Breaking the Adversarial Robustness-Performance Trade-off in Text Classification via Manifold Purification

Chenhao Dang[2], **Jing Ma**[1*]

[1]BRAIN, Renmin University of China
[2]China Electronics Technology Group Corporation 15th Research Institute
dangchenhao@std.uestc.edu.cn, majingmady@ruc.edu.cn

Abstract

A persistent challenge in text classification (TC) is that enhancing model robustness against adversarial attacks typically degrades performance on clean data. We argue that this challenge can be resolved by modeling the distribution of clean samples in the encoder's embedding manifold. To this end, we propose the Manifold-Correcting Causal Flow (MC^2F), a two-module system that operates directly on sentence embeddings. A Stratified Riemannian Continuous Normalizing Flow (SR-CNF) learns the density of the clean data manifold. It identifies out-of-distribution embeddings, which are then corrected by a Geodesic Purification Solver. This solver projects adversarial points back onto the learned manifold via the shortest path, restoring a clean, semantically coherent representation. We conducted extensive evaluations on text classification (TC) across three datasets and multiple adversarial attacks. The results demonstrate that our method, MC^2F, not only establishes a new state-of-the-art in adversarial robustness but also fully preserves performance on clean data, even yielding modest gains in Accuracy.

Introduction

The remarkable success of Pre-trained Language Models (PLMs) like BERT across numerous Natural Language Processing (NLP) tasks (Devlin et al. 2019) is shadowed by their pronounced vulnerability to adversarial attacks (Jin et al. 2020; Alzantot et al. 2018). These attacks introduce minor, often semantically imperceptible, perturbations to input text, yet can trigger catastrophic prediction errors in the model, which becomes especially critical in text classification (TC). (Vázquez-Hernández et al. 2024; Goyal et al. 2023; Li et al. 2020; Wang, Wang, and Yang 2021) In response, a wide array of defense mechanisms have been developed to enhance model robustness (Gao et al. 2023; Zheng et al. 2024; Asl et al. 2024; Zhu et al. 2019). However, their success is often marred by a critical challenge: improving resilience against adversarial inputs frequently comes at the cost of a significant drop in performance on clean, unperturbed data (Ning and Chai 2024; Wang, Wang, and Yang 2021). This robustness-accuracy challenge not only poses a significant barrier to the reliable deployment of PLMs in safety-critical applications but also highlights a fundamental gap in our understanding of their internal workings.

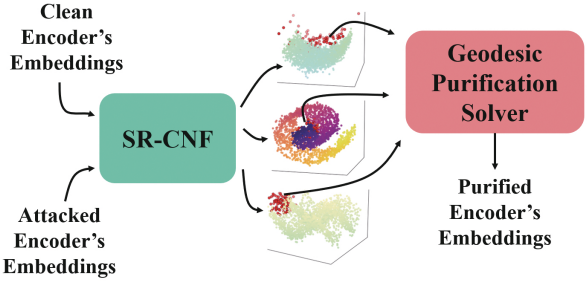

Figure 1: An overview of our proposed Manifold-Correcting Causal Flow (MC^2F) framework. The system first employs a Stratified Riemannian Continuous Normalizing Flow (SR-CNF) to model the manifold of clean data (colored points). It then identifies out-of-distribution (OOD) or adversarial samples (red points) that lie off the manifold. The Purification Solver module subsequently corrects these OOD embeddings by projecting them back onto the clean manifold along the shortest path (geodesic), a process visualized by the black arrows. In-distribution embeddings are passed through without modification.

To overcome this challenge, we move beyond heuristic defenses and investigate the geometric properties of the model's embedding space. We begin with a critical empirical observation: in the high-dimensional embedding space of a BERT-style encoder, representations of clean text are geometrically separable from their adversarial counterparts generated by prevalent attacks (e.g., TextFooler, BERT-Attack) (Subhash et al. 2023; Reif et al. 2019). This suggests that adversarial perturbations induce a measurable geometric and distributional shift, rather than creating indistinguishable overlaps. Consequently, the defense problem can be reframed: instead of a brute-force training task, it becomes a more nuanced geometric challenge of identifying and correcting embeddings that have been perturbed off the "clean" data distribution.

Building on this insight and inspired by the manifold hypothesis—which posits that high-dimensional data lies on a lower-dimensional manifold—we introduce the Manifold-

*Corresponding author.

Correcting Causal Flow (MC^2F). It is a two-module defense system that operates directly on sentence embeddings, embodying a "detect-and-correct" method to purify adversarial inputs (Li, Song, and Qiu 2023; Moraffah et al. 2024).

The first module, a Stratified Riemannian Continuous Normalizing Flow (SR-CNF), is responsible for detection. It learns a probabilistic model of the clean data manifold, leveraging a dynamically learned Riemannian metric to perform highly accurate density estimation. This allows it to effectively distinguish in-distribution ("clean") embeddings from out-of-distribution (OOD) adversarial samples based on a learned likelihood threshold.

For any embedding flagged as adversarial, the second module, a Geodesic Purification Solver, performs the correction. We formalize this process as an optimization problem: finding the shortest path (geodesic) from the anomalous point back to the learned manifold. The solver iteratively finds this path to produce a purified, semantically robust embedding, effectively neutralizing the adversarial perturbation.

The contributions of this paper are summarized as follows:

- We provide strong empirical evidence that the embeddings of clean and adversarial text are geometrically and distributionally separable in the representation space of Pre-trained Language Models. This insight allows us to reframe adversarial defense from a brute-force training problem to a geometric purification task.
- We propose the Manifold-Correcting Causal Flow (MC^2F), a novel and principled framework that operationalizes our geometric insight. MC^2F is the first system to integrate a Stratified Riemannian Continuous Normalizing Flow for out-of-distribution detection with a Geodesic Purification Solver for correction.
- Through extensive experiments across three datasets, two downstream tasks, and multiple attack methods, we demonstrate that MC^2F effectively resolves the long-standing robustness-accuracy trade-off. It achieves state-of-the-art adversarial robustness while fully preserving—and in some cases, even improving—performance on clean data.

Related Works

Our work is positioned at the intersection of adversarial defense in NLP and the geometric analysis of embedding spaces. We review the relevant literature in these areas.

Adversarial Attacks in Text Classification

The vulnerability of PLMs has been extensively demonstrated by a variety of adversarial attack methods. These methods typically generate perturbations by replacing, inserting, or deleting characters or words under certain constraints to maintain semantic similarity and fluency. Prominent word-level attack algorithms include TextFooler (Jin et al. 2020), BAE (Garg and Ramakrishnan 2020), and TextBugger (Li et al. 2018), which greedily replace words with synonyms that maximize the victim model's prediction error. More sophisticated methods, such as BERT-Attack (Li et al. 2020), leverage the victim model itself to find more effective substitutes. These attacks serve as standard benchmarks for evaluating the robustness of defense systems, including our own.

Adversarial Defense Strategies

The dominant paradigm, adversarial training (AT), augments data with adversarial examples (Zhu et al. 2019; Madry et al. 2017) yet suffers from high computational costs and accuracy trade-offs (Ning and Chai 2024; Wang, Wang, and Yang 2021). Despite improvements, an "illusion of robustness" often persists due to gradient masking (Gao et al. 2023; Wang et al. 2024; Raina et al. 2024). Alternatively, purification methods have evolved toward continuous embedding denoising (Zheng et al. 2024; Asl et al. 2024), with recent works like DAD balancing clean and robust accuracy via MMD-based detection (Zhang et al. 2025). A complementary geometric perspective relies on the manifold hypothesis (Fefferman, Mitter, and Narayanan 2016), projecting off-manifold adversarial points back to the data manifold (Minh and Tuan 2022; Yang et al. 2024). Motivated by evidence that off-manifold samples drive robustness while on-manifold data supports generalization (Altinisik et al. 2024; Dang and Zhu 2025), we pursue a formal manifold-based NLP purification framework.

Flow-based Out-of-Distribution Detection

Adversarial detection can be cast as an out-of-distribution (OOD) task, where clean and adversarial samples represent in- and out-distribution data, respectively. Generative models like Normalizing Flows (NFs) (Rezende and Mohamed 2015) and Continuous NFs (CNFs) (Chen et al. 2018) excel at this by assigning low likelihoods to OOD samples (Nalisnick et al. 2018). To overcome limitations in modeling complex manifolds, modern approaches enhance performance by applying flows in feature spaces (Cook, Lavoie, and Waslander 2024) or integrating Riemannian metrics for precise geometric density estimation (Diepeveen et al. 2024).

Preliminary Study

Our defense strategy is predicated on the hypothesis that adversarial perturbations displace text embeddings from a "clean" data manifold into distinct, separable regions of the embedding space. This section presents a rigorous empirical investigation to validate this hypothesis. We analyze the geometric and distributional properties of encoder's embeddings from a BERT-base model (Devlin et al. 2019) fine-tuned on the SST-2 dataset (Socher et al. 2013), comparing clean test set examples against their adversarial counterparts generated by TextFooler (Jin et al. 2020).

Problem Formulation

Let $f_\theta : \mathcal{X} \to \mathcal{Z}$ be a text encoder, such as BERT, mapping an input text $x \in \mathcal{X}$ to an embedding $z \in \mathcal{Z} \subseteq \mathbb{R}^d$. Let P_{clean} denote the probability distribution of embeddings $z = f_\theta(x)$ derived from a clean data distribution. An adversarial attack generates a perturbed input x_{adv} from x, resulting in an embedding $z_{adv} = f_\theta(x_{adv})$. Let P_{adv} be

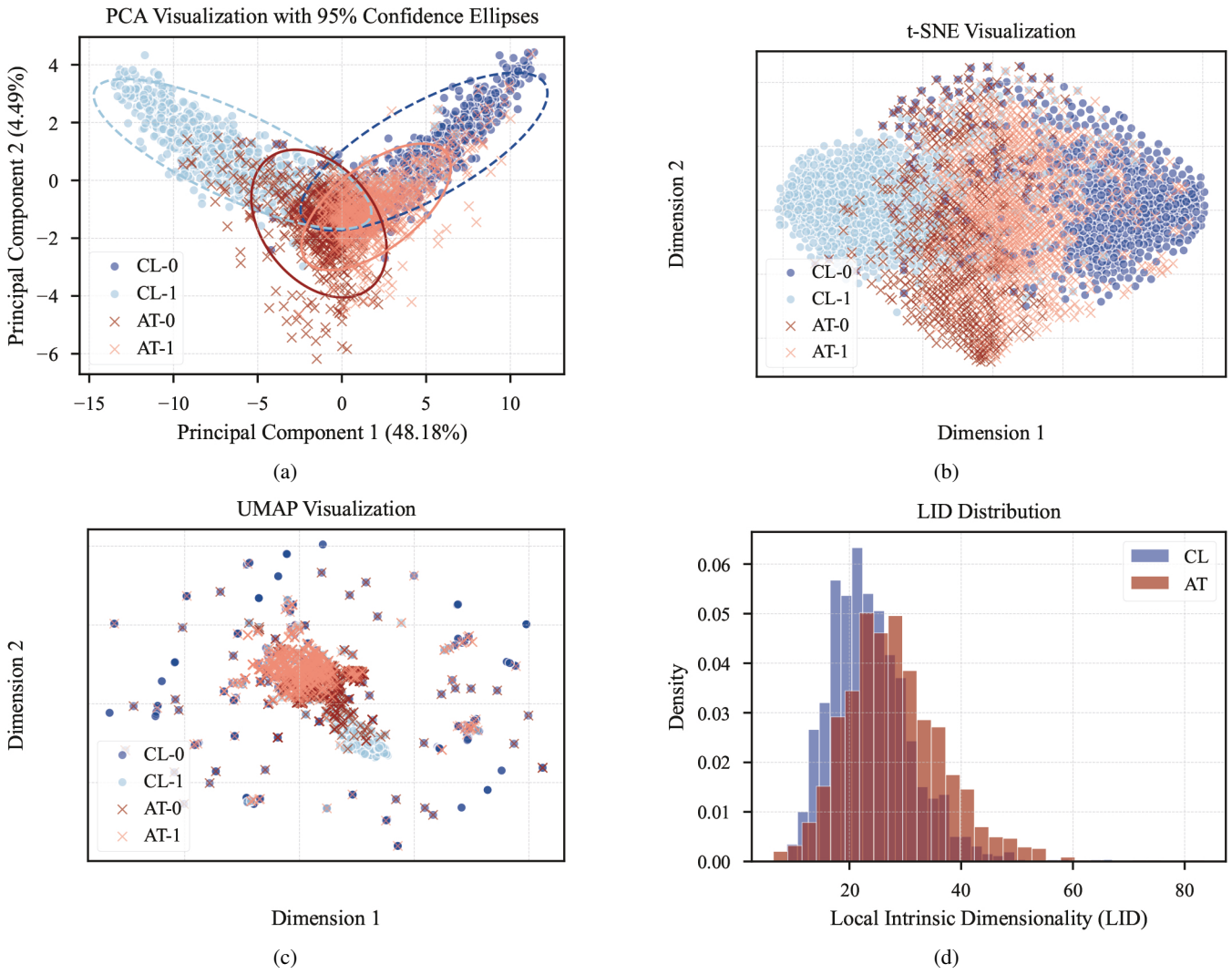

Figure 2: Geometric and distributional analysis of clean and attacked SST-2 test set embeddings. (a-c) 2D visualizations using PCA, t-SNE, and UMAP show that adversarial embeddings form clusters visibly distinct from clean embeddings. (d) The density distribution of Local Intrinsic Dimensionality (LID) values shows that attacked embeddings are visibly shifted towards higher LID values. CL-0/1: Clean samples of class 0/1; AT-0/1: Attacked samples of class 0/1.

the distribution of these adversarial embeddings. Our central premise is that P_{clean} and P_{adv} are significantly different, and more specifically, that their supports lie on separable, low-dimensional manifolds, $\mathcal{M}_{clean}$ and $\mathcal{M}_{adv}$, respectively.

Visual Evidence of Manifold Separation

To first gain a visual intuition, we project the high-dimensional embeddings into a 2D space using three different techniques: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008), and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018). As shown in Figure 2(a-c), all three methods reveal a clear separation between the clean (blue circles) and attacked (red crosses) embeddings. While PCA, a linear projection method, shows a noticeable distributional shift, the non-linear, manifold-aware methods t-SNE and UMAP illustrate a more striking separation. In the UMAP projection (Figure 2c), the clean and attacked points form well-defined, dense clusters with minimal overlap. This strong visual evidence suggests that clean and adversarial embeddings occupy distinct regions in the latent space, consistent with our manifold separation hypothesis.

Quantitative Evidence of Distributional Shift

To move beyond visual intuition, we quantify the dissimilarity between the clean and adversarial embedding distributions using three standard metrics: Maximum Mean Discrepancy (MMD) with a Gaussian kernel (Gretton et al. 2012), Jensen-Shannon Divergence (JSD) (Lin 2002), and the Wasserstein Distance. Table 1 reports these distances.

Comparison	MMD (Gaussian Kernel) ↓	JSD ↓	Wasserstein Distance ↓
Clean In-Distribution (Class 0 Half 1 vs. Half 2)	0.00068	0.00008	19.66
All Clean vs. All Attacked	**0.00254**	**0.00096**	**30.76**
Class 0: Clean vs. Attacked	**0.00353**	**0.00100**	**88.86**
Class 1: Clean vs. Attacked	**0.00183**	**0.00940**	**56.31**

Table 1: Statistical distances between embedding distributions. The distances between clean and attacked distributions are consistently and significantly larger than the in-distribution baseline, providing quantitative proof of a distributional shift caused by adversarial attacks.

For a baseline, we first measure the distance between two random halves of the clean embeddings ("Clean In-Distribution"), which is expectedly low. In stark contrast, the distances between the full set of clean and attacked embeddings ("All Clean vs. All Attacked") are substantially larger across all three metrics. This pattern holds true when examining each class individually. These results provide strong quantitative evidence that adversarial attacks induce a statistically significant distributional shift, confirming that P_{clean} and P_{adv} are indeed different distributions.

Analysis of Local Intrinsic Dimensionality

Finally, we investigate if the geometric structure of the manifolds themselves differs. We use Local Intrinsic Dimensionality (LID) (Houle 2017), which estimates the dimensionality of the data manifold in the local neighborhood of a data point. A higher LID suggests greater local complexity.

Figure 2d shows the density distributions of LID values for both clean and attacked embeddings. A clear rightward shift is visible for the attacked distribution. This observation is statistically validated in Table 2. The mean LID of attacked embeddings (28.20) is significantly higher than that of clean embeddings (23.74). A Welch's t-test confirms this difference is highly statistically significant, with a p-value approaching zero ($p \approx 10^{-43}$). This implies that adversarial perturbations do not just move embeddings to a different location, but systematically push them into regions of the embedding space that are geometrically more complex.

Foundational Hypotheses

Based on this comprehensive empirical evidence and recent advances in the study of embeddings from language models (Li and Sarwate 2025; Mamou et al. 2020), we establish two foundational hypotheses that motivate our proposed method:

1. **Manifold Separability:** The embeddings of clean text and their adversarial counterparts lie on statistically distinct and geometrically separable manifolds within the encoder's representation space.
2. **Stratified Manifold Structure:** The embedding space is not a single, uniform manifold but a stratified space composed of sub-manifolds with varying local geometric properties, such as different intrinsic dimensionalities. Adversarial examples tend to occupy regions of higher intrinsic dimensionality than clean examples.

These hypotheses suggest that an effective defense can be built by first learning the manifold of clean data and then correcting adversarial points by projecting them back onto this learned manifold, which is the core idea behind our proposed MC^2F framework.

Comparison	MLID (CL)	MLID (AT)	p-value
All Classes	23.74	28.20	8.99×10^{-43}
Class 0	23.67	25.66	2.36×10^{-5}
Class 1	21.75	26.90	5.50×10^{-20}

Table 2: Mean LID (MLID) values and Welch's t-test results. The significantly higher mean LID for attacked embeddings and the extremely low p-value suggest that adversarial perturbations move points to regions of higher geometric complexity. CL denotes clean embeddings and AT denotes attacked embeddings.

Method

Building upon our foundational hypotheses, we introduce the Manifold-Correcting Causal Flow (MC^2F). This framework is designed not only to detect adversarial samples by modeling the complex geometry of the clean data manifold, $\mathcal{M}_{clean}$, but also to purify them by projecting them back onto this manifold. MC^2F comprises two core modules: a Stratified Riemannian Continuous Normalizing Flow (SR-CNF) for detection and a Geodesic Purification Solver for correction. The MC^2F framework is summarized in Algorithm 1.

Stratified Riemannian CNF for Detection

To accurately model the probability density of clean embeddings, $p_{clean}(z)$, we must account for the stratified and non-Euclidean nature of the embedding space. We achieve this by defining a Continuous Normalizing Flow (CNF) on a learnable, data-dependent Riemannian manifold.

Learning a Stratified Riemannian Geometry Instead of assuming a fixed geometry, we learn a Riemannian metric tensor $G(z)$, a positive definite matrix that defines the local geometry at each point $z \in \mathcal{Z}$. To capture the stratified structure observed in our preliminary study, we parameterize $G(z)$ using a Mixture-of-Experts (MoE) network (Jacobs et al. 1991; Fedus, Zoph, and Shazeer 2022). This network consists of:

Algorithm 1: MC^2F: Manifold-Correcting Causal Flow

Require: Clean data embeddings Z_{clean}, downstream classifier C, hyperparameters $\lambda_{topo}, \lambda_{causal}$.
Initialize parameters Θ_{flow} for the SR-CNF map f and Θ_{metric} for the MoE metric $G(z)$.
// Training Phase
for each training batch $Z_b \subset Z_{clean}$ **do**
  Get adversarial counterparts Z_{adv} from Z_b.
  Purify Z_{adv} to get Z_{corr} using the Geodesic Purification Solver (Eq. 5) with the current metric $G(z)$.
  Compute $\mathcal{L}_{NLL}$ on Z_b (Eq. 4).
  // Compute latent counterparts by inverting the flow
  Compute $\mathcal{L}_{topo}$ on Z_b and $f^{-1}(Z_b)$ (Eq. 8).
  Compute $\mathcal{L}_{causal}$ between $C(Z_b)$ and $C(Z_{corr})$.
  Compute total loss $\mathcal{L}_{total} = \mathcal{L}_{NLL} + \lambda_{topo}\mathcal{L}_{topo} + \lambda_{causal}\mathcal{L}_{causal}$.
  Update Θ_{flow} and Θ_{metric} via gradient descent.
end for
Determine likelihood threshold τ on a validation set.
// Inference Phase
for each input embedding z_{in} to be classified **do**
  Calculate $\ell = \log p(z_{in})$ using SR-CNF.
  if $\ell < \tau$ **then**
    $z_{out} \leftarrow$ Geodesic Purification Solver(z_{in})
  else
    $z_{out} \leftarrow z_{in}$
  end if
  Obtain final prediction from $C(z_{out})$.
end for

- A gating network $g_\phi(z)$ that takes an embedding z and outputs a set of weights $\{\alpha_k(z)\}_{k=1}^K$, where K is the number of experts and $\sum_k \alpha_k(z) = 1$.
- K expert networks $\{E_{\psi_k}(z)\}_{k=1}^K$, where each expert is a neural network that outputs a matrix, specializing in the local geometry of a specific stratum.

The final metric tensor at point z is a weighted combination of the expert outputs:

$$G(z) = \sum_{k=1}^{K} \alpha_k(z) E_{\psi_k}(z). \tag{1}$$

To ensure $G(z)$ is always positive definite, each expert $E_{\psi_k}(z)$ is constructed as $L_k(z)L_k(z)^T + \epsilon I$, where $L_k(z)$ is the output of the expert network (a lower-triangular matrix) and ϵ is a small positive constant for numerical stability. This MoE structure allows the model to adaptively learn a complex, piece-wise smooth geometry across different semantic regions of the embedding space.

Riemannian Continuous Normalizing Flow With the learned Riemannian manifold $(\mathcal{M}, G)$, we define a flow as the solution to the ordinary differential equation (ODE) governed by a time-varying vector field v:

$$\frac{dz(t)}{dt} = v(z(t), t), \quad z(t_0) \sim \mathcal{N}(0, I), \tag{2}$$

where the dynamics unfold on the manifold $\mathcal{M}$. The change of variable formula for a CNF on a Riemannian manifold relates the change in log-probability to the Riemannian divergence of the vector field:

$$\frac{d \log p(z(t))}{dt} = -\text{div}_G(v(z(t), t)). \tag{3}$$

The total log-likelihood of a data point z_{in} is then obtained by integrating this quantity along the forward trajectory from its latent representation z_0 at $t = t_0$ to the data point z_{in} at $t = t_1$:

$$\log p(z_{in}) = \log p_{\mathcal{N}}(z_0) - \int_{t_0}^{t_1} \text{div}_G(v(z(t), t))dt. \tag{4}$$

Detection Mechanism: An input embedding z_{in} is classified as adversarial if its log-likelihood falls below a threshold τ, i.e., $\log p(z_{in}) < \tau$. The threshold τ is determined on a validation set to achieve a desired false positive rate.

Geodesic Purification Solver

When an embedding z_{adv} is detected as adversarial, the correction module is activated. Its goal is to find a purified point z_{corr} on the clean manifold $\mathcal{M}_{clean}$ that is "closest" to z_{adv}. In the context of our learned Riemannian geometry, this corresponds to finding the orthogonal projection of z_{adv} onto $\mathcal{M}_{clean}$. This projection point lies at the end of a geodesic (the shortest path on the manifold) starting at z_{adv} and ending on $\mathcal{M}_{clean}$.

We formulate this as an optimization problem to find the path $\gamma(t) : [0, 1] \to \mathbb{R}^d$ that minimizes the path energy functional:

$$\mathcal{L}[\gamma] = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle_{G(\gamma(t))} dt, \tag{5}$$

subject to the boundary conditions $\gamma(0) = z_{adv}$ and $\gamma(1) = z_{corr} \in \mathcal{M}_{clean}$. The optimal path γ^* is the geodesic, and its endpoint $\gamma^*(1)$ is our purified embedding z_{corr}.

This optimization problem is solved iteratively. We discretize the path $\gamma(t)$ and use gradient descent on its waypoints to minimize the energy functional $\mathcal{L}[\gamma]$ (Eq. 5), with gradients computed respecting the learned metric $G(z)$ (Eq. 1). The constraint $\gamma(1) \in \mathcal{M}_{clean}$ is operationalized by enforcing $\log p(z_{corr}) \geq \tau$ (where τ is the detection threshold) via a soft penalty term in the minimization objective.

Geometric and Topological Training Paradigm

To effectively train the MC^2F framework, we employ a multi-objective loss function that optimizes for density estimation, topological integrity, and semantic consistency simultaneously. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{NLL} + \lambda_{topo}\mathcal{L}_{topo} + \lambda_{causal}\mathcal{L}_{causal}, \tag{6}$$

where λ_{topo} and λ_{causal} are balancing hyperparameters.

Density Estimation Loss ($\mathcal{L}_{NLL}$) This is the standard negative log-likelihood loss for training normalizing flows. It drives the model to learn the distribution of clean data:

$$\mathcal{L}_{NLL} = -\mathbb{E}_{z \sim P_{clean}}[\log p(z)], \tag{7}$$

where $\log p(z)$ is computed using our SR-CNF (Eq. 4).

Dataset	Method	Clean%	BERT-Attack		TextFooler		TextBugger	
			Aua%	#Query	Aua%	#Query	Aua%	#Query
SST-2	Fine-tune	92.71	3.83	106.4	6.10	90.5	28.70	46.0
	FreeLB	92.01	23.88	174.7	29.40	132.6	49.70	53.8
	WLRE	92.11	29.80	185.4	32.80	138.4	50.10	56.4
	SD	91.36	36.46	201.2	46.30	167.3	54.50	62.3
	MC²F	**92.71**	**40.05**	**289.4**	**52.60**	**184.2**	**61.50**	**98.4**
AGNews	Fine-tune	94.68	4.09	412.9	14.70	306.4	40.00	166.2
	FreeLB	94.99	19.90	581.8	33.20	396.0	52.90	201.1
	WLRE	94.05	28.60	657.1	32.60	368.1	53.40	208.0
	SD	93.81	38.60	744.1	49.30	488.1	60.10	219.7
	MC²F	**95.13**	**45.30**	**892.5**	**53.80**	**561.4**	**64.30**	**299.0**
YELP	Fine-tune	95.19	5.40	116.2	5.20	105.4	29.60	52.6
	FreeLB	95.11	28.24	184.3	28.30	143.5	50.00	68.4
	WLRE	94.86	31.46	208.2	31.50	155.0	50.00	69.1
	SD	93.45	39.61	320.7	47.80	187.6	55.10	100.2
	MC²F	**95.26**	**48.50**	**586.4**	**54.00**	**214.3**	**63.20**	**112.4**

Table 3: Main results on adversarial robustness evaluation. We report accuracy on the clean test set (Clean%) and accuracy under attack (Aua%), along with the average number of queries per successful attack (#Query). Higher is better for all metrics. Our method achieves the best robustness across all datasets and attacks while maintaining high clean accuracy.

Topological Regularization ($\mathcal{L}_{topo}$) To ensure the learned flow (SR-CNF) f preserves the global shape of the data manifold and avoids topological tearing, we introduce a loss based on differentiable persistent homology. We compute the persistence diagrams (PD), topological summaries of data, for a batch of clean embeddings Z_b and their latent counterparts $f^{-1}(Z_b)$. The loss is the Wasserstein distance between these diagrams:

$$\mathcal{L}_{topo} = W_p(\text{PD}(Z_b), \text{PD}(f^{-1}(Z_b))). \tag{8}$$

Minimizing this loss encourages the flow to be a homeomorphism, preserving the topological features of the data manifold.

Causal & Semantic Regularization ($\mathcal{L}_{causal}$) The purification process $z_{adv} \to z_{corr}$ must be semantically consistent. We frame this from a causal perspective, where purification is an intervention to remove the confounding effect of the adversarial perturbation. To enforce this, we use a loss based on information geometry (Volpi, Thakur, and Malago 2021; Kim et al. 2022). For a clean embedding z_{clean} and its corresponding purified version z_{corr} (obtained by first attacking z_{clean} and then purifying it), we pass both through a fixed, pre-trained downstream classifier $C(\cdot)$ to get two softmax probability distributions, p_{clean} and p_{corr}. The causal loss is the Fisher-Rao distance between them:

$$\mathcal{L}_{causal} = 2 \arccos\left(\sum_i \sqrt{p_{clean,i} \cdot p_{corr,i}}\right). \tag{9}$$

This loss directly forces the purified embedding to be semantically indistinguishable from the original clean embedding from the perspective of the downstream task.

Experiments

This section presents a comprehensive set of experiments designed to rigorously evaluate the effectiveness of our proposed MC^2F framework.

Experimental Setup

Datasets. We evaluate our method on three benchmark datasets: SST-2 (Socher et al. 2013) for binary sentiment analysis on short text; AGNews (Jacovi, Shalom, and Goldberg 2018) for four-class news topic classification; and YELP (Asghar 2016), a more complex binary sentiment task involving longer reviews.

Attack Methods. We assess robustness using the TextAttack toolkit (Morris et al. 2020) with three established attacks: BERT-Attack (Li et al. 2020), which leverages BERT for context-aware substitutions; TextFooler (Jin et al. 2020), a greedy strategy using synonym replacement; and TextBugger (Li et al. 2018), which employs both word- and character-level perturbations.

Baselines. We benchmark MC^2F against a standard fine-tuned model and three competitive defense baselines:

- Fine-tune (Devlin et al. 2019): A standard BERT-base model fine-tuned for each downstream task, representing a non-robust baseline.
- FreeLB (Zhu et al. 2019): An enhanced adversarial training method that adds adversarial perturbations to word embeddings within a bounded region to improve generalization and robustness.
- WLRE (Huang et al. 2022): A defense framework that enhances robustness by introducing its own perturbations to the input during training, forcing the model to learn more invariant features.

Method	Clean%	TextFooler	
		Aua%	#Query
Fine-tune	94.68	14.7	306.4
MC²F (Full)	**95.13**	**53.8**	**561.4**
w/o $\mathcal{L}_{NLL}$	93.22	32.6	366.7
w/o $\mathcal{L}_{topo}$	93.41	32.9	375.4
w/o $\mathcal{L}_{causal}$	94.76	48.6	479.1

Table 4: Ablation study on the AGNews dataset against the TextFooler attack. The results demonstrate that each component of our multi-objective loss function is essential for achieving optimal performance.

- SD (Subspace Defense) (Zheng et al. 2024): A state-of-the-art defense that projects feature representations onto a learned "clean" subspace to discard adversarial perturbations.

Evaluation Metrics. We evaluate all methods based on three metrics:

- Clean%: The classification accuracy on the original, unperturbed test set. This metric is crucial for assessing whether a defense method compromises performance on normal data.
- Aua%: The model's accuracy on the adversarially attacked test set. This is the primary metric for evaluating adversarial robustness.
- #Query: The average number of model queries required by the attacker to successfully generate an adversarial example. A higher query count indicates a more resilient model.

Implementation Details. All experiments use BERT-base-uncased as the backbone encoder. For our MC^2F framework, we set the key hyperparameters based on validation performance: the likelihood threshold for detection $\tau = 0.3$, the weight for the topological loss $\lambda_{topo} = 0.8$, and the weight for the causal loss $\lambda_{causal} = 0.05$.

Main Results on Robustness Evaluation

The primary experimental results are summarized in Table 3. The comparison across three datasets and three attack methods reveals several key findings.

First and most importantly, MC^2F consistently and substantially outperforms all baseline methods in robust accuracy (Aua%) across all evaluated datasets and attack scenarios. For instance, under the strong BERT-Attack on the AGNews dataset, MC^2F achieves an Aua of 45.3%, a significant improvement over the next best baseline, SD, which scores 38.6%. This pattern of superior performance holds for TextFooler and TextBugger attacks as well, demonstrating the broad effectiveness of our geometric purification approach.

Second, this significant gain in robustness does not come at the cost of performance on clean data. The Clean% of MC^2F is on par with, and in the cases of AGNews and YELP, even slightly higher than the standard Fine-tune model. This successfully addresses the critical robustness-accuracy trade-off that plagues many existing defense methods like SD, which shows a noticeable drop in clean accuracy.

Third, MC^2F demonstrates heightened resilience as measured by the number of queries (#Query). Attackers consistently require more queries to find successful adversarial examples against our method compared to all baselines. On YELP against BERT-Attack, for example, an attacker needs an average of 586.4 queries for MC^2F, compared to 320.7 for SD. This indicates that our defense makes the model's decision boundaries more stable and harder to exploit.

Ablation Study

To isolate and verify the contribution of each key component in our proposed training paradigm, we conduct an ablation study on the AGNews dataset under the TextFooler attack. We systematically remove each of the three loss terms from our objective function. The results are presented in Table 4.

The analysis reveals that each component is integral to the framework's success. Removing the negative log-likelihood loss ($\mathcal{L}_{NLL}$) results in a catastrophic drop in both clean and robust accuracy. This is expected, as $\mathcal{L}_{NLL}$ is fundamental for learning the clean data distribution, which is the basis for our detection module.

Excluding the topological regularization term ($\mathcal{L}_{topo}$) causes the most significant drop in robust accuracy among the regularization terms, with Aua% falling from 53.8% to 32.9%. This result strongly supports our hypothesis that preserving the global topological structure of the data manifold is critical for preventing the model from learning a brittle representation that is easily exploited by adversaries.

Removing the causal and semantic loss ($\mathcal{L}_{causal}$) also leads to a notable decrease in robustness, with Aua% dropping to 48.6%. This demonstrates the importance of explicitly guiding the purification process to be semantically consistent with the original clean input from the perspective of the downstream classifier.

Conclusion

In this paper, we addressed the persistent trade-off between robustness and accuracy in text classification by proposing a novel defense rooted in the geometric properties of the embedding space. Our framework, MC^2F, operationalizes the manifold hypothesis by first learning a stratified Riemannian geometry of clean data with a continuous normalizing flow, and then purifying detected adversarial samples via geodesic projection. The experiments demonstrate that MC^2F not only sets a new state-of-the-art in adversarial robustness against multiple attacks but, crucially, does so without degrading performance on clean data. This work validates the power of a geometric, detect-and-correct approach, paving the way for developing more principled and robust NLP systems that resolve the long-standing robustness-accuracy dilemma.

References

Altinisik, E.; Messaoud, S.; Sencar, H. T.; Sajjad, H.; and Chawla, S. 2024. Explaining the role of Intrinsic Dimensionality in Adversarial Training. *arXiv preprint arXiv:2405.17130*.

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2890–2896.

Asghar, N. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Asl, J. R.; Panzade, P.; Blanco, E.; Takabi, D.; and Cai, Z. 2024. RobustSentEmbed: Robust Sentence Embeddings Using Adversarial Self-Supervised Contrastive Learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3795–3809.

Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Cook, E. D.; Lavoie, M.-A.; and Waslander, S. L. 2024. Feature Density Estimation for Out-of-Distribution Detection via Normalizing Flows. In *Proceedings of the Conference on Robots and Vision*. PubPub.

Dang, C.; and Zhu, Z. 2025. MERIA: Empathetic Response Generation via Parallel Disentanglement and Uncertainty-Gated Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 14021–14027.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.

Diepeveen, W.; Batzolis, G.; Shumaylov, Z.; and Schönlieb, C.-B. 2024. Score-based Pullback Riemannian Geometry: Extracting the Data Manifold Geometry using Anisotropic Flows. *arXiv preprint arXiv:2410.01950*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049.

Gao, S.; Dou, S.; Liu, Y.; Wang, X.; Zhang, Q.; Wei, Z.; Ma, J.; and Shan, Y. 2023. DSRM: Boost Textual Adversarial Training with Distribution Shift Risk Minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12177–12189.

Garg, S.; and Ramakrishnan, G. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6174–6181.

Goyal, S.; Doddapaneni, S.; Khapra, M. M.; and Ravindran, B. 2023. A survey of adversarial defenses and robustness in NLP. *ACM Computing Surveys*, 55(14s): 1–39.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The journal of machine learning research*, 13(1): 723–773.

Houle, M. E. 2017. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In *International Conference on Similarity Search and Applications*, 64–79. Springer.

Huang, P.; Yang, Y.; Jia, F.; Liu, M.; Ma, F.; and Zhang, J. 2022. Word level robustness enhancement: Fight perturbation with perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10785–10793.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jacovi, A.; Shalom, O. S.; and Goldberg, Y. 2018. Understanding Convolutional Neural Networks for Text Classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 56–65.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2020. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8018–8025.

Kim, M.; Li, D.; Hu, S. X.; and Hospedales, T. 2022. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, 11148–11161. PMLR.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6193–6202.

Li, L.; Song, D.; and Qiu, X. 2023. Text Adversarial Purification as Defense against Adversarial Attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 338–350.

Li, X.; and Sarwate, A. 2025. Unraveling the localized latents: Learning stratified manifold structures in llm embedding space with sparse mixture-of-experts. *arXiv preprint arXiv:2502.13577*.

Lin, J. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1): 145–151.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mamou, J.; Le, H.; Del Rio, M. A.; Stephenson, C.; Tang, H.; Kim, Y.; and Chung, S. 2020. Emergence of separable manifolds in deep language representations. In *Proceedings of the 37th International Conference on Machine Learning*, 6713–6723.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29).

Minh, D. N.; and Tuan, L. A. 2022. Textual manifold-based defense against natural language adversarial examples. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 6612–6625.

Moraffah, R.; Khandelwal, S.; Bhattacharjee, A.; and Liu, H. 2024. Adversarial text purification: A large language model approach for defense. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 65–77. Springer.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2018. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*.

Ning, R.; and Chai, L. 2024. Revisiting the Trade-Off Between the Performance of Adversarial and Normal Examples on NLP Few-Shot Tasks. In *International Conference on Neural Information Processing*, 47–60. Springer.

Raina, V.; Tan, S.; Cevher, V.; Rawal, A.; Zha, S.; and Karypis, G. 2024. Extreme Miscalibration and the Illusion of Adversarial Robustness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2500–2525.

Reif, E.; Yuan, A.; Wattenberg, M.; Viegas, F. B.; Coenen, A.; Pearce, A.; and Kim, B. 2019. Visualizing and measuring the geometry of BERT. *Advances in neural information processing systems*, 32.

Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International conference on machine learning*, 1530–1538. PMLR.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Subhash, V.; Bialas, A.; Pan, W.; and Doshi-Velez, F. 2023. Why do universal adversarial attacks work on large language models?: Geometry might be the answer. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Vázquez-Hernández, M.; Morales-Rosales, L. A.; Algredo-Badillo, I.; Fernández-Gregorio, S. I.; Rodríguez-Rangel, H.; and Córdoba-Tlaxcalteco, M.-L. 2024. A Survey of Adversarial Attacks: An Open Issue for Deep Learning Sentiment Analysis Models. *Applied Sciences*, 14(11): 4614.

Volpi, R.; Thakur, U.; and Malago, L. 2021. Changing the geometry of representations: α-embeddings for NLP tasks. *Entropy*, 23(3): 287.

Wang, X.; Wang, H.; and Yang, D. 2021. Measure and improve robustness in NLP models: A survey. *arXiv preprint arXiv:2112.08313*.

Wang, Y.; Liu, L.; Liang, Z.; Ye, Q.; and Hu, H. 2024. New Paradigm of Adversarial Training: Releasing Accuracy-Robustness Trade-Off via Dummy Class. *arXiv preprint arXiv:2410.12671*.

Yang, Z.; Xu, Z.; Zhang, J.; Hartley, R.; and Tu, P. 2024. Adversarial purification with the manifold hypothesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16379–16387.

Zhang, J.; Rubinstein, B. I.; Zhang, J.; and Liu, F. 2025. Ddad: A two-pronged adversarial defense based on distributional discrepancy. *arXiv preprint arXiv:2503.02169*.

Zheng, R.; Zhou, Y.; Xi, Z.; Gui, T.; Zhang, Q.; and Huang, X.-J. 2024. Subspace Defense: Discarding Adversarial Perturbations by Learning a Subspace for Clean Signals. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 15410–15421.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.