TextGuard: A Semi-Supervised Framework for Filtering Harmful Prompts

Abdulaziz Houtari Samir Shakir Linqing Mo

April 11, 2025

Abstract

This paper presents a semi-supervised anomaly detection framework for distinguishing harmful prompts from benign ones in natural language. The proposed method serves as a defensive layer for Large Language Models (LLMs), eliminating the need for expensive model retraining or fine-tuning. Our approach is formulated as a one-class classification problem, wherein an autoencoder is trained exclusively on harmful prompts represented by sentence-level transformer embeddings. By learning to reconstruct only harmful inputs, the model identifies benign prompts as out-of-distribution samples based on reconstruction error. The framework leverages pre-trained sentence transformers for creating embeddings and explores architectural variations in the autoencoder to enhance performance and robustness. We further develop an ensemble approach that combines multiple sentence transformer embedding models to improve classification stability and resilience against adversarial examples. A formal definition of the anomaly detection task is provided, along with detailed descriptions of system design and threshold calibration strategies. The implementation demonstrates high recall and precision in detecting harmful prompts while reliably flagging benign content as anomalous. The code can be found publicly at `https://github.com/azhoutari/cse895-project`.

1 Introduction and Motivation

Despite significant improvements in safety measures, Large Language Models (LLMs) remain vulnerable to harmful prompts that can bypass content filters and elicit dangerous, unethical, or illegal content. While retraining or fine-tuning LLMs to enhance safety is possible, it requires substantial computational resources and expense. Our work addresses this challenge by developing an external defensive layer that can reliably filter out harmful prompts before they reach the LLM, eliminating the need to modify the model itself. We propose a one-class anomaly detection framework using autoencoders trained on sentence embeddings of harmful prompts. The key advantage of our approach is its simplicity and efficiency—it requires only examples of harmful content without any labeled benign data, making dataset collection straightforward. This creates a practical, cost-effective shield for any LLM deployment that can be implemented with minimal computational resources and without access to the protected model's weights.

2 Related Work

Anomaly detection has been extensively studied, with foundational work like One-Class SVM [5], which estimates a high-dimensional support boundary, and Isolation Forest [6], which isolates anomalies using randomized trees.

Autoencoders have become central in high-dimensional anomaly detection. Sakurada and Yairi [10] showed how autoencoders can detect anomalies using reconstruction error, while Zhou and Paffenroth [11] proposed a robust variant using modified loss functions to improve resistance to noise.

In natural language processing, BERT [7] and Sentence-BERT [8] have enabled deep semantic understanding via transformer-based embeddings. While most harmful content detection tasks are handled with supervised learning [9], our work follows a semi-supervised route.

Recent work by Zou et al. [1] and Shen et al. [2] has focused on jailbreak prompts and adversarial attacks on large language models. Unlike their evaluation-focused contributions, our model is trained only on harmful prompts and flags benign inputs as anomalies — a less explored but promising approach.

3 Problem Definition

Let $\mathcal{P}$ denote the set of all possible text prompts, and let $\mathcal{M} \subset \mathcal{P}$ represent the subset of malicious prompts. Given access only to a collection of malicious prompts $\mathcal{M}_{\text{train}} \subset \mathcal{M}$, we aim to learn a classifier $C : \mathcal{P} \to \{0, 1\}$ that determines whether a new prompt $p \in \mathcal{P}$ belongs to $\mathcal{M}$ (malicious) or $\mathcal{P} \setminus \mathcal{M}$ (benign).

We formulate this as a one-class anomaly detection problem. Using an embedding function $\phi : \mathcal{P} \to \mathbb{R}^d$ that maps prompts to a d-dimensional space, we represent each prompt p as a vector $x = \phi(p) \in \mathbb{R}^d$. Our task is to learn a reconstruction function $f(x) = D(E(x))$ comprised of an encoder E and decoder D that minimizes the reconstruction error on malicious embeddings.

For classification, we establish a threshold τ based on the distribution of reconstruction errors on the training set. The classifier is defined as:

$$C(p) = \begin{cases} 1 \text{ (malicious)}, & \text{if } \|f(\phi(p)) - \phi(p)\|_2^2 < \tau \\ 0 \text{ (benign)}, & \text{otherwise} \end{cases}$$

This approach aims to minimize the expected misclassification rate:

$$\min_C \int_{p \in \mathcal{P}} \mathbf{1}[C(p) \neq \mathbf{1}[p \in \mathcal{M}]] \, dP(p)$$

where $\mathbf{1}[\cdot]$ denotes the indicator function and P is a probability measure over $\mathcal{P}$. The core challenge is learning a decision boundary that separates classes when training data contains only malicious examples without any benign counterexamples. This semi-supervised approach differs from supervised classification (requiring examples from all classes) and purely unsupervised anomaly detection.

4 Proposed Method

Our method consists of two primary stages:

1. **Feature Extraction:** We use pre-trained sentence transformers such as `all-MiniLM-L6-v2` to convert text prompts into dense embeddings $x \in \mathbb{R}^d$.
2. **Anomaly Detection:** We train an autoencoder comprising an encoder $E(\cdot)$ and a decoder $D(\cdot)$ on harmful prompt embeddings. The autoencoder is trained using the MSE loss:

$$\mathcal{L} = \|x - D(E(x))\|_2^2.$$

   The choice of MSE over cross entropy loss is based on the following considerations:

   - **Continuous Outputs:** Sentence transformer embeddings are high-dimensional continuous vectors. MSE loss directly quantifies the Euclidean distance between the original embedding and its reconstruction, making it well-suited to capture reconstruction fidelity.
   - **Reconstruction Objective:** Our autoencoder's goal is to accurately reconstruct the continuous embedding vector. Cross entropy is typically used for classification tasks involving probability distributions or discrete outputs, whereas MSE naturally addresses regression tasks.
   - **Empirical Performance:** In our experiments, MSE provided a robust signal for detecting deviations (i.e., benign prompts) based on reconstruction error.

   At test time, prompts with high reconstruction errors are flagged as benign.

To enhance robustness, we also developed an ensemble approach using ten different sentence transformer models. For each model (e.g., `all-MiniLM-L6-v2`), we train a separate autoencoder, resulting in ten specialized models. Each model generates a reconstruction error and binary classification based on its individual threshold. We then explore multiple ensemble strategies: (1) majority voting, where the final classification follows the majority of individual predictions; (2) weighted voting, which assigns higher weights to better-performing models; (3) average threshold, which compares the average error to the average threshold; and (4) minimum error, which classifies a prompt as malicious if any single model classifies it as such. This multi-model approach provides added resilience against adversarial examples and improves overall classification stability.

Figure 1 shows a schematic of our deep learning architecture.

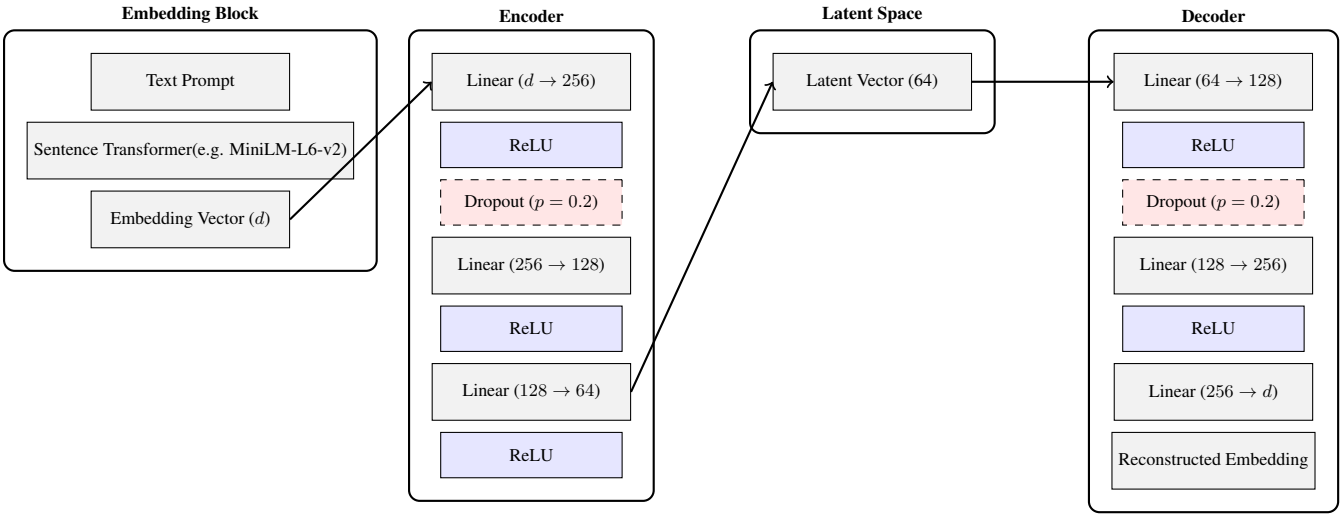


Figure 1: Architecture of the Sentence-Transformer-Based AutoEncoder. Prompts are embedded using a sentence transfromer, then passed through a fully-connected encoder-decoder architecture with ReLU activations and dropout regularization.

5 Experimental Setup

5.1 Training and Evaluation Datasets

We use malicious prompts exclusively from the following datasets:

- **AdvBench** [1]: *Universal and Transferable Adversarial Attacks on Aligned Language Models* (Zou et al., 2023).
- **In-The-Wild Jailbreak Prompts on LLMs** [2]: (Shen et al., 2024).

These datasets are split 80/20 (80% for training and 20% for evaluation).

5.2 Testing Datasets

Testing was performed on:

- **JailbreakBench/JBB-Behaviors** [3]: an open robustness benchmark for jailbreaking large language models.
- **WikiQA** [4]: where all WikiQA questions were treated as benign.

5.3 Thresholding Strategy

For evaluation, we compute the reconstruction errors on the test set and set the threshold at the 95th percentile of the training set errors. This strategy ensures that approximately 95% of the harmful prompts (training data) are reconstructed with low error, while benign prompts generate higher errors and are flagged as anomalies.

5.4 Implementation and Computational Resources

The autoencoder was implemented in Python using PyTorch and the Hugging Face Transformers library where sentence transformer embeddings were extracted using various pre-trained models. Training was executed on an NVIDIA RTX 8000 GPU, and the overall computational demand was minimal.

6 Results and Discussion

6.1 Comparative Analysis of Sentence Transformer Models

Figure 2 shows our comparative analysis of ten sentence transformer models. The best performance was observed for the RoBERTa-based model (`all-roberta-large-v1`).

6.2 Ensemble Method Comparison

Figure 3 illustrates the performance comparison among four different ensemble methods. Although the best single model yields the highest performance, the ensemble method achieves competitive, albeit slightly lower, overall metrics.

6.3 Confusion Matrix Analysis

Figure 4 combines the confusion matrices of the best-performing transformer model (all-roberta-large-v1) and the best ensemble method as subfigures. The following metrics were obtained for the best transformer model:

Table 1: Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1 Score
RoBERTa-based Transformer	98.77%	100%	91%	95.29%
Best Ensemble Method (Weighted Vote)	97.27%	100%	80%	88.89%

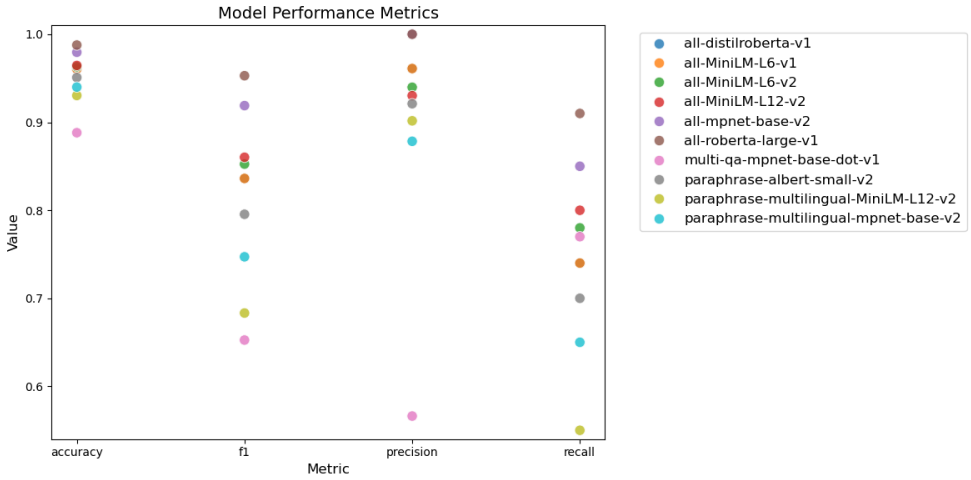


Figure 2: Comparative analysis of ten (10) different sentence transformer models used as embeddings.

6.4 Discussion

The experimental results demonstrate that our approach effectively distinguishes harmful prompts from benign ones. The RoBERTa-based model, with an accuracy of 98.77% and recall of 91%, confirms that training solely on malicious prompts allows the autoencoder to capture the nuances of harmful content. The thresholding strategy (95th percentile) proved crucial in separating benign samples, which produced higher reconstruction errors. Although the ensemble method (Wighted Vote) yields marginally lower performance (accuracy of 97.27%, precision 100%, recall 80%, and F1 of 88.89%), it validates that alternative combined strategies can be competitive. The testing on WikiQA—where all questions were treated as benign—demonstrates robust generalization of our model across diverse prompt formats.

7 Conclusion

We presented a semi-supervised one-class anomaly detection framework using pre-trained sentence transformer embeddings and autoencoders. By training solely on harmful prompts, our method successfully flags benign prompts as anomalies based on reconstruction error. Our experiments show promising recall and precision, and our work opens avenues for further refinements—including testing with alternative backbone models and expanding the dataset. Future research will explore improved architectural designs and additional ensemble techniques to further optimize performance.

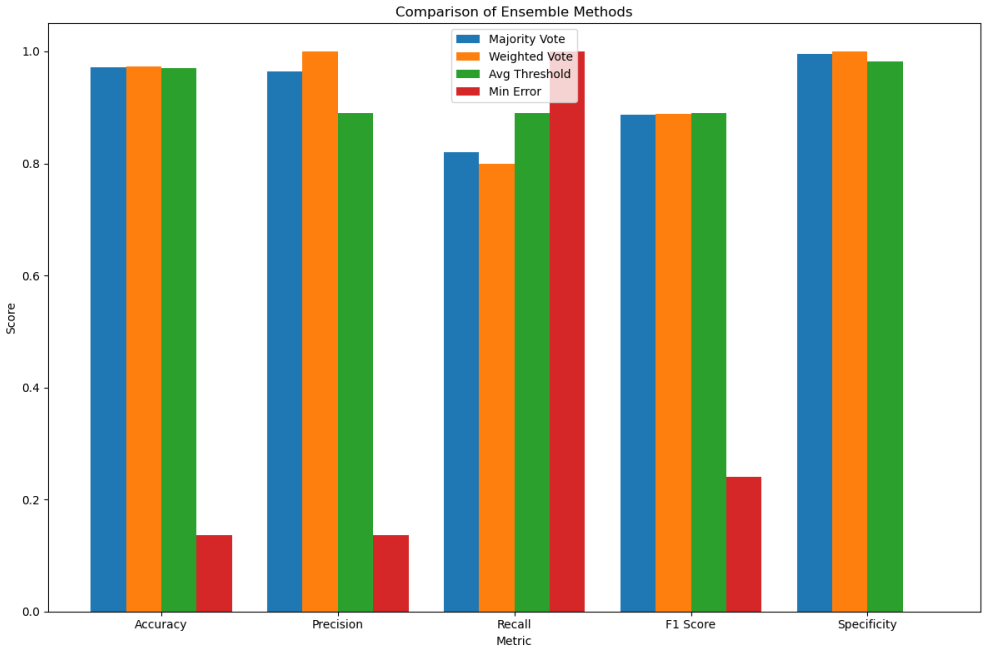
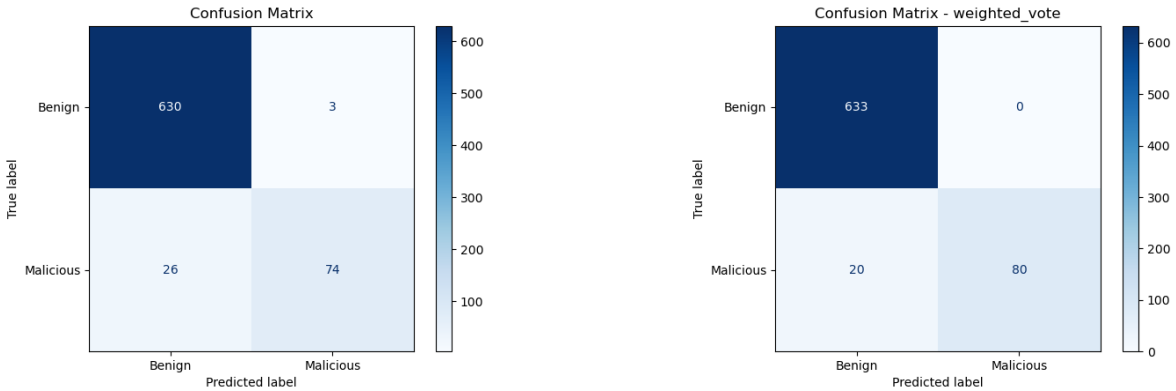


Figure 3: Performance comparison of four (4) ensemble methods.



(a) All Roberta Large V1 Model

(b) Ensemble Method (Avg Threshold)

Figure 4: Confusion matrices for the best transformer model and the best ensemble method.

References

[1] Zou, Andy, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. arXiv preprint arXiv:2307.15043, 2023.

[2] Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. ""Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, ACM, 2024.

[3] Chao, Patrick, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. *JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models*. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[4] Yang, Yi, Wen-tau Yih, and Christopher Meek. "WikiQA: A Challenge Dataset for Open-Domain Question Answering." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015. Available at: `https://aclanthology.org/D15-1237`.

[5] Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), 1443–1471.

[6] Liu, F.T., Ting, K.M., and Zhou, Z.-H. (2008). Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422).

[7] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[8] Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992).

[9] Garg, S., and Mukherjee, A. (2022). Detecting Harmful Online Content with Transformers and External Knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 5341–5355).

[10] Sakurada, M., and Yairi, T. (2014). Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis* (pp. 4–11).

[11] Zhou, C., and Paffenroth, R.C. (2017). Anomaly Detection with Robust Deep Autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 665–674).

Appendix

A Detailed Evaluation Results

A.1 Individual Model Performance

Table 2 presents the detailed performance metrics for each individual sentence transformer model evaluated in our study.

A.2 Ensemble Method Comparison

Table 3 shows the performance metrics for the four ensemble methods evaluated in our study. The results demonstrate that while ensemble methods provide competitive performance, the best individual model (all-roberta-large-v1) still outperforms all ensemble approaches.

Table 2: Performance Metrics for Individual Sentence Transformer Models

Model	Accuracy	Precision	Recall	F1	Spec.	TP	FP	TN	FN
all-MiniLM-L6-v2	0.963	0.940	0.780	0.852	0.992	78	5	628	22
all-distilroberta-v1	0.960	0.961	0.740	0.836	0.995	74	3	630	26
all-MiniLM-L12-v2	0.965	0.930	0.800	0.860	0.991	80	6	627	20
paraphrase-albert-small-v2	0.951	0.921	0.700	0.795	0.991	70	6	627	30
all-roberta-large-v1	**0.988**	**1.000**	**0.910**	**0.953**	**1.000**	91	0	633	9
all-mpnet-base-v2	0.980	1.000	0.850	0.919	1.000	85	0	633	15
multi-qa-mpnet-base-dot-v1	0.888	0.566	0.770	0.653	0.907	77	59	574	23
paraphrase-multilingual-mpnet-base-v2	0.136	0.136	1.000	0.240	0.000	100	633	0	0
all-MiniLM-L6-v1	0.940	0.878	0.650	0.747	0.986	65	9	624	35

Table 3: Performance Metrics for Different Ensemble Methods

Ensemble Method	Accuracy	Precision	Recall	F1	Spec.	TP	FP	TN	FN
Average Threshold	0.970	0.890	0.890	0.890	0.983	89	11	622	11
Majority Vote	0.971	0.965	0.820	0.886	0.995	82	3	630	18
Minimum Error	0.136	0.136	1.000	0.240	0.000	100	633	0	0
Weighted Vote	0.973	1.000	0.800	0.889	1.000	80	0	633	20