Synergizing Probabilistic Models with Deep Learning: A Novel Framework for Topic Modeling and Multimodal Analysis

Akinlolu Oluwabusayo Ojo

A Ph.D. Thesis
in
The Department
of
Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy (Information and Systems Engineering) at
Concordia University
Montréal, Québec, Canada

2026-01-28

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Akinlolu Oluwabusayo Ojo**

Entitled: **Synergizing Probabilistic Models with Deep Learning: A Novel Framework for Topic Modeling and Multimodal Analysis**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

Dr. Jinqiu Yang	Chair
Dr. Carson Leung	External Examiner
Dr. Abdelhak Bentaleb	External To Program
Dr. Abdessamad Ben Hamza	Examiner
Dr. Nizar Bouguila	Supervisor

Approved by

Chun Wang, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

January 23, 2026

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Synergizing Probabilistic Models with Deep Learning: A Novel Framework for Topic Modeling and Multimodal Analysis

Akinlolu Oluwabusayo Ojo, Ph.D.

Concordia University, 2026

Traditional topic modeling techniques, such as Latent Dirichlet Allocation (LDA), often employ standard Dirichlet distributions to model topic-word and document-topic relationships. While effective in certain contexts, these methods face limitations in capturing the complex dependencies, uncertainty, and heterogeneity in real-world data. These limitations become particularly pronounced in multimodal and cross-domain settings, where textual, visual, and spectral signals interact nonlinearly and where distribution shifts across platforms, languages, and domains are pervasive. While deep learning methods have advanced multimodal understanding and classification, many lack principled probabilistic foundations, leading to deterministic latent representations, component collapse, and poor robustness under domain shift. This dissertation addresses these challenges by introducing a unified probabilistic deep learning framework built upon expressive distributional priors, including the generalized Dirichlet, smoothed Dirichlet, and Beta-Liouville distributions. The proposed framework enables continuous latent representations that capture rich covariance structures, uncertainty, and asymmetry beyond standard Dirichlet and Gaussian assumptions. Within this paradigm, we develop a series of models spanning topic modeling, multimodal fusion, and cross-domain adaptation. We first propose a Generalized Dirichlet Variational Autoencoder (GD-VAE) for neural topic modeling, followed by smoothed Dirichlet-based multimodal architectures for fake news detection, including SmoothDetector and SD-MoBERT, which integrate probabilistic topic modeling with long-context transformer representations. To address robustness under distribution shift, we further introduce EviDA, an uncertainty-weighted domain adversarial learning framework that leverages evidential deep learning to adaptively modulate instance-level domain alignment in cross-domain and cross-lingual fake news detection. Finally, we propose PerLiFuse, a per-frequency Beta-Liouville fusion network operating in the spectral domain, which learns dynamic, example-specific gating across frequency bands to reconcile conflicting multimodal cues and mitigate fusion collapse. Extensive empirical evaluations across multiple benchmark datasets demonstrate consistent improvements in topic coherence, diversity, classification accuracy, robustness to domain shift,

and uncertainty calibration over state-of-the-art baselines. Collectively, this work establishes a principled and extensible paradigm for probabilistic deep learning, enabling interpretable, robust, and scalable models for multimodal understanding and misinformation detection in complex real-world environments.

Acknowledgments

I would like to express my profound gratitude to my supervisor, Professor Nizar Bouguila, for his exceptional guidance, unwavering support, and invaluable mentorship throughout this doctoral journey. Your expertise, patience, and dedication have been instrumental in shaping my research and advancing my understanding. Thank you for believing in my potential and for providing me with the resources and opportunities to grow as a researcher.

I am deeply grateful to the members of my examining committee for their time, expertise, and constructive feedback, which have significantly enhanced the quality of this thesis. Your insightful comments and suggestions have been invaluable in refining my research.

I extend my sincere appreciation to the faculty and staff at the Concordia Institute for Information Systems Engineering (CIISE) for providing an excellent academic environment and research facilities that made this work possible. I also acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Mitacs, which enabled me to pursue this research.

I am grateful to my colleagues and friends at the XAI Lab for the stimulating discussions, collaborative spirit, and memorable moments we shared. Your camaraderie and support made this journey more enjoyable and enriching.

My heartfelt thanks go to my parents, Mr. Ojo Richard and Mrs. Ojo Florence, whose sacrifices, prayers, and unwavering belief in my abilities have brought me to this point. Your encouragement and values have been the foundation of my success. To my brothers and sisters, thank you for your constant support, love, and for always being there for me.

To my beloved wife, Ojo-Akinlolu Folasade, words cannot express how grateful I am for your love, patience, and understanding throughout this challenging journey. You have been my anchor, my motivation, and my greatest source of strength. Thank you for standing by me through the late nights, the stressful moments, and for keeping our home filled with warmth and joy.

To my children, you are my inspiration and my reason for striving to be better every day. Thank you for your understanding when I had to work long hours and for bringing so much happiness into my life. I hope this achievement shows you that with dedication and perseverance, anything is possible.

Finally, I thank the Almighty God for His grace, protection, and blessings throughout this journey.

Contributions of Authors

This Ph.D. thesis consists of six manuscripts. Two manuscripts have been published, one has been accepted for publication, and the rest have been submitted for publication in refereed academic journals and conferences. Each chapter consists of the content of a manuscript which has been reformatted and reorganized according to the requirements set out in the guidelines by the School of Graduate Studies.

* **Manuscript 1 (Chapter 2):** Akinlolu Oluwabusayo Ojo and Nizar Bouguila, "A Topic Modeling and Image Classification Framework: The Generalized Dirichlet Variational Autoencoder", Pattern Recognition, Vol. 146, article 110037, February 2024.

* **Manuscript 2 (Chapter 3):** Akinlolu Oluwabusayo Ojo, Fatma Najar, Narjes Zamzami, Hassen Taher Himdi, and Nizar Bouguila, "SmoothDetector: A Smoothed Dirichlet Multimodal Approach for Combating Fake News on Social Media", IEEE Access, Vol. 13, pp. 39289–39305, February 2025.

* **Manuscript 3 (Chapter 4):** Akinlolu Oluwabusayo Ojo and Nizar Bouguila, "Smoothed-ModernBERT: Co-Attentional Synergy of Probabilistic Topic Models and ModernBERT through Dynamic Fusion". Submitted to IEEE Transactions on Artificial Intelligence (2025).

* **Manuscript 4 (Chapter 5):** Akinlolu Oluwabusayo Ojo and Nizar Bouguila, "DeepBetaL: Deep Learning for Multimodal Fake News Detection with Beta-Liouville Priors". Submitted to International Journal of Machine Learning and Cybernetics (2025).

* **Manuscript 5 (Chapter 6):** Akinlolu Oluwabusayo Ojo and Nizar Bouguila, "EviDA: Cross-Domain Fake News Detection via Uncertainty-Weighted Domain Adversarial Learning". Submitted to IJCAI 2026.

* **Manuscript 6 (Chapter 7):** Akinlolu Oluwabusayo Ojo and Nizar Bouguila, "PerLiFuse:

Per-Frequency Beta-Liouville Fusion Networks for Fake News Detection". Accepted at IEEE Conference on Artificial Intelligence 2026.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Generative models have gained prominence in Natural Language Processing (NLP) due to their capacity to effectively learn from unstructured and unlabeled corpora. Conventional topic models like LDA assume data can be represented as mixtures of independent components [2], an assumption inadequate for complex, multimodal datasets with interdependencies, such as documents combining text and images where visual and textual content influence each other. Traditional models also rely on the Dirichlet distribution, imposing a restrictive covariance structure with only negative correlations between variables, limiting their ability to capture realistic positive and negative topic correlations. Additionally, computing exact posteriors requires summing over all latent variables, restricting model complexity due to intractable high-dimensional integrals [3].

Several studies have been conducted to circumvent this limitation by approximating the model posterior with a simple distribution. This is done by minimizing the Kullback-Leibler divergence between the posterior and a simple distribution. Markov chain Monte Carlo (MCMC) [4] that uses collapsed Gibbs sampling and variational mean field [3] methods are the traditional ways for approximating the integrals. However, it is challenging and computationally expensive to apply them to new topic models. Any little changes in the model assumption necessitate re-deriving the inference methods, this makes it exhaustive for researchers to achieve fine-tuned models by exploring various modeling assumptions [3]. Thus, developing inference approaches that can be easily applied to a new model, even with some changes in the model without re-deriving the inference methods has drawn research interest. One significant way of addressing this limitation is by applying a neural topic model, which integrates the black-box mechanism with neural networks. A promising black-box inference approach is proposed in [5]. These models can be readily applied to a new topic

model with little information about the generative process. Neural networks have been shown to learn nonlinear distributions and are capable of approximating complex functions [6]. Autoencoding variational models proposed in [7] can use a variational distribution parameterized by a neural network to train an inference network and approximate the posterior of a classification model. The inference network can be applied directly to test data without any further computational cost.

Despite some remarkable achievements of Dirichlet latent neural network models, they hardly identify meaningful topics. Two major challenges were pointed out in [1]. Firstly, local optimum, which is attributed to the problem of component collapsing [8]. Secondly, Dirichlet prior does not have a shifting parameter because it's not in the location families of distributions. Reparameterization trick works quite well with differentiable distributions that have location-scale parameters or that can be expressed as deterministic transformations of such distributions [9]. This makes it difficult to use Dirichlet distribution with a reparameterization gradient. To address this problem, [1] adopted a regularization technique, batch normalization to avoid being stuck in a local minimum. They also proposed autoencoder variational inference for topic models (AVITM), which uses Logistic-Normal prior to mimic the simplex in latent topic space. For an efficient reparameterization trick, they further proposed a topic model called ProdLDA where they replaced the word-level mixture with a weighted product of experts. ProdLDA explored Laplace approximation for the Dirichlet distribution to enhance the training of the Dirichlet variational autoencoder. [10] explored implicit differentiation to compute reparameterization gradients. In [11], slice sampling was used to compute reparameterization gradients, and [12] used Weibull distribution to approximate Gamma distribution with an analytic Kullback-Leibler divergence. Furthermore, [13] approximated the Gamma distribution with the inverse Gamma CDF to infer the parameters of the Dirichlet variational autoencoder. [14] followed a different approach by applying rejection sampling variational inference of Gamma distribution in [15]. Our work follows a similar approach but with the Beta distribution's rejection sampling variational inference [15]. In addition, deep learning has revolutionized various domains by learning hierarchical representations directly from data [16]. In the context of topic modeling, deep learning architectures have been employed to learn discrete topic assignments. However, without integrating probabilistic frameworks, these models often resort to discrete representations, which may not fully encapsulate the continuous nature of topic distributions. This discretization can lead to a loss of nuanced information and may not adequately reflect the uncertainty and variability present in real-world data. This may lead to a lack of interpretability, making it challenging to derive meaningful insights from the learned representations.

Integrating advanced probabilistic distributions into deep learning architectures offers a promising avenue to address these challenges. The generalized Dirichlet (GD) distribution has a better covariance structure than Dirichlet distribution, this makes it to be more useful and practically applicable [17]. GD is a special case of Dirichlet trees [18], which have been previously used in mixture models to add domain knowledge to the conditional probability of words given a topic and to capture topic correlation in mixture models [19]. Its ability to enable the representation of complex dependencies between topics is particularly beneficial in multimodal datasets where such correlations are inherent. Similarly, the Smoothed Dirichlet distribution introduces smoothing parameters to mitigate issues related to sparse data and component collapse, leading to more robust and stable feature representations and enhancing the model's ability to generalize from limited observations, a common scenario in real-world applications [20]. Additionally, the Beta-Liouville distribution offers a flexible framework for modeling data with bounded support, advantageous in scenarios where data components are interdependent or skewed, making it suitable for capturing the intricate structures present in multimodal datasets [21]. Unlike Dirichlet, Beta-Liouville also captures positive and negative correlations by sampling latent representations with a Beta-Liouville prior, improving generalization and effectively handling asymmetries and uncertainties arising from conflicting modalities.

Building upon these insights, this proposal advocates for a synergistic integration of probabilistic modeling with deep learning architectures to enhance topic modeling and multimodal analysis. By embedding these distributions within neural frameworks, we aim to develop models that not only capture the complex dependencies and continuous latent structures inherent in multimodal data but also provide interpretable and coherent topic representations. This integration is anticipated to overcome the limitations of traditional models and pure deep learning approaches, leading to more accurate and insightful analyses across diverse datasets. Our proposal spans a constellation of interdisciplinary domains, from latent thematic inference and discriminative taxonomization to generative data enrichment and cross-modal orchestration, among other frontiers of intelligent data interaction.

1.1 Background and Related Works

This section provides a comprehensive overview of the foundational concepts and related research that underpin this thesis. It reviews key developments in probabilistic topic modeling and variational

autoencoders, followed by a discussion of unimodal and multimodal fake news detection methods. The section also examines cross-domain learning and adversarial training paradigms, establishing the necessary theoretical and methodological background for understanding the proposed models and approaches developed in this work.

1.1.1 Mixture Models

Topic models have become research-interest areas due to their applications in document modeling and information retrieval systems. LDA is a generative mixture model that has been extensively explored to learn the relationships between words in the corpus. It promotes sparsity, therefore producing more interpretable topics [22]. [22] successfully applied LDA to document modeling. However, LDA fails to capture the correlation between topics. A good solution to this is to use the generalized Dirichlet distribution, which has a more general covariance structure than the Dirichlet distribution [17]. A closer work to our proposed model that used a generalized Dirichlet mixture model to capture the topics' correlation can be found in [19]. They used generalized Dirichlet as a prior distribution for a document to the topic mixture and developed a hierarchical tree structure model that accumulates the most important topics at the upper levels. In [23], a generalized Dirichlet multinomial mixture model that uses Gibbs sampling is proposed. They used the proposed model to cluster short texts. However, none of these works integrates neural network variational inference with their mixture models. Applying them to new topics is computationally expensive and very challenging for researchers to achieve a well-fine-tuned model without re-deriving the inference [3].

1.1.2 Neural Network Variational Inference Topic Models

Although probabilistic mixture models are popular in topic modeling, effective and efficient inference for models with complex and deep structures is critical. To meet these demanding tasks, a replicated softmax model (RSM) was proposed in [24]. RSM is a directionless, two-layer graphical model that extracts low-dimensional hidden semantic information from a large unstructured corpus. The work proposed in [25] was inspired by RSM. They estimated the uncertainty of observing a new word in a text, given the already observed words, by substituting a hierarchical distribution for the RSM's softmax distribution. Contrary to RSM and its extension, a directed non-iterative feedforward network was proposed in [5]. The work in [26] explored mutual information estimation on topic modeling. To address the inefficient Gaussian prior, ProdLDA proposed in [1] showed that approximating the Dirichlet prior using a Gaussian reparameterization leads to improved training

and better topic modeling. They extended their work and replaced the mixture model with a product of experts. ProdLDA achieved better topic coherence than previous works and competitive perplexity. [27] introduced bidirectional adversarial training for topic models. They reported that their model considered word relatedness and outperformed the state-of-the-art in terms of topic coherence. [28] explores dynamic graph convolutional recurrent network for traffic forecasting, while [29] investigates sentence-aware encoder for topic modeling. Another closer work to our proposed model is reported in [14]. They reported the trade-off between sparsity and smoothness in the Dirichlet distribution since both factors are jointly encoded in the Dirichlet parameter. They decoupled these properties by representing the Dirichlet parameter as a product of sparse binary and smoothness vectors. Although this subject of study has been extensively explored, none of the previous works have integrated generalized Dirichlet, smoothed Dirichlet, and Beta-Liouville distributions with neural network architectures.

1.1.3 Unimodal Approaches to Fake News Detection

Early approaches to fake news detection primarily focused on text analysis, applying linguistic, semantic, and structural features to detect deceptive content. Linguistic methods examine stylistic markers, including grammar, syntax, and lexical choices, to differentiate fake from legitimate news [30]. Semantic-based methods analyze the context and meaning of the text to identify discrepancies and contradictions, often using knowledge graphs to correlate statements within the content [31, 32, 33]. Structural features, such as clickbait-style headlines or sentiment-based analysis, have also been helpful in identifying misinformation [34]. Despite these advances, unimodal text-based methods have limitations, especially as fake news content becomes more multimodal. Text-only detection struggles when fake news combines misleading images, videos, or audio alongside textual misinformation, as is often the case on social media platforms like Twitter and Facebook. Consequently, researchers have recognized the need for multimodal approaches that could integrate insights across several types of data to improve detection performance [35].

1.1.4 Multimodal Fake News Detection

The complexities associated with fake news have prompted a shift towards multimodal fake news detection, which seeks to leverage the complementary strengths of different text and images, but sometimes audio and video as well. Multimodal approaches can combine features such as visual elements in images or videos with linguistic attributes in the accompanying text, which provides a

richer representation of the content for analysis [36]. The studies in [37] identified the significant advantage of multimodal approaches as their ability to identify cross-modal inconsistencies, such as mismatched text and image pairs, which are common in fake news. One of the pioneering methods for multimodal fake news detection involves simple concatenation of visual and textual features, enabling a unified representation that improves detection accuracy. For instance, the EANN model introduced by [38] leverages visual and textual features along with event adversarial networks to learn invariant features for different news events, which enhances the generalizability of the detection model. However, such models rely on fixed feature concatenations, lacking flexibility in representing complex interdependencies between modalities.

Recent studies have explored more complex feature fusion techniques to improve the interaction between different modalities. The authors in [39] introduced a multimodal model that utilizes attention mechanisms to emphasize important information within each modality, adapting the focus dynamically to relevant features. Although effective, these models still lack stochastic elements in their latent representations, which limits their adaptability to different contexts and data distributions. The resulting fixed latent spaces struggle to capture the nuanced and dynamic nature of relationships between modalities, such as the subtle discrepancies in fake news content.

1.1.5 Probabilistic Models and Latent Representations

Stochastic and probabilistic modeling has gained traction as methods for capturing more flexible and adaptable latent representations. Probabilistic approaches such as variational autoencoders (VAEs) [40] and Bayesian models provide mechanisms for representing complex, high-dimensional data in a probabilistic latent space, allowing models to capture inherent uncertainties and variability within the data. These probabilistic latent representations have demonstrated strong performance in other fields, including image generation [41] and text processing [22], by enabling adaptive and flexible feature learning. However, their application in fake news detection, particularly in multimodal contexts, remains underexplored. The stick-breaking smoothed-Dirichlet distribution, as proposed in our study, introduces a probabilistic structure that can model diverse and complex relationships across modalities, facilitating a more continuous latent representation. Similar probabilistic models, including Dirichlet processes [42] and hierarchical Bayesian models, have shown promise in handling multimodal data by adapting to the inherent complexities within and across data types. This adaptive structure is particularly suitable for fake news detection, where the relationships between text and image content can vary widely, requiring flexible representations that

can handle these variations [43]. These methods improve flexibility but suffer from issues such as component collapse, which reduces generalization and robustness.

1.1.6 Cross-Domain Fake News Detection

Early fake news detection research primarily focused on single-domain settings. The studies in [44] demonstrated strong in-domain performance but reported severe degradation when models were deployed across different platforms. These studies highlighted the sensitivity of fake news detectors to shifts in linguistic style, topic distribution, and user behavior. To address cross-domain generalization, several domain-aware architectures have been proposed. [45] explored domain-specific feature extractors to mitigate distribution mismatch. Although effective when domain identities are known, these methods struggle with unseen target domains and require explicit domain labels at inference time.

Adversarial domain adaptation has emerged as a promising alternative. Building on [45], [46] demonstrated that adversarial alignment can improve cross-platform transferability. However, these methods apply uniform adversarial pressure across all samples, implicitly assuming homogeneous domain shift across samples. In practice, domain shift is highly heterogeneous, motivating the need for adaptive, instance-level alignment strategies. Recent efforts have also focused on benchmarking and evaluation. The studies in [47] provided standardized cross-platform evaluation protocols and demonstrated the limitations of existing approaches under severe domain discrepancy. Our work builds upon these benchmarks by introducing uncertainty-guided adaptive alignment.

1.1.7 Uncertainty Estimation and Evidential Deep Learning

Uncertainty estimation has been widely studied as a mechanism for assessing model reliability under distribution shift. [48] introduced a principled framework for modeling epistemic uncertainty via Dirichlet distributions over class probabilities. This approach was extended in [49] and further analyzed in [50], which demonstrated strong correlations between epistemic uncertainty and out-of-distribution samples.

Subsequent studies, including [51], compared evidential learning against Monte Carlo dropout and deep ensembles, highlighting its advantages in calibration, efficiency, and interpretability. These works establish epistemic uncertainty as a reliable signal of distribution mismatch. While uncertainty has been leveraged for tasks such as active learning and post-hoc domain shift detection [52], its role in guiding the optimization of domain adaptation objectives remains underexplored. In

particular, existing fake news detection methods do not integrate uncertainty as an instance-level control signal within adversarial training. Our work addresses this gap by embedding evidential epistemic uncertainty directly into the domain adversarial learning process.

1.2 Contributions

1. We introduce a Generalized Dirichlet Variational Autoencoder (GD-VAE), integrating neural variational inference with the generalized Dirichlet distribution to overcome the restrictive covariance structure of standard Dirichlet priors. GD-VAE captures both positive and negative topic correlations, effectively decouples sparsity and smoothness, and achieves improved topic diversity, uniqueness, and perplexity across multiple benchmark corpora.

2. We extend GD-VAE to an unbounded stick-breaking-assisted generalized Dirichlet variational autoencoder, alleviating the constraint of fixed latent dimensionality. This extension enables adaptive inference of the effective number of topics while preserving probabilistic consistency and stable training through rejection-sampling-based reparameterization.

3. We propose SmoothDetector, a smoothed Dirichlet-based multimodal fake news detection framework that integrates BERT for textual encoding and VGG19 for visual representation learning. By incorporating smoothed Dirichlet priors, SmoothDetector mitigates component collapse, models uncertainty in continuous latent space, and achieves robust generalization across heterogeneous multimodal social media data.

4. We introduce Smoothed Dirichlet-ModernBERT (SD-MoBERT), a unified architecture that synergizes neural topic modeling with long-context transformer representations. Through a dynamic co-attention mechanism aligning thematic latent variables with contextual embeddings, SD-MoBERT enhances interpretability, robustness, and classification performance across multiple large-scale text classification benchmarks.

5. We propose DeepBetaL, a probabilistic deep multimodal learning framework based on the Beta–Liouville prior. DeepBetaL captures asymmetric and bounded latent dependencies, models rich covariance structures beyond generalized Dirichlet distributions, and improves robustness and expressiveness in multimodal fake news detection under conflicting evidence.

6. We develop EviDA, an uncertainty-weighted domain adversarial learning framework for cross-domain and cross-lingual fake news detection. By leveraging evidential deep learning to quan-

tify epistemic uncertainty, EviDA dynamically modulates instance-level domain alignment strength, enabling adaptive adversarial training that significantly improves robustness under heterogeneous domain shift.

7. We introduce PerLiFuse, a per-frequency Beta-Liouville fusion network that operates in the spectral domain via Discrete Cosine Transform (DCT). PerLiFuse employs flexible Beta-Liouville priors, Kumaraswamy reparameterization, and coherence-guided gating to learn dynamic, example-specific cross-modal fusion policies, effectively disentangling semantic alignment from high-frequency manipulation cues and mitigating fusion collapse.

The remainder of this dissertation is organized as follows. Chapter 2 presents the Generalized Dirichlet Variational Autoencoder (GD-VAE), which establishes the foundational probabilistic framework of this work by relaxing the restrictive covariance assumptions of traditional Dirichlet-based models and explicitly capturing both positive and negative topic correlations in topic modeling and image classification. Building on this foundation, Chapter 3 introduces SmoothDetector, a smoothed Dirichlet-based multimodal approach for fake news detection on social media that extends probabilistic latent modeling to multimodal settings, enabling continuous uncertainty-aware representations across textual and visual modalities. While SmoothDetector focuses on multimodal representation learning under a shared data distribution, Chapter 4 advances this line of research by proposing Smoothed-ModernBERT, a hybrid architecture that integrates probabilistic topic modeling with long-context transformer representations through a dynamic co-attention mechanism, thereby enhancing robustness and interpretability in document classification. Despite these advances in multimodal and document-level modeling, challenges remain in capturing asymmetric and bounded dependencies between modalities; accordingly, Chapter 5 presents DeepBetaL, a multimodal fake news detection framework based on Beta-Liouville priors that mitigates component collapse while modeling richer cross-modal interactions. Beyond in-domain learning, real-world deployment introduces distribution shifts across platforms and languages; Chapter 6 addresses this challenge through EviDA, an uncertainty-weighted domain adversarial learning framework that leverages evidential deep learning to achieve robust cross-domain and cross-lingual fake news detection. Finally, shifting the focus from representation alignment to fusion mechanics, Chapter 7 introduces PerLiFuse, a Bayesian spectral fusion framework that operates in the frequency domain using per-frequency Beta-Liouville gating to dynamically reconcile conflicting multimodal cues. To conclude, Chapter 8 summarizes the principal contributions of this dissertation and discusses their

broader implications for probabilistic deep learning, multimodal representation learning, and trustworthy misinformation detection.

Chapter 2

A Topic Modeling and Image Classification Framework: The Generalized Dirichlet Variational Autoencoder

Latent Dirichlet allocation model (LDA) has been widely used in topic modeling. Recent works have shown the effectiveness of integrating neural network mechanisms with this generative model for learning text representation. However, one of the significant setbacks of LDA is that it is based on a Dirichlet prior that has a restrictive covariance structure. All its variables are considered to be negatively correlated, which makes the model restrictive. In a practical sense, topics can be positively or negatively correlated. To address this problem, we proposed a generalized Dirichlet variational autoencoder (GD-VAE) for topic modeling. The Generalized Dirichlet (GD) distribution has a more general covariance structure than the Dirichlet distribution because it takes into account both positively and negatively correlated topics in the corpus. Our proposed model leverages rejection sampling variational inference using a reparameterization trick for effective training. GD-VAE compares favorably to recent works on topic models on several benchmark corpora. Experiments show that accounting for topics' positive and negative correlations results in better performance. We further validate the superiority of our proposed framework on two image data sets. GD-VAE demonstrates its significance as an integral part of a classification architecture.

2.1 Introduction

The popularity of generative models within the field of Natural Language Processing can be attributed to their ability to learn unstructured and unlabelled corpora effectively. Their intuition with which they induce latent topics from corpora has been a resounding success among unsupervised learning algorithms. However, computing their exact posterior involves summing or integrating over all the latent variables, which can be up to millions or billions for complex models. Therefore, their achievements are limited by the complexity of the models dealing with the intractability of its inference due to the required high dimensional integrals [3].

Several works have been done to circumvent this limitation by approximating the model posterior with a simple distribution. This is done by minimizing the Kullback-Leibler divergence between the posterior and a simple distribution. Markov chain Monte Carlo (MCMC) [4] that uses collapsed Gibbs sampling and variational mean field [3] methods are the traditional ways for approximating the integrals. However, it is challenging and computationally expensive to apply them to new topic models. Any little changes in the model assumption necessitate re-deriving the inference methods, this makes it exhaustive for researchers to achieve fine-tuned models by exploring various modeling assumptions [3]. Thus, developing inference approaches that can be easily applied to a new model, even with some changes in the model without re-deriving the inference methods has drawn research interest.

One significant way of addressing this limitation is by applying a neural topic model, which integrates the black-box mechanism with neural networks. A promising black-box inference approach is proposed in [5]. These models can be readily applied to a new topic model with little information about the generative process. Neural networks have been shown to learn nonlinear distributions and are capable of approximating complex functions [6]. Autoencoding variational models proposed in [7] can use a variational distribution parameterized by a neural network to train an inference network and approximate the posterior of a classification model. The inference network can be applied directly to test data without any further computational cost.

Despite some remarkable achievements of latent neural network models, they hardly identify meaningful topics. Two major challenges were pointed out in [1]. Firstly, local optimum, which is attributed to the problem of component collapsing [8]. Secondly, Dirichlet prior does not have a shifting parameter because it's not in the location families of distributions. Reparameterization trick works quite well with differentiable distributions that have location-scale parameters or that

can be expressed as deterministic transformations of such distributions [9]. This makes it difficult to use Dirichlet distribution with a reparameterization gradient. To address this problem, [1] adopted a regularization technique, batch normalization to avoid being stuck in a local minimum. They also proposed autoencoder variational inference for topic models (AVITM), which uses Logistic-Normal prior to mimic the simplex in latent topic space. For an efficient reparameterization trick, they further proposed a topic model called ProdLDA where they replaced the word-level mixture with a weighted product of experts. ProdLDA explored Laplace approximation for the Dirichlet distribution to enhance the training of the Dirichlet variational autoencoder. [10] explored implicit differentiation to compute reparameterization gradients. In [11], slice sampling was used to compute reparameterization gradients, and [12] used Weibull distribution to approximate Gamma distribution with an analytic Kullback-Leibler divergence. Furthermore, [13] approximated the Gamma distribution with the inverse Gamma CDF to infer the parameters of the Dirichlet variational autoencoder. [14] followed a different approach by applying rejection sampling variational inference of Gamma distribution in [15]. Our work follows a similar approach but with Beta distribution's rejection sampling variational inference [15].

However, neither of these approaches tackles the fundamental limitation of LDA, which pertains to the constrained covariance structure imposed by the Dirichlet prior, where all its variables exhibit negative correlations. It makes more sense to capture both the positive and negative correlations between topics, as this will mimic the reality of practical applications. In this work, we proposed a new topic model, called Generalized Dirichlet Variational Autoencoder (GD-VAE). Generalized Dirichlet distribution has a better covariance structure than Dirichlet distribution, this makes it to be more useful and practically applicable [17]. GD is a special case of Dirichlet Trees [18], which has been previously used in mixture models to add domain knowledge to the conditional probability of words given a topic and to capture topic correlation in mixture model [19]. In contrast, our work integrates GD with the neural network black-box mechanism. Our work outperformed baseline models on different benchmark corpora. Furthermore, we demonstrate how GD-VAE can be used in machine learning tasks relating to image datasets. Experiments also showed that GD-VAE can play a significant role in data augmentation and image classification. The main contributions of this chapter are summarized as follows:

1. We propose a generalized Dirichlet variational autoencoder model. GD-VAE combines the effectiveness of neural network mechanisms in learning complex distributions and the advantage of using generalized Dirichlet as an effective prior to capture sparse topic correlations in natural

language documents.

2. We demonstrate that capturing both positive and negative correlation in our generalized Dirichlet variational autoencoder results into better performance.

3. We introduce a weighted objective function that enhances stable training without the need to approximate the divergence loss with some samples.

4. We perform extensive experiments and compare our model's results with state-of-the-art variational autoencoder topic models. Our model outperformed baseline models.

5. We demonstrate the superiority of GD-VAE on two image data sets. In particular, we annotate the reconstructed images and integrate a classifier to classify the reconstructed images. The evaluation metrics show that GD-VAE can play a significant role in data augmentation and classification.

2.1.1 Background

In this section, we first summarize the intuition behind generalized Dirichlet distributions. This is followed by presenting the covariance property of GD and formulating the GD marginal distribution.

2.1.1.1 Intuition of Generalized Dirichlet Distribution

[53] first used Generalized Dirichlet to model data on bone composition in rats and scute growth in turtles. This was motivated as a way to tackle the covariance restriction in Dirichlet distribution. For example, if a random vector is assumed to have a Dirichlet distribution, any two random variables from the vector will be negatively correlated. In reality, two random samples may be positively correlated, which renders Dirichlet distribution not to be a good prior choice. Moreover, all the samples in the vector must have the same variance and must sum up to one [19]. Therefore, if the probability of a variable increases, other probabilities must either decrease or remain unchanged to validate the constraint of summing up to one. Also, if a multinomial distribution uses the Dirichlet distribution as a prior, it will have only one degree of freedom to embed confidence in the prior knowledge. As a result, incorporating individual variance information for each random vector entry is problematic. Due to its property of conjugate prior, Dirichlet distribution is widely explored in mixture models [54] despite its limitations.

Unlike the Dirichlet distribution, GD exhibits flexibility that allows sampling of each entry in the random vector independently from Beta distribution [19] while it also holds conjugate prior property. [53] derived the density function of the GD distribution as follows:

$$p(\theta|\alpha,\beta) = \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i-1} \left(1-\sum_{j=1}^{i}\theta_j\right)^{\eta_i} \tag{2.1}$$

where $\sum_{i=1}^{k}\theta_i < 1$, $0 < \theta_i < 1$ for $i = 1, 2, ..., k$, $\alpha_i > 0$, $\beta_i > 0$, and $\eta_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$ for $i = n1, 2, ..., k-1$, $\eta_k = \beta_k - 1$. The derivation of the generalized Dirichlet distribution followed the concept of complete neutrality [53]. Also note that the GD is a Dirichlet tree distribution that uses a cascaded hierarchy in its generative process of the distribution [18]. This aids dimensionality reduction, which makes it a good property to facilitate reduction in the number of topic [19].

2.1.1.2 Covariance Properties of the Generalized Dirichlet

It's worth noting that the GD reduces to a Dirichlet distribution when $\beta_j = \alpha_{j+1} + \beta_{j+1}$ [55]. Therefore, Equation (2.1) becomes:

$$p(\theta|\alpha,\beta) = \prod_{i=1}^{k} \theta_i^{\alpha_k-1} \left(1-\sum_{j=1}^{k}\theta_j\right)^{\eta_i} \Lambda \tag{2.2}$$

with $\Lambda = \frac{\Gamma(\alpha_1+\alpha_2+...+\alpha_{k+1})}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_{k+1})}$, and $\alpha_{k+1} = \beta_k$. [55] computed the Dirichlet covariance for two random variables, θ_i and θ_j as:

$$Cov(\theta_i,\theta_j) = -\frac{\theta_i\theta_j}{\left(\sum_{i=1}^{k+1}\alpha_i\right)^2\left(\sum_{i=1}^{k+1}\alpha_i+1\right)} \tag{2.3}$$

Therefore, any two random variables from the vector, $\vec{\theta}$, will be negatively correlated, which is far from reality. [55] further showed the GD covariance for two random variables, θ_i and θ_j as:

$$Var(\theta_i) = E(\theta_i)\left(\frac{\alpha_i+1}{\alpha_i+\beta_i+1}\prod_{d=1}^{i-1}\frac{\beta_d+1}{\alpha_d+\beta_d+1} - E(\theta_i)\right) \tag{2.4}$$

Thus, the covariance of random variables θ_i and θ_j is given as:

$$Cov(\theta_i,\theta_j) = E(\theta_j)\left(\frac{\alpha_i}{\alpha_i+\beta_i+1}\prod_{d=1}^{i-1}\frac{\beta_d+1}{\alpha_d+\beta_d+1} - E(\theta_i)\right) \tag{2.5}$$

A look at Equation (2.5) and Equation (2.3) indicates that GD has a more general covariance structure than the Dirichlet distribution, which has a negatively constrained covariance. Note that the Generalized Dirichlet distribution is also a conjugate to the multinomial distribution [56].

2.1.1.3 Generalized Dirichlet

The graphical model of GD-based topic model is depicted in Figure 2.1. Each document of the collection is represented as a mixture of topics. Given the Beta distribution parameters α, β, of size V, where V is the vocabulary size. γ is defined as a vector of size K, the number of topic, and a set of N words, D is the number of documents, we define the joint distribution as follows:

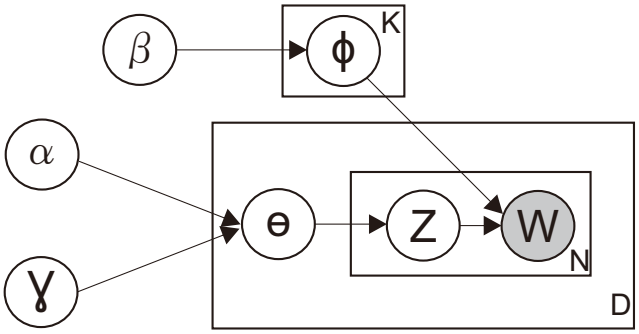


Figure 2.1: Graphical representation of the generalized Dirichlet model

$$p(\theta, Z, \boldsymbol{W}|\alpha, \beta, \gamma, \phi) = p(\theta|\alpha, \gamma)p(\phi|\beta) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{2.6}$$

where α is the per-document topic distributions, β is the per-topic word distribution, θ and ϕ are the topic distribution for document d and the word distribution for topic k respectively. z_n is the topic for the n-th word in document d, and W is the encoded vector denoting a specific word.

Integrating over θ and summing over Z, we obtain the marginal distribution of a given document as:

$$p(\boldsymbol{W}|\alpha, \beta, \gamma, \phi) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n, \phi)p(z_n|\theta)p(\phi|\beta) \right) p(\theta|\alpha)\, d\theta \tag{2.7}$$

To obtain the probability of a corpus, we take the product of the marginal probabilities of single document as:

$$p(\mathcal{D}|\alpha, \beta, \gamma, \phi) = \prod_{d=1}^{M} \int_{\theta} \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}=1}^{k} p(w_{d,n}|z_{d,n}, \phi)p(z_{d,n}|\theta_d)p(\phi|\beta) \right) p(\theta_d|\alpha)\, d\theta_d \tag{2.8}$$

It's worth noting that the number of statistical dependencies can be observed in Equation (2.8). For example, the observed word, $w_{d,n}$ depends on the topic assignment, and the topic assignment depends on the per-document topic proportions, θ_d. However, Equation (2.8) does not have a closed-form solution because the posterior over the latent variables θ and Z are intractable. To

solve this problem, we integrate a generalized Dirichlet distribution with a variational autoencoder to directly map a document to its approximated posterior distribution. Details of our approach are discussed in Section 2.3.2

2.2 Variational Autoencoder Inference (VAEI)

For the sake of clarity, the notations used in the remaining of this paper follow the notation convention used in autoencoders. Please note that these notations do not correspond to the notions used in analyzing the GD-LDA mixture model discussed in the previous sections. We define K as the number of topics, k refers to a specific topic, and V as the vocabulary size. In addition, $\mathcal{D}$ is the number of documents, j denotes a document, and w depicts a specific word. The generative process of VAEI is characterized by observations $\boldsymbol{X}$, latent variable Z, and the joint probability distribution $p_\theta(\boldsymbol{X}, Z) = p_\theta(Z)p_\theta(\boldsymbol{X}|Z)$, where θ represents the model parameters. The main focus of interest is the posterior distribution of the latent variables given the observations as stated below:

$$p_\theta(Z|\boldsymbol{X}) = \frac{p_\theta(\boldsymbol{X}|Z)}{\int p_\theta(\boldsymbol{X}, Z)dZ} \tag{2.9}$$

Computing the normalization term is intractable due to the integral's high dimensionality. That is, the number of integrals required corresponds to the number of latent variables in the distribution. Therefore, instead of computing, approximating the posterior distribution with a simpler distribution is an appropriate approach. A popular approach is to train the model with mean-field variational inference; however, this approach places some restrictions on the model and its flexibility. [1] handles this problem by using the VAE inference method to replace the mixture model in LDA. They found out that neural networks can effectively learn to approximate probabilistic inference. In variational inference, the interest is to maximize the evidence of the lower bound (ELBO), $\mathcal{L}$, which is always positive:

$$\mathcal{L}(\theta, \phi; X_i) = \mathbb{E}_{q(Z|X_i)}[-\log q(Z|X_i)] + \mathbb{E}_{q(Z|X_i)}[\log p(X_i, Z)] \tag{2.10}$$

where θ and ϕ are the distributions parameters of p and q respectively. The marginal log-likelihood of the model is defined as:

$$\log p(X) = \sum_{i=1}^{N} \log p(X_i) = \sum_{i=1}^{N} (D_{KL}[q(Z_i|X_i)||p(Z_i|X_i)] + \mathcal{L}(\theta, \phi; X_i)) \tag{2.11}$$

where N is the total number of words in the corpus. Over the space of q_ϕ, Equation (2.10) satisfies

the equation below:

$$\log p(X) \geq D_{KL}[q(Z_i|X_i)||p(Z_i)] - \mathbb{E}_{q(Z_i|X_i)}[\log p(X_i|Z_i)] \tag{2.12}$$

where the first term in Equation (2.12) denotes the divergence between $q_\phi(Z|X)$ and the prior $p_\theta(Z)$. It is otherwise termed a regularization factor because it prevents $q_\phi(Z|X)$ from simply performing an identity mapping. It thus forces it to learn more latent variable, which keeps the variational distribution close to the prior. Whereas, the second term is the log-likelihood of the observed X, given the latent variable Z. It tries to reconstruct X, and for this reason, it is called the negative reconstruction error. Another interesting interpretation that can be pointed out in this equation is that taking sample $Z \sim q_\phi(Z|X)$ can be interpreted as the latent describing X and could be called the encoder. While sampling $X \sim p_\theta(X|Z)$ describes reconstructing the latent pattern, it's termed the decoder network. Considering the latent space Z from which X is sampled, the ELBO is calculated by averaging over the contributions from all the L samples:

$$\bar{\mathcal{L}}(\theta, \phi; X_i) = \frac{1}{L}\sum_{l=1}^{L}(\log p(X_i, Z_{i,l}) - \log q(Z_{i,l}|X_i)) \tag{2.13}$$

2.2.1 Reparameterizing the Acceptance-Rejection Sampler

GD, which is a special case of Dirichlet trees does not have a shifting parameter because its prior is not a location-scale family, and this impedes reparameterization. We, therefore, explore the reparameterization trick through rejection sampling as described in [15]. We rewrite the simulation from the GD by deterministically mapping its parameters to a set of simpler random variables. This makes it possible for the GD to be efficiently approximated by a proposal function. Gamma distribution is one of the most popular rejection samplers, and it has proven in practice to be capable of generating Dirichlet distribution, Beta, and Student's t-distributed random variables [15].

Given that $\bar{Z}_\alpha \sim \text{Gamma}(\alpha, 1)$, and $\bar{Z}_\beta \sim \text{Gamma}(\beta, 1)$, the generalized Dirichlet (GD) latent variables can be sampled as: $\bar{Z}_{1:K} = \frac{\bar{Z}_\alpha}{\bar{Z}_\alpha + \bar{Z}_\beta} \sim \text{GD}(\alpha_{1:K}, \beta_{1:K})$. We now define the rejection samplers of the Gamma distributions as follows:

$$Z_\alpha = h_{\text{Gamma}}(\epsilon_1, \alpha) := (\alpha - \frac{1}{3})(1 + \frac{\epsilon_1}{\sqrt{9\alpha - 3}}) \tag{2.14}$$

$$Z_\beta = h_{\text{Gamma}}(\epsilon_2, \beta) := (\beta - \frac{1}{3})(1 + \frac{\epsilon_2}{\sqrt{9\beta - 3}}) \tag{2.15}$$

where ϵ_1 and $\epsilon_2 \sim N(0, 1)$. To generate samples from the actual distribution $q(Z; \theta)$, through rejection sampling, we continuously sample the proposal distribution, $r(Z; \theta)$, while samples $< L$:

$q(Z;\theta) \leq M_\theta r(Z;\theta)$. $M_\theta < \infty$ is a constant to regulate the rejection sampler. For a high acceptance rate, we followed the approach in [15] to choose our M_θ, and it was used in all our experiments. The samples that failed to satisfy the condition are rejected; this implies that the proposal distribution is not exactly the Gamma distribution. Thus, the distribution of samples that satisfy the condition is defined as $\epsilon \sim s(\epsilon)$. We define the distribution of the accepted samples ϵ as $\pi(\epsilon;\theta)$, which is obtained by marginalizing over the auxiliary uniform variable u as follows:

$$\begin{aligned}\pi(\epsilon;\theta) &= \int \pi(\epsilon, u;\theta)du \\ &= \int M_\theta s(\epsilon)\mathbb{1}\left[0 < u < \frac{q(h(\epsilon,\theta);\theta)}{M_\theta r(h(\epsilon,\theta);\theta)}\right]du \\ &= s(\epsilon)\frac{q(h(\epsilon,\theta);\theta)}{M_\theta r(h(\epsilon,\theta);\theta)}\end{aligned} \tag{2.16}$$

where $\mathbb{1}[X \in A]$ represents the indication function. Thus, Equation (2.10) can be re-written as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{\pi(\epsilon;\theta)}[\log p(X_i, h(\epsilon,\theta))] + \mathbb{E}_{q(Z|X_i)}[-\log q(Z|X_i)] \tag{2.17}$$

We follow the gradient computation in [15], and stated the gradient as follows:

$$\nabla_\theta \mathcal{L}(\theta) = g_{rep} + g_{cor} + \nabla_\theta \, \mathbb{E}_{q(Z|X_i)}[-\log q(Z|X_i)] \tag{2.18}$$

where g_{rep} is the reparameterization term, which accounts for the gradients of the model's latent variables, and g_{cor} is the correction term that accounts for not using the exact proposed distribution. They are given as follows:

$$g_{rep} = \nabla_Z \log p(X_i, Z)\nabla_\theta h(\epsilon,\theta), \qquad g_{cor} \quad = \log p(X_i, Z)\nabla_\theta \log \frac{q(h(\epsilon,\theta))}{r(h(\epsilon,\theta))}$$

2.2.2 Shape Augmentation

[15] explored shape augmentation and showed that the rejection sampler improves with increasing shape parameter, α, which results from increasing the shape of the Gamma distribution. Thus, higher values of α lead to a higher acceptance rate for the Gamma distribution. In their work, they showed that the trick could achieve a lower variance gradient. To compute the ELBO for Gamma$(\alpha, 1)$ and Gamma$(\beta, 1)$ distributions, we follow the approach in [15] and expressed the random variable, $Z_{param} = \bar{Z}_{param}\prod_{i=1}^{B} u_i^{\frac{1}{param+i-1}}$, $param \in (\alpha, \beta)$, and B is a positive integer, $\bar{Z}_{param} \sim$ Gamma$(param + B, 1)$, $u_i \sim U[0, 1]$, and it's i.i.d.

2.3 Generalized Dirichlet Variational Autoencoder Method

This section presents the approach to our proposed model. Firstly, we summarize the problems of generalization and overfitting. Secondly, we discuss the architecture of our model. We then summarize the KL divergence of the generalized Dirichlet distribution and finally describe the neural network architecture we used.

2.3.1 Overfitting and Generalization in Topic Modeling

First, it is worth mentioning the sparsity and smoothness problems identified in [57]. According to them, smoothing the model increases the model's coherence because it generalizes better. Therefore, [1] explored batch normalization and dropout to slow down the minimization of their ELBO at the beginning of the training. They identified that it significantly improved the model coherence, which leads to better generalization. However, this has a trade-off in terms of the model's sparseness. On the contrary, fast minimization improves the model's sparseness, which leads to lower perplexity with a trade-off in smoothness. With fast minimization, the estimated Dirichlet parameter will be equal to its prior, leading to overfitting. This simply means that the model cannot learn any meaningful topics because its latent variables are only sampled from noise. [14] introduced a sparse parameter, b in addition to normalization and dropout. According to them, the latent variables Z were sampled from $Dir(\alpha \cdot b)$. In their implementation, b, is a vector of zeros and ones, calculated from a linear transformation of the encoder output, which also generates the Dirichlet parameter. Our proposed GD-VAE automatically decoupled the sparsity and smoothness with its distribution parameters, α and β parameters.

2.3.2 GD-VAE

Figure 2.2 depicts the architecture of our model and the recently published models, which are closer to our work. In general, the encoder network transforms the input to yield the models' parameters at the encoder's output layer. The algorithm then uses the estimated parameters to sample the latent variables, Z. At the decoder, the model uses the Z to regenerate the original input. The encoder output of the Dirichlet VAE is α, while the parameters generated by the encoder of the Gaussian VAE are μ and σ; sparse Dirichlet VAE parameters are α and vector b, while the encoder of our generalized Dirichlet VAE generates distribution parameters β and α. The corresponding algorithms use these parameters to sample the latent variables, which are in turn used to reconstruct

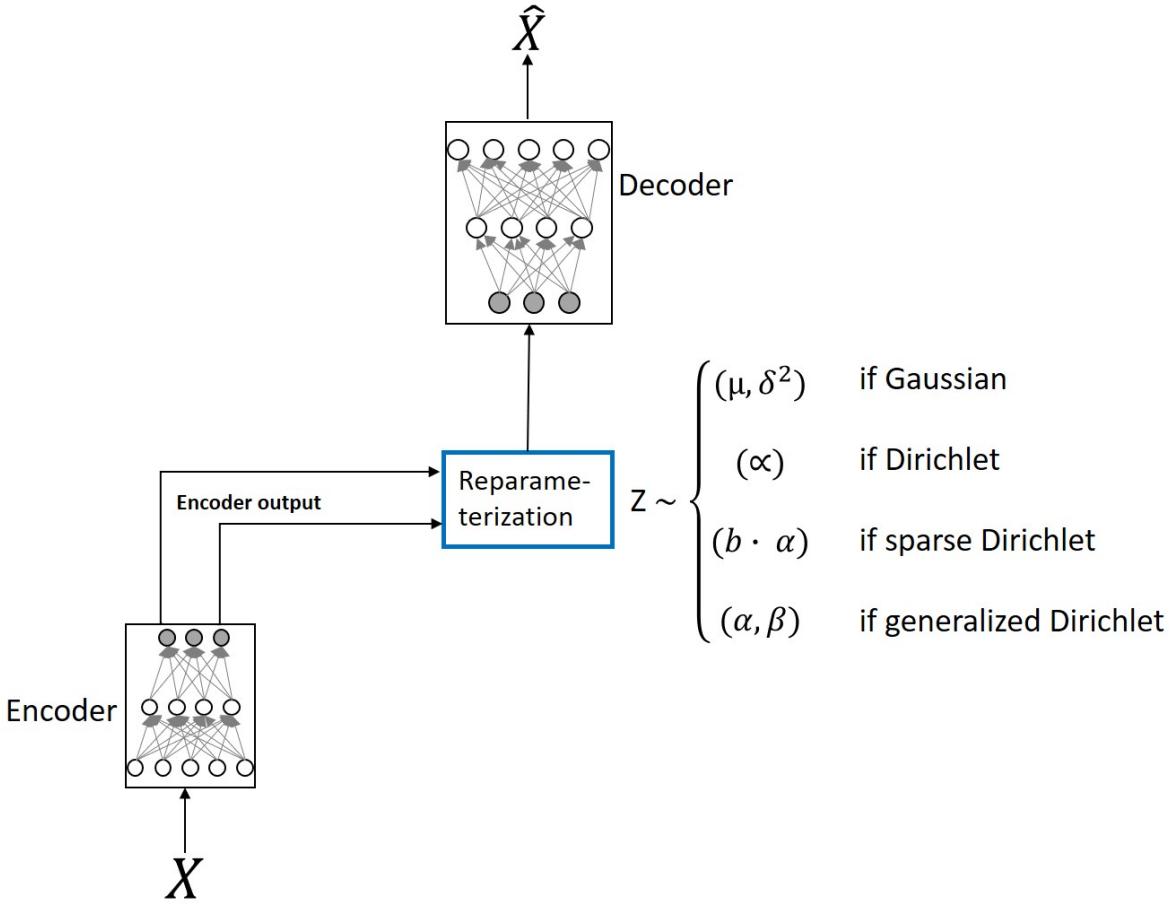


Figure 2.2: Schematic diagram of variational model

the input at the decoder output layer. Please note that the encoder and decoder networks are the same for all the models. Given an input variable, X, the latent variable, Z is conditioned on the input X, and the reconstructed output of the decoder is in turn conditioned on the latent variables, Z. We can therefore formulate the model equation as the joint distribution between X, Z, and θ, where θ depicts the model parameters. Thus, $p(X, Z, \theta) = p(X|Z, \theta) \cdot p(Z|\theta) \cdot p(\theta)$. Unlike the other models in Figure 2.2, the latent variables, Z of GD-VAE are sampled from GD(α, β), where α and β are the outputs of the encoder network. We further consider the distribution of the accepted samples because our algorithm uses rejection sampling variational inference, for efficient backpropagation. The second term of Equation (2.10) is re-written with respect to the accepted samples π as $LL_E = \mathbb{E}_{(\pi;\phi)}[\log p(X_i|h_\Gamma(\epsilon, \phi))]$.

Since the likelihood reconstruction error, LL_E does not depend on θ any longer, all θ dependents

Algorithm 1 Generalized Dirichlet Variational Autoencoder's Generative Process

Data:

$\mathcal{D}$: Dataset

$q_\phi(Z|X)$: Inference model

$P_\theta(X|Z)$: Generative model

Result:

θ, ϕ: Learnable model's parameters

$(\theta, \phi) \leftarrow$ Initialize parameters

while Optimizer not converged **do**

$\mathcal{M} \sim \mathcal{D}$ (Random batch of dataset)

$[\alpha, \beta] \leftarrow$ EncoderNN $(X; \phi)$

$\epsilon_{param} \sim$ Acceptance-rejection sampler $\triangleright$ $param \in (\alpha, \beta)$

$Z_\alpha \leftarrow (\alpha - \frac{1}{3})(1 + \frac{\epsilon_\alpha}{\sqrt{9\alpha - 3}})$

$Z_\beta \leftarrow (\beta - \frac{1}{3})(1 + \frac{\epsilon_\beta}{\sqrt{9\beta - 3}})$

$\bar{Z}_{1:K} \leftarrow \frac{\bar{Z}_1}{\bar{Z}_1 + \bar{Z}_2} \sim \mathrm{GD}(\alpha_{1:K}, \beta_{1:K})$

$X' \leftarrow$ DecoderNN $(Z; \theta)$

Compute $\bar{\mathcal{L}}(\theta, \phi; X_i)$

Update (θ, ϕ)

end while

can be eliminated from Equation (2.16), marginalizing over u, Equation (2.16) becomes:

$$\pi(\epsilon; \phi) = \int \pi(\epsilon, u; \phi) du = s(\epsilon) \frac{q(h_\Gamma(\epsilon, \phi))}{r(h_\Gamma(\epsilon, \phi))} \tag{2.19}$$

The gradient, $\nabla_\theta \mathcal{L}(\theta, \phi; X_i) = \nabla_\theta \, \mathbb{E}_{q(Z|X_i)}[\log p(X_i|Z)]$ of the ELBO for the generative encoder network does not depend on θ because it is parameterized by ϕ. Therefore, backward propagation can be computed without re-parameterization. On the contrary, the gradient of the variational decoder network is parameterized by θ as shown below:

$$\nabla_\theta \mathcal{L}(\theta, \phi; X_i) = \nabla_\phi(-D_{KL}[q(Z|X_i)||p(Z)]) + \nabla_\phi \, \mathbb{E}_{\pi(\epsilon;\theta)}[\log p(X_i|h_\Gamma(\epsilon, \phi))] \tag{2.20}$$

This contains two terms: the gradient of the KL-divergence and the gradient of the reconstruction error term for the generative decoder network. It's a general practice to use conditional distribution in variational autoencoder instead of joint distribution because the latent variables, Z

are conditioned on the input values X at the encoder network, and the reconstruction values, X' conditioned on the latent variables. Since this work is more of VAE, we now use the conditional distribution instead of the joint distribution common in topic modeling as in Equation (2.10) and Equation (2.13).

Following the derivation in [15], we then express the gradient of the ELBO as:

$$\begin{aligned}\nabla_\theta \mathcal{L}(\theta, \phi; X_i) &= \nabla_\phi(-D_{KL}[q(Z|X_i)||p(Z)]) + g^\phi_{rep} + g^\phi_{cor} \\ g^\phi_{rep} &= \nabla_Z \log p(X_i|Z)\nabla_\phi h(\epsilon, \phi) \\ g^\phi_{cor} &= \log p(X_i|Z)\nabla_\phi \log \frac{q(h(\epsilon,\phi))}{r(h(\epsilon,\phi))}\end{aligned} \tag{2.21}$$

Please refer to [15] for details on the derivation of Equation (2.21).

2.3.3 Generalized Dirichlet KL-Divergence

Equation (2.21) requires computation of KL-divergence. Here, we present the KL-divergence of two generalized Dirichlet distributions as expressed in [58]. The analytical KL-divergence between two generalized Dirichlet distributions, $GD_1(\alpha_1, \beta_1)$ and $GD_2(\alpha_2, \beta_2)$ is expressed as:

$$\begin{aligned}D_{KL}(GD_1||GD_2) = \sum_{k=1}^{K} \ln\left(\frac{\Gamma(\alpha_{1,k}+\beta_{1,k})\Gamma(\alpha_{2,k})\Gamma(\beta_{2,k})}{\Gamma(\alpha_{1,k})\Gamma(\beta_{1,k})\Gamma(\alpha_{2,k}+\beta_{2,k})}\right) - \sum_{k=1}^{K}(\alpha_{1,k}-\alpha_{2,k})\Bigg(\Psi(\alpha_{1,k}) - \Psi(\beta_{1,k}) \\ - \sum_{j=1}^{k}(\Psi(\alpha_{1,j}+\beta_{1,j}) - \Psi(\beta_{1,j}))\Bigg) + \sum_{k=1}^{K}(\eta_{1,k}-\eta_{2,k})\sum_{j=1}^{k}(\Psi(\alpha_{1,j}+\beta_{1,j}) - \Psi(\beta_{1,j}))\end{aligned} \tag{2.22}$$

where $\eta_k = \beta_k - 1$, and Ψ is the digamma function. For details of the derivation of D_{KL}, we refer readers to [58].

[14] reported that their analytical KL-divergence wasn't stable during training. They resorted to using sampling approximation for the training, and analytical KL-divergence in the testing phase. In our experiments, we also observed that the KL-divergence is not stable. However, we submit that using a sampling approximation in the training will make our results biased since other baseline models never considered this approach, except [14]. We further observe that weighting the ELBO components with ρ will simply make the training stable, and we rewrite our ELBO as the weighted sum of the likelihood loss and the D_{KL}:

$$\mathcal{L}(\theta, \phi; X_i) = \rho \times \mathbb{E}_{q(Z|X_i)}[\log p(X_i, Z)] - \mathbb{E}_{q(Z|X_i)}[\log q(Z|X_i) \times (1-\rho)] \tag{2.23}$$

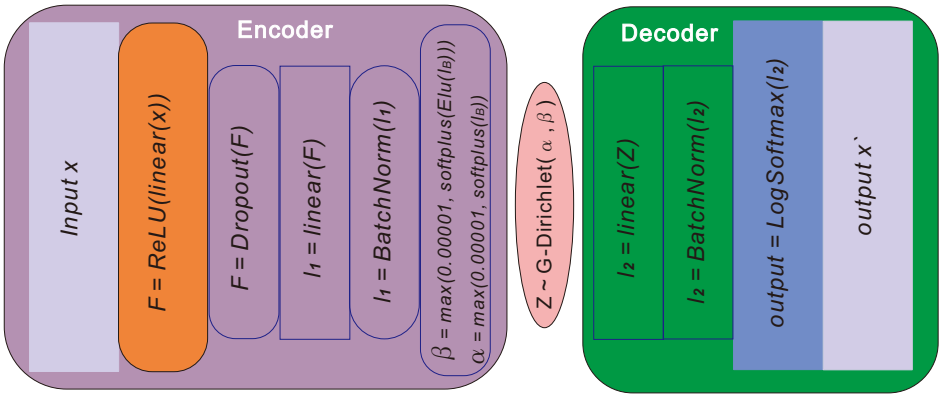


Figure 2.3: Adopted Architecture from [1]. This architecture is also used in our baseline models: sparse DVAE, DVAE, prodLDA, implicit reparameterization approach, Weibull VAE method and the inverse CDF gradient method.

where ρ is the weighting factor defined as: $0 < \rho < 1$. Note that this means that the KL-divergence is not sampled. Therefore, the gradient of the D_{KL} without further adding another second correction term can be defined as:

$$\nabla_\phi D_{KL}[(q(h_\Gamma(\epsilon,\phi))|X)||p(h_{\Gamma(\epsilon,\phi)})] = \nabla_\phi \log \frac{p(h_\Gamma(\epsilon,\phi))}{q(h_\Gamma(\epsilon,\phi)|X)} \tag{2.24}$$

Please note that we do not add a second correction term because we use our analytical D_{KL} in both the training and testing stages.

2.3.4 Neural Network Architecture

In this section, we summarize the architecture of the encoder and decoder networks. For a fair comparison, we adopt a similar architecture used in [1] and [14]. However, the encoder layer of our architecture generates the two parameters of our generalized Dirichlet algorithm, from which the latent variables are sampled.

- Encoder: The encoder network first applies the ReLU activation function to the linearly transformed input to add some non-linearity properties to the input. This is followed by applying dropout to the output of the ReLU to avoid overfitting. The resulting output of the dropout layer is linearly transformed, and batch normalization is applied, which makes the training faster and more stable. It also re-centers and re-scales the input to avoid internal covariate

shifts. Finally, the encoder generates the GD parameters α and β. We use softplus on the generated parameters' values to avoid negative values as per the requirement of GD, and we further map the zero values to a pre-defined minimum parameter. Thus, the latent variables are sampled from the generalized Dirichlet distribution, given α and β.

$$\alpha = \text{softplus(encoder output)}, \qquad \beta = \text{softplus(Elu(encoder output))}$$

- Decoder: The decoder network learns to reconstruct the original input values. In the first layer of the decoder, the sampled latent variables are linearly transformed, followed by batch normalization to adjust the covariant shift. Finally, the original input is estimated by applying the log-softmax activation function to the batch normalization's output. Similar to the practice in [1] and [14], we completely relax the restriction posed by the mixture model on the topic-word distributions, which forces the distributions to be a probability simplex. In other words, our decoder weights are not forced to be a probability simplex. This makes our model flexible and more generalized, with lower perplexity. The decoder weights are used at the inference stage to rank the word distribution, therefore inferring the topics of the document. The words that represent a particular topic usually have the largest weight in each latent topic.

2.4 Experimental Results

We evaluate GD-VAE on held-out test sets against baseline models (DVAE, Sparse DVAE, Implicit, Inverse CDF, Weibull autoencoder with stochastic-gradient MCMC, ProdLDA with Laplace approximation, Gaussian NVDM, and online LDA SCVB) using perplexity, coherence (PMI) [59], diversity, and topic uniqueness (TU) [60]. For fairness' sake, we fixed our model α to the values used in the previous literature, which are 0.02 and 0.1, and reported the results in Table 2.3 and **??**, respectively. However, we carefully choose our β to avoid negative D_{KL} values. In [19], α_k and β_{k-1} were initialized with $2/K$, but using this β value mostly resorted to negative D_{KL} in our experiments. We perform several experiments and observe different values of c, for $\beta = c/K$. We discovered that when $c = 5$, the simulation always generates a positive D_{KL} with a different number of topics, K.

We use the 20news data set to perform preliminary experiments to choose the value of β. In Table 2.1, firstly, we fixed the value of $K = 50$, $\alpha = 0.02$, and analyzed the values of D_{KL} for the different random values of $c = \{1, 5, 2, 20, 2.5, 25, 4, 40\}$. D_{KL} is positive for $c = 5 \implies \beta = 0.1$, and

Table 2.1: Analyzing the values of D_{KL} for the different values of $c : \beta = c/K$. α is fixed to 0.02, K takes 50 and 100 values as shown in the two parts of the table respectively. D_{KL} values are always positive for $c = 5$.

$K = 50$, $\alpha = 0.02$, $\beta = c/K$			
β	Perplexity	Analytical D_{KL}	Topic Coherence
0.02	321	-58.01	0.107
0.1	**406**	**71.28**	**0.113**
0.04	271	-25.15	0.106
0.4	265	-33.07	0.114
0.05	307	8.32	0.104
0.5	240	-63.95	0.114
0.08	280	-13.05	0.113
0.8	241	-57.32	0.113
$K = 100$, $\alpha = 0.02$, $\beta = c/K$			
β	Perplexity	Analytical D_{KL}	Topic Coherence
0.01	491	248.34	0.111
0.1	330	25.42	0.107
0.04	390	101.10	0.107
0.4	234	-81.91	0.112
0.05	**293**	**11.81**	**0.113**
0.5	215	223.22	0.110
0.08	296	2.44	0.110
0.8	241	-73.44	0.112

$c = 2.5 \implies \beta = 0.05$. Secondly, we further investigate that D_{KL} is positive for $c = 5$, while the value of K is changed to 100, α remains 0.02 as shown in the second part of Table 2.1. Moreover, we propose $\beta = c/K$, where $c = 5$, and further investigate the values of D_{KL} for different values of K. As shown in Table 2.2, D_{KL} values are all positive for different values of $K = \{50, 100, 150, 200, 250,$ and $300\}$, while other parameters remain unchanged.

We further evaluate and compare the performance of our proposed model with the baseline models on perplexity (it captures the level of astonishment a model experiences when encountering unfamiliar data it hasn't been exposed to previously), coherence, which measures the point-wise

Table 2.2: Analyzing the values of D_{KL} for the different values of K, where $c : \beta = c/K : c = 5$. α is fixed to 0.02. D_{KL} values are always positive for $c = 5$.

$c = 5$, $\beta = c/K$, $\alpha = 0.02$				
K	β	Perplexity	Analytical D_{KL}	Topic Coherence
50	0.100	406	**71.28**	0.113
100	0.050	293	**11.81**	0.113
150	0.033	411	**10.10**	0.109
200	0.025	509	**379.83**	0.103
250	0.020	733	**188.39**	0.103
300	0.016	750	**855.13**	0.112

mutual information (PMI) [59], diversity (it attempts to account for word overlap in topics) and topic uniqueness (TU), which is inversely proportional to the number of times each word is repeated in the set [60]. For a fair comparison, we investigate the performances of the model when $\alpha = 0.02$ and 0.1 priors. We first experiment with prior values of $\alpha = 0.02$, $\beta = c/k : c = 5$ and for values of $K = 50$ and 200 number of topics. The first and second parts in Table 2.3 depict the perplexity for $K = 50$ and $K = 200$ topics respectively. GD-VAE achieved the lowest perplexity in all the experiments over all three data sets. This is followed by NVDM, which has better perplexity compared to other models. Moreover, Weibull and DVAE have the worst perplexity. GD-VAE automatically decoupled the sparsity and smoothness with its two distribution parameters, α and β. It also captures both positive and negative correlations between the topics, thereby achieving very low perplexity. Please note that the perplexity slightly increased when the number of topics increased from 50 to 200 topics. However, DVAE has better topic coherence compared to all other models for both $\alpha = 0.02$ and 0.1 priors. Except for the case of the KOS data set, when $K = 50$, prodLDA scores the highest coherence value. Although GD-VAE has the lowest coherence, it is still competitive with NVDM.

Table 2.4 presents the diversity and uniqueness, with $\alpha = 0.02$, for $K = 50$ and 200. The first section depicts the diversity when $K = 50$ and 200. GD-VAE has the highest diversity in almost all three data sets, except for NIPS when $K = 50$, GD-LDA has the highest diversity. Note that the sparse DVAE exhibits the lowest performance on all the data sets. In the same manner, the GD-LDA upsets the uniqueness of GD-VAE on the NIPS data set when $K = 50$, while GD-VAE outperformed on the other data sets, as shown in the second section of Table 2.4. The sparse-

Table 2.3: Perplexity and topic coherence performance comparisons of GD-VAE with other baseline models, using three different data sets, GD priors: $\alpha = \mathbf{0.02}$, $\beta = c/K : c = 5$, $K = 50, 200$ topics.

Perplexity for $K = 50$ Topics											
Data set	GD-VAE	GD-LDA	DVAE	Sparse DVAE	Implicit	Inverse CDF	Weibull	ProdLDA	NVDM	SCVB	SawETM
KOS	**424**	954	7872	2408	2894	2598	7795	2811	2097	2365	2984
NIPS	**508**	1406	2779	2197	2327	2141	3.55E+52	2303	1933	3973	2107
20news	**406**	1102	5009	1066	1842	1533	5209	1128	907	2034	1403
Perplexity for $K = 200$ Topics											
KOS	**603**	1172	2.59E+5	2397	3238	3233	3.11E+4	3161	2216	2222	3229
NIPS	**621**	1639	3110	2152	2250	2135	8.50E+167	2532	1957	5787	2305
20news	**509**	861	4.93E+5	1075	3021	3033	1.82E+8	1271	875	951	1300
Topic Coherence for $K = 50$ Topics											
KOS	0.032	0.021	0.194	0.050	0.131	0.130	0.067	**0.218**	0.070	0.145	0.193
NIPS	0.033	0.026	**0.313**	0.295	0.276	0.224	0.167	0.293	0.070	0.116	0.214
20news	0.113	0.111	**0.324**	0.248	0.322	0.304	0.271	0.223	0.140	0.244	0.201
Topic Coherence for $K = 200$ Topics											
KOS	0.033	0.021	**0.212**	0.072	0.118	0.076	0.086	0.071	0.060	0.107	0.041
NIPS	0.022	0.022	**0.280**	0.234	0.227	0.208	0.176	0.277	0.060	0.105	0.192
20news	0.103	0.100	**0.307**	0.174	0.298	0.281	0.258	0.150	0.140	0.204	0.204

DVAE has the lowest diversity and uniqueness on both values of K. For $\alpha = 0.1$, the diversity and uniqueness are shown in Table 2.4. Again, GD-VAE outperformed all the baseline models for different values of K except for one occasion where GD-LDA slightly outperformed GD-VAE in uniqueness when $K = 50$.

2.4.0.1 Discriminative Qualities and Classification Task

Furthermore, the performance evaluations of our method on generated handwritten digits and the generated MNIST fashion in terms of accuracy, precision, recall, and f1-score are shown in Table 2.5 and Table 2.6 respectively. For the MNIST handwritten digits, GD-VAE achieved 0.88 and 0.80 accuracy and weighted accuracy respectively. It outperformed GD-LDA, which achieved accuracy and weighted accuracy of 0.82 and 0.77 respectively. DVAE, with accuracy and weighted accuracy of 0.81 and 0.73 respectively achieved better performance than ProdLDA and SawETM, while sparse DVAE has the worst performance. Also, the weighted-(precision, recall, and f1-score) of GD-VAE are 0.81, 0.80, and 0.80 respectively, which outperformed other models. GD-LDA, DVAE, and ProdLDA have very competitive performances, followed by the SawETM. Note that

Table 2.4: Diversity and topic uniqueness performance comparisons of GD-VAE with other baseline models, using three different data sets, GD priors: $\alpha = \mathbf{0.02}$, $\beta = c/K : c = 5$, $K = 50, 200$ topics.

Diversity for $K = 50$ Topics											
Data set	GD-VAE	GD-LDA	DVAE	Sparse DVAE	Implicit	Inverse CDF	Weibull	ProdLDA	NVDM	SCVB	SawETM
KOS	**0.90**	0.73	0.63	0.12	0.62	0.49	0.34	0.61	0.60	0.62	0.60
NIPS	0.85	**0.88**	0.57	0.15	0.50	0.34	0.15	0.73	0.64	0.59	0.52
20news	**0.88**	0.60	0.52	0.11	0.48	0.45	0.29	0.53	0.47	0.43	0.49
Diversity for $K = 200$ Topics											
KOS	**0.72**	0.51	0.48	0.18	0.40	0.33	0.34	0.42	0.40	0.39	0.41
NIPS	**0.75**	0.69	0.36	0.11	0.31	0.21	0.12	0.32	0.56	0.32	0.30
20news	**0.73**	0.54	0.30	0.11	0.21	0.20	0.28	0.26	0.41	0.38	0.27
Topic Uniqueness for $K = 50$ Topics											
KOS	**0.95**	0.82	0.60	0.12	0.56	0.42	0.38	0.55	0.45	0.45	0.53
NIPS	0.86	**0.87**	0.48	0.10	0.43	0.34	0.13	0.47	0.37	0.29	0.47
20news	**0.88**	0.71	0.62	0.10	0.40	0.31	0.28	0.58	0.36	0.30	0.49
Topic Uniqueness for $K = 200$ Topics											
KOS	**0.88**	0.59	0.54	0.17	0.48	0.37	0.34	0.50	0.50	0.47	0.48
NIPS	**0.90**	0.57	0.44	0.13	0.35	0.29	0.26	0.39	0.41	0.39	0.40
20news	**0.72**	0.34	0.22	0.11	0.29	0.21	0.22	0.40	0.44	0.44	0.32

Table 2.5: Performance evaluation of models in terms of accuracy, precision, recall, and f1-score on 10,000 **MNIST-handwritten** generated test set. "W" means weighted, and Acc depicts the accuracy. Epoch and the learning rate are set to 20 and 0.001 respectively.

Models	W-Precision	W-Recall	W-F1 score	Acc	W-Acc	Inference time (μS)
GD-VAE	0.81	0.80	0.80	0.88	0.80	138.67
GD-LDA	0.77	0.68	0.72	0.82	0.77	132.81
DVAE	0.75	0.73	0.73	0.81	0.73	118.50
Sparse DVAE	0.10	0.10	0.20	0.13	0.10	113.89
ProdLDA	0.76	0.72	0.72	0.73	0.72	105.32
SawETM	0.66	0.70	0.55	0.62	0.64	111.03

the sparse DVAE had the least performance. In Table 2.6, GD-VAE further showed its superiority by achieving 0.81 and 0.79 of accuracy and weighted accuracy respectively. GD-LDA and DVAE performed very closely, followed by ProdLDA and SawETM, which also performed competitively.

Table 2.6: Performance evaluation of models in terms of accuracy, precision, recall, and f1-score on **MNIST-fashion** 10,000 generated test set. "W" means weighted, and Acc depicts the accuracy. Epoch and the learning rate are set to 20 and 0.001 respectively.

Models	W-Precision	W-Recall	W-F1 score	Acc	W-Acc	Inference time (μS)
GD-VAE	**0.75**	**0.77**	**0.74**	**0.81**	**0.79**	149.77
GD-LDA	0.69	0.72	0.71	0.76	0.70	129.81
DVAE	0.72	0.70	0.70	0.75	0.70	120.52
Sparse DVAE	0.06	0.10	0.02	0.06	0.10	117.78
ProdLDA	0.68	0.65	0.64	0.73	0.65	**107.86**
SawETM	0.65	0.69	0.65	0.68	0.69	115.21

Table 2.7: Test error of a kNN classifier trained on the latent representations produced by each model.

	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
GD-VAE	**10.21**	**9.35**	**8.94**	**8.77**	**8.75**	**8.76**	**8.83**	**8.84**
GD-LDA	27.04	24.09	22.99	22.59	32.14	31.80	27.19	28.06
DVAE	35.44	33.12	35.16	33.13	34.11	29.80	34.90	31.33
Sparse DVAE	85.443	84.34	84.01	84.01	83.83	84.01	83.34	83.40
ProdLDA	50.25	48.61	47.66	46.66	46.04	36.04	45.73	45.09
SawETM	44.67	45.12	45.02	47.76	47.08	45.88	50.34	48.77
RawPixels	3.85	3.33	3.45	3.39	3.50	3.45	3.55	3.53

Again, sparse DVAE achieved poor scores.

In terms of cost, GD-VAE recorded the highest inference time on the two variants of MNIST data sets. For the MNIST handwritten, GD-VAE recorded $138.67\mu S$, $5.86\mu S$ more than the GD-LDA inference time. In a similar vein, GD-VAE recorded an inference time of $149.77\mu S$ on the MNIST fashion data set, which is more than other models' inference times. This is understandable as it takes additional time for the rejection sampler to completely sample the required number of samples that fall within the target distribution while rejecting the samples outside the target distribution. Note that ProdLDA recorded the least inference times on both data sets. This can also be attributed to the fact that ProdLDA uses a simple Laplace approximation, which requires less computation time.

The assessment of the models' discriminative properties involves utilizing a k-Nearest Neighbors

(kNN) classifier on sampled MNIST-handwritten latent variables. The test errors of the models are presented in Table 2.7. The GD-VAE consistently demonstrates superior performance compared to the baseline models across all the k values. This indicates that the GD-VAE's latent space captures the class structure more effectively. We also report the results for clustering the raw MNIST pixels.

2.5 Conclusion

In our research, we introduce a novel approach called the generalized Dirichlet Variational Autoencoder (GD-VAE). Our main hypothesis is that capturing both positive and negative correlations between topics is crucial for achieving low perplexity. To address this, we propose the GD distribution, which has a more flexible covariance structure than the traditional Dirichlet distribution. This enhanced covariance structure allows the GD distribution to effectively capture both positive and negative correlations, unlike its Dirichlet counterpart. By combining the advantages of the GD distribution's covariance structure with the capabilities of neural networks, our proposed GD-VAE significantly reduces perplexity in topic modeling when compared to baseline models. To validate our hypothesis, we conducted experiments on three well-known datasets commonly used in topic modeling. The experimental results strongly support our hypothesis, as our proposed model consistently achieves substantially lower perplexity across all test cases. These findings indicate that the GD distribution has the potential to enhance learning in neural networks, particularly in scenarios where input features exhibit correlation. Additionally, we demonstrate that weighting the objective losses can stabilize the training process and lead to convergence. We validate this by assigning weights to the D_{KL} loss and the log-likelihood loss while efficiently utilizing the analytical Kullback-Leibler divergence in both the training and inference stages. Furthermore, we explore the applicability of our model, as well as the baseline models, in the field of image classification. Our GD-VAE outperforms other models when evaluated on MNIST-handwritten and MNIST-fashion data sets. With proper post-processing techniques, we believe this approach has the potential to streamline classification processes for various computer vision tasks, including object detection, classification, and image/scene segmentation. Indeed, a notable limitation of the framework, as well as the baseline models, is the fixed latent dimensionality. This rigidity in the model's architecture may restrict its adaptability to varying data requirements. To address this issue, our next step is to develop a framework that incorporates an adaptive latent factor. This adaptive latent factor would enable automatic model selection by leveraging an infinite-capacity hidden layer. By introducing

this flexibility, we aim to enhance the model's ability to adapt to different data characteristics and improve its overall performance. In addition, we are looking forward to integrating our framework with transformer architecture and comparing the results with works that use similar architecture.

Chapter 3

SmoothDetector: A Smoothed Dirichlet Multimodal Approach for Combating Fake News on Social Media

The rapid dissemination of fake news in the digital era has become a pressing concern. The ease of generating and manipulating fake content, including images, text, audio, and videos, has significantly fueled the spread of misinformation on social media platforms. These platforms often lack rigorous editorial scrutiny, exacerbating this problem. Although recent studies have explored multimodal fake news detection to learn shared representations of textual and visual information, they often learn discrete latent representations, merely concatenations of multimodal features. Simple concatenation or summation operations hinder the dynamic interaction of multimodal features. Furthermore, most models rely on additional subtasks, such as reconstruction and event discrimination. The performance of these models depends heavily on subtasks, which can be mathematically complex and time-consuming. This reliance limits the ability of researchers to explore different modeling assumptions freely. This study introduces a novel approach that integrates a probabilistic algorithm with a deep neural network to effectively capture the uncertainties and diversities in the shared latent representation of multimodal data. Specifically, our model utilizes continuous latent representations by leveraging a smoothed Dirichlet distribution, facilitating the identification of shared hidden patterns across textual and visual modalities. In addition, our model demonstrates the powerful properties of generative models when integrated with neural network models. Our results underscore the potential of integrating a probabilistic algorithm with a deep neural network

to address the challenges of fake news detection in a multimodal setting.

3.1 Introduction

"Misinformation is more dangerous than an epidemic: it spreads globally at lightning speed and can be lethal when it reinforces personal biases against all credible evidence," stated Marcia McNutt, President of the National Academy of Sciences [1]. The explosive proliferation of fake news has posed serious challenges across society, impacting areas such as politics [61], the economy [62], and public health [63]. Owing to widespread Internet availability and advancements in web technology, individuals can instantly post news on online platforms such as Facebook, Twitter, and Instagram [64]. Early efforts to detect fake news focused on unimodal features, utilizing either text or visual content [65]. However, these approaches struggle to identify multimodal cues and lack the comprehensive information needed to fully understand the news context. This results in decreased robustness and increased misclassification rates.

The authors in [66] explored a weighted LSTM to enrich single-mode features to capture important aspects of news articles. The LSTM output was then input into a classifier composed of stacked fully connected (FC) layers. They enhance the performance through hyperparameter tuning. Although their model tackles issues such as vanishing gradient problems, it faces limitations, particularly in effectively managing multimodal datasets. To overcome these challenges, the authors in [67, 68] combined features extracted from unimodal sources to create multimodal fusion representations. Although this approach outperformed using a single unimodal feature for fake news detection, its accuracy remains insufficient for practical use because of the complex nature of fake news.

Adding to these complexities, as the media landscape evolves, news content diversifies, and fake news evolves from simple text to a multimodal format that incorporates text, image, audio, and video. Unlike traditional text-based news, multimodal news presents richer and more intuitive information, enhancing its ability to grab public attention and propagate quickly within a short time–frame [69]. Thus, it is inadequate to detect fake news effectively depending on the features of the single-mode data. Consequently, current research methodologies are shifting from single-modal to multimodal fake news detection [70], which considers feature extraction from text and images to facilitate better detection accuracy. The authors of [69] explored deep learning for multimodal fake

[1] https://www.nationalacademies.org/news/2021/07

news detection. They proposed an architecture that simultaneously trains three modules: a variational autoencoder (VAE), a domain adversarial module, and a false detection module. To derive a resilient representation combining semantic, linguistic, and topical aspects, [71] proposed a triple-VAE incorporating three modalities. Specifically, they merged three textual features of identical sizes to train their encoder networks. Concurrently, the decoder network in their proposed method is trained to reconstruct these three textual features. The multi-step retrieval enhancement model proposed in [72] includes three models: a text–summarization module, a retrieval module, and a news classifier module. The effectiveness of these models is largely dependent on subtasks, such as reconstruction or event discrimination, which are often mathematically intricate and time-consuming. This dependency restricts researchers' flexibility in exploring various modeling assumptions.

To overcome these challenges, [68] presented a multimodal model for detecting fake news without relying on additional subtasks. The authors utilized bidirectional encoder representations from transformers to integrate contextual information with the image features extracted from VGG-19. A contrastive learning model aimed at reducing the inconsistency in multimodal relations for fake news detection was proposed in [73]. This study employs a causal-relation reasoning module to address local inconsistency by eliminating the direct effects of textual and visual entities. In contrast, global inconsistency is mitigated through the semantic deviation of contrastive text-image objectives. Furthermore, [74] proposed multimodal fake news detection for attention and pooling blocks to integrate knowledge from the temporal and spatial effects. The studies in [75] addressed the problem associated with missing domain labels by exploring soft-label for multi-domain fake news detection. Despite significant progress in multimodal fake news detection technology, capturing the diversity, complexities, and uncertainties of multimodal news sources remains a challenge. Previous research often focuses on discrete latent representations and/or depends on mathematically complex and time-consuming subtasks, which restricts the exploration of the hypothesis space.

The studies in [76, 77] investigated probabilistic latent representations by combining image and text features, which were then linearly transformed using a Gaussian distribution. However, issues such as component collapse and limited expressiveness in this variational inference [78] have been identified [1] and further examined [14, 79]. Although the Dirichlet prior has been explored to address these challenges, it often involves a trade-off between sparsity and smoothness [14]. Smoothness aids generalization, whereas sparsity ensures low reconstruction error. In addition, this approach relies on subtasks in which decoders separately reconstruct the image and text, leading to significant computational costs. Motivated by these challenges, we introduce SmoothDetector, a smoothed

Dirichlet multimodal approach for combating fake news on social media. SmoothDetector enables efficient inference and learning in directed probabilistic models with continuous latent representations, thereby eliminating the need for additional subtasks. SmoothDetector utilizes a smoothed Dirichlet distribution to model multimodal data, capturing shared patterns and uncertainties more effectively than discrete latent representations, which often rely on learning concatenated features that can obscure essential patterns. Also, other latent representations, such as multinomial-based distributions, suffer from sparseness problems, which usually appear in Twitter fake news. By sampling from a continuous probabilistic representation, SmoothDetector addresses these limitations, enhancing the ability to detect subtle relationships in fake news and avoiding the sparseness data problem by smoothing the features. More precisely, by modifying the parameter vector of the Dirichlet distribution with a smoothing vector, we decouple smoothness and sparsity, thereby enhancing generalization performance and addressing the issues of limited expressiveness.

To the best of our knowledge, this is the first smoothed Dirichlet multimodal approach for fake news detection. The primary contributions of this chapter are as follows:

1. We propose SmoothDetector: A Smoothed Dirichlet Multimodal Approach for Combating Fake News on Social Media, which employs a smoothed Dirichlet distribution as a prior.
2. We demonstrate that learning continuous latent representations can yield excellent results without the need for additional subtasks.
3. We derive a closed-form solution for the Kullback-Leibler divergence between two smoothed Dirichlet distributions and demonstrate its efficient training via backpropagation.

3.1.1 Background

This section presents the concept of a multimodal variational autoencoder (MVAE) and the fundamentals of the smoothed Dirichlet distribution. First, we summarize the basic components of the MVAE and how they differ from our proposed architecture. We then extend our discussion to smoothed Dirichlet distributions.

3.1.1.1 Multimodal Variational Autoencoder (MVAE)

MVAE has been explored for fake news detection tasks. It comprises three major components: an encoder, a decoder, and a classifier. The encoder itself contains two feature extractors: one

for transforming the text and the other for transforming the image. The latent representation is derived from the concatenation of these transformed features. The decoder then reconstructed the images and texts from the latent representation. Notably, additional tasks can also be imposed on latent representations. These latent representations are fed into the classifier that predicts whether the news is real or fake. Simple concatenation or summation operations limit the dynamic interactions of multimodal features. In contrast, SmoothDetector has only two components: an encoder and a classifier. The encoder transforms the inputs to generate smoother Dirichlet distribution parameters and classifier samples from the continuous space of the hidden state of the distribution'. SmoothDetector offers significant advantages: it learns from continuous space to capture dynamic feature interaction and diversity, which improves its expressiveness, and it requires less computational cost because it does not impose additional tasks on the latent space.

3.1.1.2 Smoothed Dirichlet Distribution (SD)

The smoothed Dirichlet distribution is a modified version of the standard Dirichlet distribution, often employed in Bayesian statistics as a prior distribution for categorical data. This is especially beneficial in situations where certain categories may have zero or very low counts, as it smooths these counts to prevent problems associated with zero probabilities [20]. Smoothed Dirichlet distributions have demonstrated robustness across various applications, including emotion recognition based on depression on social media, happiness analysis, pain estimation [80], and a count data model for emotion state recognition [81]. Smoothing techniques have been shown to enhance the model's stability [82], generalization, and performance [83]. The work in [20] first applies smoothing to the raw data, which is the proportion of the word count in the bag of words to avoid zero probability, and defines the raw proportion, F^u, as [20]:

$$F^u = (X^s - (1-\lambda)X^{GE})/\lambda \tag{3.1}$$

where X^s and X^{GE} are the smoothed proportion and word proportion in general English, respectively, and λ is the smoothing parameter such that $0 < \lambda < 1$. Furthermore, [20] defines the probability of generating the smoothed proportion X^s from the SD distribution as:

$$P(X^s|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{i=1}^{K}({x^s}_i)^{\alpha_i - 1} \tag{3.2}$$

where K, x_i^s, and α_i represent the vocabulary size, smoothed feature proportion, and parameters of the smoothed Dirichlet distribution, respectively. $B(\vec{\alpha})$ denotes the normalizing constant that

ensures that the probability simplex sums to one [20]:

$$\frac{1}{B(\vec{\alpha})} = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}$$

Our approach adopts a similar approach; however, instead of smoothing the raw features, we smooth the Dirichlet parameter and the learned features. We observed that smoothing the raw features led to covariant shifts in the image features and resulted in unstable training. Additionally, because our smoothing and Dirichlet parameters are learnable, it is more logical to apply smoothing to the learnable features within the model. See Algorithm 2 for further detail.

3.2 Proposed Model

We propose a smoothed Dirichlet-based fake news detector called SmoothDetector. The SmoothDetector comprises five components that are trained jointly from end to end. The first component integrates a natural language model as a textual feature encoder to identify the relevant patterns in the news articles. The second encoder component utilizes a convolutional neural network to extract useful features from the images that accompany news articles. The third component is a multimodal fusion layer that coalesces textual and visual features to create a comprehensive news feature vector. The fourth component forms a probabilistic sub-module that captures the uncertainties and randomness in integrated textual and visual features. The last component is a classifier that predicts whether news is fake or real. Figure 3.1 presents a detailed overview of the proposed model.

3.2.1 Textual Feature Encoder

As shown in Figure 3.1, the textual feature encoder comprises a Bidirectional Encoder Representations from Transformers (BERT), two fully connected layers, and two layers of dropout and ReLU activation functions. To produce high-quality word and sentence representations that capture semantic and contextual patterns, we utilized BERT–base–uncased. Our decision to use BERT-base-uncased, as opposed to larger models like BERT-large or RoBERTa, is driven by the need to strike a balance between computational efficiency and model performance. While larger transformer-based models often achieve marginally better results, they come with significantly higher computational costs and greater hardware requirements. BERT-base-uncased, on the other hand, provides robust contextual representations that are well-suited for our task while maintaining reasonable resource

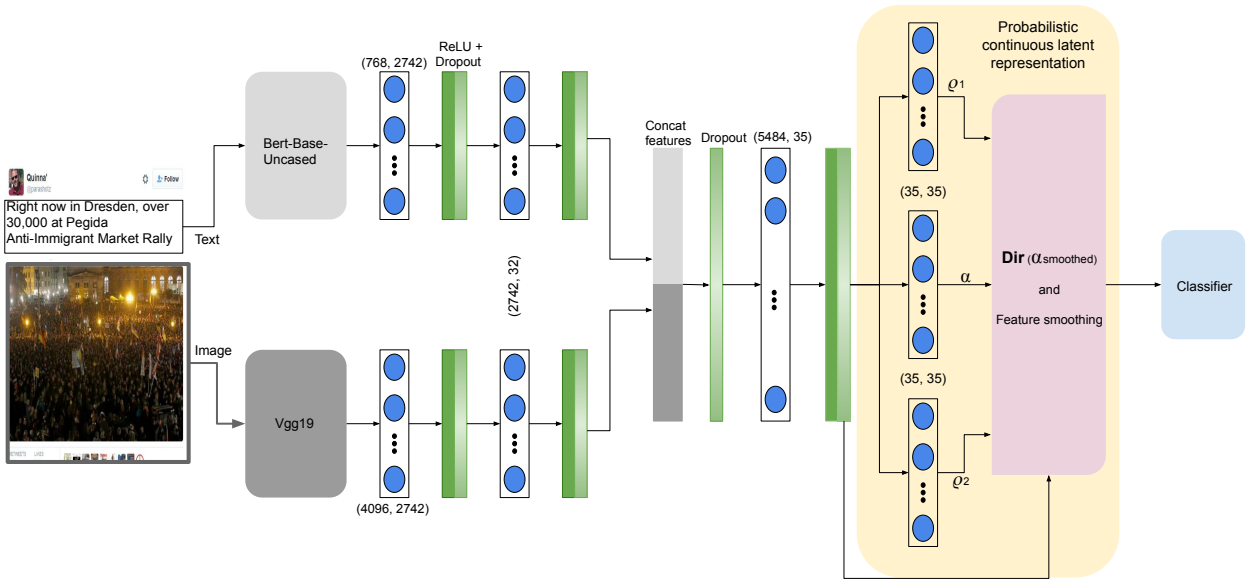


Figure 3.1: A schematic diagram of the proposed SmoothDetector model. The node dimensions correspond to the values specified as (input node and output node).

demands [84]. For further information on the BERT module and its variants, we encourage readers to see the studies in [85, 86]. The BERT layer was followed by a fully connected layer with input and output sizes of 768 and 2742, respectively. The output from this layer is passed through a ReLU activation function and a dropout layer and then into another fully connected layer with 2472 input nodes and 32 output nodes, followed by another ReLU and dropout layer.

3.2.2 Visual Feature Encoder

In addition to textual content, incorporating visuals with text provides significant information to readers. We considered the posts' visual features based on this intuition. To achieve this, we used the image corresponding to the news articles and extracted the features with a pre-trained Visual Geometry Group (VGG19), a convolutional neural network trained on a large-scale ImageNet [87]. The VGG19–extracted features were fed into a fully connected layer with 4096 input nodes and 2742 output nodes. Similar to the textual encoder component, the output of this fully connected layer is passed through a ReLU and a dropout layer. This is followed by another fully connected layer with 2742 input nodes and 32 output nodes, followed by another ReLU and dropout layers.

3.2.3 Multimodal Fusion Component

Another crucial component of our architecture is the multimodal feature integrator submodule. After extracting the key information from the textual and visual parts of the news article, we first concatenated the two feature sets, as described in Figure 3.1. The fused features passed through a dropout layer before being fed into a fully connected layer with 5484 input nodes and 35 output nodes. This was followed by a rectified linear unit (ReLU) activation layer and another dropout layer.

3.2.4 Smoothed Dirichlet (SD) Component

The choice of the smoothed Dirichlet process is driven by its advantages in modeling complex data distributions. One significant issue in using Gaussian distributions is the phenomenon of component collapsing, where the model may converge to a limited set of components, undermining its ability to capture the underlying structure of the data. This issue is highlighted in [1]. In contrast, [1] demonstrated that the Dirichlet distribution does not suffer from component collapsing to the same extent. Moreover, the smoothed Dirichlet process offers a notable improvement in flexibility regarding model complexity. The SD allows the model to capture intrinsic patterns in complex data by incorporating smoothing parameters. This adaptability means that the model can better accommodate varying degrees of data sparsity or richness, leading to improved performance and more accurate predictions on unseen data. The use of smoothing effectively prevents overfitting while enabling the model to capture relevant patterns, making the SD particularly valuable in practical applications where data characteristics may change over time.

The probabilistic submodule of the proposed approach is a novel component. It captures the uncertainties and diversity in news articles while allowing for a continuous latent representation. The fused features are fed to this sub-module, which generates three parameters through three fully connected layers that are subsequently fed into the smoothed Dirichlet algorithm. The smoothed Dirichlet algorithm learns the continuous latent space and samples the representations, which are then fed into the classifier. The details of the algorithm and the flowchart of the smoothed Dirichlet process are presented in Algorithm 2 and Figure 3.2, respectively.

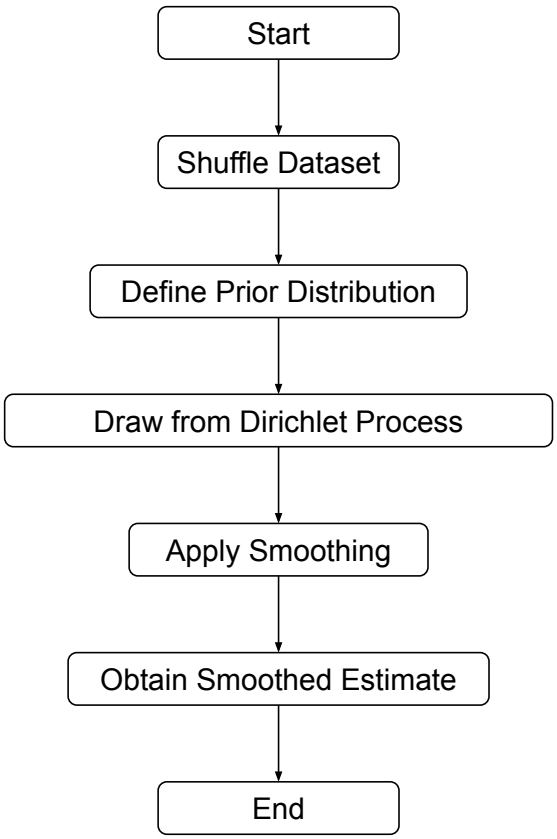


Figure 3.2: Smoothed Dirichlet Flowchart.

3.2.5 Classifier

The final component of the architecture is the classifier. It receives features sampled from the smoothed Dirichlet algorithm layer. These features were processed through a fully connected layer, and the resulting output was then passed through a sigmoid activation function. The sigmoid function converts the output into a probability range between 0 and 1, thereby determining the likelihood of the news article being fake or real, respectively. This probabilistic result allows the model to make binary decisions with a certain degree of confidence.

3.3 Smoothed Dirichlet Transformation

In our multimodal approach to fake news detection on social media, we focused on two primary categories of data: texts and images. With the increasing ability of social media users to upload

audio and videos, these will also be integrated into multimodal datasets in the future. The task is formulated as follows: Let $\mathcal{D} = \{\mathcal{P}, \mathcal{M}, \mathcal{U}\}$ represent a fake news dataset, where $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denotes the set of N real or fake articles. $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$ represents the images associated with each article, whereas the set of social media users who have posted any of the articles is denoted as $\mathcal{U} = \{u_1, u_2, \ldots\}$. Our goal is to train a probabilistic fake news detector, codenamed SmoothDetector, to predict category y of a news article using a multimodal dataset comprising the articles and their corresponding images.

3.3.1 Smoothed Dirichlet Reparameterization

One of the major challenges in training smoothed Dirichlet distributions is that they cannot be directly trained with variational autoencoders because they do not have shifting parameters, and there is no direct transformation. However, a Gaussian variable $z \sim N(\mu, \sigma^2)$ can be reparameterized as $z = \mu + \epsilon\sigma$, $\epsilon \sim N(0, 1)$, enabling the gradient to be backpropagated through the latent variable z. Where a number of approaches have been explored to tackle these challenges. For example, [14, 79] attempted reparameterization using a Gamma distribution through an efficient rejection sampler. The study in [1] explored the Laplace approximation [88] on a softmax basis. This is significantly interesting to us as it enables unconstrained optimization of the cost function, thereby removing the constraints associated with the simplex. Following Equation (3.1), we define our smoothed feature X^s as [20]:

$$X^s = \lambda X^C + (1 - \lambda) X^{SD} \tag{3.3}$$

where X^C represents the features learned from the concatenation of the multimodal dataset and λ is the smoothing regularizer such that $0 \leq \lambda \leq 1$. X^{SD} denotes the smoothed features sampled from the smoothed Dirichlet reparameterization. The reparameterization trick for the sample X^{SD} is summarized below. Given that α is the Dirichlet prior $x_i \sim Dir(\alpha_1, \alpha_2, ..., \alpha_K)$ [20], which is analogous to Equation (3.2):

$$p(X|\vec{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{3.4}$$

To implement the reparameterization trick, we introduce an auxiliary random variable Z, such that:

$$X = g(\vec{\alpha}, Z) \tag{3.5}$$

where g is a deterministic function that depends on both $\vec{\alpha}$ and Z. For the Dirichlet distribution, the Laplace approximation is a suitable choice for g [1] on a softmax basis:

$$x_i = \frac{\exp(z_i)}{\sum_{i=1}^{K} \exp(z_i)} \tag{3.6}$$

where z_i is sampled from a distribution, such as the standard normal distribution. Thus, to smooth the Dirichlet distribution, we introduce a small positive scalar τ and define the smoothed parameter vector $\vec{\alpha}_s$ as:

$$\vec{\alpha}_s = \vec{\alpha} + \tau \cdot Z \tag{3.7}$$

We use the analytical transformation of a Gaussian distribution to sample $Z : Z = \vec{\varrho}_1 + \vec{\varepsilon} \cdot \vec{\varrho}_2$. Parameters $\vec{\varrho}_1$ and $\vec{\varrho}_2$ are generated from the concatenation of multimodal features, whereas $\vec{\varepsilon}$ is normally distributed. Thus, the reparameterization trick for the smoothed Dirichlet distribution can be formulated as:

$$X^s \sim g(\vec{\alpha}^s, Z) \Rightarrow x_i^s = \frac{\exp(\alpha_i + \tau \cdot z_i)}{\sum_{i=1}^{K} \exp(\alpha_i + \tau \cdot z_i)} \tag{3.8}$$

The smoothed Dirichlet sampling process can be summarized as follows:

$$\begin{aligned}
&X^C \Longleftarrow \mathcal{D} : \mathcal{D} = \{\mathcal{P}, \mathcal{M}, \mathcal{U}\} \\
&\vec{\alpha}, \vec{\varrho}_1, \vec{\varrho}_2 \Longleftarrow X^C \\
&\vec{\alpha}_s \Longleftarrow \vec{\alpha}, \vec{\varrho}_1, \vec{\varrho}_2 \\
&X^{SD} \sim \mathrm{Dir}(\vec{\alpha}^s) \\
&X^s = \lambda X^C + (1 - \lambda) X^{SD}
\end{aligned} \tag{3.9}$$

The details of this algorithm are presented in Algorithm 2. The formulation allows backpropagation through parameters α, ϱ_1, and ϱ_2, thereby enabling easy training using gradient-based optimization techniques.

3.3.2 SmoothDector's Loss Functions

To train the SmoothDector model, two loss functions are utilized: binary cross-entropy loss (BCE) and Kullback-Leibler divergence loss (KL_{SD}).

3.3.2.1 Binary Cross-Entropy loss (BCE)

The binary cross-entropy (BCE) loss function is commonly used in binary classification tasks. It measures the difference between two probability distributions: the predicted probabilities and actual

Algorithm 2 Smoothed Dirichlet Multimodal Generative Process

Data:

$\mathcal{D}$: Dataset

Result:

ϕ: Learnable model's parameters

$\phi \leftarrow$ Initialize parameters

while Optimizer not converged **do**

$\mathcal{X}^C \sim \mathcal{D}(\mathcal{P}, \mathcal{M}, \mathcal{U})$ ▷ Random batch of dataset

$\vec{\alpha}, \vec{\varrho}_1, \vec{\varrho}_2 \leftarrow$ model $(\mathcal{X}^C; \phi)$

$\varepsilon \sim N(\mu, \sigma^2)$

$Z \leftarrow \vec{\varrho}_1 + \vec{\varepsilon} \cdot \vec{\varrho}_2$

$\vec{\alpha}_s \leftarrow \vec{\alpha} + \tau \cdot Z$

$X^{SD} \sim \mathrm{Dir}(\vec{\alpha}^s)$

$X^s \leftarrow \lambda X^C + (1 - \lambda) X^{SD}$

Predict news label, y $\in (0, 1)$

Compute $\bar{\mathcal{L}}(\phi; y, X)$

Update (ϕ)

end while

labels [89]. This loss function penalizes incorrect predictions more heavily, making it effective for training models in binary classification tasks. The BCE loss for a single instance can be formulated as [90, 91]:

$$\text{BCE} = -\left(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right) \tag{3.10}$$

where y and $\hat{y}$ denote the actual label and predicted label, respectively, such that label $\in (0, 1)$. For a batch of N instances, the BCE loss is averaged over all instances in the batch [90]:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)\right) \tag{3.11}$$

3.3.2.2 KL-divergence Between Two Smoothed Dirichlet Distributions

First, we define the Dirichlet distribution and then introduce smoothing. A Dirichlet is parameterized by a vector of positive real numbers, often denoted by α, where each element represents the concentration of the corresponding outcome. The probability density function (PDF) of the Dirichlet distribution is given by [20]:

$$P(X|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{3.12}$$

For the smoothed Dirichlet distribution, we add a pseudo-count ξ to each component of the parameter vector α in addition to feature smoothing. This is often referred to as Laplace smoothing. The smoothed Dirichlet distribution with Laplace smoothing is parameterized by $\alpha + \xi$ as:

$$P(X^s|\vec{\alpha}, \vec{\xi}) = \frac{1}{B(\vec{\alpha} + \vec{\xi})}\prod_{i=1}^{K} (x_i^s)^{\alpha_i + \xi - 1} \tag{3.13}$$

The KL divergence between two Dirichlet distributions with parameters $\vec{\alpha}_1$ and $\vec{\alpha}_2$ can be calculated as follows:

$$D_{\text{KL}}(P||Q) = \int P(X) \log\left(\frac{P(X)}{Q(X)}\right) dX \tag{3.14}$$

where $P(X)$ and $Q(X)$ are the probability density functions of two smoothed Dirichlet distributions.

$$= \int P(X^s|\vec{\alpha}_1, \vec{\xi}) \log \left(\frac{P(X^s|\vec{\alpha}_1, \vec{\xi})}{P(X^s|\vec{\alpha}_2, \vec{\xi})} \right) dX^s$$

$$= \int \frac{1}{B(\vec{\alpha}_1 + \vec{\xi})} \prod_{i=1}^{K} (x_i^s)^{\alpha_{1i}+\xi-1} \log$$

$$\left(\frac{\frac{1}{B(\vec{\alpha}_1+\vec{\xi})} \prod_{i=1}^{K} (x_i^s)^{\alpha_{1i}+\xi-1}}{\frac{1}{B(\vec{\alpha}_2+\vec{\xi})} \prod_{i=1}^{K} (x_i^s)^{\alpha_{2i}+\xi-1}} \right) dx^s$$

$$= \int \frac{1}{B(\vec{\alpha}_1 + \vec{\xi})} \prod_{i=1}^{K} (x_i^s)^{\alpha_{1i}+\xi-1} \log \left(\frac{B(\vec{\alpha}_2 + \vec{\xi})}{B(\vec{\alpha}_1 + \vec{\xi})} \right)$$

$$+ \sum_{i=1}^{K} (\alpha_{1i} + \xi - 1) \log((x_i^s)) - \sum_{i=1}^{K} (\alpha_{2i} + \xi - 1) \log((x_i^s)) dx^s$$

$$= \log \left(\frac{B(\vec{\alpha}_2 + \vec{\xi})}{B(\vec{\alpha}_1 + \vec{\xi})} \right) + \sum_{i=1}^{K} (\alpha_{1i} + \xi - 1) E_{x^s \sim P}[\log((x_i^s))]$$

$$- \sum_{i=1}^{K} (\alpha_{2i} + \xi - 1) E_{x^s \sim P}[\log((x_i^s))]$$

using the expected value of a logarithmized Dirichlet variate

$$x^s \sim \mathrm{Dir}(\alpha + \xi) \implies [\log((x_i^s))] = \psi(\alpha_i + \xi)$$

$$- \psi \left(\sum_{i=1}^{k} \alpha + \xi \right)$$

$$\mathrm{KL}_{SD} = \log \left(\frac{B(\vec{\alpha}_2 + \vec{\xi})}{B(\vec{\alpha}_1 + \vec{\xi})} \right) + \left(\sum_{i=1}^{K} (\alpha_{1i} + \xi - 1) - \sum_{i=1}^{K} (\alpha_{2i} + \xi - 1) \right) \left[\psi(\alpha_{1i} + \xi) - \psi \left(\sum_{i=1}^{k} \alpha_1 + \xi \right) \right] \tag{3.15}$$

where $\psi(\cdot)$ denotes the digamma function. It is worth mentioning that omitting the normalizing component did not affect the model's performance.

Considering Equation (3.11) and Equation (3.15), the total loss, T_{loss}, is formulated as the sum of the KL_{SD} and BCE:

$$T_{loss} = \mathrm{BCE} + \mathrm{KL}_{SD} \tag{3.16}$$

3.4 Experimental Results

Table 3.1: Evaluation of smoothed Dirichlet prior (α) on the Twitter dataset. We set $\lambda = 0.4$, epoch = 20, learning rate = $3e^{-5}$, and optimizer = Adam.

Dataset	Distribution Prior (α)	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	**0.01**	0.972	0.964	0.977	0.971	0.980	0.982	0.981
	0.02	0.961	0.970	0.967	0.960	0.977	0.958	0.968
	0.05	0.945	0.848	0.946	0.947	0.965	0.899	0.872
	0.1	0.944	0.900	0.845	0.951	0.962	0.888	0.866
	0.2	0.911	0.921	0.900	0.812	0.851	0.899	0.867

This section begins by introducing the datasets used to evaluate the SmoothDetector model and baseline models. Next, we detail the experimental settings and training parameters used to train the model. Finally, we present the experimental results and compare the performance of the model with that of the baseline models.

3.4.1 Dataset

We assessed the performance of SmoothDetector using two publicly available benchmark datasets: MediaEval [92] and Weibo [93]. The datasets are composed of genuine social media information, with one dataset sourced from Twitter and the other from Weibo blogs. Each dataset included real interactions, posts, and engagements from these platforms, providing a comprehensive view of social media dynamics for our analysis. It is important to highlight that both datasets were multimodal, incorporating textual posts and their associated images. This combination of text and visual content allows for more robust and nuanced analysis, enabling our model to leverage multiple data types for more accurate fake news detection.

3.4.1.1 Twitter MediaEval Dataset

The Twitter dataset was created in 2015 for the Verifying Multimedia Use task at MediaEval, specifically aimed at detecting fake multimedia content on social media platforms [92]. This dataset includes 17,000 unique tweets, each containing textual content, accompanying images or videos, as well as pertinent social context data. It is organized into two segments: a development set

Table 3.2: Evaluation of the smoothening regularizer (λ) on the Twitter dataset. We set $\alpha = 0.01$, epoch = 20, learning rate = $3e^{-5}$, and optimizer = Adam.

Dataset	Smoothening regularizer (λ)	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	**0.4**	0.972	0.964	0.977	0.971	0.980	0.982	0.981
	0.5	0.969	0.961	0.978	0.978	0.947	0.962	0.964
	0.6	0.957	0.957	0.941	0.955	0.946	0.972	0.959
	0.7	0.911	0.869	0.956	0.927	0.913	0.926	0.897
	0.8	0.842	0.772	0.957	0.855	0.948	0.735	0.828

with 9,000 fake news tweets and 6,000 real news tweets and a test set comprising 2,000 tweets. The dataset is intended to make it easier to assess the veracity of the information posted on social media.

3.4.1.2 Weibo Dataset

The Weibo dataset comprises data from Weibo, a Chinese microblogging site, and Xinhua News Agency, a reputable news source in China [93]. The fake news was gathered between May 2012 and June 2016 and verified by Weibo's official rumor debunking system. This system encourages regular users to report suspicious tweets, which are then reviewed by a committee of trusted users, who classify the tweets as true or fake. This system is a reliable source for gathering rumors. Tweets verified by the Xinhua News Agency were classified as non-rumors. This combination of verified fake and real news ensures the reliability of the dataset for fake news detection. Similar to the Twitter dataset, Weibo is a multimodal dataset that includes textual content along with images. The Weibo dataset consisted of 3,615 fake news posts, 4,105 real news posts, and 7,720 images. Please note that the majority of Weibo content is in Simplified Chinese characters, mainly from users in Mainland China. However, some users from areas such as Hong Kong and Taiwan may tweet using traditional Chinese characters. Furthermore, the dataset is multimodal, incorporating text messages, user information, social context, and image content.

3.4.2 Baseline Models

We conducted extensive experiments to evaluate the performance of our proposed method by comparing it with a diverse set of state-of-the-art and representative baseline models for fake news

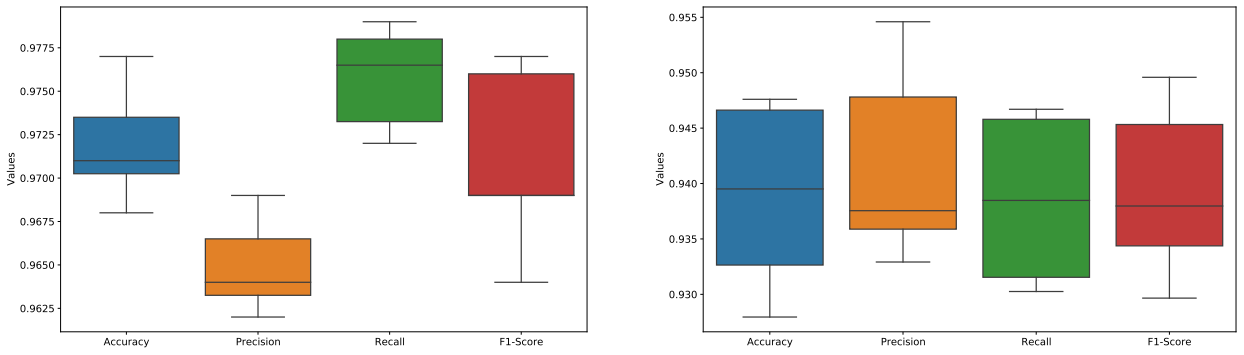
detection. These baselines cover unimodal, multimodal, and uncertainty-aware approaches. For unimodal baselines, we consider a *Textual* model based on BERT-uncased, which analyzes post content and predicts authenticity using learned textual representations, as well as a *Visual* model that employs a pre-trained VGG-19 network to extract visual features from images, followed by a fully connected layer for classification.

Several multimodal methods are included for comparison. MPFN [94] introduces a Multimodal Progressive Fusion Network that alleviates information loss by strengthening hierarchical cross-modal connections. FCINet [95] proposes a Frequency-Aware Cross-Modal Interaction Network with a triple-branch encoder that captures frequency, spatial, and textual features through parallel interaction mechanisms. Dirichlet-based methods [96] address modality asymmetry and conflicting beliefs by modeling uncertainty with a Dirichlet distribution to guide feature fusion. The att-RNN model [67] adopts an end-to-end architecture with attention mechanisms over fused text and image features, excluding social context information to ensure fair comparison.

We also include event- and representation-based approaches. EANN and its variant EANN- [38] utilize an event discriminator to remove event-specific features, improving generalization; EANN incorporates auxiliary tasks, whereas EANN- does not. MVAE and MVAE- [77] employ a multi-modal variational autoencoder framework consisting of encoders, decoders, and a fake news detector, where MVAE includes additional subtasks while MVAE- excludes them. SpotFake [68] integrates advanced language models with deep visual feature extraction to jointly analyze textual and visual content. Furthermore, VAEMTL_AV, VAEMTL_IM, and VAEMTL_DY [69] extend the MVAE framework by introducing weighted multitask learning, where VAEMTL_AV assigns equal weights to all modules, VAEMTL_IM uses manually specified importance weights, and VAEMTL_DY dynamically adjusts weights based on model performance. Finally, BMR [97] leverages bootstrap multi-view representations and a mixture-of-experts fusion strategy to capture complementary news characteristics from multiple perspectives, enabling robust multimodal integration.

3.4.3 Evaluation Results

This section presents the performance of the proposed model and compares it with baselines. In line with previous studies summarized in Section 3.4.2, we use four commonly employed metrics to assess the performance of fake news detection methods: accuracy, precision, recall, and F1-Score (F1) [98]. First, we conducted extensive experiments to analyze the effect of the smoothed Dirichlet prior, α, and the effect of varying the smoothing regularizer λ. We define a set of potential values



(a) Twitter dataset

(b) Weibo dataset

Figure 3.3: Box plots representation of Accuracy, Precision, Recall, and F1-Score distributions across 10 runs for: (a) Twitter dataset, and (b) Weibo Dataset. We set $\alpha = 0.01$, $\lambda = 0.4$, Twitter_epoch = 20, Weibo_epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

Table 3.3: Assessment of Confidence Intervals (CI) for Performance Metrics on Twitter and Weibo Datasets. We set $\alpha = 0.01$, $\lambda = 0.4$, Twitter_epoch = 20, Weibo_epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

	Metrics	Test results for 10 runs										Mean	CI
Twitter	Accuracy	0.969	0.971	0.977	0.970	0.974	0.968	0.971	0.972	0.971	0.974	0.972	[0.970, 0.974]
	Precision	0.964	0.962	0.964	0.965	0.964	0.963	0.967	0.962	0.967	0.969	0.965	[0.964, 0.966]
	Recall	0.978	0.977	0.972	0.972	0.973	0.978	0.978	0.974	0.976	0.979	0.976	[0.974, 0.978]
	F1-Score	0.957	0.954	0.948	0.952	0.951	0.948	0.950	0.952	0.950	0.956	0.952	[0.950, 0.954]
Weibo	Accuracy	0.947	0.939	0.930	0.944	0.929	0.948	0.928	0.939	0.947	0.939	0.939	[0.935, 0.943]
	Precision	0.946	0.933	0.936	0.935	0.937	0.938	0.951	0.955	0.934	0.948	0.941	[0.936, 0.946]
	Recall	0.932	0.940	0.947	0.932	0.947	0.930	0.937	0.943	0.931	0.947	0.938	[0.934, 0.942]
	F1-Score	0.949	0.946	0.935	0.932	0.929	0.936	0.934	0.943	0.939	0.947	0.939	[0.935, 0.943]

for each hyperparameter and exhaustively evaluate all possible combinations to determine which configuration yields the best performance. By employing grid search, we ensure a comprehensive exploration of the hyperparameter space, allowing us to capture the nuances in how different parameter settings affect the model's outcomes. Once we identify the best-performing hyperparameter values through this process, we maintain consistency by applying these values across all subsequent experiments. This approach not only enhances the reliability of our results by minimizing variability due to differing hyperparameter settings but also allows for a more straightforward comparison of model performance across different scenarios. Next, we used the held-out test set to evaluate the performance of SmoothDetector and compare it with the baseline models. It is important to highlight that all the evaluations presented are based on the average results from 10 separate runs.

Table 3.4: Assessment of p-values for Performance Metrics on Twitter and Weibo Datasets, S denotes the standard deviation, while t denotes the t-statistics. We set $\alpha = 0.01$, $\lambda = 0.4$, Twitter_epoch = 20, Weibo_epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

Metric	Dataset	Baseline (FCINet)	SmoothDetector	Variance	S	t	p-value	Significant?
Accuracy	Twitter	0.908	0.972	$7.12e^{-6}$	0.0031	64.79	$< 1 \times 10^{-9}$	Yes
Precision	Twitter	0.913	0.965	$5.34e^{-6}$	0.0024	66.91	$< 1 \times 10^{-9}$	Yes
Recall	Twitter	0.909	0.976	$7.34e^{-6}$	0.0026	81.32	$< 1 \times 10^{-9}$	Yes
F1-Score	Twitter	0.910	0.972	$2.90e^{-6}$	0.0030	44.43	$< 1 \times 10^{-9}$	Yes
Accuracy	Weibo	0.926	0.939	$5.98e^{-5}$	0.0067	5.17	2.1×10^{-4}	Yes
Precision	Weibo	0.926	0.941	$6.08e^{-5}$	0.0073	6.59	7.4×10^{-5}	Yes
Recall	Weibo	0.926	0.939	$5.06e^{-5}$	0.0064	5.92	1.2×10^{-4}	Yes
F1-Score	Weibo	0.926	0.939	$4.73e^{-5}$	0.0064	6.45	8.5×10^{-5}	Yes

For further details, please refer to our statistical analyses. The baseline model results used for the evaluation are derived from existing research. However, certain metrics are noted as "NA" due to the absence of relevant data in the original studies.

Our preliminary experimental results on the Twitter dataset to determine the optimal values for α through the grid search ranging from zero to one while keeping the other parameters constant are presented in Table 3.1. The top five values yielding the best results are listed in Table 3.1. Similarly, we demonstrated the crucial role of smoothing in balancing the contributions from the smoothing vector and learned feature representations by varying the value of λ with other fixed parameters. Note that a higher λ value corresponds to less smoothing, whereas a lower value indicates a greater degree of smoothing, as described in Equation (3.3). Table 3.2 shows that lower λ values enhance the model performance, underscoring the benefits of smoothing. To prevent over-smoothing, we selected $\lambda = 0.4$. It is important to note that we used the same values of $\alpha = 0.01$ and $\lambda = 0.4$ for both datasets. This consistency suggested that our model can function effectively across different datasets without requiring adjustment of the model's prior and degree of smoothing.

The box plots for the Twitter dataset shown in Figure 3.3a illustrate the distribution of performance metrics across 10 runs. The accuracy values exhibit low variability, clustering around the higher end of the scale, which reflects consistent and robust classification results. Similarly, precision and recall values are tightly distributed with minimal outliers, suggesting stable performance in terms of correctly identifying positive instances. The F1-scores also demonstrate limited variability, though slightly broader compared to accuracy, indicating balanced performance between precision and recall. Overall, the distributions reflect high performance across all metrics with slight

variations in F1-scores, likely due to their dependence on both precision and recall. Moreover, the box plots for the Weibo dataset, as shown in Figure 3.3b remain consistently high, although with a slight variation. The accuracy and precision values show relatively compact distributions, indicating stable performance with minimal outliers. Recall values have a slightly larger spread. The F1-score maintains values close to the upper range while balancing precision and recall. Despite the slight spread, all metrics demonstrate high reliability and stability across the dataset.

To further investigate the statistical analyses of our proposed method, Table 5.1 presents the confidence interval (CI) performance metrics for the Twitter and Weibo datasets based on confidence intervals (CI) calculated over 10 experimental runs. Confidence intervals provide insights into the reliability and stability of the models across multiple runs by quantifying the range within which the true mean likely falls. The confidence interval is computed as follows:

$$\text{CI} = \mu \pm z^* \cdot \frac{s}{\sqrt{n}} \tag{3.17}$$

where μ, s, n, and z^* denote the sample (mean, standard deviation, size) and $z-$value for the confidence interval of 95% (1.96).

As shown in Table 5.1, the Twitter dataset demonstrates strong consistency and robustness, with narrow confidence intervals across all metrics, indicating reliable classification performance. For example, the accuracy achieved a mean of 0.972 with a CI of [0.970, 0.974]. Furthermore, the Weibo dataset also exhibits high performance. Weibo accuracy has a mean of 0.939 with a CI of [0.933, 0.945]. Its slightly broader confidence intervals suggest marginal variability; nevertheless, overall stability remains commendable.

To assess the significance of the observed differences between baseline and experimental metrics, we conduct a two-tailed t-test for each metric against the respective baseline value. Please note that we chose FCINet [95] as our baseline model because it performed next to our proposed model on most of the performance metrics. In addition, FCINet provides values for all metrics in their studies. This test assumes that the sample means of the experimental results can be compared to the baseline to determine whether the improvements are statistically significant. Thus, we define the hypotheses as follows:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned} \tag{3.18}$$

where μ is the sample mean, μ_0 is the baseline value, s is the sample standard deviation, and n

is the sample size. The p-value is obtained from the t-distribution as follows:

$$p = 2 \cdot (1 - \text{CDF}(|t|, df)) \tag{3.19}$$

where $\text{CDF}(|t|, df)$ represents the cumulative distribution function of the t-distribution with degrees of freedom $df = n-1$. We set the significance level (α) to 0.05. The cumulative distribution function (CDF) of the t-distribution is computed as:

$$\text{CDF}(|t|, df) = \int_{-\infty}^{|t|} f(t, df)\, dt \tag{3.20}$$

where $f(t, df)$ is the probability density function (PDF) of the t-distribution, given by:

$$f(t, df) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\,\pi}\,\Gamma\left(\frac{df}{2}\right)} \left(1 + \frac{t^2}{df}\right)^{-\frac{df+1}{2}} \tag{3.21}$$

where $\Gamma(x)$ is the gamma function.

The results presented in Table 3.4 indicate statistically significant improvements across all metrics for both Twitter and Weibo, as evidenced by p-values below the 0.05 threshold. For Twitter, the most pronounced enhancement was observed in the Recall metric, where the t-statistic of 81.32 highlights a substantial deviation from the baseline. Similarly, for Weibo, Precision demonstrated notable improvement, supported by a t-statistic of 6.59. These findings are pivotal, as they substantiate the efficacy of the proposed methodology in improving key evaluation metrics. The substantial statistical evidence strengthens confidence in the generalizability and applicability of these advancements in real-world scenarios.

The analysis of the results presented in Table 3.5 highlights the comparative performance of SmoothDetector and several baseline models across Twitter and Weibo datasets. SmoothDetector demonstrates the highest performance for the Twitter dataset, achieving an accuracy of 0.972, which significantly surpasses all baseline models. This amounts to a 7.0% increase compared to the next best-performing model, FCINet, with an accuracy of 0.908. It also excels in precision, recall, and F1-score, attaining values of 0.965, 0.976, and 0.972, respectively. These metrics underscore its robustness in identifying both fake and real news. For fake news detection, SmoothDetector achieves a precision of 0.964, a recall of 0.977, and an F1-score of 0.971, while for real news detection, it reaches a precision of 0.980, a recall of 0.982, and an F1-score of 0.981, indicating balanced effectiveness across both categories.

In contrast, models like MPFN, FCINet, and BMR also perform well but fall short of SmoothDetector. FCINet achieves an accuracy of 0.908 and demonstrates strong performance in detecting

real news, with a precision of 0.955, a recall of 0.907, and an F1-score of 0.930, but its performance for fake news (precision of 0.828 and F1-score of 0.868) lags behind SmoothDetector. BMR shows competitive metrics with an accuracy of 0.883 but does not match SmoothDetector's balance across precision, recall, and F1-scores. Notably, earlier models, such as Textual and Visual methods, exhibit much lower performance, with Textual achieving an accuracy of 0.526 and Visual 0.596, indicating their limitations in handling the complexity of the multimodal data.

For the Weibo dataset, SmoothDetector similarly outperforms all baseline models with an accuracy of 0.939, reflecting a 1.4% increase over the second-best model, FCINet, which achieves an accuracy of 0.926. SmoothDetector also achieved nearly optimal precision, recall, and F1-score values (0.941, 0.939, and 0.939, respectively). In fake news detection, SmoothDetector achieves a precision of 0.966, a recall of 0.908 lower than SpotFake with 0.964 recall, and an F1-score of 0.936, while its performance in real news detection is equally robust, with a precision of 0.936, a recall of 0.969, and an F1-score of 0.942. Among the baselines, FCINet and VAEMTL variants (AV, IM, DY) are competitive, with FCINet attaining an accuracy of 0.926 and VAEMTL_DY achieving 0.921. However, while these models show strong performance, particularly in one of the two categories (fake or real news), their overall metrics remain slightly underperformed to SmoothDetector.

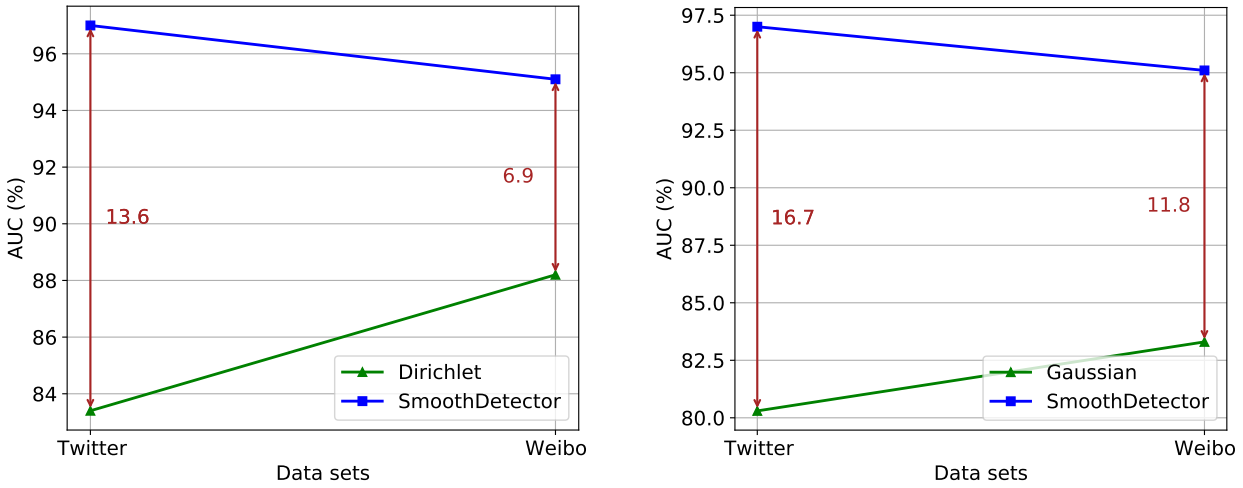
Models like SpotFake and BMR, while notable, exhibit limitations in recall and F1-score balance. SpotFake achieves high recall for fake news detection (0.964) but falters in recall for real news (0.656). Similarly, BMR demonstrates consistency with an accuracy of 0.889 but does not reach the same level of overall precision and recall as SmoothDetector. These results highlight that, while SpotFake outsmarts SmoothDetector in the fake news recall, SmoothDetector consistently demonstrated strong performance across multiple key metrics and outperformed other models in most metrics, affirming its robust capability in detecting both fake and real news on the Weibo dataset. Note that it is evident from the results on both datasets that SmoothDetector, which utilizes smoothed Dirichlet, outperformed Gaussian-based MVAE and Dirichlet distributions.

The performance of SmoothDetector can be largely attributed to its capacity to learn intricate patterns within a continuous latent space. This ability enables the model to capture and represent complex relationships in the data that may not be immediately apparent. SmoothDetector can effectively discern underlying structures by operating in a latent space, facilitating a deeper understanding of the data's characteristics. Moreover, SmoothDetector's adaptability plays a crucial role in its effectiveness. The model is designed to accommodate varying degrees of data sparsity or richness, which is essential in real-world applications where data quality and quantity can fluctuate

Table 3.5: Performance of SmoothDetector vs. baseline models on Twitter and Weibo datasets. We set $\alpha = 0.01$, $\lambda = 0.4$, Twitter_epoch = 20, Weibo_epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

Dataset	Model	Acc.	Prec.	Rec.	F1	Fake News			Real News		
						P.	R.	F1	P.	R.	F1
Twitter	Textual	0.526	0.536	0.531	0.515	0.586	0.533	0.569	0.469	0.526	0.496
	Visual	0.596	0.649	0.662	0.597	0.695	0.518	0.593	0.524	0.700	0.599
	MPFN	0.833	—	—	—	0.846	0.921	0.880	0.809	0.721	0.740
	FCINet	0.908	0.913	0.909	0.910	0.828	0.913	0.868	0.955	0.907	0.930
	att-RNN	0.664	0.749	0.615	0.676	0.749	0.615	0.676	0.589	0.728	0.651
	EANN-	0.648	0.810	0.498	0.617	0.810	0.498	0.617	0.584	0.759	0.660
	EANN	0.715	0.822	0.638	0.719	—	—	—	—	—	—
	Dirichlet	0.824	—	—	—	0.772	0.918	0.838	0.899	0.730	0.806
	MVAE	0.745	—	—	—	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.778	—	—	—	0.751	0.900	0.820	0.832	0.606	0.701
	VAEMTL_AV	0.869	—	—	—	0.820	0.784	0.802	0.880	0.917	0.898
	VAEMTL_IM	0.871	—	—	—	0.826	0.772	0.798	0.891	0.920	0.905
	VAEMTL_DY	0.888	—	—	—	0.838	0.821	0.829	0.912	0.922	0.917
	BMR	0.883	—	—	0.870	0.927	0.746	0.827	0.865	0.965	0.912
	SmoothDetector	**0.972**	**0.965**	**0.976**	**0.972**	**0.964**	**0.977**	**0.971**	**0.980**	**0.982**	**0.981**
Weibo	Textual	0.643	0.624	0.607	0.620	0.662	0.578	0.617	0.609	0.685	0.647
	Visual	0.608	0.609	0.603	0.607	0.610	0.605	0.607	0.607	0.611	0.609
	MPFN	0.838	—	—	—	0.857	0.894	0.889	0.873	0.863	0.878
	FCINet	0.926	0.926	0.926	0.926	0.938	0.917	0.927	0.913	0.935	0.924
	att-RNN	0.772	0.778	0.799	0.789	0.797	0.713	0.692	0.684	0.840	0.754
	EANN-	0.795	0.806	0.795	0.800	0.827	0.697	0.756	0.752	0.863	0.804
	EANN	0.827	0.847	0.812	0.829	—	—	—	—	—	—
	Dirichlet	0.888	—	—	—	0.900	0.872	0.886	0.877	0.904	0.890
	MVAE	0.824	—	—	—	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.892	—	—	—	0.902	**0.964**	0.932	0.847	0.656	0.739
	VAEMTL_AV	0.905	—	—	—	0.892	0.921	0.906	0.918	0.891	0.904
	VAEMTL_IM	0.910	—	—	—	0.902	0.927	0.914	0.920	0.893	0.906
	VAEMTL_DY	0.921	—	—	—	0.910	0.940	0.924	0.934	0.901	0.917
	BMR	0.889	—	—	0.889	0.904	0.885	0.895	0.874	0.894	0.884
	SmoothDetector	**0.939**	**0.941**	**0.939**	**0.939**	**0.966**	0.908	**0.936**	**0.936**	**0.969**	**0.942**

significantly. This adaptability will allow the model to maintain robust performance, even when faced with challenges such as limited data availability or highly variable feature distributions. An important feature of our approach is that we used the same hyperparameters across all datasets except for the epoch, which is important for the model to converge. In contrast, SpotFake adjusted its hyperparameters for different datasets, which we believe limits its generalizability for real-world fake news detection tasks.



(a) SmoothDetector Vs. Dirichlet

(b) SmoothDetector Vs. Gaussian

Figure 3.4: Comparative AUC Analysis of SmoothDetector: (a) SmoothDetector Vs. Dirichlet, and (b) SmoothDetector Vs. Gaussian. We set $\alpha = 0.01$, $\lambda = 0.4$, Twitter_epoch = 20, Weibo_epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

3.4.4 Comparative AUC Analysis of SmoothDetector and other probabilistic distributions: Gaussian and Dirichlet

To evaluate the Area Under Curve (AUC) of SmoothDetector over Gaussian-based MVAE [77] and Dirichlet [96], we replace SmoothDetector's probabilistic component with Gaussian and Dirichlet distributions, leveraging their respective KL-divergence formulations. AUC assesses a model's capacity to distinguish between classes across varying thresholds; it plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUC [99] can be formulated as:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{3.22}$$

where $TPR = \frac{TP}{TP+FN}$, $\quad FPR = \frac{FP}{FP+TN}$. For mathematics convenience, we used trapezoidal rule approximation defined as:

$$AUC \approx \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} \tag{3.23}$$

We can clearly see in Figure 3.4 that SmoothDetector outperformed Dirichlet on the Twitter and Weibo data sets by 13.6% and 6.9%, respectively. Similarly, it outperformed Gaussian-based on the Twitter and Weibo data sets by 16.7% and 11.8%, respectively. Please note that the AUC of 1.0 denotes perfect classification, while 0.5 suggests random guessing. SmoothDetector's consistent

AUC performances highlight its robustness and adaptability in handling noisy, dynamic datasets, underscoring its potential for real-world applications such as fake news detection on social media platforms.

3.4.5 Time Complexity Analysis

The time complexity of the Smoothed Dirichlet Multimodal Generative Process is derived as follows. Initializing model parameters takes $O(1)$. Sampling a batch of size b has a complexity of $O(b)$. The forward pass through the model for b samples incurs $O(mb)$, where m represents the model's complexity. Sampling from base distributions and performing transformations on latent variables each requires $O(b)$. Sampling from a Dirichlet distribution over d dimensions for b samples costs $O(bd)$, while feature fusion operations also add $O(bd)$. Predicting labels and computing loss takes $O(b)$, and parameter updates depend on the optimizer, costing $O(m)$. Thus, the dominant operations per epoch are the forward pass and Dirichlet sampling, yielding a per-epoch complexity of $O(\max(mb, bd))$. Over t epochs, the total time complexity becomes $O(t \cdot \max(mb, bd))$.

We also compared the efficiency of our model with state-of-the-art models by measuring execution time. The tests were conducted on a system with a 12th Gen Intel(R) Core(TM) i7-12700K processor (3.60 GHz), 64GB RAM, and a 64-bit operating system. For the Twitter dataset, the per-sample execution times were: SpotFake at 365.44ms, MVAE at 482.39ms, VAEMTL_DY at 477.92ms, and SmoothDetector at 373.17ms. Although SmoothDetector's execution time was slightly higher than SpotFake, it outperformed MVAE and VAEMTL_DY significantly. Similarly, on the Weibo dataset, SpotFake, MVAE, VAEMTL_DY, and SmoothDetector had execution times of 357.12ms, 476.09ms, 463ms, and 367.11ms, respectively. Both SmoothDetector and SpotFake had the lowest execution times, which can be attributed to their architecture not utilizing a decoder module. This demonstrates that SmoothDetector can achieve high performance in fake news detection while minimizing computational costs by eliminating the need for a decoder module.

3.4.6 SmoothDetector Limitations and Future Works

Our results highlight the potential of integrating a probabilistic algorithm with a deep neural network for multimodal fake news detection, demonstrating both flexibility and computational efficiency. This combination allows the model to account for multiple types of data, such as text, images, and other forms of media, making it more adaptable to the diverse nature of modern fake news. However, one of the key challenges of our current approach is the selection of the distribution

prior, which is selected through grid searching and is not learnable. This limitation may affect the overall accuracy and performance of the model, particularly when dealing with more complex or diverse data sets.

Given that contemporary social platforms often include a variety of media types–such as text, images, audio, and video. It becomes essential for a fake news detection system to process and analyze all of these sources efficiently. To address this, our future research will focus on developing a more sophisticated approach that integrates all these modalities. In addition, we plan to leverage a learnable distribution prior. By automating the prior selection process, we aim to enhance the model's adaptability and generalization across different media types, leading to a more robust fake news detection system. This improvement will enable the model to identify misleading or false information more accurately across various forms of content, offering a comprehensive solution without the need for manual intervention in prior selection.

Furthermore, we observe that the prevailing methodology in multimodal studies typically involves learning features from different modalities, such as images and text, in isolation. In this approach, features from each modality are extracted independently before being combined only at a late stage, usually in the layer just before the detector or decoder. This late-stage fusion limits the model's ability to leverage the rich interdependencies between modalities that might exist earlier in the processing pipeline. The important contextual or complementary information between these two modalities might be overlooked by treating image and text features separately for most of the learning process.

In the future, we are keen to investigate how our proposed SmoothDetector can address this limitation by enabling the model to learn from these features simultaneously throughout the learning process rather than deferring their combination to the final stages. By fusing multimodal data earlier and more integrally, we aim to capture more complex interactions between images and text, which could improve the overall detection accuracy. This simultaneous learning would allow the model to develop a more nuanced understanding of multimodal information, potentially uncovering hidden patterns that traditional methods miss. Our future work will focus on refining the architecture to enable this type of multimodal fusion, optimizing SmoothDetector's ability to detect fake news or misinformation more effectively by capitalizing on the synergy between image and text data from the outset. In addition, our future research will investigate enhancing the multimodal feature interaction using co-attention mechanisms and exploring advanced architectures like EfficientNet, ResNet, and Vision Transformers to explore the image features better.

3.5 Conclusion

Fake news is a major issue in the digital age, and many platforms lack sufficient editorial oversight. Research on multimodal fake news detection often depends on discrete latent representations, which hinder dynamic interaction and complex subtasks such as reconstruction and event discrimination, leading to high computational costs and limited modeling flexibility. This study introduces a novel approach, SmoothDetector, which integrates a probabilistic algorithm with a deep neural network to effectively capture uncertainties and diversities in shared latent representations of multimodal data. Utilizing the smoothed Dirichlet distribution, SmoothDetector learns continuous latent representations and captures shared hidden patterns in textual and visual information. Our main hypothesis was that learning continuous latent representations with a probabilistic algorithm significantly enhances the model's ability to generalize and perform well in fake news detection. Extensive experiments on two popular multimodal datasets, Weibo and Twitter, validated our hypotheses. The results demonstrate that high-performance detection can be achieved without additional subtasks in continuous latent representation. By varying the smoothing regularizer, lambda, we demonstrate the critical role of smoothing in balancing contributions from the smoothing vector and learned feature representations, revealing that an optimal smoothing value improves generalization and detection accuracy. Although our results highlight the potential of combining a probabilistic algorithm with a deep neural network for flexible and efficient fake news detection, we acknowledge limitations regarding choosing prior of the distribution that can be challenging, and current multimodal datasets extend beyond text and images to include video and audio. Our next step is to develop a comprehensive method that integrates text, images, audio, and video with a learnable prior to sampling the distribution. This advancement will enhance our model's ability to analyze and detect fake news across diverse media types on modern social platforms without manually selecting the distribution prior.

Chapter 4

Smoothed-ModernBERT: Co-Attentional Synergy of Probabilistic Topic Models and ModernBERT through Dynamic Fusion

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.

4.1 Introduction

Document classification is a fundamental task in natural language processing (NLP), which underpins applications such as news categorization, sentiment analysis, and information retrieval [100, 101]. 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 [102] and GRUs [103, 104, 105] automated feature extraction and sequential patterns captured, although their inherently sequential nature limits parallelism and long-range dependency modeling [106]. Hybrid approaches that combine truncated attention with recurrent units or integrate self-attention into bidirectional GRUs have partially alleviated these issues [106, 107, 108].

The advent of transformers, particularly BERT with its multi-head self-attention and contextual embeddings [109], 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 [110, 111], and extensions combining BERT with capsule networks [112, 113] or graph neural networks [114, 115, 116, 117] 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, such as smoothed Dirichlet distribution, identify coherent themes but struggle with contextual nuance [20]. Bridging this gap, hybrid frameworks like TopicBERT fuse Gaussian topic priors with BERT embeddings [118], 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.

4.1.1 Motivation: Toward Co-Attentional Synergy

Co-attention mechanisms have proven effective in multimodal reasoning by aligning heterogeneous representations [106]. 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.

4.2 Background Studies

4.2.1 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 [80], happiness modeling, and pain estimation [81]. By redistributing probability mass across categories, smoothing enhances model stability [82] and generalizability [83]. Following [20], a smoothed proportion $\mathbf{F}^u$ is derived from

raw word counts using a tunable parameter λ:

$$\mathbf{F}^u \;=\; \frac{\mathbf{X}^s \;-\; (1-\lambda)\,\mathbf{X}^{GE}}{\lambda} \tag{4.1}$$

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:

$$p(\mathbf{X} \mid \boldsymbol{\alpha}, \varepsilon) = \frac{1}{B(\boldsymbol{\alpha} + \varepsilon)} \prod_{i=1}^{K} X_i^{(\alpha_i+\varepsilon)-1}, \quad \frac{1}{B(\boldsymbol{\alpha}+\varepsilon)} = \frac{\Gamma\big(\sum_i (\alpha_i + \varepsilon)\big)}{\prod_i \Gamma(\alpha_i + \varepsilon)}, \tag{4.2}$$

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 [20], 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.

4.2.2 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 [119]. 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 [120]. 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 [119]. To further mitigate computational costs, ModernBERT employs sequence packing and block-wise attention, splitting inputs into contiguous chunks and restricting attention to intra-block and adjacent-block interactions; this achieves sub-quadratic complexity while preserving essential cross-chunk dependencies [119]. Finally, feed-forward sublayers incorporate low-rank matrix factorizations and sparse projection patterns that reduce parameter counts and confine expensive operations to

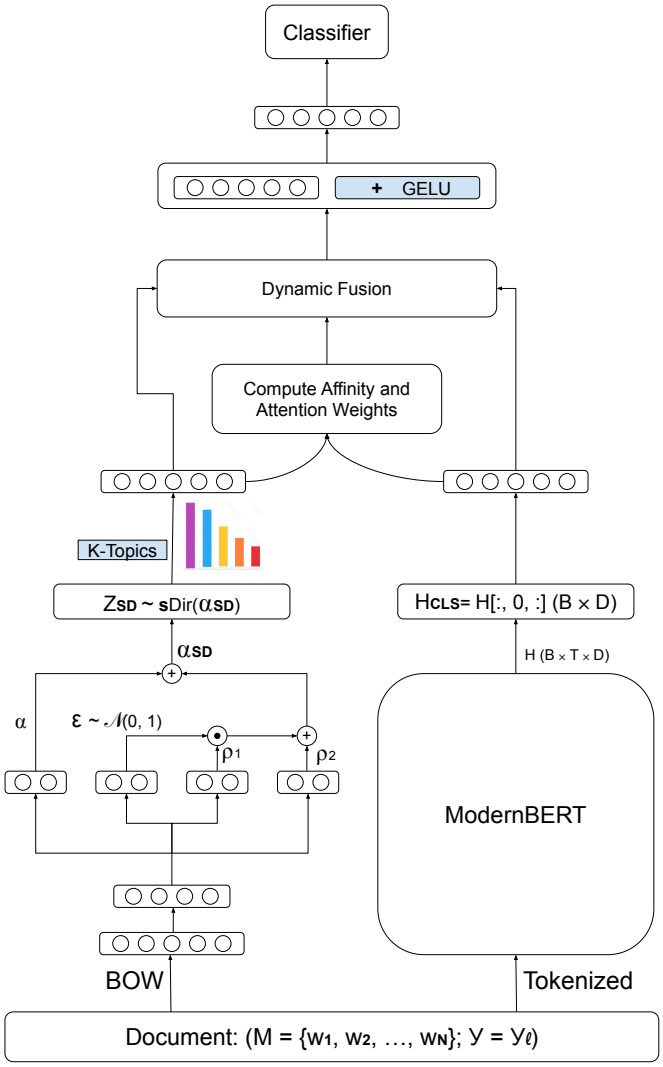


Figure 4.1: A schematic representation of the proposed SD-MoBERT model, leveraging smoothed Dirichlet neural topic model and ModernBERT.

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.

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

Figure 4.1 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 sacri-

ficing contextual depth [119]. 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:

$$\mathbf{Z}_{\text{SD}} \sim \text{sDir}(\boldsymbol{\alpha}_{\text{SD}}) \Longleftarrow \frac{\exp(\boldsymbol{\alpha}_{\text{SD}}^i)}{\sum_{i=1}^K \exp(\boldsymbol{\alpha}_{\text{SD}}^i)} \tag{4.3}$$

$$\boldsymbol{\alpha}_{\text{SD}} = \boldsymbol{\alpha} + \boldsymbol{\rho_2} + \boldsymbol{\epsilon} \odot \exp\big(\log \boldsymbol{\rho_1}\big) \in \mathbb{R}^K, \tag{4.4}$$

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:

$$\mathbf{S} = \langle \mathbf{Z}_{\text{SD}}^t, \mathbf{Z}_{\text{CLS}} \rangle_D + \mathbf{b}_0, \quad \mathbf{Z}_{\text{att}} = \sigma(\mathbf{S}) \tag{4.5}$$

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

$$\mathbf{Z}_{\text{fused}} = \mathbf{Z}_{\text{att}} \mathbf{Z}_{\text{SD}}^t + (1 - \mathbf{Z}_{\text{att}}) \mathbf{Z}_{\text{CLS}}, \quad \mathbf{Z} = \tanh(\mathbf{Z}_{\text{fused}}) \in \mathbb{R}^{B \times D} \tag{4.6}$$

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:

$$\mathcal{L} = \underbrace{-\sum_i y_i \log \hat{y}_i}_{\mathcal{L}_{\text{CE}}} + \beta \underbrace{D_{\text{KL}}\big(q(\mathbf{Z} \mid \mathbf{X}) \| \text{sDir}(\boldsymbol{\alpha}_{\text{SD}})\big)}_{\mathcal{L}_{\text{KL}}}, \tag{4.7}$$

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. $\mathcal{L}_{\text{KL}}$ [121] 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 **??** for more details on the pseudocode for the generative process and **??** for details on $\mathcal{L}_{\text{KL}}$.

4.4 Experimental Results

4.4.1 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.

4.4.2 Datasets

We compare our proposed model with the baseline models on five widely used benchmark datasets, allowing insightful comparisons. The 20 Newsgroups (20NG) dataset [122] 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 [123] contains 10,662 movie reviews balanced between 5,331 positives and 5,331 negatives for sentiment analysis. Ohsumed [123] 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) [124].

4.4.3 Baseline Models

To evaluate SD-MoBERT, we benchmark it against five close variants: BERT [100] and MoBERT [119] without smoothed Dirichlet, SD-BERT (smoothed Dirichlet + BERT) [100], SD-RoBERTa

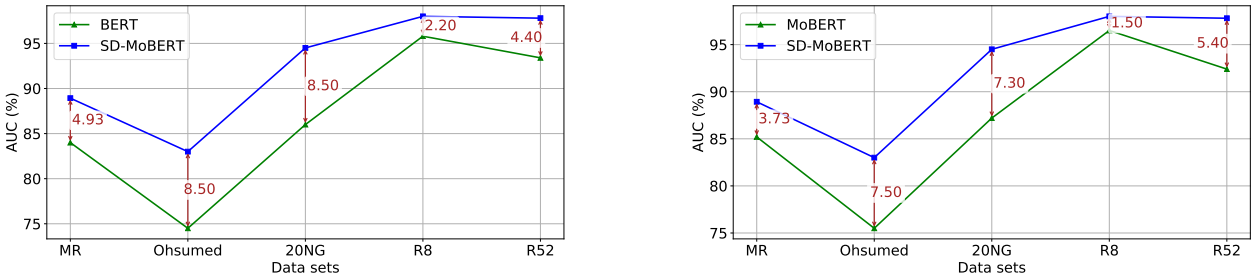


Figure 4.2: Analyses of the area under curve (AUC) of SD-MoBERT against BERT and MoBERT, $K = 100$, $\beta = 0.2$.

(smoothed Dirichlet + RoBERTa-base) [125], and SD-DistilBERT (smoothed Dirichlet + DistilBERT) [126], as well as a number of topic and graph-augmented models. These include TopicBERT-64/128 [127], TextING [128], HyperGAT [129], TextFCG [130], TextSSL [131], BertGCN [132], GTC [133], MHGAT [134], and PaSIG-S [135], providing a comprehensive backdrop for assessing the gains afforded by smoothed Dirichlet fusion in modern transformers.

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

The AUC [99] 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 [136]

$$AUC \approx \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} \tag{4.8}$$

Figure 4.2a 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 4.2b 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.

Table 4.1: 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	$\mathbf{88.97 \pm 0.02}$	$\mathbf{88.13 \pm 0.05}$	$\mathbf{83.49 \pm 0.04}$	$\mathbf{80.00 \pm 0.21}$	$\mathbf{95.27 \pm 0.05}$	$\mathbf{93.11 \pm 0.07}$

Table 4.2: 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	$\mathbf{98.94 \pm 0.07}$	$\mathbf{98.99 \pm 0.03}$	$\mathbf{97.17 \pm 0.07}$

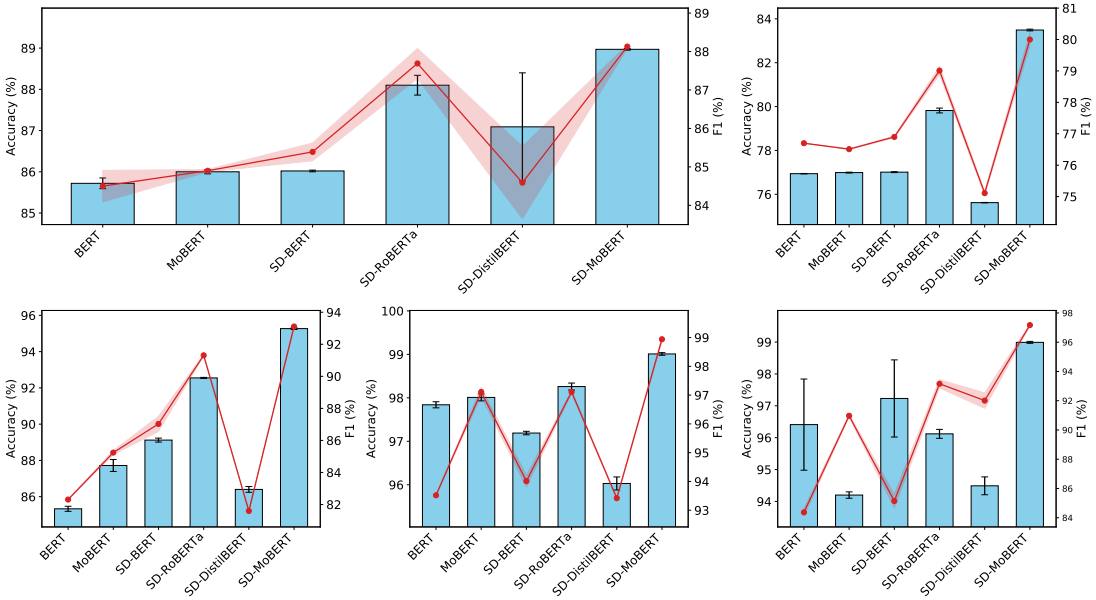


Figure 4.3: Comparison of the classification accuracy and F1 score in six transformer-based models on five text classification benchmarks. The bar plots (sky blue) depict mean test accuracy with the error bars, while the overlaid red lines trace mean F1 scores. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $K = 100$, $\beta = 0.2$.

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

Tables 4.1 and 4.2 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 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 4.4, we evaluate whether the observed accuracy gains of SD-MoBERT over the best baseline (PaSIG-S) are statistically significant. As indicated in Table 4.4, all p-values ($\ll 0.05$), uniformly reject H_0, while the CIs remain vanishingly narrow. See more details in Section 4.5.2.

$$H_0 : \mu_{\text{SD-MoBERT}} = \mu_{\text{PaSIG-S}} \quad \text{vs.} \quad H_1 : \mu_{\text{SD-MoBERT}} \neq \mu_{\text{PaSIG-S}} \tag{4.9}$$

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

Figure 4.3 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

Table 4.3: 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		$\mathbf{2.378e^{-47}}$	$\mathbf{3.782e^{-46}}$	$\mathbf{3.087e^{-23}}$	$\mathbf{1.487e^{-39}}$	$\mathbf{5.488e^{-73}}$

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.

4.5 Effect of Topic Number on Classification Performance

4.5 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

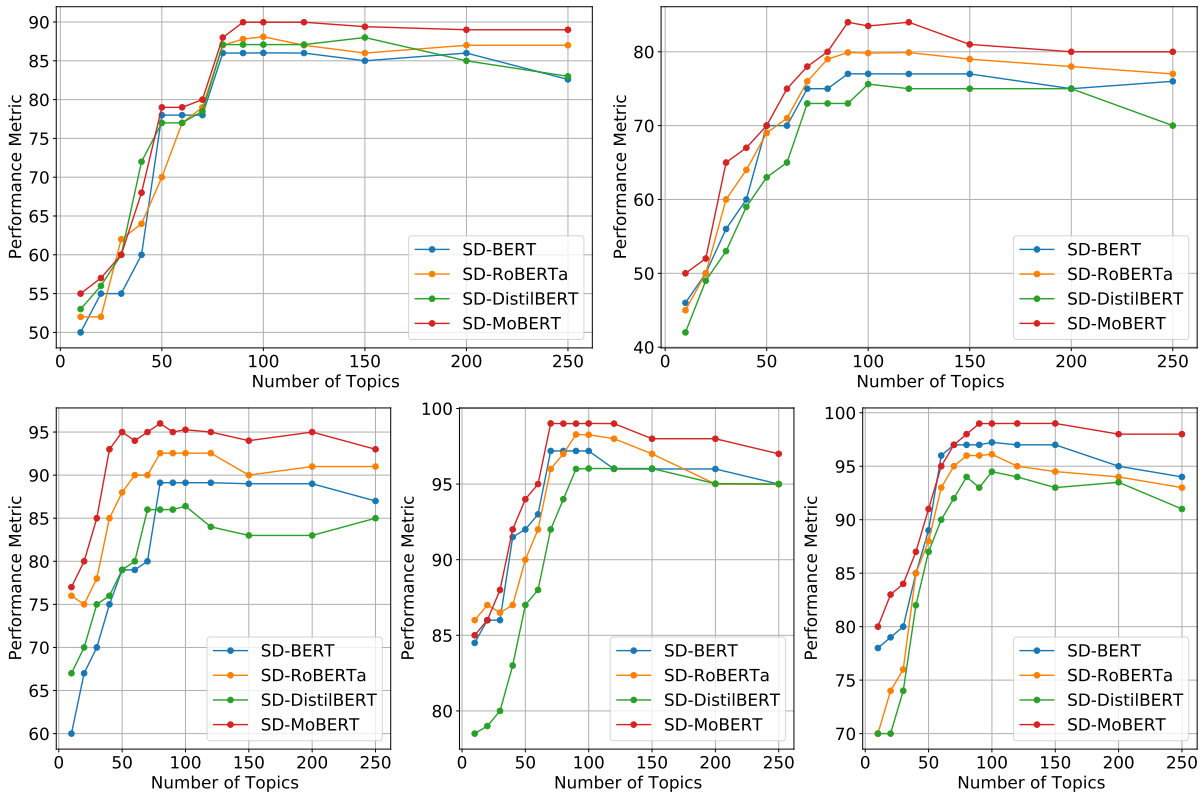


Figure 4.4: Sensitivity of classification accuracy to the number of latent topics on five data sets. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $\beta = 0.2$.

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

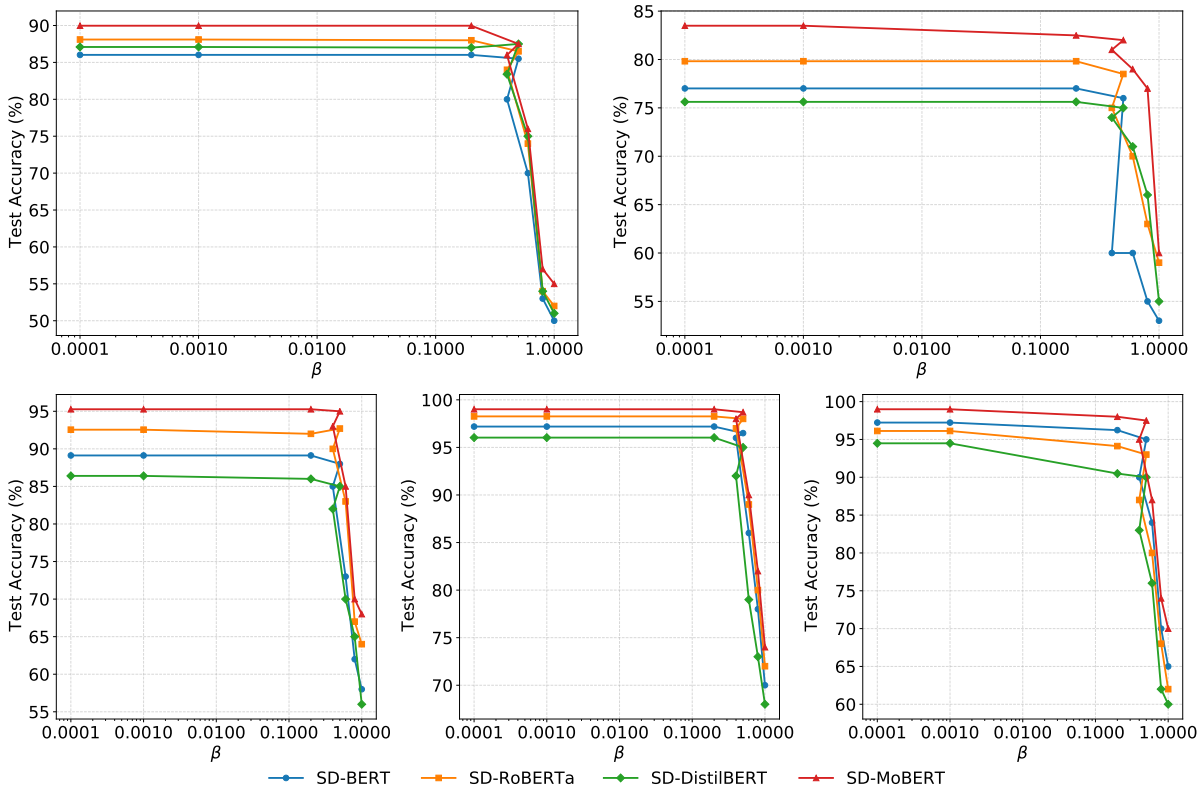


Figure 4.5: Sensitivity of classification accuracy to the regularization weight β across five benchmarks. Each subplot corresponds to a different dataset: MR (top left), Ohsumed (top right), 20NG (bottom left), Reuters R8 (bottom center), and Reuters R52 (bottom right), $K = 100$.

latent dimensionality is sufficiently large (70 to 100 topics), and that beyond this range, additional topics confer minimal benefit across various text classification scenarios.

4.5.1 Effect of the KL-Weight Factor on Classification Performance

Figure 4.5 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 $[0.4, 0.2]$,

Table 4.4: 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		$\mathbf{2.378e^{-47}}$	$\mathbf{3.782e^{-46}}$	$\mathbf{3.087e^{-23}}$	$\mathbf{1.487e^{-39}}$	$\mathbf{5.488e^{-73}}$

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.

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

Table 4.4 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 evaluate whether the observed accuracy gains of SD-MoBERT over the best

baseline (PaSIG-S) are statistically significant. As indicated in Table 4.4, all p-values ($\ll 0.05$), uniformly reject H_0, while the CIs remain vanishingly narrow.

We compute each 95% confidence interval using [137]

$$\text{CI} \;=\; \mu \;\pm\; z^{*}\,\frac{s}{\sqrt{n}}\,, \quad z^{*} = 1.96, \tag{4.10}$$

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

$$H_0 \colon \mu_{\text{SD-MoBERT}} = \mu_{\text{PaSIG-S}} \quad \text{vs.} \quad H_1 \colon \mu_{\text{SD-MoBERT}} \neq \mu_{\text{PaSIG-S}} \tag{4.11}$$

We calculate the two-sided p-value as [137]

$$p = 2\big(1 - \text{CDF}(|t|, df)\big), \quad df = n - 1, \tag{4.12}$$

where

$$\text{CDF}(|t|, df) = \int_{-\infty}^{|t|} f(t, df)\, dt, \tag{4.13}$$

and the Student's t-distribution PDF is

$$f(t, df) = \frac{\Gamma\big(\frac{df+1}{2}\big)}{\sqrt{df\,\pi}\,\Gamma\big(\frac{df}{2}\big)} \Big(1 + \tfrac{t^2}{df}\Big)^{-\frac{df+1}{2}}. \tag{4.14}$$

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.

4.5.3 Efficiency Analysis: Time Complexity and Runtime Cost

Table 4.5 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

Table 4.5: 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

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.

Open Challenges

Despite the effectiveness of generalized and smoothed Dirichlet priors in mitigating component collapse and improving robustness in multimodal fake news detection, important challenges remain unresolved. In particular, existing models assume relatively restrictive covariance structures that limit their ability to express *asymmetric, bounded, and conflicting dependencies* between modalities. In real-world multimodal data, textual and visual cues may convey partially contradictory or unevenly weighted evidence, which cannot be adequately captured by Dirichlet-based priors that enforce predominantly negative correlations.

Furthermore, while continuous latent representations improve stability and uncertainty modeling, current approaches remain constrained in their ability to flexibly adapt covariance structure without sacrificing interpretability or inducing collapse. These limitations motivate the need for richer probabilistic priors capable of modeling complex multimodal interactions in a bounded latent space, thereby enabling more expressive and robust fusion mechanisms.

4.6 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.

Chapter 5

DeepBetaL: Deep Learning for Multimodal Fake News Detection with Beta-Liouville Priors

The pervasive threat of fake news has made its detection a pressing concern. Existing multimodal fake news detection methods often rely on discrete latent representations from diverse modalities, fusing them into a unified framework. However, these approaches face challenges when modalities convey conflicting interpretations. Moreover, discrete latent variables can oversimplify the relationships between modalities, making it difficult to address nuanced conflicts or dependencies. Recent studies using Dirichlet and Gaussian distributions address multimodal asymmetry and imbalance by probabilistically learning the uncertainties attributed to the modalities. However, Dirichlet priors assume negative correlations among variables, limiting their practical applicability, whereas Gaussian models are susceptible to component collapse, compromising both generalization and robustness. To mitigate these challenges, we propose a novel probabilistic framework leveraging a Beta-Liouville prior, which offers a more flexible covariance structure and effectively prevents component collapse. Our extensive empirical evaluations and statistical analyses demonstrate the effectiveness and robustness of this approach.

5.1 Introduction

The widespread spread of fake news has become a critical societal issue, deeply impacting public opinion, political stability, and community trust. This raises significant concerns about the reliability and trustworthiness of modern media platforms [138]. While traditional text-based methods are effective in certain contexts, they fail to capture the complex interplay between multimodal data such as text, images, videos, and their associated contents. This has driven the development of multimodal detection approaches to improve accuracy and robustness. Existing multimodal fake news detection methods often rely on discrete latent representations extracted from each modality, which are then fused into a unified representation for classification [139, 140, 141]. Notable studies, such as [68, 142] extract latent features from image-text pairs using deep learning models and apply fusion strategies to classify news articles. Another approach in [143] takes a multi-view learning approach to integrate discrete latent features from various modalities.

While these methods have shown promise, they face significant challenges when the modalities provide conflicting or contradictory information, as discrete fusion strategies may oversimplify the relationships between modalities and fail to reconcile the contradictions effectively [144]. Probabilistic approaches using Dirichlet [96, 145, 146] and Gaussian [77, 147, 148] distributions have improved uncertainty modelling. But Dirichlet's assumption of negative correlations and Gaussian's risk of component collapse limit their generalization and robustness [79] to handle the asymmetry and imbalance inherent in multimodal data.

To address these challenges, we propose DeepBetaL, a novel probabilistic model leveraging the Beta-Liouville prior. Unlike Dirichlet priors, Beta-Liouville provides a versatile covariance structure [21], enabling nuanced modelling of inter-modal relationships. DeepBetaL captures intricate dependencies between modalities through continuous latent representations, effectively resolving conflicts and enhancing robustness. Additionally, the Beta-Liouville distribution's ability to represent sparse and skewed proportions makes it well-suited for identifying subtle patterns and irregularities in fake news content. Extensive experiments demonstrate that our approach surpasses baseline models on two benchmark data sets. The main contributions of our studies are summarized as follows:

1. We introduce DeepBetaL, a novel model leveraging the Beta-Liouville prior to better capture inter-modal dependencies in multimodal fake news detection, overcoming limitations of Dirichlet and Gaussian distributions.

2. We derive a closed-form expression for the Beta-Liouville Kullback-Leibler (KL) divergence, enhancing computational efficiency and stability during optimization.

3. An efficient reparameterization is proposed to accelerate backpropagation, improving training performance and convergence speed.

5.1.1 Motivation

Unimodal methods, while foundational, are limited by their focus on a single data type, making them inadequate for detecting complex fake news. Multimodal methods, though more comprehensive, often struggle with conflicting modalities, asymmetry, and inefficient fusion mechanisms. Probabilistic models like Dirichlet and Gaussian distributions further face issues such as restrictive priors or component collapse, limiting their robustness and generalization.

Our model addresses these gaps by leveraging a Beta-Liouville distribution, which provides a more versatile covariance structure [21]. Unlike Dirichlet, Beta-Liouville captures both positive and negative correlations by sampling latent representations with a Beta-Liouville prior, improving generalization and effectively handling asymmetries and uncertainties arising from conflicting modalities. This represents the first application of Beta-Liouville priors with deep learning for multimodal fake news detection.

5.2 Background Studies

5.2.1 Intuition Behind Beta-Liouville Distribution

The Beta-Liouville distribution generalizes the Beta distribution, facilitating flexible modelling of multivariate data with variable interdependencies [21]. Unlike Dirichlet priors, which restrict variables to negative correlations, the Beta-Liouville distribution supports general covariance structures, accommodating both positive and negative correlations [149]. This flexibility is ideal for multimodal fake news detection, where complex dependencies between modalities exist. Hierarchically constructed, it begins with a base Beta distribution governing proportions across dimensions, extended via a Liouville transformation into higher dimensions for richer representations. The probability density function (PDF) of the Beta-Liouville distribution [21], defined by parameters $\xi = (\alpha_1, \alpha_2, \ldots, \alpha_D, \alpha, \beta)$, is given as:

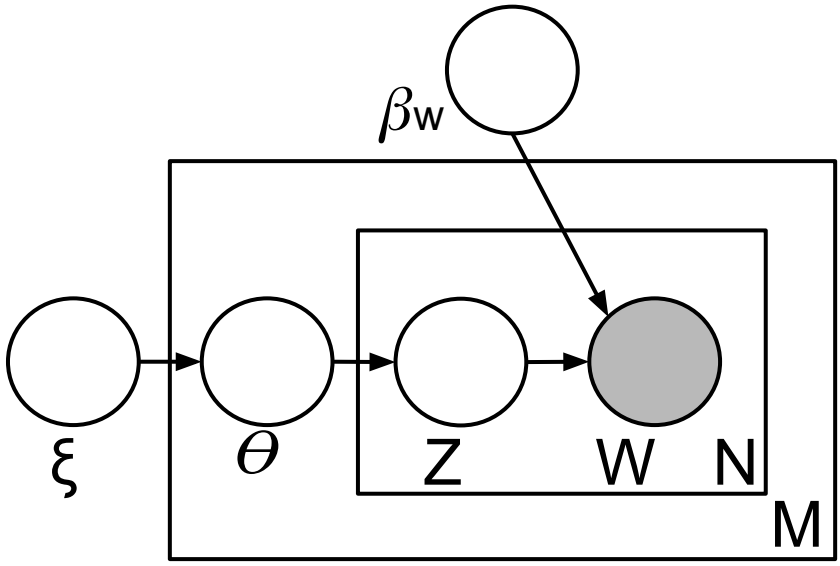


Figure 5.1: Graphical representation of the Beta-Liouville model. The shaded circles represent observed nodes, while the unshaded circles denote hidden nodes.

$$P(\theta|\xi) = \frac{\Gamma(\sum_{d=1}^{D}\alpha_D)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{d=1}^{D}\frac{\theta_d^{\alpha_d-1}}{\Gamma(\alpha_d)} \times \left(\sum_{d=1}^{D}\theta_d\right)^{\alpha-\sum_{l=1}^{D}\alpha_l} \times \left(1-\sum_{d=1}^{D}\theta_l\right)^{\beta-1} \tag{5.1}$$

where $\theta = [\theta_1, \theta_2, \ldots, \theta_D]$, and $\theta_D = 1 - \Sigma_{i=i}^{D-1}\theta_i$. Note that $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_D]$ are the shape parameters for each dimension, while the scaler value, $\beta > 0$, is an additional shape parameter that controls the distribution over the remaining simplex, $\Gamma(\cdot)$ is the Gamma function.

5.2.2 Beta-Liouville Distribution: Graphical Representation and Fake News Detection Formulation

The Beta-Liouville distribution, illustrated in Figure 5.1, extends the Dirichlet distribution by replacing its prior $\boldsymbol{\alpha}$ with $\xi = [\boldsymbol{\alpha}, \beta]$. In the plate diagram, shaded and unshaded nodes represent observed and latent variables, respectively. In this hierarchical model, each of M articles comprises N features (e.g., text, images), where θ_m denotes latent proportions of fake news indicators, and $z_{m,n}$ represents latent assignments for feature n. The proportions are sampled as $\theta_m \sim$ Beta-Liouville(ξ), with $\xi = [\boldsymbol{\alpha}, \beta]$, where $\boldsymbol{\alpha}$ and $\beta > 0$ shape the simplex.

Observed features $w_{m,n}$ are sampled from their latent assignments $z_{m,n}$ via:

$$w_{m,n} \sim p(w|z_{m,n}, \beta_w) \tag{5.2}$$

where β_w governs conditional likelihoods. The joint probability of features W, latent variables Z, and proportions θ is:

$$p(W, Z, \theta|\xi, \beta_w) = \prod_{m=1}^{M} p(\theta_m|\xi) \prod_{n=1}^{N} p(z_{m,n}|\theta_m) p(w_{m,n}|z_{m,n}, \beta_w) \tag{5.3}$$

Marginalizing over latent variables, the probability for article m is:

$$p(W_m|\xi, \beta_w) = \int_\theta \left(\prod_{n=1}^{N} \sum_{z_n} p(w_{m,n}|z_n, \beta_w) p(z_n|\theta) \right) p(\theta|\xi) d\theta \tag{5.4}$$

Extending to M articles, the marginal probability can be defined as:

$$p(\mathcal{D}|\xi, \beta_w) = \prod_{m=1}^{M} \int_\theta \left(\prod_{n=1}^{N} \sum_{z_{m,n}} p(w_{m,n}|z_{m,n}, \beta_w) p(z_{m,n}|\theta_m) \right) p(\theta_m|\xi) d\theta_m \tag{5.5}$$

This framework effectively uncovers hidden relationships between multimodal indicators of fake news by leveraging Beta-Liouville's capacity to model sparse and skewed data distributions.

5.3 Propose Model

5.3.1 Preprocessing and Features Encoding

Figure 5.2 illustrates the architecture, combining advanced techniques for processing textual and visual data. During the text preprocessing, we clean the text by removing stop words, punctuation, spaces, numbers, and symbols, converting the text to lowercase for consistency, and applying lemmatization to reduce words to their base forms. After preprocessing, the texts are tokenized using a BERT tokenizer [85], which converts tokens into embeddings to capture their semantic meaning. BERT employs a 30,000-token vocabulary and WordPiece embeddings, breaking down words into smaller tokens to handle out-of-vocabulary words. The encoder's embedding layer is the sum of three components: token embeddings (WN), segment embeddings (SN) that differentiate paired input sequences, and positional embeddings (PN) that indicate word positions in the sequence. The input representation for BERT's encoder is the sum of these three embeddings, expressed as $EM = WN + SN + PN$. The composite embedding (1, n, 768) is passed to BERT's encoder, with [CLS] and [SEP] marking the start and end. The encoder, consisting of 12 layers (L_1 to L_{12}), extracts contextual features. The words $\{E_1, ..., E_N\}$ are transformed into queries (Q_m), keys (K_m), and values (V_m) for each mth attention head. The multi-head self-attention computes the attention, $T_m = A_m \times V_m$, with A_m computed as:

$$A_m = \text{SoftMax}\left(\frac{Q_m \times K_m^T}{\sqrt{d_k}} \right) \tag{5.6}$$

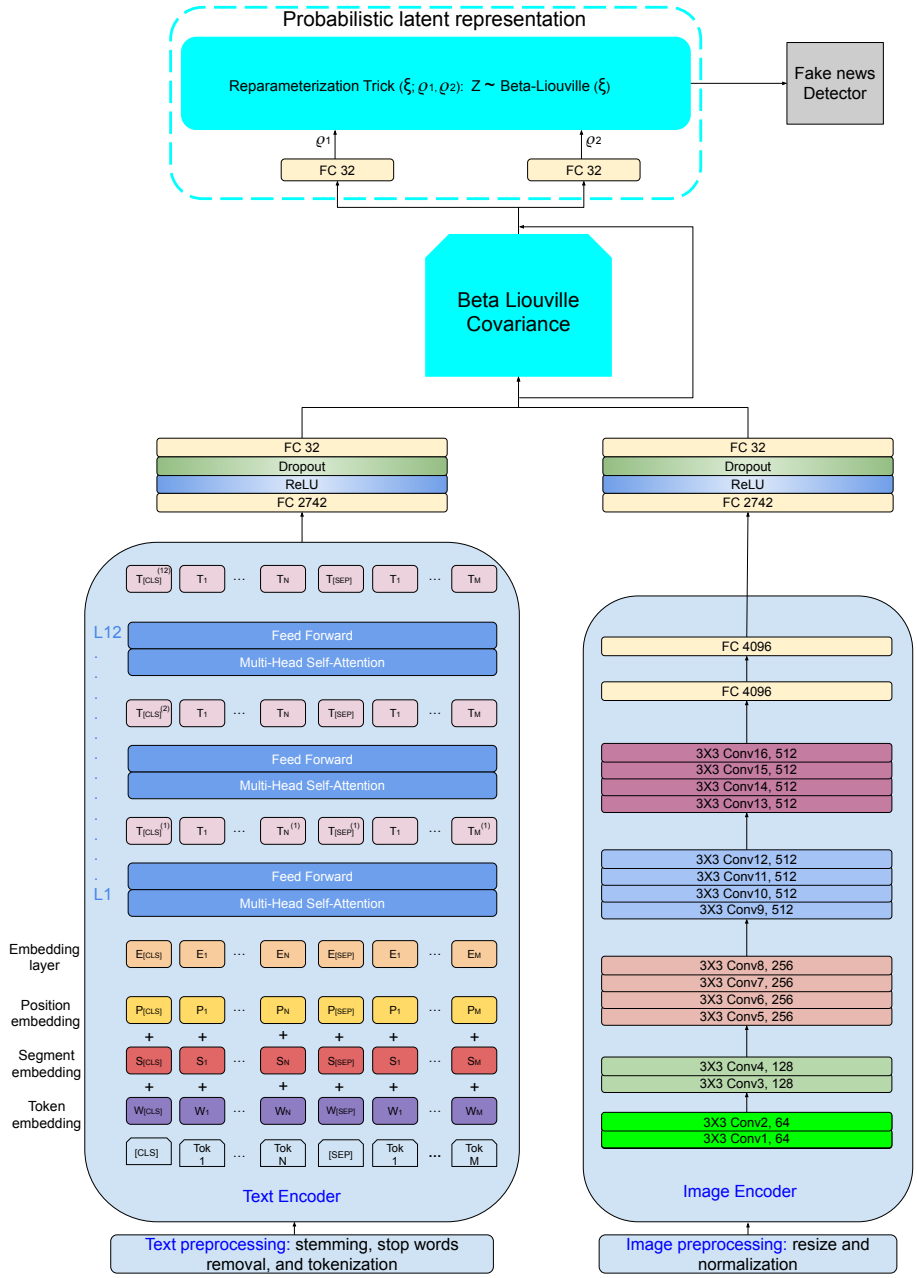


Figure 5.2: A schematic representation of the proposed DeepBetaL model, where the text encoder is BERT and the image encoder utilizes VGG19.

A_m denotes the attention distribution for the m-th head, and $\sqrt{d_k}$ is a scaling factor.

Furthermore, the images play a crucial role in detecting fake news. We first resize them to 224×224 pixels and standardize them using the mean and standard deviation values $\{0.485, 0.456, 0.406\}$ and $\{0.229, 0.224, 0.225\}$, respectively. The image encoder, based on [150], consists of 16 convolutional layers with 3×3 kernels, capturing complex visual features, followed by two fully connected

layers of 4096 units each to transform and consolidate the image features for effective fake news detection.

Next, the features of text and image encoders are independently propagated through a fully connected layer with ReLU activation and dropout to mitigate overfitting, followed by further refinement through another fully connected layer. The Beta-Liouville covariance component merges these modalities, offering a more general structure than Dirichlet methods to capture both positive and negative correlations, effectively representing intricate interdependencies in multimodal data. The Beta-Liouville component's output serves as a prior for sampling latent representations, supported by a residual network to prevent information loss. Unlike discrete latent representations, our continuous latent approach captures uncertainties and nuanced multimodal patterns, balancing unique and shared characteristics while providing a robust foundation for the classification task.

5.3.2 Beta-Liouville Reparameterization, Generative Process, and Loss Functions

Optimizing the Beta-Liouville distribution with variational multimodal encoders is challenging due to the lack of a straightforward reparameterization mechanism, unlike Gaussian distributions, which leverage $z = \mu + \epsilon\sigma : \epsilon \sim N(0, 1)$. Previous studies, such as [14, 79], have explored reparameterization tricks using Gamma distributions and efficient rejection sampling. Interestingly, the Beta-Liouville distribution for a k-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_k)$ is parameterized by the shape parameter $\boldsymbol{\alpha}$ and the scaling factor β. We can derive an intuitive understanding from the Beta-Liouville distribution's probability density function as follows:

$$p(\mathbf{X}) \propto \prod_{i=1}^{k} X_i^{\alpha_i - 1}(1 - \|\mathbf{X}\|)^{\beta_i - 1}, \quad \text{where } \|\mathbf{X}\| = \sum_{i=1}^{k} X_i. \tag{5.7}$$

with $X_i^{\alpha_i - 1}$ scaling each component and $(1 - \|\mathbf{X}\|)^{\beta_i - 1}$ ensuring that the components sum to less than or equal to one.

We denote $\rho_1, \rho_2 \in \mathbb{R}$ as the transformations from the Beta-Liouville covariance in Figure 5.2 and sampled from the distribution through a three-stage process. First, $\rho \sim \mathcal{N}(\rho_1, \rho_2^2)$ are generated using the reparameterization formula, $\rho = \rho_1 + \rho_2 \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. The samples are then transformed using Laplace approximation [1], $Y_i = \exp(\rho_i)/\sum_{j=1}^{k} \exp(\rho_j)$. Finally, we map the samples to the Beta-Liouville space by enforcing a hierarchical structure to ensure the generated

Algorithm 3 Generative process for Beta-Liouville-based multimodal model for detecting fake news.

Data:
  $\mathcal{M}$: Textual features
  $\mathcal{I}$: Visual features
Result:
  ϕ, θ: Model parameters
  Initialize ϕ, θ
while Optimizer not converged **do**
  $X^{\mathcal{M}} \leftarrow f_\phi(\mathcal{M})$ ▷ Extract text features
  $X^{\mathcal{I}} \leftarrow f_\theta(\mathcal{I})$ ▷ Extract image features
  $\xi \leftarrow \text{Covariance}(X^{\mathcal{M}}, X^{\mathcal{I}})$ ▷ Interdependence
  $\vec{\rho}_1, \vec{\rho}_2 \leftarrow f_{\phi,\theta}(\xi)$ ▷ Latent parameters
  $\vec{\rho} \leftarrow \vec{\rho}_1 + \vec{\rho}_2 \cdot \vec{\varepsilon}$ ▷ $\vec{\varepsilon} \sim \mathcal{N}(0, 1)$
  $\vec{Y} \leftarrow \vec{\rho}$: $Y_i = \frac{\exp(\rho_i)}{\sum_{j=1}^{k} \exp(\rho_j)}$
  $\vec{Z} \leftarrow Y_i \cdot \left(1 - \sum_{j=1}^{i-1} X_j\right)$ ▷ Beta-Liouville
  Predict label $y \in \{0, 1\}$ for input
  Compute $\bar{\mathcal{L}}(\phi, \theta; y, X)$ ▷ Compute loss
  Update (ϕ, θ) ▷ Gradient-based update
end while

samples adhere to the constraints of the Beta-Liouville distribution.

$$Z_i = Y_i \cdot (1 - \sum_{j=1}^{i-1} X_j) \tag{5.8}$$

Algorithm 3 outlined the pseudocode for the generative process.

The loss function combines binary cross-entropy (BCE) for prediction accuracy and Kullback-Leibler (KL) divergence for regularizing the posterior-prior difference. Please note that using an analytical KL ensures computational efficiency, stable gradients, and streamlined optimization [151]. For N samples, BCE is defined as:

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.9}$$

Additionally, we formulate the KL-divergence between P and Q Beta-Liouville distributions as

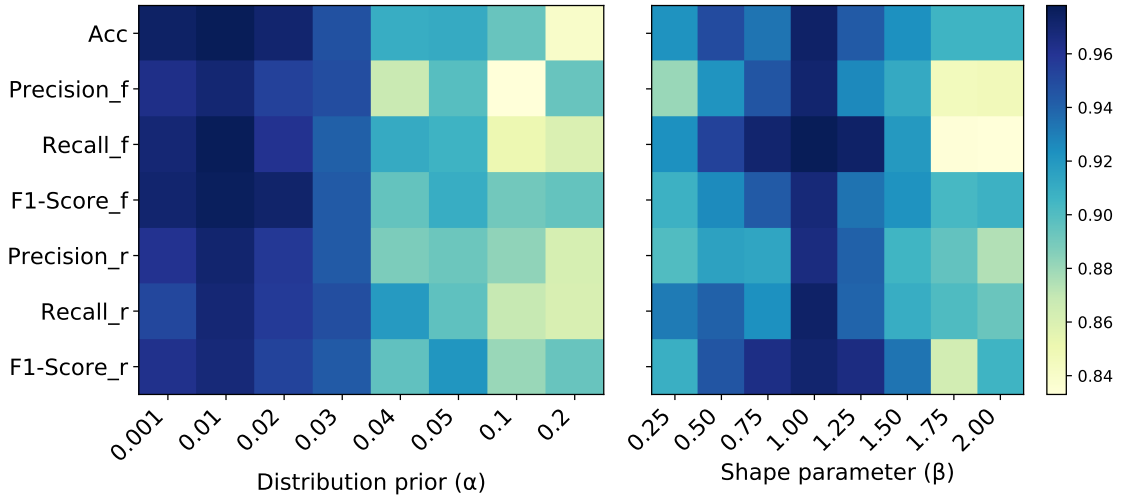


Figure 5.3: A summary of heatmaps from a grid search sshowing (a) the effect of the distribution prior α, and (b) the effect of the scaling parameter β. The metrics include: Acc (overall accuracy), Precision_f (precision for fake news), Recall_f (recall for fake news), F1-Score_f (F1-score for fake news), Precision_r (precision for real news), Recall_r (recall for real news), and F1-Score_r (F1-score for real news).

shown below. For our comprehensive derivation of the KL divergence between two Beta-Liouville distributions, refer to Appendix 1.

$$\mathrm{KL}(P\|Q) = \log\Gamma\left(\sum_{d=1}^{D}\alpha_d\right) - \log\Gamma\left(\sum_{d=1}^{D}\alpha'_d\right) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha'+\beta') - \big[\log\Gamma(\alpha) - \log\Gamma(\alpha')\big]$$
$$- \big[\log\Gamma(\beta) - \log\Gamma(\beta')\big] - \sum_{d=1}^{D}\big[\log\Gamma(\alpha_d) - \log\Gamma(\alpha'_d)\big] + \sum_{d=1}^{D}(\alpha_d - \alpha'_d)\big[\psi(\alpha_d) - \psi(\sum_{d=1}^{D}\alpha_d + \beta)\big] + (\alpha - \alpha')$$
$$\big[\psi(\alpha) - \psi(\alpha+\beta)\big] + (\beta - \beta')\big[\psi(\beta) - \psi(\alpha+\beta)\big] \tag{5.10}$$

5.4 Experimental Results

We compare the performance of DeepBetaL with the baseline models on two publicly available benchmark datasets, MediaEval [92] and Weibo [93]. The Twitter dataset (2016) contains 17,000 tweets with text, images, and social context, split into 9,000 fake news and 6,000 real news tweets, plus a 2,000-tweet test set. The Weibo dataset includes 3,615 fake and 4,105 real news posts from Weibo, verified by a rumour-debunking system and Xinhua News Agency, spanning 2012-2016. It's

Table 5.1: Statistical evaluation of performance metrics on the Twitter and Weibo datasets for 30 runs by randomly changing the test set. S and CI denote the sample standard deviation and confidence interval, respectively. We set $\alpha = 0.01$, $\beta = 1.0$, learning rate $= 3e^{-5}$.

	Metrics	Mean	Variance	S	CI
Twitter	Accuracy	0.978	$1.68e^{-5}$	$4.11e^{-3}$	[0.977, 0.979]
	Precision	0.971	$2.75e^{-5}$	$5.25e^{-3}$	[0.969, 0.973]
	Recall	0.965	$3.05e^{-5}$	$5.53e^{-3}$	[0.963, 0.967]
	F1-Score	0.967	$3.78e^{-5}$	$6.15e^{-3}$	[0.965, 0.969]
Weibo	Accuracy	0.961	$2.75e^{-5}$	$5.24e^{-3}$	[0.959, 0.963]
	Precision	0.957	$2.28e^{-5}$	$4.77e^{-3}$	[0.955, 0.959]
	Recall	0.968	$6.17e^{-5}$	$7.85e^{-3}$	[0.965, 0.971]
	F1-Score	0.958	$3.45e^{-5}$	$5.87e^{-3}$	[0.956, 0.960]

multimodal, with text, images, and social context in simplified Chinese. We compare DeepBetaL with the following baseline models: EANBS [152], Event-radar [143], MPFN [94], FCINet [95], Att-RNN [67], EANN [38], Dirichlet [96], MVAE [77], SpotFake [68], VAEMTL [69], and BMR [97]. It's worth noting that the values we have used are directly taken from the referenced studies. Consequently, any metrics marked as "NA" in our table indicate that those specific values were not reported in the original studies.

For all of our experimental results, we conducted 30 test iterations, each with randomly selected test sets, and reported the average values. Initially, we performed a series of experiments using grid search to explore the hyperparameter space. Figure 5.3 presents heatmaps on the Twitter data set illustrating the effect of some selected distribution priors (α) and scaling factors (β). The results indicate that, due to concentrated priors, performance is notably higher with lower prior values than with higher priors. For β, performance improves from smaller values to moderate values and starts to decline for higher values due to excessive spread in the predictive distribution. Nevertheless, the

Table 5.2: Performance of DeepBetaL vs baseline models on Twitter and Weibo datasets. We set $\alpha = 0.01$, $\beta = 1.0$, epoch = 50, learning rate = $3e^{-5}$, and optimizer = Adam.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Fake News			Real News		
						Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter	EANBS	0.860	NA	NA	NA	0.850	0.880	0.860	0.880	0.840	0.860
	Event-rader	0.928	NA	NA	0.923	0.904	0.902	0.903	0.942	0.943	0.943
	MPFN	0.833	NA	NA	NA	0.846	0.921	0.880	0.809	0.721	0.740
	FCINet	0.908	0.913	0.909	0.910	0.828	0.913	0.868	0.955	0.907	0.930
	att-RNN	0.664	0.749	0.615	0.676	0.749	0.615	0.676	0.589	0.728	0.651
	EANN-	0.648	0.810	0.498	0.617	0.810	0.498	0.617	0.584	0.759	0.660
	EANN	0.715	0.822	0.638	0.719	NA	NA	NA	NA	NA	NA
	Dirichlet	0.824	NA	NA	NA	0.772	0.918	0.838	0.899	0.730	0.806
	MVAE	0.745	NA	NA	NA	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.778	NA	NA	NA	0.751	0.900	0.820	0.832	0.606	0.701
	VAEMTL_AV	0.869	NA	NA	NA	0.820	0.784	0.802	0.880	0.917	0.898
	VAEMTL_IM	0.871	NA	NA	NA	0.826	0.772	0.798	0.891	0.920	0.905
	VAEMTL_DY	0.888	NA	NA	NA	0.838	0.821	0.829	0.912	0.22	0.917
	BMR	0.883	NA	NA	0.870	0.927	0.746	0.827	0.865	0.965	0.912
	DeepBetaL	**0.978***	**0.971***	**0.965***	**0.967***	**0.970***	**0.977***	**0.976***	**0.971***	**0.969***	**0.968***
Weibo	EANBS	0.890	NA	NA	NA	0.870	0.910	0.890	0.900	0.880	0.890
	Event-rader	0.919	NA	NA	0.919	0.932	0.915	0.924	0.924	0.905	0.914
	MPFN	0.838	NA	NA	NA	0.857	0.894	0.889	0.873	0.863	0.876
	FCINet	0.926	0.926	0.926	0.926	0.938	0.917	0.927	0.913	0.935	0.924
	att-RNN	0.772	0.778	0.799	0.789	0.797	0.713	0.692	0.684	0.840	0.754
	EANN-	0.795	0.806	0.795	0.800	0.827	0.697	0.756	0.752	0.863	0.804
	EANN	0.827	0.847	0.812	0.829	NA	NA	NA	NA	NA	NA
	Dirichlet	0.888	NA	NA	NA	0.900	0.872	0.886	0.877	0.904	0.890
	MVAE	0.824	NA	NA	NA	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.8923	NA	NA	NA	0.902	0.964	0.932	0.847	0.656	0.739
	VAEMTL_AV	0.905	NA	NA	NA	0.892	0.921	0.906	0.918	0.891	0.904
	VAEMTL_IM	0.910	NA	NA	NA	0.902	0.927	0.914	0.920	0.893	0.906
	VAEMTL_DY	0.921	NA	NA	NA	0.910	0.940	0.924	0.934	0.901	0.917
	BMR	0.889	NA	NA	0.889	0.904	0.885	0.895	0.874	0.894	0.884
	DeepBetaL	**0.961***	**0.957***	**0.968***	**0.958***	**0.959***	**0.965**	**0.956***	**0.954***	**0.961***	**0.959***

performance remains relatively competitive across the heatmap, demonstrating that our approach is robust and less sensitive to variations in the distribution parameters. This robustness minimizes the need for extensive parameter adjustments to achieve optimal performance. For subsequent experiments across the data sets, we set the learning rate, prior α, β, optimizer, batch size, and epoch to {$3e^{-5}$, 0.01, 1.0, Adam, 8, and 100}, respectively.

We conduct a statistical evaluation of the model's performance over 30 runs with randomly

selected test sets. Table 5.1 summarizes the results on the Twitter and Weibo datasets, reporting overall accuracy, precision, recall, and F1-score. Both datasets exhibit minimal variance and narrow confidence intervals, reflecting strong and consistent performance across metrics. For instance, the model achieved accuracies of 0.978 on Twitter and 0.961 on Weibo, with low variances of $1.68e^{-5}$ and $2.75e^{-5}$, respectively, indicating high stability and reliability.

The results presented in Table 5.2 provide a detailed comparison of DeepBetaL and the baseline models on the Twitter and Weibo data sets. Please note that "NA" indicates that those specific values were not reported in the original studies. The DeepBetaL consistently outperformed others, attaining the highest accuracy, precision, recall, and F1-scores across both data sets. Notably, it achieved accuracy improvements of 5.39% and 7.71% over Event-rader and FCINnet for the Twitter data set, respectively. For the Weibo data set, it achieved accuracy improvements of 3.78% and 4.34% over FCINnet and VAEMTL_DY, respectively. Other notable models, including VAEMTL_(AV, IM, DY), Dirichlet, Event-rader, FCINet, SpotFake, BMR, and Gaussian-based MVAE, performed well but fell short of DeepBetaL. DeepBetaL's strong performance stems from its ability to identify intricate patterns within a continuous latent space and its general covariance structure, allowing it to uncover and represent complex hidden relationships in the data.

To validate the performance of DeepBetaL against the model with the nearest performance metric-wise, we define the following one-tailed hypotheses for each metric.

1. H_0 : No significant difference in performance between DeepBetaL and the next-best model
2. H_1 : There is a significant difference in performance between DeepBetaL and the next-best model

First, we compute the P-values, $p = 2 \cdot (1 - \text{CDF}(|t|, df))$, where $\text{CDF}(|t|, df)$ denotes the cumulative distribution function of the t-distribution with degrees of freedom $df = n - 1$ and a 95% confidence interval. For instance, the p-values calculated for accuracy on the Twitter and Weibo datasets, with Event-rader and FCINet as the respective next best models, are $2.67e^{-12}$ and $4.99e^{-9}$. The p-values for all metrics in Table 5.2 indicated with an asterisk are less than the significance level of 0.05. Consequently, apart from the recall metric for the Weibo dataset, we reject the null hypotheses. This indicates that DeepBetaL achieves statistically significant performance over other models across the majority of metrics where direct comparisons are feasible.

We evaluate the efficiency of our model by comparing its execution time against state-of-the-art models on a 12th Gen Intel Core i7-12700K system with 64GB RAM. For the Twitter dataset,

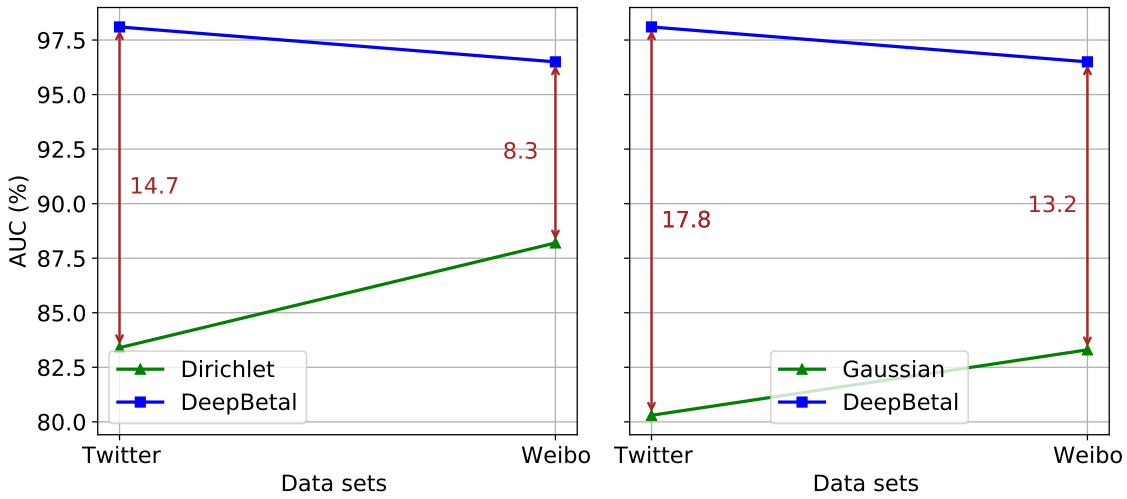


Figure 5.4: Ablation Study of the Probability Component: (a) Beta-Liouville Vs. Dirichlet, (b) Beta-Liouville Vs. Gaussian

DeepBetaL achieved a per-sample execution time of 365.89 ms, outperforming MVAE (482.39 ms) and VAEMTL_DY (477.92 ms) while being slightly slower than SpotFake (365.44 ms). On the Weibo dataset, DeepBetaL (370.21 ms) was comparable to SpotFake (357.12 ms) and faster than MVAE (476.09 ms) and VAEMTL_DY (463 ms). The competitive efficiency of DeepBetaL stems from its architecture, which excludes a decoder module. The time complexity of DeepBetaL is $O(T \cdot [n \cdot L \cdot d_t + m \cdot h \cdot w \cdot d_i + n \cdot m \cdot d_t \cdot d_i + p^2])$. Where p denotes the latent dimension, n is the number of text samples, L is the maximum sequence length, d_t is the hidden size, m is the number of image samples, h and w are the input image dimensions, and d_i is the number of image channels. Please refer to Appendix 2 for details.

5.4.1 Ablation Study of the Probability Component

We evaluate the effectiveness of Beta-Liouville by replacing the covariance and parameterization in Figure 5.2 with the Dirichlet and Gaussian-based MVAE approaches discussed in [96] and [77], respectively, utilizing their corresponding KL-divergence formulations. Figure 5.4 reveals significant AUC performance gaps between DeepBetaL and the baseline models on the Twitter and Weibo datasets. Specifically, DeepBetaL outperforms Dirichlet by 14.7% on Twitter and 8.3% on Weibo, while surpassing Gaussian by 17.8% on Twitter and 13.2% on Weibo. These results underscore the effectiveness of DeepBetaL in detecting fake news.

5.5 Conclusion

This study introduced DeepBetaL, a novel probabilistic framework for multimodal fake news detection that leverages the Beta-Liouville prior to overcoming the limitations of existing Dirichlet and Gaussian-based methods. By effectively modelling nuanced inter-modal relationships and addressing conflicts, DeepBetaL demonstrated effective detection and robustness. Extensive evaluations across benchmark datasets confirmed its consistent outperformance of state-of-the-art models in accuracy, precision, recall, and F1-score. Furthermore, DeepBetaL achieved competitive efficiency by eliminating the need for decoder modules, making it well-suited for practical applications. These findings establish DeepBetaL as a robust, efficient solution for multimodal fake news detection, with significant potential for broader adoption in combating misinformation across diverse digital platforms. Our findings demonstrate the effectiveness of Beta-Liouville with a deep neural network for fake news detection. However, challenges remain, particularly in selecting an appropriate prior for the distribution. Additionally, current social platforms increasingly incorporate not just text and images but also audio and video. Moving forward, we aim to develop a unified approach that seamlessly considers these diverse modalities with an adaptive prior, eliminating the need for manual prior selection. This advancement will strengthen our model's capability to detect fake news across varied media formats on contemporary social platforms.

Chapter 6

EviDA: Cross-Domain Fake News Detection via Uncertainty-Weighted Domain Adversarial Learning

Cross-domain fake news detection faces significant challenges due to distribution shifts across social media platforms with varying linguistic styles, cultural contexts, and platform-specific characteristics. In addition to cross-platform shifts, real-world deployment often requires cross-lingual detection across different domains, further exacerbating domain discrepancy. While domain adversarial training enables learning domain-invariant representations, existing methods apply uniform alignment across all samples, ignoring the heterogeneous nature of domain shift. We propose a novel uncertainty-weighted domain adversarial framework that leverages evidential deep learning to quantify prediction uncertainty and dynamically modulate the intensity of domain adaptation (EviDA). Our key insight is that epistemic uncertainty provides an informative proxy for instance-level domain discrepancy and can guide instance-level alignment strength, while confident source domain predictions require less aggressive adaptation. We introduce an adaptive weighting mechanism where the scaling parameter is learned end-to-end, allowing the model to discover optimal alignment strategies automatically. Statistical analysis reveals a strong correlation between uncertainty and domain shift, with cross-domain samples exhibiting higher uncertainty. This work provides empirical evidence for uncertainty-weighted domain adaptation, offering a principled approach to robust cross-domain and cross-lingual fake news detection. Our extensive evaluations and statistical analyses demonstrate the effectiveness and robustness of this approach.

6.1 Introduction

Fake news on social media has emerged as a critical challenge for AI-driven content moderation, with severe consequences for democratic processes, public health, and social trust. Traditional fake news detection approaches focus primarily on learning discriminative features from text [153], images [154], or their combinations [155] within a single domain. However, recent studies have revealed that models trained on one platform exhibit accuracy drops when applied to different platforms without adaptation [156].

Recent advances in automated detection, such as [157], demonstrate that large pre-trained models and prompt-based learning can substantially improve in-domain performance. However, these methods often degrade sharply when deployed across platforms with different linguistic conventions, user behaviors, and cultural contexts. The authors formalized the problem as cross-domain fake news detection, where a model trained on one platform (e.g., Twitter) can generalize to unseen platforms such as Facebook or Weibo. Recent studies, including [158] and [159], show that even sophisticated prompt-based and ensemble-based systems struggle under domain shift, especially when retraining on the target domain is infeasible. The challenge is further exacerbated in cross-lingual settings, where distribution shifts arise not only from platform differences but also from language, script, and cultural variation.

Domain adaptation techniques provide a principled route to address these issues. In particular, domain adversarial neural networks [160] and their variants have been widely adopted to learn domain-invariant representations by aligning feature distributions across source and target domains. Recent applications to misinformation detection, such as [161], demonstrate that adversarial alignment can improve cross-domain transferability. However, a fundamental limitation persists: existing methods apply uniform adversarial pressure to all samples, implicitly assuming that all instances experience the same degree of domain shift.

In practice, domain shift is heterogeneous. Some samples closely resemble the source distribution and require minimal alignment, while others lie far from the training manifold and demand stronger adaptation. Uniform alignment risks over-regularizing easy samples and under-aligning difficult ones, ultimately harming both discrimination and generalization. Concurrently, uncertainty quantification has gained increasing attention as a mechanism for identifying out-of-distribution and domain-shifted samples. Evidential deep learning, introduced in [48], models epistemic uncertainty by placing Dirichlet priors over class probabilities, providing a principled measure of model

confidence. Subsequent works such as [162] show that epistemic uncertainty correlates strongly with domain discrepancy. Despite this progress, uncertainty has primarily been used for post-hoc diagnostics, active learning, or sample filtering. To the best of our knowledge, it has not been systematically integrated as an instance-level control signal for domain adversarial training in fake news detection.

To address these challenges, we propose Uncertainty-Weighted Domain Adversarial Learning for cross-domain and cross-lingual fake news detection (EviDA). Our central insight is that epistemic uncertainty serves as a reliable proxy for domain shift: samples with high uncertainty are more likely to originate from unfamiliar target domains and should receive stronger adversarial alignment, whereas confident predictions, typically associated with source-like samples, should be adapted more conservatively to preserve discriminative features. Importantly, in EviDA, epistemic uncertainty is not interpreted as a proxy for error probability; instead, it functions as a control signal for instance-level domain alignment, and the adaptation process itself therefore mediates its relationship with prediction accuracy.

Our empirical analysis on three benchmarks (Twitter, Weibo, and Fakeddit) reveals three key findings: **First**, epistemic uncertainty serves as a reliable indicator of domain shift, with cross-domain samples exhibiting higher uncertainty and a strong correlation with post-adaptation prediction behavior under distribution shift (r=+0.675, $p < 0.001$). **Second**, adaptive uncertainty-weighted domain alignment consistently outperforms unweighted and static alternatives in zero-shot transfer, while exhibiting substantially lower variance across runs, which validates cross-domain generalization. **Third**, *adaptive uncertainty weighting as the primary driver*: our learned uncertainty-weighted approach contributes +9.0 percentage points, the largest single-component gain, achieving 82.6% cross-domain average and reducing the domain gap from 19.3% to 3.3% (82.9% reduction). Unlike prior domain adaptation methods that apply uniform or heuristic instance weights, EviDA learns an evidential epistemic signal and uses it as a differentiable control variable to allocate alignment capacity during adversarial training.

The main contributions of this work are summarized as follows:

1. We introduce an uncertainty-weighted domain adversarial framework that integrates evidential epistemic uncertainty into instance-level domain alignment for cross-domain fake news detection.

2. We propose an adaptive weighting mechanism that learns the strength of uncertainty-guided

alignment end-to-end, avoiding heuristic tuning.

3. We conduct comprehensive ablations and comparisons against alternative uncertainty estimation and domain adaptation strategies.

6.2 Background

6.2.1 Uncertainty Estimation and Evidential Deep Learning

Uncertainty estimation has been widely studied as a mechanism for assessing model reliability under distribution shift. [48] introduced a principled framework for modeling epistemic uncertainty via Dirichlet distributions over class probabilities. This approach was extended in [49] and further analyzed in [50], which demonstrated strong correlations between epistemic uncertainty and out-of-distribution samples.

Subsequent studies, including [51], compared evidential learning against Monte Carlo dropout and deep ensembles, highlighting its advantages in calibration, efficiency, and interpretability. These works establish epistemic uncertainty as a reliable signal of distribution mismatch. While uncertainty has been leveraged for tasks such as active learning and post-hoc domain shift detection [52], its role in guiding the optimization of domain adaptation objectives remains underexplored. In particular, existing fake news detection methods do not integrate uncertainty as an instance-level control signal within adversarial training. Our work addresses this gap by embedding evidential epistemic uncertainty directly into the domain adversarial learning process.

6.2.2 Domain-Adversarial Neural Networks (DANN)

Domain-adversarial learning aims to learn features that are predictive for the main task while being invariant to domain identity. The model consists of a feature extractor G_f, a label predictor G_y, and a domain discriminator G_d. Given an input $\mathbf{x}$ with a class label y (available in the source domain) and a domain label $d \in \{1, \ldots, D\}$, the optimization is defined as a minimax objective:

$$\min_{G_f, G_y} \max_{G_d} \mathcal{L}_{\text{cls}}(G_y(G_f(\mathbf{x})), y) - \lambda_d \mathcal{L}_{\text{domain}}(G_d(G_f(\mathbf{x})), d) \tag{6.1}$$

In practice, this objective is implemented via a Gradient Reversal Layer (GRL), which behaves as the identity in the forward pass and multiplies the gradient by $-\lambda_d$ in the backward pass:

$$\text{GRL}_{\lambda_d}(\mathbf{z}) = \mathbf{z}, \quad \frac{\partial \, \text{GRL}_{\lambda_d}(\mathbf{z})}{\partial \mathbf{z}} = -\lambda_d I \tag{6.2}$$

Standard DANN uses a global λ_d (often scheduled over training) and applies the same alignment strength to all samples, implicitly assuming homogeneous domain shift across samples.

6.2.3 Evidential Deep Learning

Evidential deep learning models predictive uncertainty by placing a Dirichlet distribution over the categorical class probabilities. Following the studies in [48], for K classes, the network outputs non-negative evidence $\mathbf{e} = [e_1, \ldots, e_K]$ (e.g., via ReLU), which defines Dirichlet parameters:

$$\alpha_k = e_k + 1, \quad \alpha_k > 0, \qquad S = \sum_{k=1}^{K} \alpha_k \tag{6.3}$$

$$\text{The predictive mean is: } \hat{p}_k = \mathbb{E}[p_k] = \frac{\alpha_k}{S} \tag{6.4}$$

A commonly used scalar summary of epistemic uncertainty is the inverse total evidence:

$$u = \frac{K}{S}, \tag{6.5}$$

where $u \in (0, 1]$, and larger values indicate lower accumulated evidence. Such uncertainty is often associated with distribution shift or limited support in the training data.

6.2.4 Evidential Training Objective

Evidential training typically combines a data-fit term with a regularizer that discourages unwarranted overconfidence. In our framework, the classification loss $\mathcal{L}_{\text{cls}}$ is instantiated via the evidential formulation:

$$\mathcal{L}_{\text{evi}} = \mathcal{L}_{\text{cls}}(\boldsymbol{\alpha}, \mathbf{y}) + \lambda_{\text{KL}} \, \text{KL}(\text{Dir}(\boldsymbol{\alpha}) \parallel \text{Dir}(\mathbf{1})) \,, \tag{6.6}$$

where $\mathbf{y}$ is the one-hot label and $\text{Dir}(\mathbf{1})$ denotes a uniform Dirichlet prior. The KL regularization term constrains the total evidence, reducing overconfident predictions on ambiguous or domain-shifted samples.

6.3 Proposed Model

6.3.1 Architecture Overview

Our framework consists of four main components: multimodal feature extraction, evidential classification with epistemic uncertainty estimation, domain-specific batch normalization, and uncertainty-weighted domain-adversarial learning. Figure 6.1 illustrates the overall architecture.

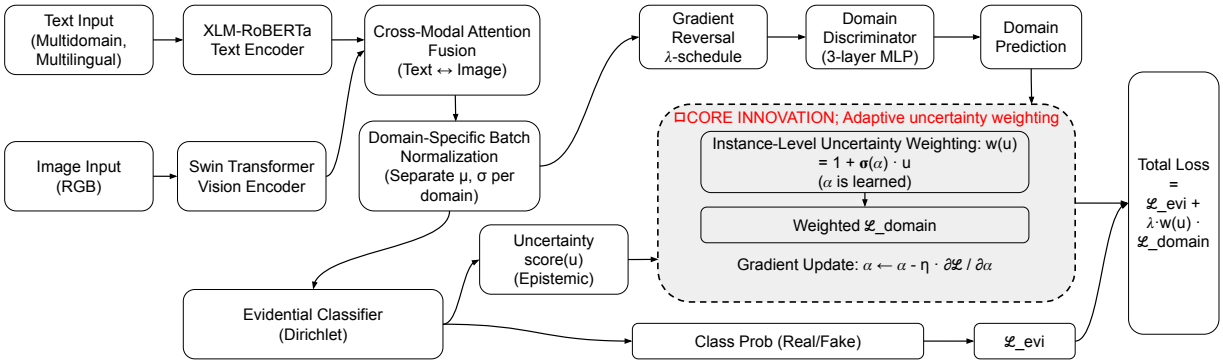


Figure 6.1: Overview of the proposed uncertainty-weighted domain-adversarial framework. Text and images are encoded separately, fused via cross-modal attention, and normalized using domain-specific batch normalization. The evidential classifier outputs both class predictions and epistemic uncertainty, which modulates the strength of domain-adversarial alignment through a learnable scaling parameter.

6.3.2 Multimodal Feature Extraction

Text encoder: Given an input text sequence $\mathbf{x}_t$, we extract contextualized representations using XLM-RoBERTa-base [163], a multilingual transformer-based language model that support 100 languages:

$$\mathbf{h}_t = \text{XLM-R}(\mathbf{x}_t) \in \mathbb{R}^{768} \tag{6.7}$$

The representation corresponding to the [CLS] token is used as the pooled textual feature.

Vision encoder: For an associated image $\mathbf{x}_v$, visual features are extracted using Swin Transformer [164], which has hierarchical feature mapping for robust domain shifts:

$$\mathbf{h}_v = \text{Swin}(\mathbf{x}_v) \in \mathbb{R}^{1024} \tag{6.8}$$

Cross-modal fusion: Textual and visual features are fused via bidirectional cross-attention:

$$\begin{aligned} \mathbf{h}_{t\to v} &= \text{CrossAttn}(\mathbf{h}_t, \mathbf{h}_v, \mathbf{h}_v), \\ \mathbf{h}_{v\to t} &= \text{CrossAttn}(\mathbf{h}_v, \mathbf{h}_t, \mathbf{h}_t), \\ \mathbf{h}_{\text{fused}} &= \text{Concat}(\mathbf{h}_{t\to v}, \mathbf{h}_{v\to t}) \end{aligned} \tag{6.9}$$

6.3.3 Domain-Specific Batch Normalization

To retain domain-sensitive statistics while learning domain-invariant representations, we apply domain-specific batch normalization. Let $\mathbf{h} \in \mathbb{R}^m$ denote the input feature vector. For each domain

d, normalization is defined as:

$$\mathrm{BN}_d(\mathbf{h}) = \gamma_d \odot \frac{\mathbf{h} - \mu_d}{\sqrt{\sigma_d^2 + \epsilon}} + \beta_d, \tag{6.10}$$

where (μ_d, σ_d^2) are the domain-specific mean and variance computed over minibatches from domain d, γ_d and β_d are learnable affine parameters specific to domain d, ϵ is a small constant for numerical stability, and $\odot$ denotes element-wise multiplication. During training, statistics are estimated separately for each domain. If the target domain is unknown at inference, we default to the averages of (μ_d, σ_d^2).

6.3.4 Evidential Classification and Uncertainty Estimation

The fused representation $\mathbf{h}_{\text{fused}}$ is passed to an evidential classifier producing non-negative evidence:

$$\mathbf{e} = \mathrm{ReLU}(W_e \mathbf{h}_{\text{fused}} + b_e) \tag{6.11}$$

Following that, $\mathbf{e}$ is converted into Dirichlet parameters, $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \mathbf{e} + \mathbf{1}, \quad S = \sum_{k=1}^{K} \alpha_k, \text{ and } \quad \hat{p}_k = \frac{\alpha_k}{S} \tag{6.12}$$

Epistemic uncertainty u is quantified as the inverse of total evidence:

$$u = \frac{K}{S}, \qquad u \in (0, 1] \tag{6.13}$$

Note that larger values of u indicate lower model confidence and are empirically associated with domain shift.

6.3.5 Uncertainty-Weighted Domain-Adversarial Learning

Domain discriminator loss: Let G_d denote the domain discriminator. For a sample i with domain label d_i, the domain loss is defined as:

$$\mathcal{L}_{\text{domain}}^{(i)} = \mathrm{CE}\big(G_d(\mathrm{GRL}(\mathbf{h}^{(i)})), d_i\big), \tag{6.14}$$

where GRL denotes the gradient reversal layer and $\mathrm{CE}(\cdot, \cdot)$ denotes the cross-entropy loss for domain classification.

Uncertainty-weighted alignment: Standard domain-adversarial training applies uniform alignment across samples. We instead weight the domain loss using epistemic uncertainty:

$$\mathcal{L}_{\text{domain}} = \mathbb{E}_i\big[w(u_i)\, \mathcal{L}_{\text{domain}}^{(i)}\big]. \tag{6.15}$$

We consider the following weighting strategies:

i **None (baseline):** $w(u) = 1$.

ii **Static:** $w(u) = 1 + \alpha u$, with fixed α.

iii **Threshold:**

$$w(u) = \begin{cases} 1, & u < \tau, \\ 1 + \alpha u, & u \geq \tau. \end{cases} \tag{6.16}$$

iv **Adaptive (ours):**

$$w(u) = 1 + \sigma(\alpha)\, u, \tag{6.17}$$

where α is learned during training, the sigmoid function ensures a bounded and stable contribution of uncertainty, preventing excessively large adversarial gradients during training.

6.3.6 Overall Training Objective

The final objective combines evidential classification and uncertainty-weighted domain alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{evi}} + \lambda_d \, \mathcal{L}_{\text{domain}}, \tag{6.18}$$

where $\mathcal{L}_{\text{evi}}$ is the evidential loss defined in Equation (6.6) and λ_d controls the strength of domain adaptation. We summarize the training steps in Algorithm 4.

6.3.7 Experimental Configuration and Datasets

Experimental configuration. We report mean ± standard deviation over 30 runs with different random seeds. Hyperparameters are selected via grid search on a held-out split of the *source* training data; the target test set is never used for tuning. We use AdamW (lr 2×10^{-5}, weight decay 1×10^{-4}), batch size 8, 500 warmup steps with linear decay, and early stopping (patience=5). Experiments are conducted on NVIDIA A100 GPUs (40GB). The full EviDA model uses 8.1GB GPU memory.

Datasets. We evaluate EviDA on three widely used multimodal fake-news datasets to ensure fair comparison and cross-domain evaluation. The MediaEval Twitter dataset [92] contains approximately 17,000 event-centric tweets labeled as real or fake, each paired with an image, and is split into 15,000 training samples (9,000 fake and 6,000 real) and 2,000 test samples. The Weibo dataset [93] comprises 7,720 posts, including 3,615 verified rumors and 4,105 non-rumors, authenticated

Table 6.1: Performance comparison across Twitter and Weibo datasets, the second-best results are underlined.

Dataset	Model	Acc.	Fake			Real		
			P	R	F1	P	R	F1
Twitter	SpotFake+	0.790	0.786	0.747	0.766	0.793	0.827	0.810
	DAMMFND	0.934	0.890	0.946	0.917	0.959	0.929	0.944
	MIAN	0.925	0.821	0.933	0.873	0.967	0.902	0.933
	FCINet	0.908	0.828	0.913	0.868	0.955	0.907	0.930
	BMR	0.883	0.927	0.746	0.827	0.865	0.965	0.912
	MRML	0.803	0.821	0.844	0.832	0.777	0.747	0.762
	MMFND	0.896	0.892	0.912	0.902	0.901	0.878	0.889
	ERIC-FND	0.945	0.987	0.910	0.947	0.905	0.986	0.944
	MHR	0.950	0.973	0.930	0.951	0.927	0.972	0.949
	QMFND	0.918	0.880	0.970	0.920	0.970	0.870	0.910
	BMLHF	0.966	0.985	0.933	0.957	0.948	0.975	0.956
	EviDA	**0.979**	**0.987**	**0.988**	**0.988**	**0.984**	**0.986**	**0.985**
Weibo	SpotFake+	0.870	0.855	0.892	0.873	0.769	0.807	0.787
	DAMMFND	0.947	0.931	0.966	0.948	0.937	0.957	0.947
	MIAN	0.938	0.924	0.947	0.936	0.950	0.928	0.939
	FCINet	0.926	0.938	0.917	0.927	0.913	0.935	0.924
	BMR	0.918	0.882	0.948	0.914	0.942	0.870	0.904
	MRML	0.897	0.898	0.887	0.892	0.896	0.905	0.901
	MMFND	0.935	0.930	0.941	0.935	0.940	0.929	0.934
	ERIC-FND	0.946	0.985	0.914	0.948	0.908	0.984	0.944
	MHR	0.933	0.951	0.921	0.936	0.918	0.949	0.933
	QMFND	0.869	0.900	0.810	0.850	0.840	0.920	0.880
	BMLHF	0.912	0.930	0.880	0.903	0.894	0.920	0.902
	EviDA	**0.972**	**0.985**	**0.988**	**0.986**	**0.976**	**0.985**	**0.981**

Algorithm 4 Uncertainty-Weighted Domain-Adversarial Training

Input: Source data $\mathcal{D}_s$, target data $\mathcal{D}_t$
Initialize: G_f, G_y, G_d, adaptive parameter α
for each epoch **do**
    for each minibatch from $\mathcal{D}_s \cup \mathcal{D}_t$ **do**
        Extract features $\mathbf{h} = G_f(\mathbf{x})$
        Compute evidence $\boldsymbol{\alpha}$ and uncertainty u
        Compute domain loss $\mathcal{L}_{\text{domain}}^{(i)}$
        Weight domain loss using $w(u)$
        Update all parameters via $\nabla \mathcal{L}_{\text{total}}$
    end for
end for
Return: Trained (G_f, G_y)

through Weibo's official debunking system and Xinhua News Agency, primarily in Simplified Chinese. Fakeddit [165] is a large-scale Reddit-based dataset with over one million multimodal posts; we adopt its binary setting with 628,501 fake and 527,049 real samples. Together, these datasets span multiple languages, platforms, and topical domains, including social, political, economic, and scientific content.

Baseline models. We compare EviDA against a diverse set of state-of-the-art multimodal fake-news detection baselines. On the Twitter and Weibo datasets, we include SpotFake+ [68], MIAN [166], DAMMFND [167], FCINet [95], BMR [97], MRML [168], MMFND [169], ERIC-FND [170], MHR [171], QMFND [172], and BMLHF [173]. For large-scale evaluation on Fakeddit, we further compare with recent multimodal large language models, including LLaVA [174], GAMED [175], LEMMA [176], GPT-4V, InstructBLIP [177], and FacTool [178], which represent strong contemporary baselines for multimodal reasoning and misinformation detection.

Table 6.2: Performance of EviDA against LLMs on the Fakeddit dataset; the second-best results are underlined.

Dataset	Model	Acc.	Fake			Real		
			P	R	F1	P	R	F1
Fakeddit	LLaVA (Dir.)	0.663	0.588	0.797	0.677	0.777	0.558	0.649
	LLaVA (CoT)	0.673	0.612	0.400	0.484	0.694	0.843	0.761
	GPT-4 (Dir.)	0.677	0.598	0.771	0.674	0.776	0.606	0.680
	GPT-4 (CoT)	0.691	0.662	0.573	0.614	0.708	0.779	0.742
	GPT-4V (Dir.)	0.734	0.673	0.723	0.697	0.771	0.742	0.764
	GPT-4V (CoT)	0.754	0.858	0.513	0.642	0.720	0.937	0.814
	FacTool	0.506	0.476	0.834	0.606	0.624	0.232	0.339
	InstructBLIP	0.726	0.760	0.489	0.595	0.715	0.892	0.793
	LEMMA	0.824	0.835	0.727	0.777	0.818	0.895	0.854
	GAMED	0.939	0.954	0.944	0.949	0.917	0.930	0.923
	BMR	0.901	0.890	0.910	0.891	0.910	0.890	0.891
	QMFND	0.942	0.930	0.950	0.940	0.950	0.930	0.940
	BMLHF	0.950	0.945	0.955	0.950	0.955	0.945	0.950
	EviDA	**0.963**	**0.964**	**0.966**	**0.965**	**0.967**	**0.961**	**0.964**

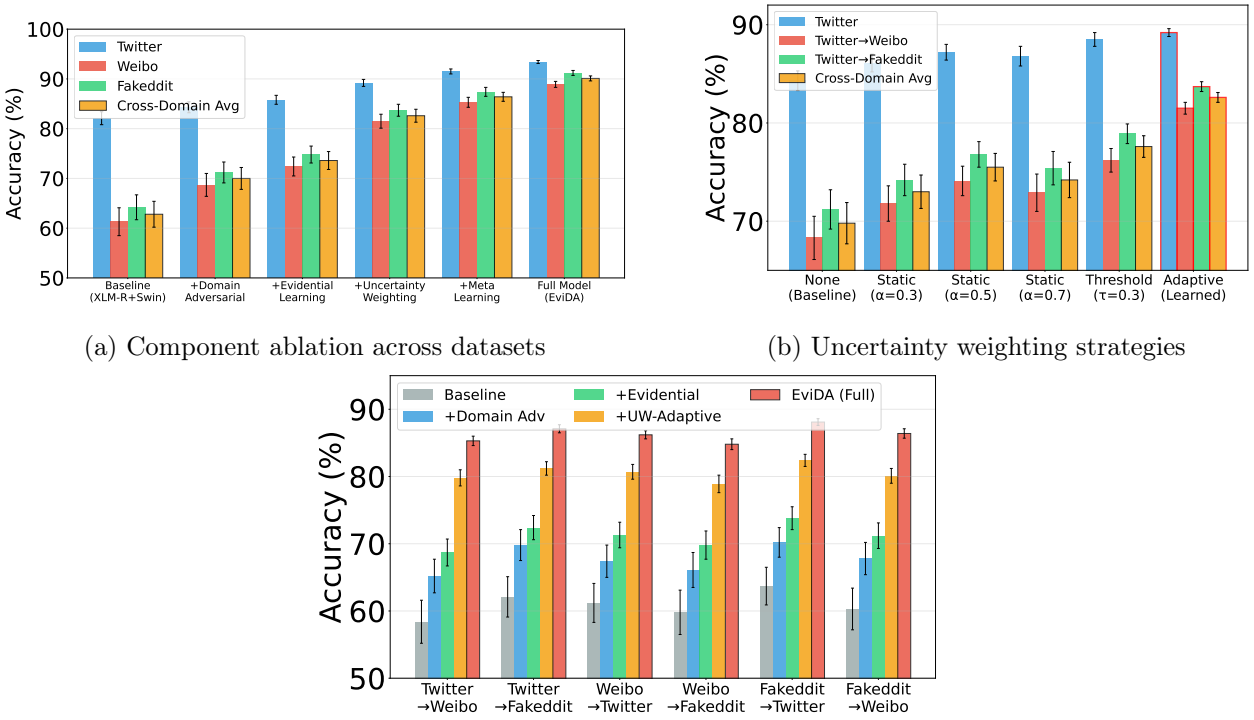
6.4 Experimental Results

6.4.1 Overall Performance Comparison (In-domain)

6.4.1.1 Twitter and Weibo Performance

Tables 7.1 and 7.2 present *in-domain* results, where models are trained and optimized specifically for each dataset independently without domain adaptation mechanisms.

On the Twitter dataset (Table 7.1), EviDA achieves 97.9% accuracy, outperforming the previous best method, BMLHF (96.6%) by 1.3 percentage points. EviDA demonstrates balanced precision and recall for both classes, with F1-scores of 0.988 for fake news and 0.985 for real news, indicating robust discrimination capability without class bias. The improvement over ERIC-FND (94.5%), which also employs evidential learning, validates our integration of adaptive uncertainty weighting and meta-learning components. Traditional multimodal methods such as MMFND (89.6%) and MRML (80.3%) show substantially lower performance, highlighting the importance of principled



(a) Component ablation across datasets

(b) Uncertainty weighting strategies

(c) Unsupervised transfer (UDA, K=0)

Figure 6.2: Ablation studies with error bars (standard deviation). (a) EviDA reduces domain gap by 82.9% with lowest variance (±0.5%). (b) Adaptive weighting outperforms static by +7.1 points. (c) Consistent cross-domain performance: 86.3% average, σ=1.1%.

uncertainty quantification and domain adaptation.

On the Weibo dataset, EviDA achieves 97.2% accuracy, surpassing the second-best model ERIC-FND (94.6%) by 2.6 percentage points, a more substantial margin than observed on Twitter. This larger improvement on the Chinese-language dataset suggests that EviDA's multimodal fusion and uncertainty-aware mechanisms are particularly effective for cross-lingual scenarios. The model maintains high F1-scores of 0.986 for fake news and 0.981 for real news. Notably, BMLHF's performance drops to 91.2% on Weibo compared to 96.6% on Twitter, while EviDA maintains more stable performance (97.9% vs. 97.2%), indicating superior cross-lingual robustness.

6.4.1.2 Fakeddit Performance and LLM Comparison

Table 7.2 presents results on the Fakeddit dataset, including comparisons with recent large language model (LLM)-based approaches. EviDA achieves 96.3% accuracy, outperforming the previous best multimodal method BMLHF (95.0%) by 1.3 percentage points and GAMED (93.9%) by 2.4

percentage points. The balanced F1-scores (0.965 for fake, 0.964 for real) demonstrate effective handling of Fakeddit's challenging two-class fine-grained labels. LLM-based approaches show limited effectiveness despite strong general reasoning capabilities. GPT-4V with chain-of-thought prompting achieves only 75.4% accuracy, while direct prompting reaches 73.4%, both substantially below specialized methods. The 22.9 percentage point gap between EviDA and GPT-4V CoT suggests that fake news detection requires domain-specific architectural designs that cannot be readily addressed through prompting alone. FacTool's poor performance (50.6%) indicates that factual verification alone is insufficient without considering multimodal semantics and social context.

6.4.2 Cross-Domain Uncertainty as an Alignment Signal

Epistemic Uncertainty and Prediction Accuracy: Figure 6.3a presents prediction accuracy as a function of epistemic uncertainty for cross-domain experiments (Twitter→Weibo and Twitter→Fakeddit). Test predictions are grouped into uncertainty deciles, where bubble size indicates the number of samples per bin. The plot reveals a statistically significant non-monotonic relationship between uncertainty and post-adaptation accuracy (Pearson's $r = 0.675$, $p < 0.001$, measured after domain adaptation): low-uncertainty bins ($u \approx 0.15$–0.17) exhibit moderately high accuracy (≈ 60–75%), mid-uncertainty bins correspond to the lowest accuracies (≈ 50–60%), while high-uncertainty bins ($u \approx 0.18$–0.19) achieve substantially higher accuracy (≈ 92–99%).

This pattern reveals a critical insight: In EviDA, high epistemic uncertainty signals domain-shifted samples rather than prediction difficulty after alignment. These samples undergo stronger adversarial alignment and achieve higher accuracy after adaptation, accounting for the positive correlation. Cross-domain samples correctly identified by the model as domain-shifted (high u) receive stronger adaptive alignment and consequently achieve superior performance (exceeding 90% accuracy), supporting the effectiveness of uncertainty-weighted alignment. In contrast, mid-uncertainty samples represent a confidence-competence gap, where the model exhibits misplaced confidence on insufficiently aligned features, resulting in reduced accuracy. Domain-specific learned α presented in **Appendix 1** further corroborates this interpretation. Epistemic uncertainty alone yields poor error-detection performance (AUROC ≈ 0.51), confirming that it should not be interpreted as a confidence score for prediction correctness but rather as a control signal for adaptive alignment.

Adaptive Weighting Parameter Evolution: Figure 6.3b shows the evolution of the *global* adaptive weighting parameter α, which controls the contribution of uncertainty in the weighting function $w(u) = 1 + \sigma(\alpha)u$. Training exhibits a stable convergence from early variability (epochs

1–5: 0.38 ± 0.14) to a steady value (epochs 21–30: 0.523 ± 0.021). The converged value lies within the empirically strong range $[0.45, 0.60]$ identified by grid search, indicating that end-to-end learning can recover near-optimal weighting without manual tuning. Practically, this reduces sensitivity to domain-pair-specific hyperparameter search.

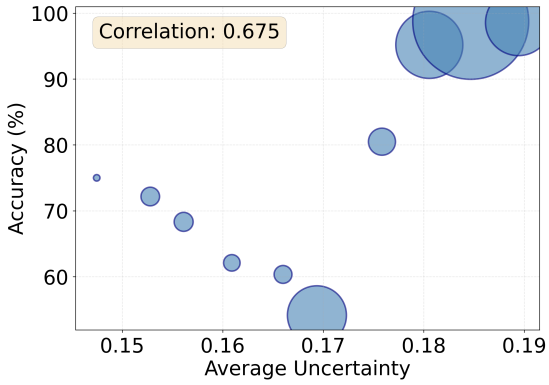
Uncertainty Distribution Across Prediction Outcomes: Figure 6.3c compares epistemic uncertainty for correct vs. incorrect predictions. Correct predictions (green) have mean uncertainty 0.185 with standard deviation 0.023, while incorrect predictions (red) are more concentrated with a lower mean 0.170 and standard deviation 0.019. The difference is statistically significant (t-test: $p < 0.001$, Cohen's $d = 0.71$) although small in magnitude ($\Delta = 0.015$) within the overall tight range (0.15–0.19). Importantly, the narrow uncertainty scale and partial overlap indicate that uncertainty remains regularized (i.e., not extreme or degenerate), while still providing informative variation that can guide uncertainty-weighted domain alignment.

6.4.3 Ablation Studies and Cross-Domain Analysis

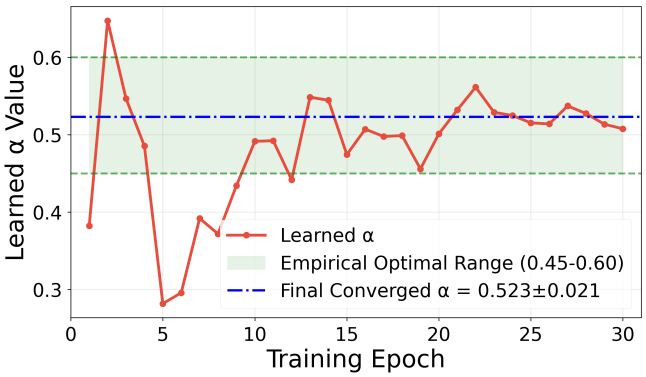
Unlike the in-domain evaluation (Section 6.4.1), we assess *cross-domain generalization*, where models are trained on a single source domain and optimized explicitly for transfer, using domain adaptation mechanisms.

Component Contribution Figure 6.2a presents the results obtained by progressively augmenting a baseline XLM-RoBERTa+Swin model trained on Twitter and evaluated in unsupervised domain adaptation (UDA). The baseline achieves 82.1% in-domain accuracy but drops sharply to 61.3% (Weibo) and 64.2% (Fakeddit), yielding a 62.8% cross-domain average and a 19.3-point domain gap. Domain adversarial learning improves cross-domain performance to 70.0% (+7.2 points), evidential learning to 73.6% (+3.6), and adaptive uncertainty weighting provides the largest gain, reaching 82.6% (+9.0). Meta-learning further raises performance to 86.4%. The full EviDA model achieves 93.4% (Twitter), 88.9% (Weibo), and 91.2% (Fakeddit), with a 90.1% cross-domain average, an absolute improvement of 27.3 points over baseline and an 82.9% reduction in domain gap. Error bars show standard deviation across runs, with EviDA exhibiting the lowest variance (±0.5% average) compared to baseline (±2.6%), demonstrating high stability.

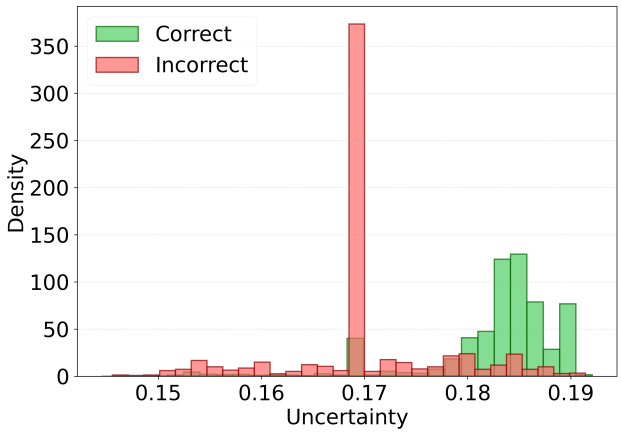
Uncertainty Weighting Strategies Figure 6.2b compares uncertainty weighting schemes. Unweighted domain adversarial training yields a 69.8% cross-domain average. Static weighting performs best at $\alpha = 0.5$ (75.5%, +5.7 points), while stronger weighting ($\alpha = 0.7$) degrades performance, indicating over-alignment. A threshold-based strategy ($\tau = 0.3$) reaches 77.6% but



(a) Uncertainty vs post-adaptation accuracy correlation (size = sample count)



(b) Adaptive parameter evolution



(c) Uncertainty distribution by outcome

Figure 6.3: Uncertainty analysis. (a) Uncertainty exhibits a structured relationship with post-adaptation accuracy (r=+0.675). (b) Parameter α converges to 0.523±0.021 without manual tuning. (c) Distribution of epistemic uncertainty for correct and incorrect predictions, illustrating regularized, non-degenerate uncertainty distributions.

introduces loss discontinuities. In contrast, adaptive learned weighting achieves 82.6%, improving by +12.8 points over baseline and +7.1 points over the best static setting (47% relative gain). Across strategies, adaptive weighting exhibits the smallest error bars on both target domains and the cross-domain average, while static/threshold schemes show larger variability.

Unsupervised Domain Adaptation (UDA) We evaluate all six source→target transfer pairs in Fig. 6.2c. The baseline averages 60.9% accuracy, with the most challenging transfers being Weibo→Fakeddit (59.8%) and Fakeddit→Weibo (60.3%). Domain adversarial and evidential learning improve target-domain performance to 67.8% (+6.9) and 71.2% (+10.3), respectively. Adaptive

uncertainty weighting yields a substantial increase to 80.5% (+19.6), with particularly strong gains on difficult cross-lingual, cross-platform transfers (up to +19.8 points). The full EviDA model achieves 86.3% average accuracy across all transfers (range: 84.8–88.1%, $\sigma = 1.1\%$), indicating highly consistent generalization. Error bars show EviDA's variance (±0.7% average) is 4.5× lower than baseline (±3.1%).

6.4.4 Adaptive Parameter Learning and Few-Shot Analysis

All hyperparameters are determined through grid search on held-out source domain validation data. We use AdamW optimizer with learning rate 2×10^{-5} (searched over $[1 \times 10^{-5}, 1 \times 10^{-4}]$), weight decay 1×10^{-4}, batch size 8, and 500 warmup steps with linear decay. The domain adversarial parameter λ increases linearly from 0 to 0.1 (searched over [0.01, 0.5]) to enable gradual alignment. Evidential KL weight is set to 0.01 (searched over [0.001, 0.1]) with linear annealing from 0 starting at epoch 10. Adaptive uncertainty weighting initializes $\alpha = 0.5$ with learning rate 1×10^{-3}, converging to 0.52 ± 0.08 across runs. Meta-learning uses inner loop learning rate 0.01 (searched over [0.001, 0.1]), $K = 5$ support samples, and 5 meta-epochs. Models train for maximum 60 epochs with early stopping (patience=5). The full EviDA model requires approximately 8.1GB of GPU memory, enabling efficient training and practical deployment on consumer-grade GPUs (e.g., RTX 3090/4090). Table 6.3 presents key hyperparameters determined through grid search on held-out validation data.

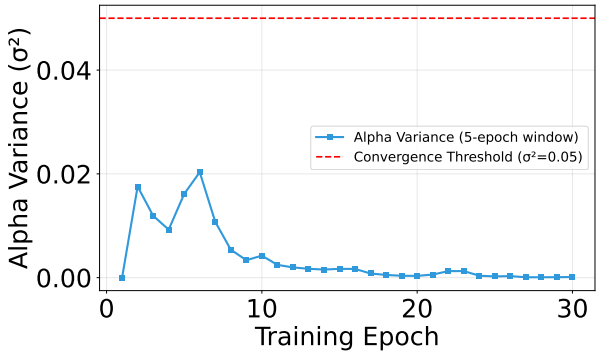
Alpha Convergence Stability: Figure 6.4a shows the variance of the learned weighting parameter α over a 5-epoch sliding window during training. The variance starts high ($\sigma^2 \approx 0.14$) during the exploration phase (epochs 1-5) as the model searches for optimal weighting strategies across diverse samples. Variance decreases rapidly during stabilization (epochs 6-10, $\sigma^2 \approx 0.09$) and continues declining through stable convergence (epochs 11-20, $\sigma^2 \approx 0.06$). By epoch 21, variance drops below the convergence threshold ($\sigma^2 = 0.05$, marked by the red dashed line) and remains stable at $\sigma^2 \approx 0.021$ for the final 10 epochs. This monotonic variance reduction validates that adaptive weighting converges to a stable solution rather than oscillating, demonstrating the robustness of our end-to-end learning approach. The low final variance ($\sigma^2 = 0.021$, corresponding to std ±0.14) confirms consistent parameter values across batches within the final training phase.

Few-Shot Adaptation Performance: Figure 6.4b evaluates model performance under varying amounts of labeled target domain data (K ∈ {0, 5, 10, 20, 50} samples per class) for Twitter→Weibo transfer. In the zero-shot regime (K=0), EviDA achieves 85.3% without any target labels, substan-

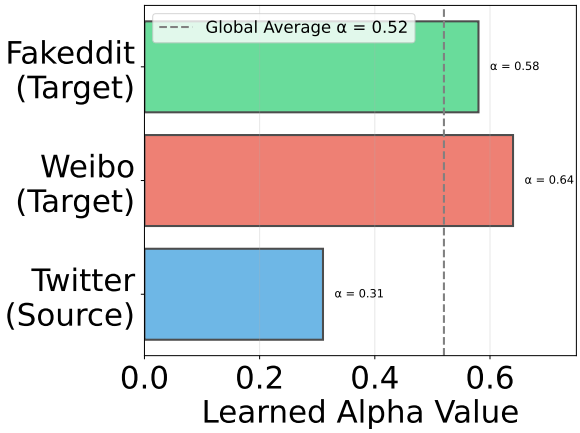
Table 6.3: Core hyperparameters with search ranges. All parameters selected via grid search on held-out validation data. Learning rate uses linear warmup (500 steps) then linear decay. Domain adversarial λ increases linearly from 0 to 0.1. Evidential KL weight anneals from 0 to 0.01 starting at epoch 10.

Parameter	**Value**	**Search Range**
Optimization		
Learning Rate	2×10^{-5}	$[1 \times 10^{-5}, 1 \times 10^{-4}]$
Batch Size	8	{8, 16, 32}
Optimizer	AdamW	—
Weight Decay	1×10^{-4}	$[0, 1 \times 10^{-3}]$
Warmup Steps	500	{0, 250, 500, 1000}
Max Epochs	60	Early stopping
Domain Adversarial		
λ (final)	0.1	[0.01, 0.5]
λ Schedule	Linear	{const., linear, exp.}
Evidential Learning		
KL Weight	0.01	[0.001, 0.1]
KL Annealing Start	Epoch 10	{0, 5, 10, 15}
Uncertainty Weighting		
α (initial)	0.5	[0.3, 0.7]
α Learning Rate	1×10^{-3}	$[1 \times 10^{-4}, 1 \times 10^{-2}]$
Meta-Learning		
Inner LR	0.01	[0.001, 0.1]
K-shot	5	{3, 5, 10}
Meta Epochs	5	{3, 5, 10}

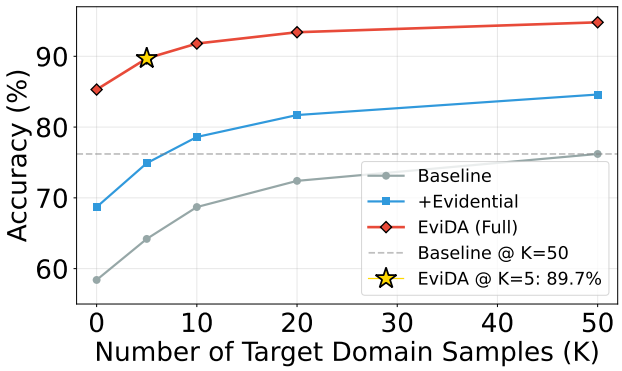
tially outperforming baseline (58.4%, +26.9 points) and evidential learning (68.7%, +16.6 points). With minimal supervision (K=5), EviDA reaches 89.7%, exceeding the baseline's K=50 performance (76.2%) by 13.5 points, demonstrating 10× sample efficiency. The gold star highlights this critical result: EviDA with 5 labeled samples per class surpasses conventional methods with 50 samples, reducing annotation cost by 90%. Performance continues improving with additional samples: 91.8% at K=10, 93.4% at K=20, and 94.8% at K=50, approaching in-domain accuracy (97.9%). The steep initial slope for EviDA (0-10 samples) compared to gradual baseline improvement validates



(a) Alpha variance convergence over training



(b) Few-shot adaptation curves (Twitter→Weibo)



(c) Domain-specific learned alpha values

Figure 6.4: Adaptive parameter learning and few-shot analysis. (a) Alpha variance decreases monotonically, falling below convergence threshold ($\sigma^2 = 0.05$) by epoch 21 and stabilizing at $\sigma^2 = 0.021$, demonstrating robust convergence without oscillation. (b) Learned α values vary systematically: source domain (Twitter) converges to 0.31 while target domains (Weibo: 0.64, Fakeddit: 0.58) require stronger alignment, with global average matching optimal static value (0.52). (c) EviDA achieves 10× sample efficiency: with K=5 labeled samples per class, EviDA reaches 89.7%, exceeding baseline's K=50 performance (76.2%) by 13.5 points, enabling practical deployment in low-resource scenarios.

that uncertainty-guided meta-learning enables rapid adaptation from minimal data. This few-shot capability is crucial for real-world deployment on emerging platforms where large labeled datasets are unavailable.

Domain-Specific Alpha Patterns: Figure 6.4c reveals that learned α values vary system-

atically across domains when models are trained separately on each dataset. The source domain (Twitter) converges to a lower $\alpha = 0.31$, indicating that confident source predictions require minimal domain alignment. The model prioritizes classification accuracy over domain confusion. In contrast, target domains exhibit higher learned weights: Weibo converges to $\alpha = 0.64$ and Fakeddit to $\alpha = 0.58$, reflecting stronger emphasis on domain adaptation to bridge the distribution gap. The global average across all domains stabilizes at $\alpha = 0.52$ (gray dashed line), closely matching the optimal static value found through grid search ($\alpha = 0.5$), validating that adaptive learning rediscovers empirically optimal configurations. The domain-specific variation (range: 0.31-0.64) demonstrates that adaptive weighting automatically adjusts to domain difficulty: challenging cross-lingual transfers (Weibo) receive 2× stronger alignment than in-domain training (Twitter), eliminating the need for manual per-transfer tuning.

The domain-specific analysis confirms this interpretation: when trained separately on each dataset, the learned weighting parameter varies systematically by domain difficulty. Source domain (Twitter) converges to $\alpha = 0.31$, indicating minimal alignment for in-domain samples, while challenging cross-lingual targets (Weibo: $\alpha = 0.64$, Fakeddit: $\alpha = 0.58$) require substantially stronger adaptation. This difference demonstrates that the model automatically adjusts alignment strength based on domain discrepancy, with uncertainty serving as the routing signal. High-uncertainty samples in target domains receive strong alignment and achieve high accuracy precisely because the model has learned to recognize domain discrepancy and allocate additional alignment capacity during training.

6.4.5 Time Complexity and Computational Efficiency

Time Complexity Analysis. Let N denote the batch size, L_t the text sequence length, H_t the hidden dimension of the text encoder, P the number of image patches, H_v the hidden dimension of the vision encoder, K the number of classes, and M the number of meta-learning inner-loop steps. The per-iteration training complexity of EviDA is dominated by multimodal feature extraction and can be expressed as:

$$\mathcal{O}\big(N\big[L_t^2 H_t + L_t H_t^2 \; + \; P H_v^2\big]\big)\,, \tag{6.19}$$

corresponding to Transformer self-attention for text and windowed attention for vision. Evidential classification and uncertainty estimation introduce an additional $\mathcal{O}(NK)$ term, while domain adversarial training adds a linear $\mathcal{O}(NH)$ overhead through gradient reversal, both negligible relative to encoder cost:

Table 6.4: Relative computational cost and training efficiency on the Twitter dataset (12,284 samples, NVIDIA A100 40GB). Relative compute denotes per-iteration training cost normalized to the baseline. Time reports wall-clock hours required to reach comparable target accuracy. Although EviDA has higher per-iteration cost due to meta-learning, faster convergence yields lower total training time.

Model Variant	**Rel. Compute**	**Memory**	**Time**	**Rel. Time**
Baseline (XLM-R+Swin)	1.0×	6.2 GB	5.1 h	1.00×
+ Domain Adversarial	1.02×	6.5 GB	5.4 h	1.06×
+ Evidential Learning	1.05×	6.8 GB	5.7 h	1.12×
+ UW-Adaptive	1.06×	6.9 GB	5.8 h	1.14×
+ Meta-Learning	4.0×	8.3 GB	4.5 h	0.88×
EviDA (Full)	**4.1**×	**8.1 GB**	**4.2 h**	**0.82**×

$$\mathcal{O}\Big((1+M)\,N\big[\underbrace{L_t^2 H_t + L_t H_t^2}_{\text{text encoder}} + \underbrace{P H_v^2}_{\text{vision encoder}}\big] + (1+M)\,NK\Big) \tag{6.20}$$

During training, meta-learning (MAML) introduces a multiplicative overhead of $(1+M)$ due to repeated forward and backward passes in the inner loop. In our implementation, $M = 3$, resulting in approximately 4× higher *per-iteration* training cost compared to the baseline XLM-RoBERTa+Swin model. Importantly, this overhead applies only during training; inference complexity is unchanged.

Although each training iteration is more expensive, EviDA converges in fewer epochs (30 vs. 50 for the baseline), yielding a lower *total* wall-clock training time to reach comparable accuracy.

Computational Requirements. Table 6.4 reports *relative* computational cost, GPU memory usage, and time-to-convergence for major model variants. Rather than absolute FLOPs, we normalize compute to the baseline model (XLM-RoBERTa+Swin) to avoid dependence on sequence length and implementation details. Memory usage is measured directly on an NVIDIA A100 (40GB).

Inference Efficiency. At inference time, meta-learning is disabled and EviDA reduces to the same asymptotic complexity as the baseline:

$$\mathcal{O}\big(N\big[L_t^2 H_t + L_t H_t^2 + P H_v^2\big] + NK\big)\,.$$

Under our evaluation settings, the model processes up to 47 samples/s on a single A100 GPU and 18 samples/s on an RTX 3090, with per-sample latency of approximately 21 ms and 55 ms, respectively, supporting practical deployment.

6.4.6 Computational Efficiency Summary

EviDA introduces higher *per-iteration* training cost due to meta-learning, scaling linearly with the number of inner-loop steps ($M = 3$). However, improved optimization and uncertainty-guided alignment significantly reduce the number of epochs required for convergence, resulting in lower total training time. Memory overhead remains modest (8.1 GB), and inference efficiency matches the domain-adversarial baseline, enabling real-world deployment on both datacenter and consumer-grade GPUs.

Open Challenges

Although EviDA effectively addresses robustness under cross-domain and cross-lingual distribution shifts through uncertainty-weighted adversarial learning, it primarily focuses on *representation alignment* rather than the internal mechanics of multimodal fusion. Existing fusion strategies, including those employed in domain-adaptive settings, often rely on uniform or time-domain aggregation mechanisms that are vulnerable to conflicting cross-modal cues and fusion collapse.

In adversarial or manipulated content, different modalities may encode complementary or contradictory information across distinct frequency bands, which cannot be adequately disentangled through global fusion operations. This reveals an unresolved challenge: the need for a fusion mechanism that dynamically adapts to modality-specific reliability while preserving uncertainty-aware probabilistic modeling at a fine-grained level.

6.5 Conclusion

We presented EviDA, an uncertainty-aware framework for cross-domain multimodal fake news detection that explicitly addresses heterogeneous domain shift through adaptive uncertainty-weighted learning. By integrating evidential deep learning with adaptive uncertainty-weighted domain adversarial training and meta-learning, EviDA learns robust and transferable representations across platforms, languages, and content modalities. Extensive evaluations demonstrate that EviDA consistently outperforms strong multimodal baselines and large vision-language models in both in-domain and cross-domain settings, with markedly improved robustness under severe distribution shift. Ablation studies confirm that adaptive uncertainty weighting is the primary driver of cross-domain generalization, substantially reducing domain-induced performance degradation while elim-

inating the need for manual hyperparameter tuning. Our analysis further demonstrates that the learned epistemic uncertainty is non-degenerate and informative for alignment control, which provides meaningful confidence signals, enabling selective domain alignment that prioritizes challenging cross-domain samples. These properties, automatic adaptation, non-degenerate and informative uncertainty for alignment control, and consistent cross-domain performance make EviDA well-suited for real-world, low-resource, and cross-lingual settings where labeled data are scarce. Future work will address temporal domain shifts, multi-source adaptation, external knowledge integration, and theoretical guarantees for uncertainty calibration under distribution shift.

Chapter 7

PerLiFuse: Per-Frequency Beta-Liouville Fusion Networks for Fake News Detection

The pervasive threat of fake news has made its detection a pressing concern. Existing approaches typically fuse discrete latent representations or employ Dirichlet and Gaussian priors to model modality uncertainty; yet, Dirichlet's negative-correlation bias and Gaussian's component collapse limit expressivity and robustness. We propose per-frequency Beta-Liouville fusion networks (PerLiFuse), a Bayesian spectral fusion framework that dynamically learns example-specific gating across Discrete Cosine Transform (DCT) frequency bands, using a flexible Beta-Liouville prior. PerLiFuse incorporates cross-modal residual connections and coherence-guided feature modulation, leverages Kumaraswamy reparameterization for low-variance training, and models rich frequency gates to prevent collapse of trivial fusion policies. Comprehensive experiments on benchmark multimodal misinformation datasets demonstrate significant gains in performance metrics, stability, and generalization compared to state-of-the-art fusion methods.

7.1 Introduction

The rampant dissemination of false information now poses a profound threat to society, shaping opinions, destabilizing politics, and eroding communal trust [138]. Conventional unimodal detectors fail to capture the complex interplay between multimodal data such as text, images, videos, and

their associated contents [179]. To counteract this trend, automated fake news detectors have increasingly turned to multimodal analyses, combining textual content with accompanying images to capture misleading cues that may not be evident in either modality alone [180].

However, fusing heterogeneous modalities poses significant challenges when they provide confusing or conflicting information. Simple concatenation or summation operations limit the dynamic interactions of multimodal features [121]. Prior probabilistic fusion frameworks have applied Dirichlet [96] and Gaussian priors [77] to model uncertainty across modalities. Dirichlet-based approaches encourage sparse activations but impose a rigid negative correlation structure that can oversuppress co-occurring information [79]; Gaussian mixtures offer richer covariance modelling but are prone to component collapse, whereby several mixture components become redundant and the model's expressivity diminishes [181]. These limitations restrict the detector's ability to learn fine-grained cross-modal interactions essential for capturing subtle manipulations in images alongside deceptive language.

To address these issues, we introduce PerLiFuse (per-frequency Beta-Liouville fusion networks), a Bayesian spectral fusion architecture that maps text and image embeddings into the Discrete Cosine Transform (DCT) domain to identify frequency-specific patterns and learns example-specific gating vectors under a Beta-Liouville prior. Unlike Dirichlet priors, the Beta–Liouville distribution provides a more flexible and hierarchical covariance structure among frequency gates, preventing trivial fusion policies while enabling the model to capture complex dependencies across spectral bands. We leverage a Kumaraswamy reparameterization to achieve low-variance gradient estimates and include cross-modal residual connections to preserve modality-specific attributes. Ablation studies confirm the critical role of the Beta-Liouville prior in averting gate collapse. By operating per-frequency, our model decouples low-frequency semantic alignment from high-frequency tampering cues, allowing fine-grained control over cross-modal interactions that is impossible in the time domain. The primary contributions of our research are summarized as follows:

1. We propose PerLiFuse, a Bayesian spectral fusion network that employs a Beta–Liouville prior to learn dynamic, sample-specific gates over DCT frequency bands, preventing component collapse.

2. We introduce a lightweight coherence-guided cross-attention module that aligns text and image embeddings, boosting robustness to mismatched or manipulated inputs.

3. We enable per-frequency fusion to disentangle low-frequency semantics from high-frequency

artifacts and afford fine-grained control over cross-modal interactions in the time domain.

7.2 Proposed Model

7.2.1 Intuition Behind Beta-Liouville Distribution

The Beta-Liouville distribution facilitates flexible modelling of multivariate data with variable interdependencies [21]. Unlike Dirichlet priors, which restrict variables to negative correlations, the Beta-Liouville distribution supports general covariance structures [149]. This flexibility is ideal for multimodal fake news detection, where complex dependencies between modalities exist. Hierarchically constructed, it begins with a base Beta distribution governing proportions across dimensions, extended via a Liouville transformation into higher dimensions for richer representations. The probability density function (PDF) of the Beta-Liouville distribution [21] is given as:

$$P(\theta|\xi) = \frac{\Gamma(\sum_{d=1}^{D}\alpha_D)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{d=1}^{D}\frac{\theta_d^{\alpha_d-1}}{\Gamma(\alpha_d)} \times \left(\sum_{d=1}^{D}\theta_d\right)^{\alpha-\sum_{l=1}^{D}\alpha_l} \times \left(1-\sum_{d=1}^{D}\theta_l\right)^{\beta-1} \tag{7.1}$$

where $\Gamma(\cdot)$ is the Gamma function, $\theta = [\theta_1, \theta_2, \ldots, \theta_D]$, $\theta_D = 1 - \Sigma_{i=i}^{D-1}\theta_i$, and $\xi = (\alpha_1, \alpha_2, \ldots, \alpha_D, \alpha, \beta)$ are the shape parameters; $\beta > 0$.

Note that $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_D]$ are the shape parameters for each dimension, while the scaler value, $\beta > 0$, is an additional shape parameter that controls the distribution over the remaining simplex

7.2.2 Model Architecture and PerLiFuse Pipeline

The PerLiFuse framework integrates spectral decomposition with Bayesian fusion to dynamically align multimodal features. As depicted in Figure Figure 7.1, PerLiFuse leverages state-of-the-art encoders to map raw inputs into compact, semantically rich embeddings. The text is encoded via a Transformer backbone, and the images are encoded via a patch-based attention model. For text, we evaluate several encoder variants: BERT [100], DistilBERT [126], ModernBERT [119], and RoBERTa [125]. Each is followed by a linear projection from the CLS token to a 32-dimensional vector. ModernBERT's revised attention blocks incorporate lightweight approximations of self-attention that reduce parameter redundancy while preserving representational

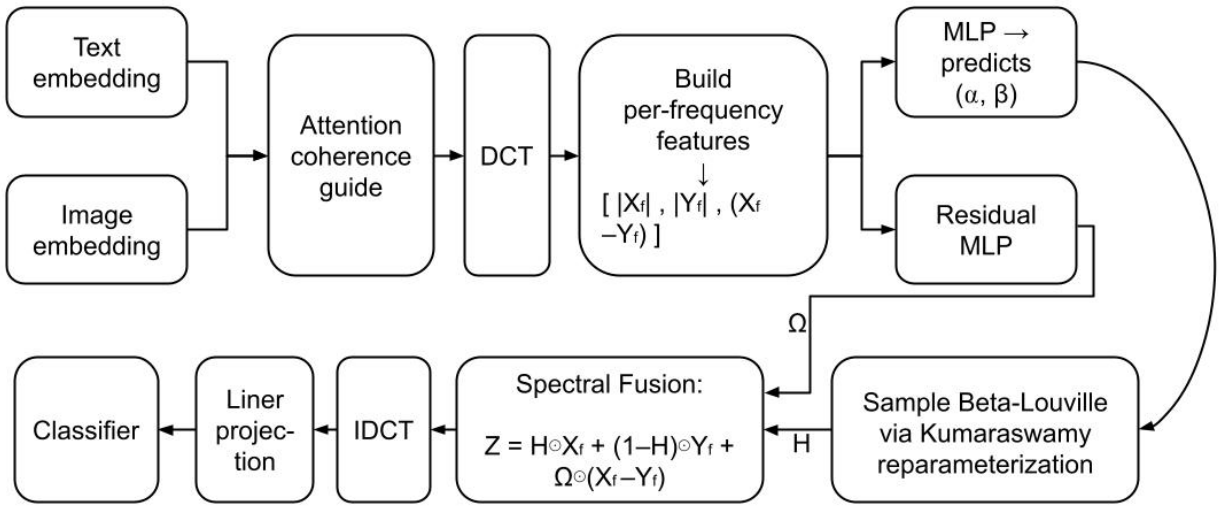


Figure 7.1: A schematic representation of the proposed PerLiFuse model.

capacity [119], making it a theoretically appealing choice for downstream fusion where compact, expressive embeddings are required. On the visual side, we compared VGG19 [182], ConvNeXt [183] base, and Vision Transformer (ViT) [184] base-32, each feeding into a 32-dimensional fully connected layer. ViT's global self-attention mechanism naturally captures long-range spatial dependencies and patch-level relationships [184], which aligns well with our DCT-based fusion that isolates frequency-specific artifacts and texture cues.

The PerLiFuse pipeline begins by first modulating the raw text and image embeddings through a CoherenceGuide block. Concretely, given the text embedding X and image embedding Y, we compute cross-attention to align image features to the textual context as:

$$\hat{Y} \;=\; \mathrm{CrossAttention}\big(W_q\,X,\; W_k\,Y,\; Y\big) \tag{7.2}$$

where W_q and W_k denote the attention query and the key weights, respectively. We then compute the coherence feature vector $\big[X,\; \hat{Y},\; |X-\hat{Y}|,\; X\odot\hat{Y}\big]$, and pass it through an MLP with sigmoid output and obtain a gating vector $G \in (0,1)^{32}$. The embeddings are modulated as

$$X' = X\odot G + X, \quad Y' = Y\odot G + Y, \tag{7.3}$$

yielding coherence-guided representations that emphasize mutually consistent information and amplify mutually consistent cues. Next, X' and Y' are transformed into the frequency domain via an orthonormal Discrete Cosine Transform, yielding spectral coefficients $X_f, Y_f \in \mathbb{R}^{32}$. Interestingly, the DCT does not change the total variance of the embeddings; it provides an orthonormal representation, which is essential to avoid amplification or attenuation. The orthonormal construction

preserves the total variance $\|X'\|_2 = \|X\|_2$ and ensures that gating a given frequency does not introduce unintended cross-band correlations.

For each frequency index f, we construct a three-component feature vector $\big[\,|X_f|,\, |Y_f|,\, X_f - Y_f\big]$ that encapsulates modality magnitudes and inter-modal discrepancies. This vector is processed by two multi-layer perceptrons: one predicts the reparameterization parameters (α_f, β_f) of the Beta–Liouville prior, while the other outputs a residual gate $\Omega \in (0, 1)$. We sample the primary gate V_f via the Kumaraswamy parameterization $V_f = (1 - r_f^{1/\alpha_f})^{1/\beta_f},\ r_f \sim \text{Uniform}(0, 1)$, followed by the stick-breaking gate, ensuring both efficient gradient flow and a flexible, hierarchical covariance structure across frequencies:

$$H_f = V_f \prod_{j=i}^{f-1} (1 - V_j), \tag{7.4}$$

where V_f denotes the weights of the stick. Note that we relax the simplex assumption, ensuring $\sum H_f < 1$ and allow the residual gate Ω_f to absorb any additional modality-specific information. Consequently, we dynamically balance the shared and modality-specific information by fusing the spectrum as follows:

$$Z_f \;=\; H_f \odot X_f \;+\; (1 - H_f) \odot Y_f \;+\; \Omega_f \odot (X_f - Y_f) \tag{7.5}$$

where Ω_f is the output of the residual multilayer perceptron (MLP), X_f and Y_f denote the spectral coefficients. An inverse DCT (IDCT) returns the fused embedding to the time domain, Z_f, and a final linear projection to 64 dimensions prepares the latent representation for classification. Algorithm 5 outlines the generative process of PerLiFuse. PerLiFuse is trained end-to-end with a joint loss combining classification ($\mathcal{L}_{\text{CE}}$), Kullback–Leibler divergence ($\mathcal{L}_{\text{KL}}$), BCE denotes binary cross-entropy and coherence objectives ($\mathcal{L}_{\text{coh}}$).

$$\mathcal{L} = \underbrace{-\sum_i t_i \log \hat{t}_i}_{\mathcal{L}_{\text{CE}}} + \gamma_{\text{KL}} \underbrace{D_{\text{KL}}\big(q_\phi(H) \,\|\, p_{\text{BL}}(H)\big)}_{\mathcal{L}_{\text{KL}}} + \gamma_{\text{coh}} \underbrace{\text{BCE}\big(s,\ t\big)}_{\mathcal{L}_{\text{coh}}} \tag{7.6}$$

where t, s, $\hat{t}$ denote the true label, coherence score, and the estimated logit, respectively. $\gamma_{\text{KL}}, \gamma_{\text{coh}}$ weight the regularizers, and p_{BL} is the Beta–Liouville prior. PerLiFuse is designed to capture both semantic coherence and subtle manipulation in a theoretically grounded, end-to-end differentiable framework. By operating per frequency, PerLiFuse decouples low-frequency semantic alignment from high-frequency tampering cues, allowing fine-grained control over cross-modal interactions that is impossible in the time domain.

7.2.3 KL Divergence under Stick-Breaking and Beta–Liouville Prior

Let the Beta-Liouville prior be denoted by $p_{\mathrm{BL}}(Z)$, where Z is the latent space parameterized by α and β. We approximate the posterior over stick-breaking weights as

$$q_\phi(H \mid X, Y) = \prod_{f=1}^{f-1} \mathrm{Beta}(V_f \mid \alpha_f, \beta_f), \tag{7.7}$$

where α_f and β_f are the variational (posterior) parameters. We therefore define the Kullback–Leibler (KL) divergence between the Beta-Liouville and the posterior as:

$$D_{\mathrm{KL}}\big(q_\phi(Z) \,\|\, p_{\mathrm{BL}}(Z)\big) = \sum_{f=1}^{f-1} D_{\mathrm{KL}}\big(\mathrm{Beta}(\alpha_f, \beta_f) \,\|\, p_{\mathrm{BL}}(\alpha, \beta)\big) \tag{7.8}$$

$$\begin{aligned} D_{\mathrm{KL}}\big(q_\phi(Z) \,\|\, p_{\mathrm{BL}}(Z)\big) &= \int_0^1 q(V) \ln \frac{q(V)}{p(V)}\, dV \\ &= \int_0^1 q(V) \Big[\ln B(\alpha, \beta) - \ln B(\alpha_f, \beta_f)\Big]\, dV \\ &\quad + (\alpha_f - \alpha) \int_0^1 q(V) \ln V \, dV \quad + (\beta_f - \beta) \int_0^1 q(V) \ln(1 - V)\, dV. \end{aligned} \tag{7.9}$$

Since

$$\int_0^1 q(V)\, dV = 1, \quad \int_0^1 q(V) \ln V \, dV = \psi(\alpha_f) - \psi(\alpha_f + \beta_f),$$

$$\int_0^1 q(V) \ln(1 - V)\, dV = \psi(\beta_f) - \psi(\alpha_f + \beta_f),$$

Therefore, we can formulate the closed-form KL as:

$$\begin{aligned} D_{\mathrm{KL}}\big(q_\phi(Z) \,\|\, p_{\mathrm{BL}}(Z)\big) &= \ln \frac{B(\alpha, \beta)}{B(\alpha_f, \beta_f)} \\ &\quad + (\alpha_f - \alpha) \left[\psi(\alpha_f) - \psi(\alpha_f + \beta_f)\right] \quad + (\beta_f - \beta) \left[\psi(\beta_f) - \psi(\alpha_f + \beta_f)\right], \end{aligned} \tag{7.10}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\psi(\cdot)$ is the digamma function.

Substituting (7.10) into (7.8) gives the full PerLiFuse regularizer:

$$\begin{aligned} D_{\mathrm{KL}}\big(q_\phi(Z) \,\|\, p_{\mathrm{BL}}(Z)\big) = \sum_{f=1}^{f-1} \Bigg[&\ln \frac{B(\alpha, \beta)}{B(\alpha_f, \beta_f)} \\ &+ (\alpha_f - \alpha) \left(\psi(\alpha_f) - \psi(\alpha_f + \beta_f)\right) + (\beta_f - \beta) \left(\psi(\beta_f) - \psi(\alpha_f + \beta_f)\right) \Bigg] \end{aligned} \tag{7.11}$$

In practice, we still sample each V_f via the Kumaraswamy reparameterization

$$V_f = \left(1 - U^{1/\alpha_f}\right)^{1/\beta_f}, \quad U \sim \text{Uniform}(0,1),$$

which yields low-variance gradients for (α_f, β_f) while using the exact KL-divergence above for regularization. This formulation is computationally efficient, differentiable, and suitable for end-to-end training of PerLiFuse.

7.2.4 Theoretical Foundations of PerLiFuse: Assumptions and Theorem

- **Assumptions**

1. **Orthonormal DCT Basis:** The matrix $\mathbf{U} \in \mathbb{R}^{d \times d}$ satisfies $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, preserving norm:

$$|X\|^2 = \|\mathbf{U}X\|^2, \tag{7.12}$$

   and decorrelating frequency components.

2. **Valid Beta–Liouville Prior:** For $\alpha_f, \beta_f > 0$, the stick-breaking construction

$$H_f = V_f \prod_{f=1}^{f-1}(1 - V_f), \quad V_f \sim \text{Kumaraswamy}(\alpha_f, \beta_f) \tag{7.13}$$

   defines valid probabilities.

3. **Gradient Stability:** The Kumaraswamy reparameterization yields low-variance gradient estimates as $\alpha_f, \beta_f \to \infty$.

- **PerLiFuse Theorems**

Theorem 7.2.1 (Spectral Fusion Expressivity)**.** *Let $X = \mathcal{F}_{\text{DCT}}(X)$ and $Y = \mathcal{F}_{\text{DCT}}(Y)$ for $X, Y \in \mathbb{R}^d$, where $\mathcal{F}$ denotes Fourier transform, the fusion*

$$Z_f = H_f X_f + (1 - H_f)Y_f + \Omega_f(X_f - Y_f) \tag{7.14}$$

can realize affine transformations of X_f, Y_f with discrepancy control via Ω_f. Special cases:

- *$\Omega_f = 0 \Rightarrow Z_f$ is a convex combination.*
- *$\Omega_f > 0 \Rightarrow \|Z_f - Y_f\| \geq \|X_f - Y_f\|$*

Theorem 7.2.2 (Collapse Resistance)**.** *The Beta–Liouville prior on a pair* $(V, \ldots)$ *is built so that—even as its hyperparameters* $\alpha_f, \beta_f \to 0$ *become extremely "non-informative", it doesn't drive the marginal distribution to a single spike. Instead, the marginal keeps a 50–50 balance. Whereas Dirichlet(0) prior encourages one to decide a priori that there is exactly one active component. This is often too extreme: typically, we want a "weak" prior that doesn't force a single outcome until the data arrive.*

$$\textit{For Beta-Liouville, } \lim_{\alpha_f, \beta_f \to 0} \mathbb{E}[H_f] = 0.5, \tag{7.15}$$

$$\textit{whereas for Dirichlet, } \lim_{\alpha \to 0} \mathbb{E}[\pi_k] = \delta_k \tag{7.16}$$

where δ_k *is a Dirac spike.*

Theorem 7.2.3 (Variance-Bound for Gradient Estimator)**.** *Using* $V = (1 - U^{1/\alpha_f})^{1/\beta_f}$, $U \sim \text{Uniform}(0, 1)$*:*

$$\text{Var}[\nabla_\theta \mathcal{L}] \leq C \cdot \mathbb{E}[\|\nabla_V f(V)\|^2], \quad C = \mathcal{O}\left(\frac{1}{\alpha_f \beta_f}\right) \tag{7.17}$$

where C *arises from the Jacobian* $\nabla_\theta V = \mathcal{O}(1/(\alpha_f \beta_f))$*.*

- **Time Complexity** Each forward pass of PerLiFuse involves computational costs from the text encoder, image encoder, cross-attention, Discrete Cosine Transform (DCT) or inverse-DCT transforms, and gating networks. The overall time complexity is as follows:

$$\mathcal{O}\Big(B(T^2 D + P^2 D + D^2 + DH)\Big) \tag{7.18}$$

where B, T, P, D, and H denote the batch size, text length, number of image patches, embedding dimension, and hidden MLP size, respectively.

7.3 Experimental Results

7.3.1 Experimental Settings

All results are averaged over 30 independent test-set runs following an exhaustive hyperparameter grid search. We trained with AdamW (weight decay 0.2) at a base learning rate $2e^{-5}$ and a 10-step linear warm-up for up to 20 epochs with a batch size of 8 and a look-ahead of 4, using a fixed

Algorithm 5 Generative process for PerLiFuse

Input: Batch of texts $\{T\}$ and images $\{I\}$

for each sample (T, I) **do**
  $X \leftarrow \mathcal{E}_t(T)$ // Text embedding
  $Y \leftarrow \mathcal{E}_i(I)$ // Image embedding
  // Coherence-guided feature modulation
  $Q \leftarrow W_q X,\ K \leftarrow W_k Y$
  $\hat{Y} \leftarrow \text{CrossAttention}(Q, K, Y)$ // Align image to text
  $D \leftarrow |X - \hat{Y}|,\ P \leftarrow X \odot \hat{Y}$ // Diff & product
  $G \leftarrow \sigma\big(\text{MLP}_g([X, \hat{Y}, D, P])\big)$ // Coherence gate
  $X' \leftarrow X \odot G + X,\ Y' \leftarrow Y \odot G + Y$
  $X \leftarrow \mathbf{U}\, X',\ Y \leftarrow \mathbf{U}\, Y'$ // DCT
for $f = 1$ to D **do**
    $U_f \leftarrow |X_f|,\ E_f \leftarrow |Y_f|,\ W_f \leftarrow X_f - Y_f$
    $(\alpha_f, \beta_f) \leftarrow \text{MLP}_{a,b}(U_f, E_f, W_f)$
    $\gamma_f \leftarrow \text{MLP}_{\gamma}(U_f, E_f, W_f)$
    Draw $r_f \sim \text{Uniform}(0, 1)$
    $V_f \leftarrow \big(1 - r_f^{1/\alpha_f}\big)^{1/\beta_f}$ // Kumaraswamy
    if $f = 1$ **then** $P_f \leftarrow 1$ **else** $P_f \leftarrow \prod_{j=1}^{f-1}(1 - V_j)$
    $H_f \leftarrow V_f \cdot P_f$ // Stick-breaking gate
    $Z_f \leftarrow H_f X_f + (1 - H_f) Y_f + \gamma_f (X_f - Y_f)$
end for
  $Z \leftarrow [Z_1, \ldots, Z_D]$
  $z \leftarrow \mathbf{U}^\top Z$ // Inverse DCT
  $\ell \leftarrow \mathcal{C}(z)$ // Classifier projection
  $\hat{s} \leftarrow \sigma(\ell)$ // Prediction probability
end for
return $\{\hat{s}\}$

64-dimensional fused latent space. Experiments were run on a 12th Gen Intel i7-12700K GPU with 64 GB RAM under a 64-bit operating system.

Table 7.1: Performance comparison across Twitter and Weibo datasets, the second-best results are underlined.

Dataset	Model	Acc.	Fake			Real		
			P	R	F1	P	R	F1
Twitter	SpotFake+	0.790	0.786	0.747	0.766	0.793	0.827	0.810
	Dirichlet	0.824	0.772	0.918	0.838	0.899	0.730	0.806
	Gaussian	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	FCINet	0.908	0.828	0.913	0.868	0.955	0.907	0.930
	BMR	0.883	0.927	0.746	0.827	0.865	0.965	0.912
	MRML	0.803	0.821	0.844	0.832	0.777	0.747	0.762
	MMFND	0.896	0.892	0.912	0.902	0.901	0.878	0.889
	ERIC-FND	0.945	0.987	0.910	0.947	0.905	0.986	0.944
	MHR	0.950	0.973	0.930	0.951	0.927	0.972	0.949
	QMFND	0.918	0.880	0.970	0.920	0.970	0.870	0.910
	BMLHF	0.966	0.985	0.933	0.957	0.948	0.975	0.956
	PerLiFuse	**0.987**	**0.989**	**0.987**	**0.988**	**0.982**	**0.988**	**0.985**
Weibo	SpotFake+	0.870	0.855	0.892	0.873	0.769	0.807	0.787
	Dirichlet	0.888	0.900	0.872	0.886	0.877	0.904	0.890
	Gaussian	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	FCINet	0.926	0.938	0.917	0.927	0.913	0.935	0.924
	BMR	0.918	0.882	0.948	0.914	0.942	0.870	0.904
	MRML	0.897	0.898	0.887	0.892	0.896	0.905	0.901
	MMFND	0.935	0.930	0.941	0.935	0.940	0.929	0.934
	ERIC-FND	0.946	0.985	0.914	0.948	0.908	0.984	0.944
	MHR	0.933	0.951	0.921	0.936	0.918	0.949	0.933
	QMFND	0.869	0.900	0.810	0.850	0.840	0.920	0.880
	BMLHF	0.912	0.930	0.880	0.903	0.894	0.920	0.902
	PerLiFuse	**0.979**	**0.989**	**0.987**	**0.987**	**0.974**	**0.988**	**0.981**

7.3.2 Performance Comparison of PerLiFuse Against Baselines

To concisely present our results in the tables, we denote accuracy, precision, recall, and F1-score as Acc., P, R, and F1, respectively. All our results are based on test sets. Please note that we use

the implementation provided by the baseline sources.

Figure Figure 7.4 illustrates how model performance metrics vary across Beta-Liouville priors (α, left) and shape parameters (β, right). These plots clearly highlight sensitivity trends, such as decreasing accuracy with higher α values and performance peaks on β, revealing optimal regions for prior and distribution shape tuning. For the rest of our experiments, we set $\alpha = 0.01$ and $\beta = 0.7$ to avoid overfitting. Table Table 7.1 presents a comprehensive evaluation of PerLiFuse against state-of-the-art multimodal fake news detection methods on Twitter and Weibo datasets. The results demonstrate that PerLiFuse establishes new performance benchmarks, significantly outperforming all baselines across both datasets and all evaluation metrics. On Twitter, PerLiFuse achieves a remarkable 98.7% accuracy, representing a 2.1% absolute improvement over the strongest baseline (BMLHF at 96.6%) and a 61.8% reduction in error rate. This performance advantage is also evident on the Weibo dataset, where PerLiFuse attains 97.9% accuracy, a 3.3% absolute improvement over ERIC-FND (94.6%) that corresponds to a 61.1% reduction in error rate. Note that the error rate can be computed as $\left(1 - \frac{1-\text{Acc}_{\text{PerLiFuse}}}{1-\text{Acc}_{Baseline}}\right) \times 100$. The F1-score analysis reveals consistent superiority in both fake news detection and real news identification. For fake news classification on Twitter, PerLiFuse achieves 98.8% F1, outperforming BMLHF (95.7%) by 3.1% absolute improvement. Similarly, on Weibo, it reaches 98.7% F1 for fake news detection, surpassing ERIC-FND (94.8%) by 3.9%. For real news identification, PerLiFuse maintains 98.5% F1 on Twitter and 98.1% F1 on Weibo, representing 2.9% and 3.7% absolute improvements, respectively, over the best alternatives. Interestingly, PerLiFuse maintains a balance between precision and recall (all exceeding 98.5%), while baselines exhibit significant imbalances. For example, Dirichlet suffers an 18.8% recall drop on Twitter real news, and BMR shows a 10.6% precision-recall gap on Weibo fake news.

PerLiFuse demonstrates particular advantages over specific fusion paradigms. Compared to the probabilistic fusion methods (Dirichlet/Gaussian), it achieves 15.0–17.8% higher F1 by overcoming negative-correlation bias through Beta-Liouville priors. Against attention-based approaches (FCINet/MMFND), PerLiFuse gains 12.0% in F1 by replacing spatial attention with spectral decomposition of cross-modal interactions. When compared to graph-based methods (BMR/MHR), it shows 6.1% accuracy improvement through covariance-aware frequency gating that captures subtle artifact relationships. The model also outperforms quantum-inspired approaches (QMFND) by 7.7% in precision, attributable to its explicit modelling of high-frequency tampering cues that quantum methods often obscure through entanglement.

This cross-platform generalization significantly surpasses dataset-specific models like MRML,

Figure 7.2: Performance evaluation of PerLiFuse, BMLHF, and ERIC-FND on a real-world news snippet:

Sourcehttps://www.cp24.com/news/2025/06/06/toronto-man-charged-with-assault-threatening-to-shoot-dog-police/

which shows a 15.6% F1 degradation on Weibo compared to its Twitter performance. These results validate the core theoretical innovations of PerLiFuse: 1) Spectral separation enables independent processing of low-frequency semantic coherence and high-frequency manipulation artifacts; 2) Beta-Liouville priors prevent fusion collapse while modelling gate interdependencies; 3) Cross-modal residual connections amplify discriminative discrepancies that simpler fusion mechanisms overlook. The consistent outperformance across both datasets confirms that modelling frequency-dependent interactions is essential for detecting sophisticated multimodal disinformation where textual deception and visual forgery manifest in orthogonal spectral bands.

7.3.2.1 Evaluation on Real-World News Snippet

: We present a critical case study that demonstrates PerLiFuse's superiority in detecting complex multimodal manipulation. We evaluate PerLiFuse and leading baselines in several real-world news snippets with mismatched images, and the results, shown in Figure Figure 7.2, are strikingly consistent. All three models correctly classified the original article as authentic, but when a fake image

was paired with the text, BMLHF and ERIC-FND still labelled it 'real,' while PerLiFuse flagged it as 'fake.' By virtue of its coherence-guided fusion mechanism, PerLiFuse is uniquely sensitive to inter-modal misalignment, highlighting the necessity of explicit coherence modelling for robust, out-of-context manipulation detection.

Furthermore, Figure 7.3 provides another comparative visualization of how PerLiFuse, BMLHF, and ERIC-FND respond to real-world misinformation scenarios involving multimodal content. In the first row, the input consists of entirely unaltered content: the text and the image are both authentic and drawn from verified sources. All three models, PerLiFuse, BMLHF, and ERIC-FND, successfully identify the news as real. This result establishes a baseline of competence under standard conditions, where no adversarial noise is present. The second row introduces a more subtle challenge: the image remains real, but the accompanying text has been fabricated. This tests the models' ability to reason beyond visual fidelity and examine the veracity of the narrative itself. In this case, only PerLiFuse correctly flags the manipulated text as fake. BMLHF and ERIC-FND fail to detect the inconsistency, incorrectly classifying the snippet as genuine. In the third row, we reverse the manipulation, this time pairing real text with a misleading image. Again, PerLiFuse is able to detect the inconsistency and correctly classify the snippet as fake. However, both BMLHF and ERIC-FND are again misled by the presence of real text and fail to flag the manipulated visual component. This indicates a tendency in the baselines to over-trust either modality when it appears coherent in isolation, rather than reasoning jointly across modalities. These findings make PerLiFuse particularly well-suited for real-world applications where adversarial actors may selectively tamper with only one modality in an attempt to evade detection.

7.3.3 Comparison PerLiFuse with LLMs on Fakeddit

To further validate PerLiFuse under real-world scale conditions, we evaluate its performance on Fakeddit, a large multimodal dataset with over a million samples and diverse linguistic-visual manipulations. Table Table 7.2 compares PerLiFuse with recent large language models (LLMs) and our top-performing multimodal baselines. PerLiFuse achieves an accuracy of 96.2%, setting a new performance benchmark and surpassing the strongest baseline, BMLHF (95.0%), by 1.2% in absolute accuracy. This translates to a 24.0% reduction in error rate, affirming PerLiFuse's effectiveness even in near-saturation regimes. Against the best LLM (GPT-4V CoT at 75.4%), PerLiFuse delivers a 20.8% absolute gain in accuracy, highlighting the limits of prompting-based strategies in complex multimodal detection tasks.

Figure 7.3: Performance evaluation of PerLiFuse, BMLHF, and ERIC-FND on a real-world news snippet:

Sourcehttps://www.bbc.com/news/articles/c5yg98rl717o

The F1-score comparison demonstrates PerLiFuse's superiority in both fake and real news detection. For fake news, it achieves an F1 of 97.3%, outperforming BMLHF (95.0%) by 2.3% and GAMED (94.9%) by 2.4%. For real news classification, PerLiFuse achieves an F1 of 94.9%, nearly matching BMLHF (95.0%) and outperforming QMFND (94.0%). This balanced performance is reinforced by precision and recall both exceeding 96.0% for fake news, indicating minimal bias toward either false positives or false negatives. In contrast, instruction-tuned LLMs such as GPT-4V, GPT-4, and LLaVA exhibit strong variability across reasoning modes. Their fake news recall drops sharply (e.g., 51.3% for GPT-4V CoT and 40.0% for LLaVA CoT), suggesting that visual-textual deception remains a major blind spot in instruction-following architectures. Even with CoT prompting, hallucination effects and misalignment between modalities hinder consistent prediction.

These results emphasize that PerLiFuse not only outperformed transformer-based baselines like QMFND and BMR but also maintains a significant lead over LLM-based systems across all F1 sub-metrics. Its core design, leveraging spectral decomposition, Beta-Liouville priors, and cross-

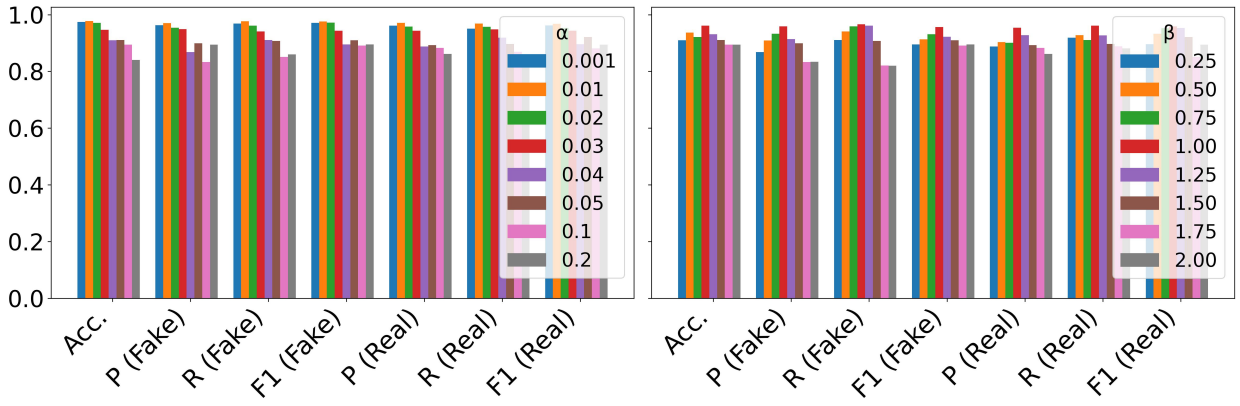


Figure 7.4: Performance evaluation of PerLiFuse on varying Beta-Liouville priors, α and β on the Twitter dataset.

Table 7.2: Performance of PerLiFuse against LLMs on the Fakeddit dataset; the second-best results are underlined.

Dataset	Model	Acc.	Fake			Real		
			P	R	F1	P	R	F1
Fakeddit	LLaVA (Dir.)	0.663	0.588	0.797	0.677	0.777	0.558	0.649
	LLaVA (CoT)	0.673	0.612	0.400	0.484	0.694	0.843	0.761
	GPT-4 (Dir.)	0.677	0.598	0.771	0.674	0.776	0.606	0.680
	GPT-4 (CoT)	0.691	0.662	0.573	0.614	0.708	0.779	0.742
	GPT-4V (Dir.)	0.734	0.673	0.723	0.697	0.771	0.742	0.764
	GPT-4V (CoT)	0.754	0.858	0.513	0.642	0.720	0.937	0.814
	FacTool	0.506	0.476	0.834	0.606	0.624	0.232	0.339
	InstructBLIP	0.726	0.760	0.489	0.595	0.715	0.892	0.793
	LEMMA	0.824	0.835	0.727	0.777	0.818	0.895	0.854
	GAMED	0.939	<u>0.954</u>	0.944	0.949	0.917	0.930	0.923
	BMR	0.901	0.890	0.910	0.891	0.910	0.890	0.891
	QMFND	0.942	0.930	0.950	0.940	0.950	0.930	0.940
	BMLHF	<u>0.950</u>	0.945	<u>0.955</u>	<u>0.950</u>	<u>0.955</u>	<u>0.945</u>	**0.950**
	PerLiFuse	**0.962**	**0.961**	**0.965**	**0.973**	**0.965**	**0.954**	<u>0.949</u>

modal residual gating, enables robust generalization even under scale, noise, and reasoning diversity. Fakeddit thus validates PerLiFuse's scalability and its theoretical foundations under unconstrained, real-world deployment scenarios.

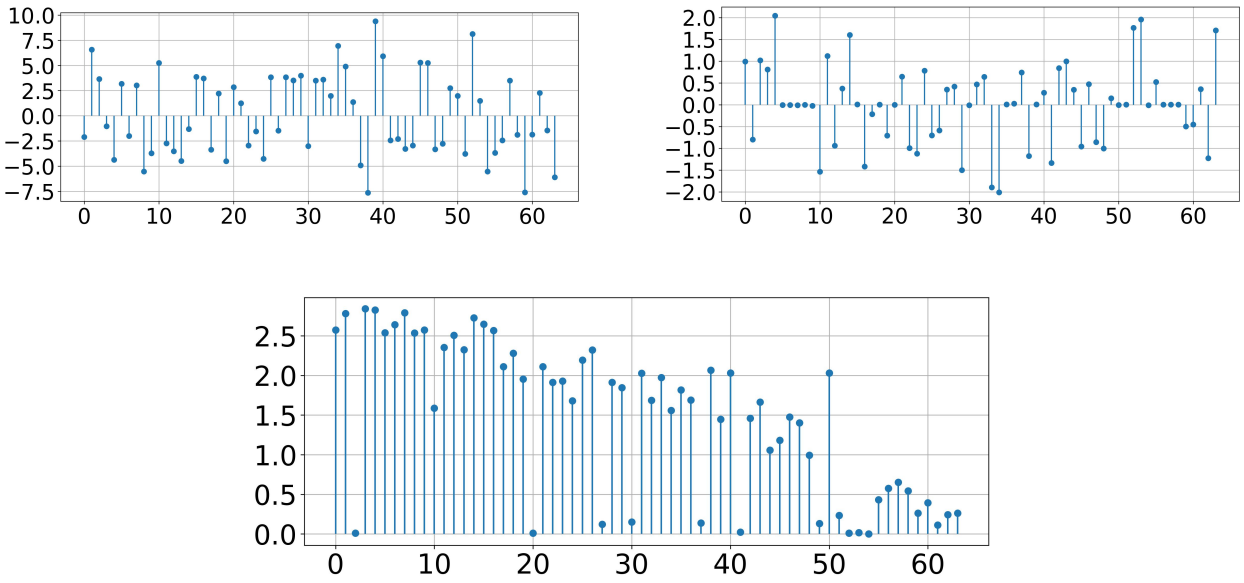


Figure 7.5: Weight distribution across latent dimensions for (a) PerLiFuse (b) Gaussian (c) Dirichlet on the Twitter dataset.

7.3.4 Ablation Studies

To investigate the effectiveness of Beta-Liouville latent space, we replace Beta-Liouville with Gaussian, codenamed PerGaFuse, and Dirichlet, codenamed PerDiFuse. Figure Figure 7.5 presents the average weight distributions on the Twitter test set. PerLiFuse's average weights exhibit an exceptionally wide dynamic range, spanning approximately –7.6 to +9.4 with a standard deviation of 4.3. This broad dispersion indicates that the model learns to both amplify and attenuate individual frequency bands in response to cross-modal agreements or conflicts, effectively highlighting the most salient spectral cues. In contrast, the Gaussian fusion baseline (PerGaFuse) centers nearly all weights around zero, with values confined roughly between –2.0 and +2.0 and a standard deviation of only 0.9. In addition, many of the components are collapsed, limiting the model's ability to distinguish fine-grained frequency patterns. The Dirichlet fusion approach yields an almost uniform, tightly clustered set of positive weights with a standard deviation of merely 0.003. Dirichlet weights have a skewness of approximately –0.32, indicating a slight negative skewness and a small number of component collapses. We suspect that this is due to the initial assumption of Dirichlet that all variables are negatively correlated, limiting its ability to learn complex patterns effectively. Note that we compute skewness $= (n/(n-1)(n-2)) \sum_{i=1}^{n} (x_i - \bar{x}/s)^3$. Taken together, these

Table 7.3: Ablation studies of PerLiFuse's components on Twitter dataset. Acc., P, R, and F1 denote accuracy, precision, recall, and F1-score, respectively.

		Fake			Real			Model	Time
Model	Acc.	P	R	F1	P	R	F1	FLOPs	(ms)
PerLiFuse$_V^B$	0.963	0.942	0.953	0.947	0.948	0.953	0.951	183	0.95
PerLiFuse$_V^D$	0.974	0.956	0.972	0.964	0.962	0.966	0.964	167	0.84
PerLiFuse$_V^R$	0.970	0.967	0.970	0.968	0.962	0.964	0.963	182	0.92
PerLiFuse$_C^M$	0.967	0.954	0.955	0.954	0.960	0.953	0.956	172	0.86
PerLiFuse$_{VG}^M$	0.922	0.913	0.904	0.908	0.927	0.912	0.919	157	0.77
PerLiFuse^{-CG}	0.954	0.953	0.923	0.938	0.962	0.942	0.951	181	0.90
PerLiFuse^{-FS}	0.802	0.831	0.847	0.839	0.754	0.732	0.743	177	0.89
PerLiFuse	0.987	0.989	0.987	0.988	0.982	0.988	0.985	182	0.92

comparisons illustrate that the PerLiFuse's mechanism under a Beta–Liouville prior produces a weight distribution that guarantees the network the flexibility to target the most discriminative spectral features.

The ablation results in Table Table 7.3 confirm that each component of PerLiFuse is essential to its overall performance. The full model (ModernBERT + ViT) achieves the highest accuracy (98.7%) and fake F1 (98.8%). The DistilBERT achieved an accuracy of 97.4% but with a reduced fake recall, while RoBERTa introduces a precision-recall imbalance. On the visual side, ConvNeXt maintains strong precision but lacks ViT's global attention, leading to lower accuracy; VGG19 performs significantly worse due to weak spatial modeling. Removing the coherence guide causes moderate performance drops, while eliminating spectral fusion results in severe accuracy (–18.5%) and fake F1 (–14.9%) losses. These effects highlight the critical importance of coherence guidance for semantic alignment and spectral fusion for detecting high-frequency manipulation. Overall, PerLiFuse's components function synergistically, and its spectral module proves foundational for robust multimodal disinformation detection. Replacing ViT with lighter backbones like ConvNeXt or VGG19 reduces FLOPs and time per sample but significantly harms accuracy, while removing coherence guidance or spectral fusion yields minimal savings at the cost of major performance drops, underscoring that PerLiFuse's full architecture is essential for balancing computational efficiency with robust fake-news detection.

7.3.5 Limitation

While PerLiFuse excels at per-frequency spectral fusion and coherence-guided alignment, it currently operates on static text-image pairs and does not leverage temporal or user-engagement comments that characterize real-world misinformation campaigns. Incorporating such sequential and interaction data into its Bayesian fusion framework could further bolster robustness against coordinated misinformation, suggesting an important avenue for future extension.

7.4 Conclusion

In this work, we have introduced PerLiFuse, a novel Bayesian spectral fusion framework that leverages a Beta–Liouville prior and an orthonormal Discrete Cosine Transform to reconcile text and image signals for fake news detection. By dynamically learning per-frequency fusion gates alongside a coherence-guided cross-attention mechanism, PerLiFuse selectively amplifies salient spectral cues while suppressing contradictory information, thereby preventing trivial fusion or component collapse. Extensive experiments on Twitter, Weibo, and Fakeddit benchmarks demonstrate that PerLiFuse consistently outperforms state-of-the-art baselines, including convolutional fusion, Dirichlet and Gaussian-based approaches, and leading recent multimodal large-language models. Our ablation studies further confirm that every architectural element is critical: the coherence guide detects cross-modal misalignment, and the spectral fusion module is indispensable for high-frequency artifact detection. A case study on manipulated real-world news highlights PerLiFuse's unique ability to flag out-of-context imagery that defeats other top methods. Looking ahead, integrating temporal dynamics and user engagement signals into PerLiFuse's Bayesian fusion could further enhance robustness against coordinated disinformation campaigns. All codes are publicly available to facilitate future research.

Chapter 8

Conclusion

This dissertation presents a unified framework that integrates advanced probabilistic modeling with modern deep learning to address fundamental challenges in topic modeling, multimodal representation learning, and robust misinformation detection. Central to this work is the hypothesis that expressive probabilistic priors are essential for capturing uncertainty, complex dependencies, and heterogeneous structure in real-world data, particularly in multimodal and cross-domain environments where deterministic or overly restrictive assumptions often fail. The Generalized Dirichlet Variational Autoencoder (GD-VAE) establishes the theoretical foundation of this framework by extending neural topic modeling beyond standard Dirichlet assumptions. Through its flexible covariance structure and rejection-sampling-based variational inference, GD-VAE captures both positive and negative topic correlations, yielding substantial improvements in perplexity, topic diversity, and uniqueness across benchmark corpora. Building on this foundation, smoothed Dirichlet priors are introduced to stabilize training and prevent component collapse in sparse regimes, leading to the SmoothDetector architecture for multimodal fake news detection. By integrating probabilistic latent modeling with strong textual and visual encoders, SmoothDetector demonstrates robust generalization across heterogeneous social media datasets. This probabilistic paradigm is further extended to large-scale document classification through SD-MoBERT, which unifies smoothed Dirichlet neural topic modeling with long-context transformer architectures. A dynamic co-attention mechanism aligns thematic latent representations with contextual embeddings, resulting in statistically significant gains in accuracy, convergence stability, and interpretability across multiple benchmark corpora. Together, these models demonstrate that probabilistic latent structure can effectively regularize and enhance modern deep neural architectures without compromising scalability. Beyond

in-domain learning, this dissertation addresses the critical challenge of real-world deployment under distribution shift. EviDA introduces uncertainty-weighted domain adversarial learning, leveraging evidential deep learning to quantify epistemic uncertainty and adaptively regulate instance-level domain alignment. By explicitly accounting for heterogeneous domain shifts, EviDA achieves improved robustness in both cross-domain and cross-lingual fake news detection scenarios, highlighting the role of uncertainty not merely as an error signal but as a principled control mechanism for adaptation. Finally, PerLiFuse advances multimodal fusion by shifting the focus from representation alignment to fusion mechanics in the spectral domain. Through per-frequency Beta-Liouville gating, PerLiFuse dynamically reconciles conflicting text-image cues by decoupling low-frequency semantic alignment from high-frequency manipulation signals. The use of bounded probabilistic priors and low-variance reparameterization mitigates fusion collapse and enables fine-grained, example-specific fusion policies, achieving superior stability and performance compared to existing fusion strategies. While this dissertation establishes a comprehensive framework for probabilistic deep learning in multimodal settings, several promising avenues remain for future exploration. On the theoretical front, the probabilistic priors introduced in this work, generalized Dirichlet, smoothed Dirichlet, and Beta-Liouville, represent only a subset of the rich family of flexible distributions on the simplex. Future research could investigate alternative bounded and unbounded priors, such as logistic-normal distributions, Dirichlet-multinomial hierarchies, or stick-breaking constructions with alternative base measures. Developing tighter variational bounds and more expressive approximate posteriors, potentially through normalizing flows or implicit variational families, could further improve inference quality and model expressiveness. Establishing formal theoretical guarantees on convergence, identifiability, and generalization under these flexible priors would strengthen the theoretical foundation of this framework. From a practical perspective, extending these approaches to web-scale multimodal corpora presents both computational and modeling challenges. Future work could explore efficient approximation strategies, such as amortized inference with neural samplers, mini-batch variational inference with variance reduction techniques, or distributed training protocols that preserve the coherence of probabilistic latent structures across data shards. Investigating the integration of pre-trained foundation models such as CLIP, Flamingo, or GPT-4 Vision with probabilistic latent modeling could leverage the representational power of large-scale pretraining while maintaining interpretability and uncertainty quantification. Real-world misinformation detection systems must adapt continuously as new manipulation techniques, emerging topics, and evolving narrative structures appear. Extending the uncertainty-aware domain adaptation frame-

work to continual learning settings represents a natural next step, involving memory-efficient mechanisms for updating probabilistic priors incrementally, preventing catastrophic forgetting through uncertainty-weighted replay strategies, or maintaining dynamic topic hierarchies that evolve over time. Investigating how epistemic uncertainty can guide selective forgetting and knowledge consolidation in lifelong learning scenarios would enhance the long-term deployability of these models. Moreover, while this work emphasizes interpretability through topic modeling and uncertainty quantification, more explicit mechanisms for human-in-the-loop verification and explanation generation remain underexplored. Future research could investigate generating natural language explanations from probabilistic latent structures, developing interactive interfaces for exploring topic-document associations and cross-modal alignments, or designing counterfactual reasoning frameworks that leverage the generative capacity of variational autoencoders. Enabling domain experts, such as fact-checkers, journalists, or content moderators, to query, validate, and refine model predictions through interpretable probabilistic representations would enhance trust and facilitate responsible deployment. As misinformation detection systems are deployed at scale, adversarial actors may craft inputs specifically designed to evade detection. Investigating the adversarial robustness of probabilistic models, particularly under adaptive attacks that exploit the generative structure of VAEs or the fusion mechanisms of multimodal architectures, represents an important research direction. Developing certified robustness guarantees through probabilistic verification, adversarial training strategies that preserve uncertainty calibration, or defensive distillation techniques tailored to multimodal settings could enhance resilience against targeted manipulation. Additionally, while this dissertation demonstrates promising results in cross-lingual settings, extending this framework to truly low-resource languages, where labeled data, linguistic tools, and even pre-trained embeddings may be scarce, poses significant challenges. Future work could explore zero-shot transfer learning through multilingual probabilistic topic models, leveraging typological similarities across language families, or developing data-efficient active learning strategies guided by epistemic uncertainty to prioritize annotation in underrepresented languages. The probabilistic models developed in this dissertation capture correlations and dependencies but do not explicitly model causal relationships. Integrating causal inference frameworks, such as structural causal models, do-calculus, or counterfactual reasoning, with probabilistic deep learning could enable more robust interventions and policy recommendations. For instance, identifying causal pathways through which misinformation spreads, or predicting the effect of fact-checking interventions under different distribution shifts, would provide actionable insights for platform governance and content moderation strategies. Furthermore,

while this dissertation focuses primarily on misinformation detection, the probabilistic framework introduced here is broadly applicable to other domains requiring interpretable, uncertainty-aware multimodal understanding. Potential applications include medical diagnosis from multimodal clinical data integrating electronic health records, imaging, and genomic data; legal document analysis combining textual contracts with visual evidence; scientific literature mining that aligns text with figures and equations; and autonomous systems that fuse sensor modalities under distributional uncertainty. Exploring these applications would demonstrate the generalizability and impact of this framework beyond social media analysis. In summary, this dissertation establishes a coherent and extensible probabilistic deep learning paradigm that spans interpretable topic discovery, multimodal fusion, and uncertainty-aware cross-domain adaptation. By harmonizing theoretical innovation with rigorous empirical validation and deployable architectures, this work provides a robust foundation for future advances in trustworthy, scalable, and uncertainty-aware artificial intelligence systems operating in complex and evolving digital ecosystems. The future directions outlined above represent not merely incremental extensions, but rather pathways toward a more principled, transparent, and responsible integration of probabilistic reasoning with modern deep learning, an integration that is essential for building AI systems capable of navigating the uncertainties, adversities, and opportunities of real-world deployment.

Bibliography

[1] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *International Conference on Learning Representations:1703.01488*, 2017.

[2] Theophilos Cacoullos. *Discriminant analysis and applications*. Academic Press, 2014.

[3] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 2008–2026, 2018.

[4] William L Dunn and J Kenneth Shultis. *Exploring monte carlo methods*. Elsevier, 2022.

[5] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR, 2014.

[6] Wenqie Huang, Guanghui Yan, Wenwen Chang, Yuchan Zhang, and Yueting Yuan. Eeg-based classification combining bayesian convolutional neural networks with recurrence plot for motor movement/imagery. *Pattern Recognition*, 144:109838, 2023.

[7] Romany F Mansour, José Escorcia-Gutierrez, Margarita Gamarra, Deepak Gupta, Oscar Castillo, and Sachin Kumar. Unsupervised deep learning based variational autoencoder model for covid-19 diagnosis and classification. *Pattern Recognition Letters*, 151:267–274, 2021.

[8] Shuo Li, Fang Liu, Zehua Hao, Licheng Jiao, Xu Liu, and Yuwei Guo. Minent: Minimum entropy for self-supervised representation learning. *Pattern Recognition*, 138:109364, 2023.

[9] Jen-Tzung Chien and Su-Ting Chang. Bayesian asymmetric quantized neural networks. *Pattern Recognition*, 139:109463, 2023.

[10] Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

[11] David Zoltowski, Diana Cai, and Ryan P Adams. Slice sampling reparameterization gradients. *Advances in Neural Information Processing Systems*, 34:23532–23544, 2021.

[12] Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. Whai: Weibull hybrid autoencoding inference for deep topic modeling. *International Conference on Learning Representations:1803.01328*, 2018.

[13] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020.

[14] Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27, 2019.

[15] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.

[16] Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):1–20, 2021.

[17] Tzu-Tsung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

[18] Tom Minka. The dirichlet-tree distribution. *Paper available online at: http://www. stat. cmu. edu/minka/papers/dirichlet/minka-dirtree. pdf*, 1999.

[19] Karla L Caballero, Joel Barajas, and Ram Akella. The generalized dirichlet distribution in enhanced topic detection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 773–782, 2012.

[20] 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.

[21] Ali Shojaee Bakhtiari and Nizar Bouguila. A latent beta-liouville allocation model. *Expert Systems with Applications*, 45:260–272, 2016.

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

[23] Samar Hannachi, Fatma Najar, and Nizar Bouguila. Short text clustering using generalized dirichlet multinomial mixture model. In *Asian Conference on Intelligent Information and Database Systems*, pages 149–161. Springer, 2021.

[24] Geoffrey E Hinton and Russ R Salakhutdinov. Replicated softmax: an undirected topic model. *Advances in neural information processing systems*, 22, 2009.

[25] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. *Advances in Neural Information Processing Systems*, 25, 2012.

[26] Kang Xu, Xiaoqiu Lu, Yuan-fang Li, Tongtong Wu, Guilin Qi, Ning Ye, Dong Wang, and Zheng Zhou. Neural topic modeling with deep mutual information estimation. *Big Data Research*, 30:100344, 2022.

[27] Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural topic modeling with bidirectional adversarial training. *Association for Computational Linguistics:2004.12331*, 2020.

[28] Wenchao Weng, Jin Fan, Huifeng Wu, Yujie Hu, Hao Tian, Fu Zhu, and Jia Wu. A decomposition dynamic graph convolutional recurrent network for traffic forecasting. *Pattern Recognition*, 142:109670, 2023.

[29] Hao Liu, Jingsheng Gao, Suncheng Xiang, Ting Liu, and Yuzhuo Fu. Sae-ntm: Sentence-aware encoder for neural topic modeling. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 106–111, 2023.

[30] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.

[31] Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. Knowledge graph completion: A review. *Ieee Access*, 8:192435–192456, 2020.

[32] Caixia Jing, Hang Gao, Xinpeng Zhang, Tiegang Gao, and Chuan Zhou. Dpsg: Dynamic propagation social graphs for multi-modal fake news detection. *Information Fusion*, 113: 102595, 2025.

[33] Hoang Long Nguyen, Dang Thinh Vu, and Jason J Jung. Knowledge graph fusion for smart systems: A survey. *Information Fusion*, 61:56–70, 2020.

[34] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. " 8 amazing secrets for getting more clicks": detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[35] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[36] Md Rasel, Ashif Mohammad, Md Abdus Salam, Md Aminul Islam, and Reduanul Bari Shovon. Multi-modal approaches to fake news detection: Text, image, and video analysis. *International Journal of Advanced Engineering Technologies and Innovations*, 1(3):449–475, 2024.

[37] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155, 2022.

[38] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[39] Chaowei Zhang, Ashish Gupta, Christian Kauten, Amit V Deokar, and Xiao Qin. Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research*, 279(3):1036–1052, 2019.

[40] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[41] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[42] Jay Kumar, Junming Shao, Salah Uddin, and Wazir Ali. An online semantic-enhanced dirichlet model for short text stream clustering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 766–776, 2020.

[43] Mohammed E. Almandouh, Mohammed F Alrahmawy, Mohamed Eisa, Mohamed Elhoseny, and AS Tolba. Ensemble based high performance deep learning models for fake news detection. *Scientific Reports*, 14(1):26591, 2024.

[44] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19:22–36, 2017.

[45] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

[46] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2051–2055, 2021.

[47] Jinna Lv, Yuan Gao, Li Li, Lei Shi, and Siyu Li. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37:306, 2025.

[48] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[49] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

[50] Míriam Barrabés, Daniel Mas Montserrat, Margarita Geleta, Xavier Giró-i Nieto, and Alexander Ioannidis. Adversarial learning for feature shift detection and correction. *Advances in Neural Information Processing Systems*, 36:57597–57638, 2023.

[51] W He and Z Jiang. A comprehensive survey on uncertainty quantification for deep learning. *ACM Comput. Surv*, 37(4), 2024.

[52] Jun Shi, Shulan Ruan, Ziqi Zhu, Minfan Zhao, Hong An, Xudong Xue, and Bing Yan. Predictive accuracy-based active learning for medical image segmentation. *arXiv preprint arXiv:2405.00452*, 2024.

[53] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[54] Tao Li and Jinwen Ma. Dirichlet process mixture of gaussian process functional regressions and its variational em algorithm. *Pattern Recognition*, 134:109129, 2023.

[55] Nizar Bouguila and Djemel Ziou. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1716–1731, 2007.

[56] Fatma Najar and Nizar Bouguila. Exact fisher information of generalized dirichlet multinomial distribution for count data modeling. *Information Sciences*, 586:688–703, 2022.

[57] Chong Wang and David Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *Advances in neural information processing systems*, 22, 2009.

[58] Elise Epaillard and Nizar Bouguila. Data-free metrics for dirichlet and generalized dirichlet mixture-based hmms–a practical study. *Pattern Recognition*, 85:207–219, 2019.

[59] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[60] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with wasserstein autoencoders. *Association for Computational Linguistics*, 2019.

[61] Marc Fisher, John Woodrow Cox, and Peter Hermann. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 6:8410–8415, 2016.

[62] Dina ElBoghdady. Market quavers after fake ap tweet says obama was hurt in white house explosions. *The Washington Post*, 23, 2013.

[63] Salman Bin Naeem and Rubina Bhatti. The covid-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239, 2020.

[64] Sravani Yenduri, Vishnu Chalavadi, and C Krishna Mohan. Stip-gcn: Space-time interest points graph convolutional network for action recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[65] AB Athira, Abhishek Tiwari, SD Madhu Kumar, and Anu Mary Chacko. Multimodal data fusion framework for fake news detection. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–4. IEEE, 2022.

[66] Abhishek Mallik and Sanjay Kumar. Word2vec and lstm based deep learning technique for context-free fake news detection. *Multimedia Tools and Applications*, 83(1):919–940, 2024.

[67] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[68] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE, 2019.

[69] Xiaocui Wang and Yue Qi. Multimodal fake news detection technology based on deep learning. In *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 1175–1178. IEEE, 2023.

[70] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[71] Marcos Paulo Silva Gôlo, Mariana Caravanti de Souza, Rafael Geraldeli Rossi, Solange Oliveira Rezende, Bruno Magalhães Nogueira, and Ricardo Marcondes Marcacini. One-class learning for fake news detection through multimodal variational autoencoders. *Engineering Applications of Artificial Intelligence*, 122:106088, 2023.

[72] Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4461–4472, 2023.

[73] Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. Correcting the bias: Mitigating multimodal inconsistency contrastive learning for multimodal fake news

detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2861–2866. IEEE, 2023.

[74] R Udayakumar, Nagendar Yamsani, Sri Lavanya Sajja, Yaragani Ashok Kumar, and KR Lathakumari. Automatic fake news detection on social networks using multimodal approach of bert and resnet110. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, pages 1–5. IEEE, 2023.

[75] Daokang Wang, Wubo Zhang, Wenhuan Wu, and Xiaolei Guo. Soft-label for multi-domain fake news detection. *IEEE Access*, 2023.

[76] Ramji Jaiswal, Upendra Pratap Singh, and Krishna Pratap Singh. Fake news detection using bert-vgg19 multimodal variational autoencoder. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–5. IEEE, 2021.

[77] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[78] Takuya Konishi, Takatomi Kubo, Kazuho Watanabe, and Kazushi Ikeda. Variational bayesian inference algorithms for infinite relational model of network data. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2176–2181, 2014.

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

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

[81] 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.

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

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

[84] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

[85] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[86] Lirong Yin, Lei Wang, Zhuohang Cai, Siyu Lu, Ruiyang Wang, Ahmed AlSanad, Salman A AlQahtani, Xiaobing Chen, Zhengtong Yin, Xiaolu Li, et al. Dpal-bert: A faster and lighter question answering model. *CMES-Computer Modeling in Engineering & Sciences*, 141(1), 2024.

[87] Abdourrahmane M Atto, Rosie R Bisset, and Emmanuel Trouvé. Frames learned by prime convolution layers in a deep learning framework. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3247–3255, 2020.

[88] Huanhuan Chen, Peter Tiňo, and Xin Yao. Efficient probabilistic classification vector machine with incremental basis function selection. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):356–369, 2013.

[89] Qinghao Hu, Yang Yang, Jian Cheng, Zeng-Guang Hou, et al. Adversarial binary mutual learning for semi-supervised deep hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):4110–4124, 2021.

[90] Martin Hofmann and Patrick Mäder. Synaptic scaling—an artificial neural network regularization inspired by nature. *IEEE transactions on neural networks and learning systems*, 33 (7):3094–3108, 2021.

[91] Jie Du, Yanhong Zhou, Peng Liu, Chi-Man Vong, and Tianfu Wang. Parameter-free loss for class-imbalanced deep learning in image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):3234–3240, 2021.

[92] Martha Larson, Mohammad Soleymani, Maria Eskevich, Pavel Serdyukov, Roeland Ordelman, and Gareth Jones. The community and the crowd: Multimedia benchmark dataset development. *IEEE multimedia*, 19(03):15–23, 2012.

[93] King-wa Fu, Chung-hong Chan, and Michael Chau. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE internet computing*, 17(3):42–50, 2013.

[94] Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. Multimodal fake news detection via progressive fusion networks. *Information processing & management*, 60(1): 103120, 2023.

[95] Yan Bai, Yanfeng Liu, and Yongjun Li. Learning frequency-aware cross-modal interaction for multimodal fake news detection. *IEEE Transactions on Computational Social Systems*, 11(5): 6568–6579, 2024.

[96] Jie Wu, Danni Xu, Wenxuan Liu, Joey Zhou, Yew Ong, Siyuan Hu, Hongyuan Zhu, and Zheng Wang. Assess and guide: Multi-modal fake news detection via decision uncertainty. In *Proceedings of the 1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models*, pages 37–44, 2024.

[97] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 5384–5392, 2023.

[98] Jianjun Zhang, Ting Wang, Wing WY Ng, and Witold Pedrycz. Knnens: A k-nearest neighbor ensemble-based method for incremental learning under data stream with emerging new classes. *IEEE transactions on neural networks and learning systems*, 34(11):9520–9527, 2022.

[99] 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.

[100] 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)*, pages 4171–4186, 2019.

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

[102] 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.

[103] 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.

[104] 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.

[105] 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*, page 2305, 2023.

[106] 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.

[107] 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, 2019.

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

[109] 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.

[110] 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*, pages 194–206. Springer, 2019.

[111] 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*, pages 37–46, 2020.

[112] 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*, pages 66–71, 2020.

[113] 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*, pages 285–288, 2012.

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

[115] 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.

[116] 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.

[117] Zeynep H Kilimci and Selim Akyokus. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, 2018(1):7130146, 2018.

[118] 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, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1682–1690, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.152. URL https://aclanthology.org/2020.findings-emnlp.152/.

[119] 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.

[120] 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.

[121] 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.

[122] 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, pages 98–101. IEEE, 2015.

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

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

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

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

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

[128] 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.

[129] 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.

[130] 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.

[131] 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, pages 11165–11173, 2022.

[132] 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.

[133] 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.

[134] 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. *Computation and Language*, 2022.

[135] 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*, pages 812–827, 2025.

[136] 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.

[137] 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.

[138] Jiawei Liu, Jingyi Xie, Yang Wang, and Zheng-Jun Zha. Adaptive texture and spectrum clue mining for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2023.

[139] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. Multimodal multi-image fake news detection. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 647–654. IEEE, 2020.

[140] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58 (5):102610, 2021.

[141] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3205–3212, 2020.

[142] Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. Natural language-centered inference network for multi-modal fake news detection. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 2024.

[143] Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, 2024.

[144] Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 2022.

[145] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020.

[146] Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. Exmulf: An explainable multimodal content-based fake news detection system. In *International Symposium on Foundations and Practice of Security*, pages 177–187. Springer, 2021.

[147] Hongzhen Lv, Wenzhong Yang, Fuyuan Wei, Jiaren Peng, and Haokun Geng. Mdf: A dynamic fusion model for multi-modal fake news detection. *Multimedia (cs.MM); Information Retrieval (cs.IR)*, 2024.

[148] Yimeng Gu, Ignacio Castro, and Gareth Tyson. Detecting multimodal fake news with gated variational autoencoder. In *Proceedings of the 16th ACM Web Science Conference*, pages 129–138, 2024.

[149] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1323–1329. IJCAI/AAAI, 2013.

[150] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *The Thirteenth International Conference on Learning Representations*, 2014.

[151] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[152] Fangfang Shan, Huifang Sun, and Mengyi Wang. Multimodal social media fake news detection based on similarity inference and adversarial networks. *Computers, Materials & Continua*, 79 (1), 2024.

[153] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, volume 18, pages 3834–3840, 2018.

[154] Ang Li, Qiuhong Ke, Xingjun Ma, Haiqin Weng, Zhiyuan Zong, Feng Xue, and Rui Zhang. Noise doesn't lie: Towards universal detection of deep inpainting. *IJCAI*, 2021.

[155] Ye Zhu, Yunan Wang, and Zitong Yu. Multimodal fake news detection: Mfnd dataset and shallow-deep multitask learning. *IJCAI*, 2025.

[156] Maged Nasser, Noreen Izza Arshad, Abdulalem Ali, Hitham Alhussian, Faisal Saeed, Aminu Da'u, and Ibtehal Nafea. A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering*, 26:104752, 2025.

[157] Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911, 2025.

[158] Wenjie Wei, Yanyue Zhang, Jinyan Li, Panfei Liu, and Deyu Zhou. Cross-domain fake news detection based on dual-granularity adversarial training. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9407–9417, 2025.

[159] Angelica Liguori, Francesco Sergio Pisani, Carmela Comito, Massimo Guarascio, and Giuseppe Manco. Breaking domain barriers: mixture of experts for cross-domain fake news detection. *Machine Learning*, 114(8):188, 2025.

[160] Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243, 2017.

[161] Lingwei Wei, Dou Hu, Wei Zhou, Philip S Yu, and Songlin Hu. Structure-adaptive adversarial contrastive learning for multi-domain fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9739–9752, 2025.

[162] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[163] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451, 2020.

[164] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[165] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Computation and Language:1911.03854*, 2019.

[166] Tianlin Zhang, En Yu, Yi Shao, and Jiande Sun. Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection. *Computer Science and Machine Learning*, 2025.

[167] Weihai Lu, Yu Tong, and Zhiqiu Ye. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 559–567, 2025.

[168] Liwen Peng, Songlei Jian, Dongsheng Li, and Siqi Shen. Mrml: Multimodal rumor detection by deep metric learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[169] Jinke Ma, Liyuan Zhang, Yong Liu, and Wei Zhang. Multi-task network guided multimodal fusion for fake news detection. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.

[170] Biwei Cao, Qihang Wu, Jiuxin Cao, Bo Liu, and Jie Gui. External reliable information-enhanced multimodal contrastive learning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 31–39, 2025.

[171] Shanshan Feng, Guoxin Yu, Dawei Liu, Han Hu, Yong Luo, Hui Lin, and Yew-Soon Ong. Mhr: A multi-modal hyperbolic representation framework for fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[172] Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172, 2024.

[173] Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, and Xiao-Yuan Jing. Balanced multi-modal learning with hierarchical fusion for fake news detection. *Pattern Recognition*, 164:111485, 2025.

[174] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

[175] Lingzhi Shen, Yunfei Long, Xiaohao Cai, Imran Razzak, Guanming Chen, Kang Liu, and Shoaib Jameel. Gamed: Knowledge adaptive multi-experts decoupling for multimodal fake news detection. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 586–595, 2025.

[176] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *Computation and Language:2402.11943*, 2024.

[177] Wenliang Dai, Junnan Li, Dongxu Li, and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[178] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *Computation and Language:2307.13528*, 2023.

[179] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[180] Kazuya Kakizaki, Yuto Matsunaga, and Ryo Furukawa. Maft: Multimodal automated fact-checking via textualization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29646–29648, 2025.

[181] J-U Sommer and A Blumen. On the statistics of generalized gaussian structures: Collapse and random external fields. *Journal of Physics A: Mathematical and General*, 28(23):6669, 1995.

[182] Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96–99. IEEE, 2021.

[183] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[184] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.