning through topic modeling. *International conference on learning representations*, 2022.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6626–6637, 2020.

Neethu Elizabeth Michael, Ramesh C Bansal, Ali Ahmed Adam Ismail, A Elnady, and Shazia Hasan. A cohesive structure of bi-directional long-short-term memory (bilstm)-gru for predicting hourly solar radiation. *Renewable Energy*, 222:119943, 2024.

Farhad Mortezapour Shiri, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *Journal on Artificial Intelligence 2024 Vol. 6 Issue 1 Pages 301-360*, pp. 2305, 2023.

Alessandro Moschitti and Roberto Basili. Complex linguistic features for text classification: A comprehensive study. In *European conference on information retrieval*, pp. 181–196. Springer, 2004.

Fatma Najar and Nizar Bouguila. Smoothed generalized dirichlet: A novel count-data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, 3(5):685–698, 2021.

Fatma Najar and Nizar Bouguila. Emotion recognition: A smoothed dirichlet multinomial solution. *Engineering Applications of Artificial Intelligence*, 107:104542, 2022.

Ramesh Nallapati, Thomas Minka, Hugo Zaragoza, and Stephen Robertson. The smoothed-dirichlet distribution: Explaining kl-divergence based ranking in information retrieval. *def*, 2:32, 2007.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Advances in neural information processing systems*, 30, 2017.

Akinlolu Oluwabusayo Ojo and Nizar Bouguila. A topic modeling and image classification framework: The generalized dirichlet variational autoencoder. *Pattern Recognition*, 146:110037, 2024.

Akinlolu Oluwabusayo Ojo, Fatma Najar, Nuha Zamzami, Hanen T Himdi, and Nizar Bouguila. Smoothdectector: A smoothed dirichlet multimodal approach for combating fake news on social media. *IEEE Access*, 13:39289–39305, 2025.

Yinhua Piao, Sangseon Lee, Dohoon Lee, and Sun Kim. Sparse structure learning via graph neural networks for inductive document classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11165–11173, 2022.

Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)*, pp. 200–204. IEEE, 2019.

Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, and Abdulwahab Ali Almazroi. A fine-tuned bert-based transfer learning approach for text classification. *Journal of healthcare engineering*, 2022(1):3498123, 2022.

Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102, 2018.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Hugging Face*, 2019.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pp. 194–206. Springer, 2019a.

Xia Sun, Yi Gao, Richard Sutcliffe, Shou-Xi Guo, Xin Wang, and Jun Feng. Word representation learning based on bidirectional grus with drop loss for sentiment classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(7):4532–4542, 2019b.

Zoubir Talai and Nada Kherici. Compact cnn-based architecture for text classification and sentiment analysis. *Revue de l'Information Scientifique et Technique*, 27(2):50–55, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Daniel Voskergian, Rashid Jayousi, and Malik Yousef. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. *Scientific Reports*, 14 (1):23516, 2024.

Congcong Wang, Paul Nulty, and David Lillis. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th international conference on natural language processing and information retrieval*, pp. 37–46, 2020a.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. Representing mixtures of word embeddings with mixtures of topic embeddings. *International conference on learning representations*, 2022.

Shiyu Wang, Gang Zhou, Jicang Lu, Jing Chen, and Ningbo Huang. Pre-trained semantic interaction based inductive graph neural networks for text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 812–827, 2025.

Yizhao Wang, Chenxi Wang, Jieyu Zhan, Wenjun Ma, and Yuncheng Jiang. Text fcg: Fusing contextual information via graph learning for text classification. *Expert Systems with Applications*, 219:119658, 2023.

Ziniu Wang, Zhilin Huang, and Jianling Gao. Chinese text classification method based on bert word embedding. In *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, pp. 66–71, 2020b.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Association for Computational Linguistics: Anonymous submission*, 2024.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18, 2024.

Shi-Tao Yeh et al. Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annual SAS® User Group International (SUGI'02)*, 4:1, 2002.

Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. *Advances in neural information processing systems*, 31, 2018.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. Every document owns its structure: Inductive text classification via graph neural networks. *Association for Computational Linguistics*, 2020.

A Appendix

B Pseudocode for SD-MoBERT Generative Process

In this section, we present the high-level algorithmic steps of SD-MoBERT's generative process. Algorithm 1 outlines how each document is first encoded via a bag-of-words topic model and a ModernBERT backbone, then dynamically fused through a smoothed-Dirichlet co-attention mechanism, and finally passed through a classification head. By iterating thematic nuances for the smoothed-Dirichlet prior and jointly optimizing the transformer and topic parameters, SD-MoBERT learns to leverage both topical and contextual information in a unified training loop.

Algorithm 1 Generative process for SD-MoBERT

Data:
  M: document tokens $\{w_1, \ldots, w_N\}$
  y: true label
Parameters:
  Transformer weights (ModernBERT)
  Topic MLP weights $\{W^{(i)}, b^{(i)}\}$,
    $W_\mu, b_\mu, W_{\log\sigma}, b_{\log\sigma}, W_\alpha, b_\alpha$
  Fusion weights $W_t, b_t, W_c, b_c, b_0, W_1, b_1, W_2, b_2$
Result:
  Logits $\mathbf{o}$, loss $\mathcal{L}_{\text{total}}$
Initialize all parameters
while not converged **do**
  // 1. Prepare Inputs
  $\mathbf{X} \leftarrow \mathrm{BoW}(M)$ ▷ size V
  $\{t_1, \ldots, t_T\} \leftarrow \mathrm{Tokenize}(M)$ ▷ $T \leq 8192$
  $\mathbf{E} \leftarrow \mathrm{ModernBERT}(\{t_i\})$ ▷ size $T \times D$
  $\mathbf{h}_{\mathrm{CLS}} \leftarrow \mathbf{E}[0]$ ▷ D-dim

  // 2. Neural Topic Model Inference
  $\boldsymbol{\pi} \leftarrow \mathrm{MLP}(\mathbf{X})$ ▷ size H
  $\boldsymbol{\mu} \leftarrow W_\mu\, \boldsymbol{\pi} + b_\mu$
  $\log \boldsymbol{\sigma} \leftarrow W_{\log\sigma}\, \boldsymbol{\pi} + b_{\log\sigma}$
  $\boldsymbol{\alpha} \leftarrow W_\alpha\, \boldsymbol{\pi} + b_\alpha$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
  $\boldsymbol{\alpha}_{SD} \leftarrow \boldsymbol{\alpha} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \exp(\log \boldsymbol{\sigma})$
  $\mathbf{Z}_{\mathrm{SD}} \leftarrow \mathrm{softmax}(\boldsymbol{\alpha}_{SD})$

  // 3. Co-Attention Fusion
  $\mathbf{Z}^t_{\mathrm{SD}} \leftarrow \mathrm{GELU}(W_t\, \mathbf{Z}_{\mathrm{SD}} + b_t)$
  $\mathbf{Z}_{\mathrm{CLS}} \leftarrow \mathrm{GELU}(W_c\, \mathbf{h}_{\mathrm{CLS}} + b_c)$
  $S \leftarrow \langle \mathbf{Z}^t_{\mathrm{SD}}, \mathbf{Z}_{\mathrm{CLS}} \rangle + b_0$
  $\mathbf{Z}_{\mathrm{att}} \leftarrow \sigma(S)$
  $\mathbf{Z}_{\mathrm{fused}} \leftarrow \mathbf{Z}_{\mathrm{att}}\, \mathbf{Z}^t_{\mathrm{SD}} + (1 - \mathbf{Z}_{\mathrm{att}})\, \mathbf{Z}_{\mathrm{CLS}}$
  $\mathbf{Z} \leftarrow \tanh(\mathbf{Z}_{\mathrm{fused}})$

  // 4. Classification Head
  $\mathbf{h}_1 \leftarrow \mathrm{GELU}(W_1\, \mathbf{Z} + b_1)$
  $\mathbf{o} \leftarrow W_2\, \mathbf{h}_1 + b_2$

  // 5. Loss
  $\mathcal{L}_{\mathrm{CE}} \leftarrow \mathrm{CrossEntropy}(\mathbf{o}, y)$
  $\mathcal{L}_{\mathrm{KL}} \leftarrow \mathrm{KL\_Dirichlet}(\tilde{\boldsymbol{\alpha}} \| \alpha_0 \mathbf{1})$
  $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{CE}} + \lambda\, \mathcal{L}_{\mathrm{KL}}$
  Update parameters via optimizer
end while

B.1 SMOOTHED DIRICHLET THEMATIC LOSS: KULLBACK-LEIBLER DIVERGENCE (KL)

Below we extend the standard Dirichlet-to-Dirichlet KL proof to the case where both distributions include an additive smoothing parameter $\varepsilon > 0$. In essence, if

$$P = \mathrm{Dir}(\boldsymbol{\alpha} + \varepsilon\,\mathbf{1}), \quad Q = \mathrm{Dir}(\boldsymbol{\beta} + \varepsilon\,\mathbf{1}), \tag{10}$$

then the KL divergence $\mathrm{KL}[P\|Q]$ takes exactly the same closed-form as for the unsmoothed case, with each α_i and β_i replaced by $\alpha_i + \varepsilon$ and $\beta_i + \varepsilon$, respectively.

The Dirichlet density with smoothing ε is defined for $\mathbf{X} \in \Delta^{K-1}$ by

$$p(\mathbf{X} \mid \boldsymbol{\alpha}, \varepsilon) = \frac{1}{B(\boldsymbol{\alpha} + \varepsilon\,\mathbf{1})} \prod_{i=1}^{K} X_i^{(\alpha_i+\varepsilon)-1}, \tag{11}$$

where

$$B(\boldsymbol{\alpha} + \varepsilon) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i + \varepsilon)}{\Gamma\big(\sum_{i=1}^{K} (\alpha_i + \varepsilon)\big)} \tag{12}$$

is the multivariate Beta function Nallapati et al. (2007). The standard KL divergence between two densities p and q is

$$\mathrm{KL}[P\|Q] = \int p(\mathbf{x})\ \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x} \tag{13}$$

$$\text{Let} \quad P(\mathbf{x}) = \mathrm{Dir}(\boldsymbol{\alpha} + \varepsilon\,\mathbf{1}), \quad Q(\mathbf{x}) = \mathrm{Dir}(\boldsymbol{\beta} + \varepsilon\,\mathbf{1}). \tag{14}$$

Substituting both into the KL definition and bringing out constant terms yields

$$\begin{aligned} \mathrm{KL}[P\|Q] &= \int P(\mathbf{x}) \Big[\log \tfrac{B(\boldsymbol{\beta}+\varepsilon)}{B(\boldsymbol{\alpha}+\varepsilon)} + \sum_{i=1}^{K} \big((\alpha_i + \varepsilon) - (\beta_i + \varepsilon)\big)\ \log x_i\Big]\, d\mathbf{x} \\ &= \log \tfrac{B(\boldsymbol{\beta}+\varepsilon)}{B(\boldsymbol{\alpha}+\varepsilon)} + \sum_{i=1}^{K} (\alpha_i - \beta_i)\, \mathbb{E}_P[\log x_i], \end{aligned} \tag{15}$$

For a smoothed Dirichlet with parameters $\alpha_i' = \alpha_i + \varepsilon$, the moment is

$$\mathbb{E}[\log x_i] = \psi(\alpha_i + \varepsilon)\ -\ \psi\Big(\textstyle\sum_{j=1}^{K} (\alpha_j + \varepsilon)\Big), \tag{16}$$

where ψ is the digamma function. Writing the log-ratio of Beta functions in terms of Gamma yields

$$\log \tfrac{B(\boldsymbol{\beta}+\varepsilon)}{B(\boldsymbol{\alpha}+\varepsilon)} = \sum_{i=1}^{K} \big[\log\Gamma(\beta_i+\varepsilon) - \log\Gamma(\alpha_i+\varepsilon)\big] + \log\Gamma\Big(\textstyle\sum_i \alpha_i + K\,\varepsilon\Big) - \log\Gamma\Big(\sum_i \beta_i + K\,\varepsilon\Big) \tag{17}$$

Combining the above parts, the KL divergence between two smoothed Dirichlet distributions is

$$\begin{aligned} \mathcal{L}_{\mathrm{KL}} \implies \mathrm{KL}[\mathrm{Dir}(\boldsymbol{\alpha} + \varepsilon) \parallel \mathrm{Dir}(\boldsymbol{\beta} + \varepsilon)] &= \sum_{i=1}^{K} \Big[\log\Gamma(\beta_i + \varepsilon) - \log\Gamma(\alpha_i + \varepsilon)\Big] \\ &+ \log\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\,\varepsilon\Big) - \log\Gamma\Big(\sum_{i=1}^{K} \beta_i + K\,\varepsilon\Big) \\ &+ \sum_{i=1}^{K} (\alpha_i - \beta_i)\Big[\psi(\alpha_i + \varepsilon) - \psi\big(\textstyle\sum_{j=1}^{K} \alpha_j + K\,\varepsilon\big)\Big] \end{aligned} \tag{18}$$

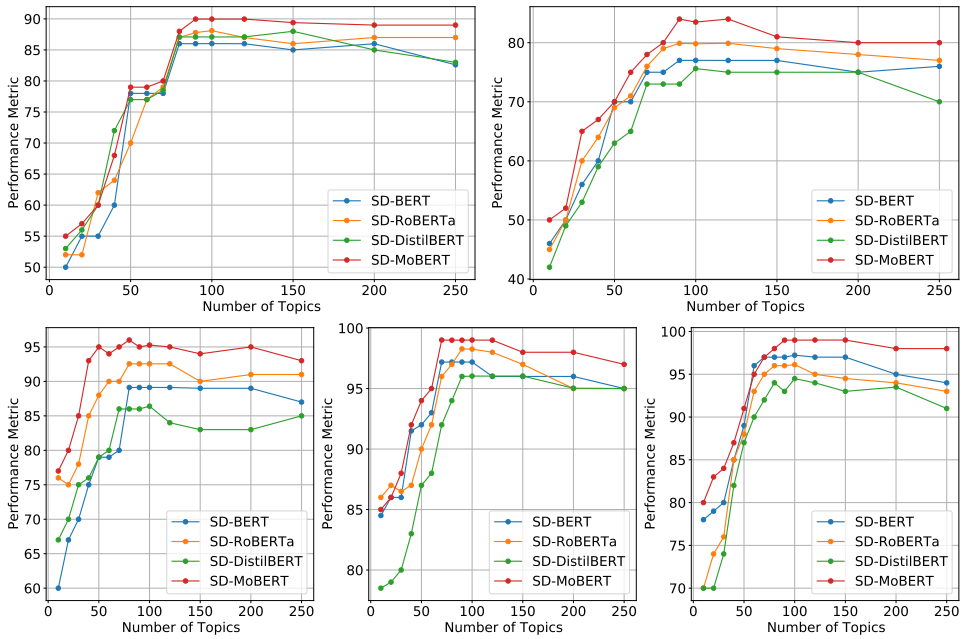

Figure 4: Sensitivity of classification accuracy to the number of latent topics on five data sets. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $\beta = 0.2$.

Note that this expression reduces to the standard Dirichlet KL divergence when $\varepsilon \to 0$.

C Effect of Topic Number on Classification Performance

C depicts how the variation in the number of latent topics influences the classification accuracy in five benchmark datasets (MR, Ohsumed, 20 Newsgroups, R8, and R52) for four smoothed-Dirichlet variants: SD-BERT, SD-RoBERTa, SD-DistilBERT, and SD-MoBERT. In all cases, performance increases when the topic number increases from very low values (10-40), reflecting the transition from an overly coarse to a sufficiently expressive latent representation. Beyond approximately 70-100 topics, gains begin to plateau or even fluctuate slightly, indicating diminishing returns from further topic subdivisions.

On the Movie Review (MR) dataset, SD-MoBERT achieves the highest peak accuracy of roughly 90% at 80 topics, outperforming its counterparts by 2-4%, while all models converge around 86-88% for larger topic numbers. A similar pattern emerges on Ohsumed: SD-MoBERT reaches about 84% at 90–100 topics, whereas the other transformers level off around 77-80%. In the more fine-grained 20 Newsgroups setting, SD-MoBERT again leads with nearly 96% at 80 topics, compared to 92–93% for SD-RoBERTa and SD-BERT, and slightly lower performance for the DistilBERT variant. For the more specialized Reuters subsets R8 and R52, the advantage of SD-MoBERT is most pronounced. On R8, SD-MoBERT rapidly climbs to over 99% accuracy at 80 topics and sustains this around 98-99% as topics increase. The other models attain roughly 96-98% in the same range, with SD-DistilBERT typically the lowest. On R52, SD-MoBERT surpasses 99% by 100 topics, while SD-RoBERTa and SD-BERT stabilize around 95-97%, and SD-DistilBERT around 91–94%.

In general, these plots demonstrate that integrating ModernBERT with a smoothed Dirichlet topic prior (SD-MoBERT) consistently yields better classification performance, especially once the latent dimensionality is sufficiently large (70 to 100 topics), and that beyond this range, additional topics confer minimal benefit across various text classification scenarios.

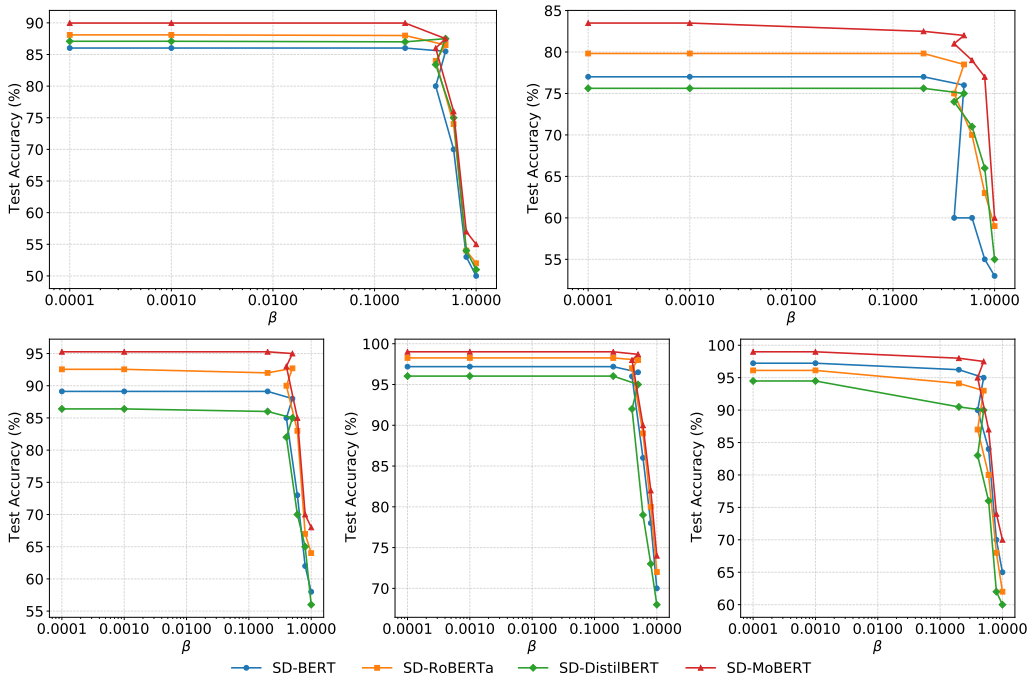

Figure 5: Sensitivity of classification accuracy to the regularization weight β across five benchmarks. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $K = 100$.

D Effect of the KL-Weight Factor on Classification Performance

Figure 5 depicts the test-accuracy plots to visualize how the balance between cross-entropy loss and the KL divergence (controlled by the regularization coefficient β in $\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta\,\mathcal{L}_{\text{KL}}$) affects classification accuracy on five benchmark corpora (MR, Ohsumed, 20NG, R8 and R52).

Across the five benchmarks, we observe the following consistent pattern: when β is large ($\beta \geq 0.6$), the models under-emphasize the cross-entropy term and suffer in accuracy. For example, on the MR, all four methods plateau around 70-80 % at $\beta \geq 0.6$. As β decreases into the range $[0.4, 0.2]$, the accuracy rises, indicating that the KL regularization has been sufficiently relaxed to allow the classifier to leverage discriminative features while still benefiting from topic-based smoothing. In particular, $\beta = 0.2$ yields near-peak performance for every dataset. SD-BERT achieves 86.02 % on MR and 89.12 % on 20NG, SD-RoBERTa reaches 88.10 % and 92.55 %, SD-DistilBERT attains 87.09 % and 86.40 %, and SD-MoBERT tops out at 89.97 % and 95.27 %, respectively—while further reductions of β below 0.2 produce only marginal gains or slight degradations.

On the Ohsumed, R8, and R52 corpora a similar "elbow" appears at $\beta = 0.2$: performance rises from the mid-70s to the high-70s or low-80s as β falls from 0.6 to 0.2, then asymptotes or even dips slightly for $\beta < 0.2$. This behaviour confirms that $\beta = 0.2$ achieves the optimal trade-off between enforcing the consistency of the topic model (via $\mathcal{L}_{\text{KL}}$) and preserving classification accuracy (via $\mathcal{L}_{\text{CE}}$) across all settings. We therefore fix $\beta = 0.2$ in subsequent experiments, as it uniformly delivers near-best or best accuracy with robust stability across data sets and model backbones.

E Hypothesis Testing: Statistical Comparison of SD-MoBERT and PaSIG-S

Table 3 summarizes the F1 mean, variance, standard deviation (S), 95% confidence intervals (CI), and two-sided p-values for SD-MoBERT versus the best baseline (PaSIG-S) across the five benchmarks.

We compute each 95% confidence interval using Greenland et al. (2016)

$$\mathrm{CI} \;=\; \mu \;\pm\; z^{*}\,\frac{s}{\sqrt{n}}\,, \quad z^{*} = 1.96, \tag{19}$$

where n is the number of evaluation runs. For example, on the MR dataset with $\mu = 88.13$, $S = 0.029$, and 30 trials, the resulting interval is [88.120 - 88.140].

To test whether SD-MoBERT's mean F1 differs from PaSIG-S, we formulate

$$H_0 \colon \mu_{\text{SD-MoBERT}} = \mu_{\text{PaSIG-S}} \quad \text{vs.} \quad H_1 \colon \mu_{\text{SD-MoBERT}} \neq \mu_{\text{PaSIG-S}} \tag{20}$$

We calculate the two-sided p-value as Greenland et al. (2016)

$$p = 2\big(1 - \mathrm{CDF}(|t|, df)\big), \quad df = n - 1, \tag{21}$$

where

$$\mathrm{CDF}(|t|, df) = \int_{-\infty}^{|t|} f(t, df)\, dt, \tag{22}$$

and the Student's t-distribution PDF is

$$f(t, df) = \frac{\Gamma\big(\frac{df+1}{2}\big)}{\sqrt{df\,\pi}\,\Gamma\big(\frac{df}{2}\big)} \Big(1 + \tfrac{t^2}{df}\Big)^{-\frac{df+1}{2}}. \tag{23}$$

where $df = n - 1$ denotes the degree of freedom and Γ represents the Gamma function.

All five datasets yield $p < 0.05$, thus, we reject the NULL hypothesis H_0 and accept the alternative hypothesis H_1. The extremely small p-values (e.g. 2.38×10^{-47} on MR) and tight confidence intervals demonstrate that SD-MoBERT's improvements over PaSIG-S are both statistically significant and consistently observed.

F Efficiency Analysis: Time Complexity and Runtime Cost

Table 4 compares six transformer-based classifiers in terms of their time complexity, approximate floating-point operations per token (FLOPs), and measured CPU inference time on a single Reuters R8 document. All experiments are conducted on a 12th Gen Intel(R) Core(TM) i7-12700K processor (3.60 GHz), 64GB RAM, and a 64-bit operating system. The baseline BERT and its long-context variant MoBERT both exhibit the familiar $\mathcal{O}(b \cdot L \cdot T^2 \cdot D)$ complexity, where b denotes the batch size, L the number of transformer layers, T the sequence length, and D the hidden dimension. BERT incurs approximately 148 GFLOPs per token and requires 0.74 ms to process a single R8 document, whereas MoBERT's optimizations reduce this to 118 GFLOPs and 0.59 ms.

Incorporating the smoothed-Dirichlet topic model adds an $\mathcal{O}(b\,(V \cdot H + H \cdot K))$ term (with vocabulary size V, topic-MLP hidden size H, and K topics). Thus SD-BERT's complexity becomes $\mathcal{O}(b\,(L\,T^2 D + V\,H + H\,K))$, raising FLOPs to 158 GFLOPs and inference time to 0.79 ms. SD-RoBERTa, which uses a larger embedding dimension D_{large}, further increases cost to 220 GFLOPs and 1.10 ms. DistilBERT's lighter backbone ($L' < L$) yields the fastest pure transformer variant: SD-DistilBERT achieves only 84 GFLOPs and 0.42 ms despite the same topic-model overhead. Finally, SD-MoBERT combines ModernBERT's quantization advantages with a small co-attention fusion ($\mathcal{O}(b\,(D'\,H'))$), resulting in $\mathcal{O}(b\,(L\,T^2 D + V\,H + H\,K + D'\,H'))$, 126 GFLOPs, and 0.63 ms. D' denotes the fusion layer output dimensionality, D_{large} is the larger embedding dimension in RoBERTa-base, and H' is the hidden layer size in the classification head.

Overall, MoBERT and SD-MoBERT strike the best balance between high capacity and low latency, while SD-DistilBERT offers the most lightweight option when computational resources are constrained.

Table 4: Comparison of time complexity, per-token FLOPs, and CPU inference latency on the Reuters R8 dataset (single document) for BERT, MoBERT, and their smoothed-Dirichlet variants.

Model	Time complexity	FLOPs	CPU Time (ms)
BERT	$\mathcal{O}(b \cdot L \cdot T^2 \cdot D)$	148 GFLOPs	0.74
MoBERT	$\mathcal{O}(b \cdot L \cdot T^2 \cdot D)$	118 GFLOPs	0.59
SD-BERT	$\mathcal{O}(b \cdot (L \cdot T^2 \cdot D + V \cdot H + H \cdot K))$	158 GFLOPs	0.79
SD-RoBERTa	$\mathcal{O}(b \cdot (L \cdot T^2 \cdot D_{large} + V \cdot H + H \cdot K))$	220 GFLOPs	1.1
SD-DistilBERT	$\mathcal{O}(b \cdot (L' \cdot T^2 \cdot D + V \cdot H + H \cdot K))$	84 GFLOPs	0.42
SD-MoBERT	$\mathcal{O}(b \cdot (L \cdot T^2 \cdot D + V \cdot H + H \cdot K + D' \cdot H'))$	126 GFLOPs	0.63

Listing 1 Smoothed-Dirichlet MLP forward function; see SMDIRICHLET class below.

```
    def forward(self, input_bows):
        # Run BOW through MLP
        pi = self.mlp(input_bows)

        # Use this to get rho1, log_rho2 for Dirichlet
        rho1 = self.rho1(pi)
        logrho2 = self.rho2(pi)
        alpha = self.alpha(pi)

        epsilons = torch.normal(0, 1, size=(
            input_bows.size()[0], self.num_topics)).to(input_bows.device)

        sample, alpha_smoothed  = self.reparameterize(alpha, rho1, logrho2, epsilons)


        logits = self.log_softmax(self.dec_projection(sample))


        kld = self.kld(alpha_smoothed , prior_alpha = torch.tensor(0.01), epsilon=torch.tensor

        rec_loss = -1 * torch.sum(logits * input_bows, 1)
        loss_nvdm_lb = torch.mean(rec_loss + kld)

        return sample, logits, torch.mean(kld), loss_nvdm_lb
```

Listing 2 Smoothed-Dirichlet MLP (SMDIRICHLET class)

```python
import torch
import torch.nn as nn
import torch.nn.functional as F

class SMDIRICHLET(nn.Module):

    @staticmethod
    def _param_initializer(module):
        if isinstance(module, nn.Linear):
            nn.init.xavier_normal_(module.weight)

        if isinstance(module, nn.Linear) and module.bias is not None:
            module.bias.data.zero_()

    def __init__(self, vocab_size, num_topics=10, hidden_size=256, hidden_layers=1, nonlineari
        super().__init__()
        self.num_topics = num_topics
        self.vocab_size = vocab_size

        # First MLP layer compresses from vocab_size to hidden_size
        mlp_layers = [nn.Linear(vocab_size, hidden_size), nonlinearity()]
        # Remaining layers operate in dimension hidden_size
        for _ in range(hidden_layers - 1):
            mlp_layers.append(nn.Linear(hidden_size, hidden_size))
            mlp_layers.append(nonlinearity())

        self.mlp = nn.Sequential(*mlp_layers)
        self.mlp.apply(SMDIRICHLET._param_initializer)

        # Create linear projections for Dirichlet params (rho1 & rho2)
        self.rho1 = nn.Linear(hidden_size, num_topics)
        self.rho1.apply(SMDIRICHLET._param_initializer)

        # Custom initialization for rho2
        self.rho2 = nn.Linear(hidden_size, num_topics)
        self.rho2.bias.data.zero_()
        self.rho2.weight.data.fill_(0.)

        # create linear projrction for alpha
        self.alpha = nn.Linear(hidden_size, num_topics)
        self.alpha.apply(SMDIRICHLET._param_initializer)

        self.dec_projection = nn.Linear(num_topics, vocab_size)
        self.log_softmax = nn.LogSoftmax(-1)

    def reparameterize(self, alpha, rho1, logrho2, eps):
        rho2 = torch.exp(logrho2)
        #eps = torch.randn_like(std)
        alpha_smoothed = alpha + eps * rho2 + rho1

        Z_sd =  F.softmax(alpha_smoothed, dim=1)

        return Z_sd, alpha_smoothed

    def kld(self, model_alpha, prior_alpha, epsilon):

        model_alpha = torch.max(torch.tensor(0.0001), model_alpha).to(model_alpha.device)
        alpha = prior_alpha.expand_as(model_alpha)
        sum1 = torch.sum((model_alpha + epsilon - 1) * torch.digamma(model_alpha + epsilon), d

        sum2 = torch.sum((alpha + epsilon - 1) * torch.digamma(alpha + epsilon), dim=1)
        kl_loss = torch.mean(sum1 - sum2)

        return kl_loss
```

Listing 3 Smoothed Dirichlet ModernBERT

```
    '''This module contains the SD-ModernBERT model with a co-attention mechanism and dynamic

import torch
import torch.nn as nn
from transformers import ModernBertModel
from models.smdirichlet import SMDIRICHLET


class TopicBERT(nn.Module):
    '''This module contains the SD-ModernBERT model with a co-attention mechanism and dynamic t
    def __init__(self, vocab_size, num_labels, alpha=0.9, dropout=0.1):
        super().__init__()
        self.encoder = ModernBertModel.from_pretrained('answerdotai/ModernBERT-base')
        self.smdirichlet = SMDIRICHLET(vocab_size)

        # Co-attention projection layers
        hidden_size = self.encoder.config.hidden_size
        topic_dim = self.smdirichlet.num_topics
        self.co_attn_b = nn.Linear(hidden_size, hidden_size, bias=False)
        self.co_attn_t = nn.Linear(topic_dim, hidden_size, bias=False)
        self.attn_bias = nn.Parameter(torch.zeros(1))

        # Combine co-attended representation
        self.combine_proj = nn.Linear(hidden_size, hidden_size)

        # Classification head
        self.projection = nn.Sequential(
            nn.Dropout(dropout),
            nn.Linear(hidden_size, hidden_size, bias=False),
            nn.GELU(),
            nn.Linear(hidden_size, num_labels)
        )
        self.projection.apply(TopicBERT._get_init_transformer(self.encoder))

        self.bert_loss = nn.CrossEntropyLoss(reduction='mean')

    @staticmethod
    def _get_init_transformer(transformer):
        def init_transformer(module):
            if isinstance(module, (nn.Linear, nn.Embedding)):
                module.weight.data.normal_(mean=0.0, std=transformer.config.initializer_range)
            elif isinstance(module, nn.LayerNorm):
                module.bias.data.zero_()
                module.weight.data.fill_(1.0)
            if isinstance(module, nn.Linear) and module.bias is not None:
                module.bias.data.zero_()
        return init_transformer

```

Listing 4 Smoothed Dirichlet ModernBERT forward function

```
    def forward(self, input_ids, attention_mask, bows, labels):
        # BERT encoding
        hiddens_last = self.encoder(input_ids, attention_mask=attention_mask)[0]
        embs = hiddens_last[:, 0, :]  # [CLS] token embeddings

        # Topic model encoding
        h_tm, _, kld, loss_diri = self.smdirichlet(bows)

        # Co-attention
        proj_b = self.co_attn_b(embs)             # (batch_size, hidden_size)
        proj_t = self.co_attn_t(h_tm)            # (batch_size, hidden_size)
        # Compute affinity and attention weight
        scores = torch.sum(proj_b * proj_t, dim=1, keepdim=True) + self.attn_bias  # (batch_si
        alpha = torch.sigmoid(scores)            # (batch_size, 1)
        #alpha = self.softmax(scores)
        # Fuse representations
        joint = alpha * proj_b + (1 - alpha) * proj_t
        co_emb = torch.tanh(self.combine_proj(joint))  # (batch_size, hidden_size)

        # Classification
        logits = self.projection(co_emb)

        # Loss computation
        loss_bert = self.bert_loss(logits, labels.max(1).indices)
        loss_total = loss_bert + kld * 0.2
        return logits, loss_total, kld
```