Class Distillation with Mahalanobis Contrast: An Efficient Training Paradigm for Pragmatic Language Understanding Tasks

Anonymous ACL submission

Abstract

Detecting deviant language such as sexism, or nuanced language such as metaphors or sarcasm, is crucial for enhancing the safety, clarity, and interpretation of social interactions. While existing classifiers deliver strong results on these tasks, they often come with significant computational cost and high data demands. In this work, we propose **Cla**ss **D**istillation (ClaD), a novel training paradigm that targets the core challenge: distilling a small, well-defined target class from a highly diverse and heterogeneous background. ClaD integrates two key innovations: (i) a loss function informed by the structural properties of class distributions, based on Mahalanobis distance, and (ii) an interpretable decision algorithm optimized for class separation. Across three benchmark detection tasks – sexism, metaphor, and sarcasm – ClaD outperforms competitive baselines, and even with smaller language models and orders of magnitude fewer parameters, achieves performance comparable to several large language models. These results demonstrate ClaD as an efficient tool for pragmatic language understanding tasks that require gleaning a small target class from a larger heterogeneous background.

1 Introduction

The widespread adoption of social media and the polarized nature of online discourse have amplified the need for improved communication dynamics, fostering research aimed at promoting safety and mutual respect. A critical part of this effort involves detecting complex linguistic phenomena such as figurative speech – such as sarcasm and metaphor (Riloff et al., 2013; Oraby et al., 2016; Ghosh et al., 2020; Ge et al., 2023) – as well as harmful language like aggression or sexism (Safi Samghabadi et al., 2020; Samory et al., 2021). These tasks present significant urgency and challenges due to the nuances of figurative speech and the variability of deviant language.

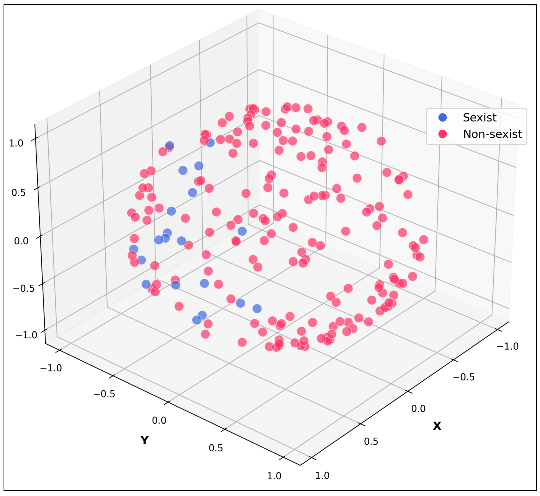


Figure 1: The minority target class representing deviant language (●) versus a highly diverse and heterogeneous non-target class of *everything else* (●). This t-SNE (van der Maaten and Hinton, 2008) visualization (where lighter shades indicate instances located further away in 3-D) displays a representative sample from the "Call me sexist but ..." corpus (Samory et al., 2021).

Most prior research (see §6) has approached these tasks as traditional binary classification problems, utilizing ground truth labels provided in various datasets. Despite varying degrees of success, this formulation has overlooked a crucial commonality: the objective in such tasks is often to isolate a minority target class, characterized more by its pragmatic function in natural language than its semantics, from the much larger and incredibly diverse negative class encompassing *everything else*. The complexities arising from the somewhat nebulous dichotomy,[1] alongside the diverse and heterogeneous nature of the predominant non-target class, can be gleaned from Figure 1. In this context, accurate binary classification proves challenging due to the immense linguistic diversity encompassed

[1] There is a sizeable body of legal and linguistic scholarship on the boundaries of unwarranted and figurative language. See, for example, Rosenfeld (2002); Kiska (2012); Kasparian (2013); Athanasiadou (2024).

Figurative	f_1)	My alarm clock makes sure I love every Monday morning!	⊘
	f_2)	My alarm clock makes sure that I dread every Monday morning!	●
	f_3)	My alarm clock always wakes me up on Monday mornings.	●
	f_4)	A white carpet is a great choice when you have messy kids.My alarm clock always wakes me up on Monday mornings.	⊘
	f_5)	A white carpet is an engaging choice when you have messy kids, if you take extra care.My alarm clock always wakes me up on Monday mornings.	●
		$f_1 \overset{0.47}{\sim} f_2$ $f_1 \overset{0.91}{\sim} f_3$ $f_2 \overset{0.62}{\sim} f_3$ $f_4 \overset{0.78}{\sim} f_5$	
Deviant	d_1)	A female astronaut, because they need sandwiches up there.	⊘
	d_2)	An astronaut needs sandwiches up there.	●
		$d_1 \overset{0.76}{\sim} d_2$	

Table 1: Instances of figurative (f_i, sarcasm detection) and deviant (d_i, sexism detection) language. Similarity scores are based on the `stsb-roberta-large` cross-encoder model fine-tuned on the STS benchmark introduced by Cer et al. (2017). These scores reveal a deeper problem: a target class instance (⊘) may be highly *dis*-similar to several non-target instances (●) while also being very similar to other non-target instances.

within the non-target category.[2] The model must be adept at learning a complex decision boundary without succumbing to overfitting on the training corpus. As such, the most effective solutions often employ sophisticated deep neural models or ensemble methods, which (a) demand large amounts of training data, (b) require significant computational resources for training and inference, (c) lack interpretability, and (d) depend on careful regularization and hyperparameter tuning.

Since the target class has a well-defined function in terms of natural language pragmatics, we conjecture that a detailed study of the target and non-target class distributions will reveal structural differences that can be leveraged to design a model training paradigm better suited for **cla**ss **d**istillation (ClaD, discussed in § 2), and therefore, superior in terms of (a) inferring test instances of the target class, (b) the demands it places on computational resources during training, and (c) interpretability. We test this conjecture on multiple tasks and benchmark datasets (§ 3), starting with statistical analyses to glean the **structural properties and distributional differences between the target and non-target classes** (§ 3.1). With these insights, **we develop novel contrastive loss functions** derived from Mahalanobis (1936) distance (§ 4.1), which leverage intra-class covariance to contrast the target class against the diverse and heterogeneous collection of negative samples. We then introduce an **interpretable decision algorithm** based on the normalized squared Mahalanobis distance (§ 4.2) to identify target instances.

Our results (§ 5) demonstrate superior inference and resource efficiency across all tasks. We raise two vital questions in § 5.2, investigating how small language models with ClaD, given limited task-specific training, compares to LLMs in low-resource transfer learning. We also examine the extent to which increasing LLM size improves performance with identical training data. Recent findings, such as those from DeepSeek (Liu et al., 2024; Guo et al., 2025), underscore the need for such emphasis on economical training and efficient inference. Further, we present ablation experiments to discern (1) the impact of our decision algorithm, (2) the effect of our novel loss function, and (3) whether traditional one-class classification is comparable to ClaD with Mahalanobis contrast.

2 ClaD: A Whiteboard Discussion

Class **D**istillation (ClaD) is a specialized training paradigm for binary classification, emphasizing the separation of a distinct category from a diverse and often disproportionately larger non-target background. Non-target instances frequently include expressions that are semantically similar to the target class, while also encompassing elements with no syntactic, semantic, or pragmatic resemblance to each other. This dual challenge leads to significant ambiguity and overlap, making accurate classification particularly difficult.

The predicament is not specific to a single task, as Table 1 shows with instances from figurative (sarcasm) and deviant (sexism) language use. These examples highlight the limitations of relying solely on simple prompting with large language models for effective inference, as decisions are easily confounded by the diversity of the non-target class. Further, we show in § 5.3 that applying straightforward semantic similarity measures fails to capture the nuanced characteristics defining the target class.

Drawing insight from Figure 1, the target instances are not uniformly distributed across the fea-

[2] The target class exhibits rich syntactic and semantic variations while serving a specific pragmatic function, whereas the non-target class presents even greater semantic variety with no common pragmatic function. *E.g.*, instances of sexism specifically discriminate on the basis of sex, while the non-target class is unified only by the absence of such hostility.

ture space; rather, they appear to form a structured subset that can be viewed as a manifold within a higher-dimensional space. To better understand the shape and properties of this manifold, we analyze the structural and distributional characteristics of the target and non-target classes, which provides foundational insights into the class geometries and informs our formulation of novel loss functions (§ 4.1) and ClaD's decision algorithm (§4.2).

A visual approach to unveiling the distributional characteristics is relegated to Appendix A, while our systematic analysis of the datasets and target class' geometric properties is presented next.

3 Tasks and Datasets

We concentrate on three tasks for our analyses and experiments: two types of figurative language (sarcasm and metaphors) and one form of deviant language (sexism), utilizing a dedicated benchmark corpus for each to illustrate that the patterns we uncover and the class distillation paradigm we propose are broadly applicable across such tasks.

1. Sarcasm Headlines (SH) is a curated dataset comprising professionally crafted headlines from The Onion and HuffPost (Misra and Arora, 2019, 2023). Notably free of spelling errors and informal language, it offers high-quality labels and self-contained headlines. Compared to social media datasets, it is a clean and reliable resource that precludes the need to worry about spurious data correlations arising from viral social media trends (Gururangan et al., 2018; Bender et al., 2021).

2. Trope Finder (TroFi) (Birke and Sarkar, 2006) is built to distinguish between literal and non-literal verb usage. It leverages the '88-'89 Wall Street Journal (WSJ) Corpus and enhances it with WordNet, databases of idioms and metaphors, and tags from advanced taggers. TroFi improves metaphor detection by minimizing unverified literal uses and addressing the scarcity of non-literal instances.

3. Call Me Sexist But . . . (CMSB) is an innovative corpus designed to detect sexism, comprising tweets that explicitly use the titular phrase to voice potential sexism (Samory et al., 2021). It is enhanced with synthetic adversarial modifications to challenge machine learning models.

3.1 Statistical Tests of Normality

To systematically analyze the geometric properties of target class representations, we evaluate the normality of the embedding distributions in reduced dimensionality space. Specifically, we apply the Henze and Zirkler (HZ) (1990) test to assess multivariate normality across three t-SNE dimensions for BERT and SimCSE,[3] examining three pragmatic language detection tasks: metaphor, sarcasm, and sexism. We complement this with Anderson and Darling (AD) (1952) tests to evaluate univariate normality along individual dimensions. The results (Table 2) offer a rigorous statistical characterization of the manifold structure, beyond the Q-Q plot inspections shown in Appendix A.

Model	Class	Empirical normality test statistics			
		HZ	Anderson-Darling		
			d_1	d_2	d_3
Task + Corpus: Metaphor detection on *TroFi*					
BERT	*Metaphor*	5.14	1.53	2.59	1.96
	Other	5.93	1.83	1.71	3.40
SimCSE	*Metaphor*	4.32	2.83	2.11	0.70
	Other	5.19	2.94	2.79	1.54
Task + Corpus: Sarcasm detection on *Sarcasm Headlines*					
BERT	*Sarcasm*	29.94	13.67	10.14	24.99
	Other	33.24	13.87	23.95	5.76
SimCSE	*Sarcasm*	21.49	19.00	16.38	23.68
	Other	24.82	12.21	32.65	42.99
Task + Corpus: Sexism detection on *Call Me Sexist But*					
BERT	*Sexism*	7.48	0.97	1.79	9.27
	Other	34.08	6.70	40.34	23.96
SimCSE	*Sexism*	6.38	2.34	2.99	2.08
	Other	21.21	17.07	1.77	9.11

Table 2: Empirical results on how well the target and non-target classes fit (a) multivariate normality, using the Henze-Zirkler (HZ) statistic, and (b) univariate normality on the three t-SNE dimensions d_1, d_2, d_3, using the Anderson-Darling statistic. For both tests, larger numbers indicate greater deviation from normality.

The HZ tests consistently show lower values for target data (metaphor, sarcasm, and sexism) compared to non-target data, indicating that target data are closer to a multivariate normal distribution compared to their non-target counterparts. This is further supported by the results of the AD tests along each dimension. Reduced deviation from the theoretical distribution suggests that the target data

[3] We analyze distributional properties (not downstream performance) using BERT and SimCSE as foundational bidirectional Transformer and contrastive sentence embedding models, respectively. Their selection aligns with established probing protocols prioritizing consistency and transferability across tasks (Reimers and Gurevych, 2019; Rogers et al., 2021; Gao et al., 2021). Findings generalize to architectures like ALBERT and DistilBERT, while rare outliers (GPT-2 and Phi) reflect pretraining misalignment (Ethayarajh, 2019), rather than methodological drawbacks.

exhibits a more homogeneous manifold structure. In contrast, the non-target data manifests greater diversity and complexity. Thus, in line with the argument presented earlier with illustrative examples (§ 2), it is indeed less likely that the they possess discernible common traits beyond their opposition to the target class. Hence, we hypothesize that a loss function ought to be designed primarily around the target class. The consistency and regularity of the target class' distribution provide a more reliable foundation for learning stable predictors, in contrast to the somewhat more chaotic diversity observed in the distribution of the non-target class.

4 Training and Inference

Mahalanobis (1936) distance is ideal for data approximating a multivariate normal distribution, as it accounts for the manifold structure of the target class, rendering the distance measure scale-invariant. By incorporating the variance and correlations among variables, it accurately reflects the underlying distribution of the data and thereby improves discrimination in detecting non-target instances by robust identification of outliers, reducing false positives.[4] Accordingly, we explore Mahalanobis distance in formulating the loss function.

4.1 Mahalanobis Loss

Let $\mathcal{X} = \{x_i\}, \mathcal{Y} = \{y_j\}$ denote n target and m non-target training samples (resp.). Further, let $f : \mathcal{X} \cup \mathcal{Y} \mapsto \mathbb{R}^d$ denote a representation function mapping these instances to d dimensions. For a given instance $x \in \mathcal{X}$, we randomly select $x^+ \in \mathcal{X} \setminus \{x\}$, and $y^- \in \mathcal{Y}$. These random selections are employed to learn a representation that minimizes (maximizes) the similarity between x and y^- (x^+). We achieve this with **Mahalanobis loss**:

$$\mathcal{L}_{\text{MAH}} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{\text{sim}_{\text{MAH}}(x, y^-)}{\text{sim}_{\text{MAH}}(x, x^+) + \text{sim}_{\text{MAH}}(x, y^-)} \tag{1}$$

where $\text{sim}_{\text{MAH}}(x, y)$ is defined using the covariance matrix Σ of the set $\{f(x_i)\}$

$$\text{sim}_{\text{MAH}}(x, y) = \exp\left\{ -\frac{(f(x) - f(y))^T \Sigma^{-1} (f(x) - f(y))}{d} \right\}.$$

Alternatively, **Mahalanobis mean loss** uses the mean μ of $\{f(x_i)\}$:

$$\mathcal{L}_{\text{MAH},\mu} = -\frac{1}{|\mathcal{X}|} \sum_{(x,y^-) \in \mathcal{X}} \Big[\log\big(\text{sim}_{\text{MAH}}(\mu, x)\big) + \log\big(1 - \text{sim}_{\text{MAH}}(\mu, y^-)\big) \Big]. \tag{2}$$

It maximizes the similarity between a target instance x and the mean representation of the target class (making the class more compact), and minimizes the similarity between a negative example y^- and the mean (increasing inter-class margin).

Algorithm 1 Mahalanobis β-decision algorithm

```
Require: New instance $X = X^*$
  $X_{n+1} \leftarrow X^*$
  $\hat{\mu} \leftarrow \frac{1}{n+1} \sum_{i=1}^{n+1} X_i$
  Compute $\hat{\Sigma}$ as the sample covariance matrix of $X_1, \ldots, X_{n+1}$
  $d^2_{n+1}(\hat{\mu}, \hat{\Sigma}) \leftarrow (X_{n+1} - \hat{\mu})^T \hat{\Sigma}^{-1} (X_{n+1} - \hat{\mu})$
  $T \leftarrow \frac{n+1}{n^2} d^2_{n+1}(\hat{\mu}, \hat{\Sigma})$
  Compute the critical value $v_\beta$ for Beta$\left(\frac{d}{2}, \frac{n-d-1}{2}\right)$
  if $T < v_\beta$ then
      $X \leftarrow 1$                         ▷ Target class
  else
      $X \leftarrow 0$                         ▷ Non-target class
  end if
```

4.2 Inference and Decision Algorithm

The inference task is, fundamentally, identical to that of any supervised binary classifier: ascertain if a test instance belongs to the target class. Given representations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ with sample mean $\hat{\mu}$ and covariance matrix Σ adhering to a multivariate normal distribution, the squared Mahalanobis distance for a specific observation $\mathbf{x}_i$ is given by

$$d_i^2(\hat{\mu}, \Sigma) = (\mathbf{x}_i - \hat{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \hat{\mu}), \tag{3}$$

which follows the Beta distribution (Wilks, 1962; Ververidis and Kotropoulos, 2008):

$$\frac{n}{(n-1)^2} d_i^2(\hat{\mu}, \Sigma) \sim \text{Beta}\left(\frac{d}{2}, \frac{n-d-1}{2}\right) \tag{4}$$

This insight informs the design of the **Mahalanobis β-decision algorithm** (Algorithm 1), to test an instance for class membership by comparing its normalized squared Mahalanobis distance to critical values of the corresponding Beta distribution.[5]

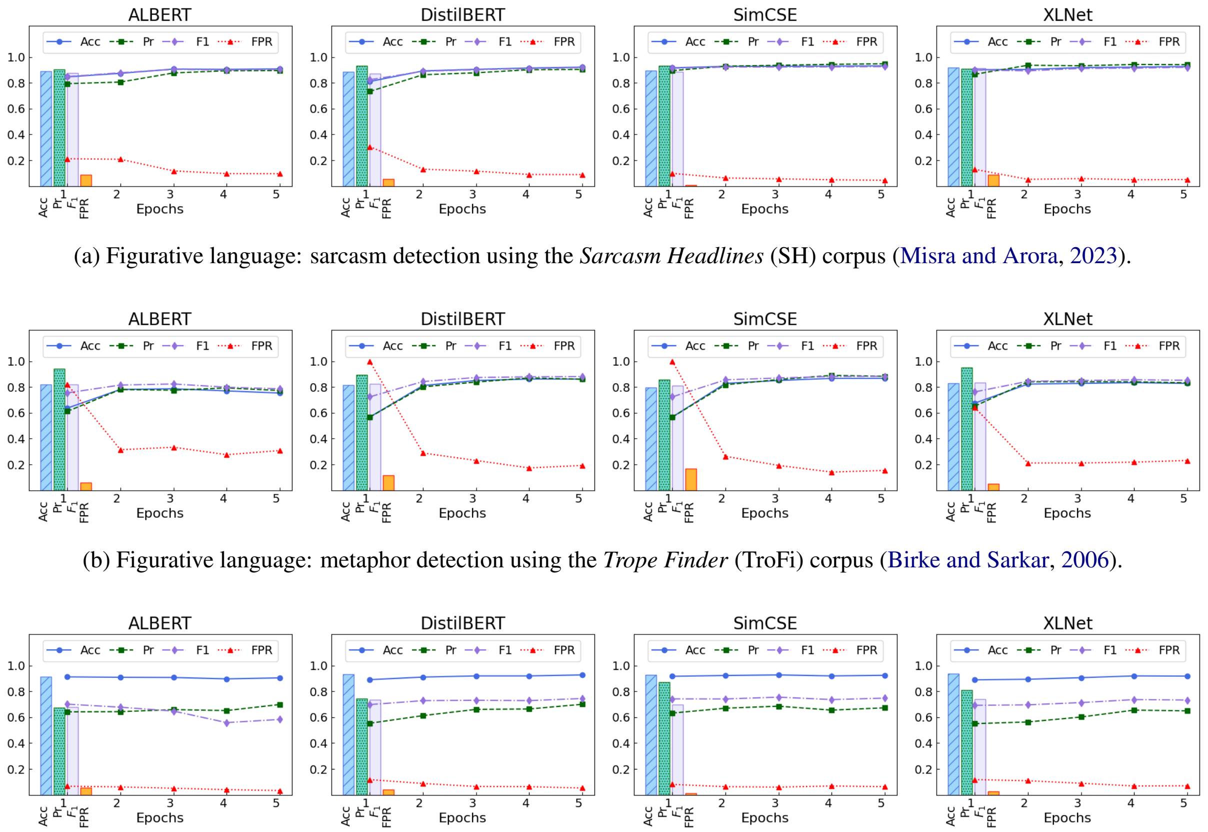
5 Experiments and Results

We empirically evaluate[6] ClaD across two challenging categories of language understanding tasks: detecting figurative (metaphor and sarcasm) and harmful (sexism) language. ClaD leverages the Mahalanobis mean loss, $\mathcal{L}_{\text{MAH},\mu}$ (Eq. 2), to fine-tune pretrained embeddings, followed by inference with the Mahalanobis β-decision algorithm (Alg. 1).

[4] Reducing false positives in these tasks is particularly important in several applications. For example, in social media moderation, so as to not penalize users for innocuous remarks.

[5] The critical threshold value is determined based on development data, ensuring optimal calibration for inference.

[6] Appendix B contains implementation details (§ 5.1-§ 5.2).



(a) Figurative language: sarcasm detection using the *Sarcasm Headlines* (SH) corpus (Misra and Arora, 2023).

(b) Figurative language: metaphor detection using the *Trope Finder* (TroFi) corpus (Birke and Sarkar, 2006).

(c) Deviant language: sexism detection using the *Call me sexist but . . .* (CMSB) corpus (Samory et al., 2021).

Figure 2: Comparison of ClaD across three detection tasks (from top to bottom) – (a) sarcasm, (b) metaphors, and (c) sexism – against four transfer learning baseline results where Transformer-based models are fine-tuned on task-specific data: (from left to right) ALBERT, DistilBERT, SimCSE, and XLNet.

ClaD is benchmarked against two modern language model paradigms: (a) specialized encoder(-decoder) architectures optimized for language understanding: SimCSE (Gao et al., 2021), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019),[7] and (b) large language models (LLMs), primarily decoder-only architectures distinguished by their scale. We evaluate all models on 80/10/10 splits for train/dev/test, primarily focusing on the false positive rate (FPR) and F_1 score for the target class.[8]

5.1 Comparison with Encoder-based Models

Figure 2 illustrates ClaD's distinctive advantage: achieving performance competitive with or superior to these established models in just one training epoch. In contrast, transfer learning with these models typically requires 3-5 epochs to attain similar metrics across all three tasks. ClaD's rapid convergence yields substantial computational savings without compromising detection quality. For instance, in sarcasm detection, ClaD achieves a lower FPR after one epoch than most baselines do after five.[9] Similar patterns emerge across all tasks and metrics, where ClaD's single-epoch training matches or outperforms multi-epoch training of the transfer-learning baseline models. The results suggest that ClaD's geometric approach enables efficient adaptation to task-specific features, a finding further supported by our ablation study (§ 5.3).

5.2 Comparison with Large Language Models

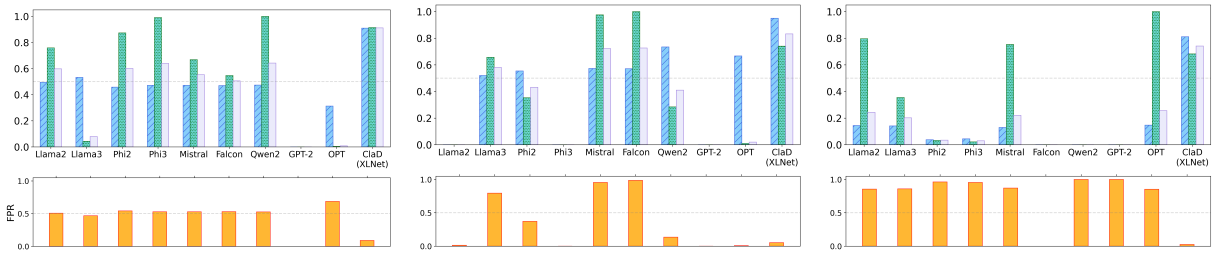
Next, we evaluate ClaD against a suite of recent large language models (LLMs): OPT, GPT, Phi, Llama, Mistral, Qwen, and Falcon. As ClaD is a training paradigm, and not a model, these evaluations are geared to answer two research questions:

Q1. *Is limited task-specific ClaD-training with*

[7] DeBERTa (He et al., 2021) performs much like ALBERT. So, for architectural diversity, we include XLNet instead.

[8] Reducing false positives in these tasks is particularly important in several applications such as social media moderation, so as to not penalize users for innocuous remarks

[9] Only XLNet marginally surpasses ClaD after epoch 2, with a difference of 0.036. A bootstrap analysis reveals this as statistically insignificant: ClaD's FPR (8.87%) is well within the 95% CI (5.05%, 9.93%) of XLNet's mean FPR.



(a) Figurative language: sarcasm (b) Figurative language: metaphors (c) Deviant language: sexism

Figure 3: Comparison of 5-shot evaluation of a suite of nine large language models (left to right): Llama2, Llama3, Phi2, Phi3, Mistral-7B, Falcon, Qwen2, GPT-2, and OPT, against ClaD's single-epoch training (rightmost).

small language models better than low-resource transfer-learning with LLMs?

Q2. *With identical training data, how much larger are the LLMs (if any) that outperform ClaD-training with small language models?*

Against all LLMs in zero-shot (Appendix C) and few-shot scenarios (discussed next), ClaD demonstrates consistently superior performance.

Few-shot Classification: As shown in Figure 3, ClaD retains its substantial advantage over few-shot classification (five instances) with LLMs, which achieve markedly lower F_1 scores: from 0.0% (GPT-2) to 64.2% (Qwen-2) on *SH*; 0.0% (GPT-2) to 72.6% (Falcon) on *TroFi*, and 0.0% (Falcon) to 25.6% (OPT) on *CMSB*. Models exhibit distinct characteristics in each task: for example, in metaphor detection, GPT-2, Llama2, and Phi3 completely avoid positive predictions, while Falcon predicts aggressively with perfect recall. The most dramatic changes are seen in deviant language detection, with more models completely avoiding the target class (Falcon, Qwen2, and GPT-2). OPT, on the other hand, exhibits perfect recall. Despite accuracy ranges similar to zero-shot, models show more extreme precision-recall trade-offs, with AUC scores stalled near 0.5 and persistently high FPR.

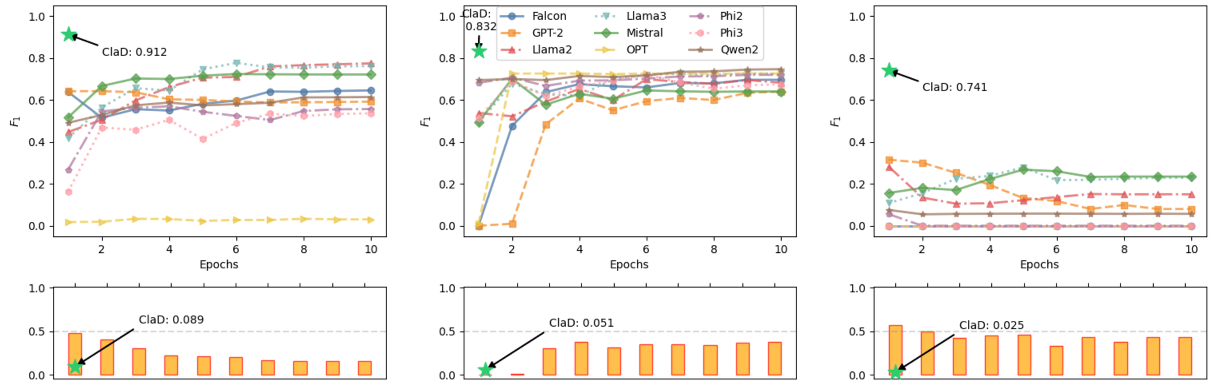
LLMs thus exhibit notable limitations and variability across all tasks, likely stemming from insufficient feature learning in low-resource scenarios (reflected in the AUC stagnation), causing them to fall back on their pretrained biases, particularly for subtle, context-dependent linguistic cues. The erratic behavior changes between zero- and few-shot settings also suggest unstable optimization paths, possibly due to the large parameter count in these models, stochasticity of gradient updates, and insufficient regularization.

Low-resource Training: We extend our analysis to low-resource training (with 100 instances provided to the LLMs) to examine whether this limited increase in data improves decision stability, precision-recall trade-offs, and task adaptation. ClaD's single-epoch training continues to outperform all LLMs in terms of both F_1 score and FPR, except in metaphor detection. There, although GPT-2, Falcon, and OPT report lower FPRs for the first two epochs, their F_1 scores are nearly zero as they completely avoid false positives (which comes at the cost of failing to avoid *any* positives). Figure 4 reveals clear patterns across model scales: while the smaller XLNet-based ClaD achieves superior performance within one epoch, larger models like Llama3 (8B parameters) require multiple epochs to reach their peak performance (*e.g.*, 77.6% F_1 at epoch 6 for sarcasm detection). Model size significantly impacts learning trajectories: large models (7B-8B parameters) show rapid initial improvements, mid-size models (2B-4B parameters) plateau early with suboptimal performance, and smaller models (124M-350M parameters) struggle to learn effectively. Sexism detection remains the most challenging task, with all LLMs showing conservative labeling of the target class and an inability to learn from limited data. In sarcasm detection, the only task where LLMs perform significantly better than chance, FPR correlates inversely with model size, ranging from 15.4% (Llama3, epoch 8) to 48.1% (GPT-2, epoch 10).

Identical Task-specific Training Data: To address our second research question, we compare ClaD's single-epoch training with smaller models against the suite of LLMs (Figure 5).[10]

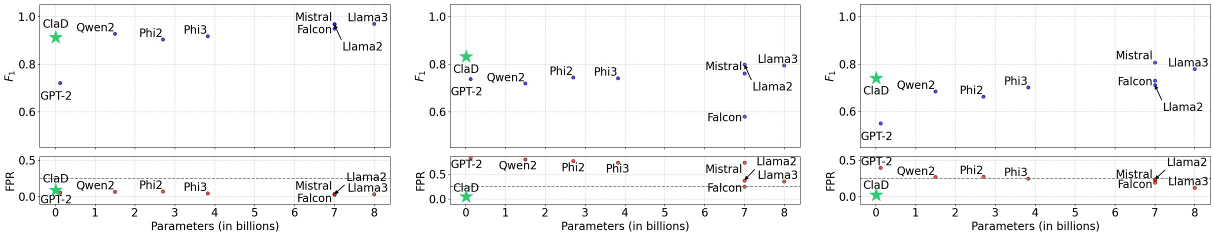
The relationship between model size and performance varies significantly across tasks. While larger models (7-8B parameters) generally perform well in sarcasm detection (F_1: 0.96-0.97), this advantage diminishes in metaphor detection (F_1:

[10] OPT markedly underperformed across all tasks and evaluation metrics, and is thus excluded in the comparison.



(a) Figurative language: sarcasm (b) Figurative language: metaphors (c) Deviant language: sexism

Figure 4: Comparison of ClaD across the three detection tasks against nine large language models (LLMs) in a limited data regime. The LLMs are trained on 100 instances over 10 epochs. Results shown for the target class are: (top) the F_1 scores; and (bottom) the false positive rates (FPR) for the best-performing LLM.



(a) Figurative language: sarcasm (b) Figurative language: metaphors (c) Deviant language: sexism

Figure 5: Comparison of ClaD across the three detection tasks against LLMs, with identical training data: all models utilize the entire training set for a single epoch. F_1 scores (*top*) show ClaD being competitive with most LLMs, and outperforming a few others, while false positive rates (FPR) (*bottom*) show ClaD remaining superior to the LLMs.

0.71-0.81) and almost disappears in sexism detection, where many smaller models achieve competitive performance. This suggests that larger models do not consistently translate to better performance across various language understanding tasks.

More often than not, performance improves with increased model size, but with diminishing returns. A striking pattern emerges in the false positive rates, however: while larger models show very low FPR in sarcasm detection (~ 0.03), their FPR varies widely in other tasks, sometimes performing worse than smaller models. Particularly interesting is deviant language (sexism) detection, where even the largest models struggle with high FPR. The consistent performance of XLNet-based ClaD, with only 110M params, especially in maintaining lower FPR across tasks while remaining competitive on F_1 scores, suggests that efficiency derived from a geometric understanding eclipses model size.

5.3 Ablation Experiments

We conduct systematic ablation experiments to evaluate the individual contributions of ClaD's core components: the novel loss function, the decision algorithm, and the Mahalanobis contrast mechanism. We present comparisons using SimCSE as the base model, as it achieved the lowest false positive rate in sarcasm detection.[11] The impact of our novel loss function is evident in Table 3. Compared to standard loss functions in task-specific fine-tuning, F_1 scores improve by 26.5%, 6%, and 13% for sarcasm, metaphor, and sexism detection, respectively (with corresponding proportionate decreases seen in FPR: 24.6%, 39.1%, and 15.3%). On the other hand, replacing our β-decision algorithm with a 3-layer fully connected feed-forward network for classification results in the F_1 scores dropping by 58.5%, 80.8%, and 46.0% on these tasks, respectively. Finally, we show in Appendix D that traditional one-class classification and anomaly detection methods do not perform well in these pragmatic language tasks where the minority target class requires gleaning from a large heterogeneous non-target majority.

[11] ClaD (XLNet) reports marginally better F_1, and experiments with XLNet as the base model yield similar results.

	Acc ↑	Pr ↑	FPR ↓	F_1 ↑
Sarcasm detection on *Sarcasm Headlines*				
$\mathcal{L}_{\text{MAH}}$ + β-decision	0.896	0.931	0.009	0.885
$\mathcal{L}_{\text{COSINE}}$ + β-decision	0.492	0.479	0.255	0.620
$\mathcal{L}_{\text{MAH}}$ + MLP	0.521	0.482	0.209	0.300
Metaphor detection on *TroFi*				
$\mathcal{L}_{\text{MAH}}$ + β-decision	0.794	0.857	0.167	0.808
$\mathcal{L}_{\text{COSINE}}$ + β-decision	0.675	0.667	0.558	0.748
$\mathcal{L}_{\text{MAH}}$ + MLP	0.433	0.000	0.000	0.000
Sexism detection on *Call Me Sexist But*				
$\mathcal{L}_{\text{MAH}}$ + β-decision	0.928	0.870	0.012	0.696
$\mathcal{L}_{\text{COSINE}}$ + β-decision	0.832	0.433	0.165	0.566
$\mathcal{L}_{\text{MAH}}$ + MLP	0.134	0.134	1.000	0.236

Table 3: Ablation study comparing ClaD's components: $\mathcal{L}_{\text{MAH}}$ (our Mahalanobis contrast loss), $\mathcal{L}_{\text{COSINE}}$ (the standard cosine similarity loss), and β-decision (Algorithm 1). The combination of $\mathcal{L}_{\text{MAH}}$ + β-decision achieves superior performance across all metrics, particularly in reducing false positive rates (FPR) while maintaining high F_1 and target-class precision.

6 Related Work

Extensive research aims to model pragmatic language nuances for respectful communication, advancing the identification of figurative language (Chakrabarty et al., 2022; Saakyan et al., 2022; Wachowiak and Gromann, 2023; Lai and Nissim, 2024) and deviant content (Fortuna and Nunes, 2018; Yin and Zubiaga, 2021; Guest et al., 2021; Bose and Su, 2022). Most leverage BERT-based supervised learning: e.g., BERT-BiLSTM for hate speech (Bose and Su, 2022), dual BERT models for metaphors (Wan et al., 2021), and BERT-LSTM for sarcasm (Kumar and Anand, 2020). Enhancements include syntactic (Wan et al., 2020) or semantic (Zhou et al., 2021) feature integration and multi-task frameworks (Safi Samghabadi et al., 2020). However, generalization remains limited, and ensemble methods (Lemmens et al., 2020; Gregory et al., 2020) trade interpretability for computational cost. Our Class Distillation (ClaD) paradigm addresses these gaps via an interpretable decision algorithm and novel loss function.

ClaD shares similarities with one-class classification, which detects anomalies by focusing on the target class. Common methods include one-class SVM (Schölkopf et al., 2001; Noumir et al., 2012), DeepSVDD (Ruff et al., 2018), and adversarial one-class classifiers (Sabokrou et al., 2018), but they often struggle with domain generalization, overfitting, and nuanced data – key challenges in pragmatic language tasks. ClaD, leveraging Mahalanobis contrast, effectively addresses these issues, demonstrated by ablation results in Appendix D.

LLMs like GPT-2 (Radford et al., 2019), Phi-2 (Javaheripi et al., 2023), and OPT (Zhang et al., 2022) excel in text classification but are computationally expensive (Wang et al., 2023). Some, like GPT-3, reportedly struggle with nuanced tasks like metaphor detection (Wachowiak and Gromann, 2023), while others face reasoning limitations and token constraints in in-context learning (Sun et al., 2023). Unlike recent efforts to address these issues, ClaD combines Mahalanobis contrast with *smaller* models, to *efficiently* learning task manifolds.

7 Conclusion

This work challenges a fundamental implicit assumption in modern NLP: that scale (in models, pretraining data, or fine-tuning) guarantees superior downstream performance. Through rigorous empirical analysis, we demonstrate that our geometrically grounded training paradigm surpasses state-of-the-art LLMs by significant margins in low-data regimes, achieving superior results in a single epoch where larger models plateau after several. Notably, ClaD matches or exceeds the performance of models nearly two orders of magnitude larger, *even with identical task-specific training*.

Our findings align with broader trends toward efficiency, spurred by DeepSeek's compute-optimal scaling (Liu et al., 2024; Guo et al., 2025), and reveal a novel insight: **architectural minimalism**, coupled with **geometric alignment to task manifolds**, can unlock capabilities previously thought to require massive scale. While recent work optimizes *how* to scale, we demonstrate that *whether to scale* depends critically on data geometry. ClaD's innovations – manifold-aware training, Mahalanobis contrast, and the decision algorithm – prove that for nuanced language understanding tasks, modeling latent structure trumps brute-force scaling.

Our work does not negate scaling, but expands the efficiency frontier, showing that geometric principles can supplant scale and provide a complementary pathway for real-world applications. As AI research increasingly prioritizes *efficiency* alongside performance – whether through scaling laws, sparsity, or geometric learning – our findings position the geometric understanding of data as a foundational pillar of sustainable NLP.

Limitations

While our proposed **Cla**ss **D**istillation (ClaD) paradigm demonstrates consistently strong performance and efficiency across sarcasm, metaphor, and sexism detection tasks – outperforming smaller Transformer models with equal training, and LLMs with in limited resouce settings – several limitations should be acknowledged. First, although we tested ClaD on diverse tasks encompassing figurative and deviant language, the chosen benchmarks (Sarcasm Headlines, TroFi, and CMSB) may not fully capture the richness of real-world scenarios, and more domain-specific or multilingual tasks could present additional linguistic and cultural nuances not addressed in our current evaluation. Whether our approach can be generalized to specialized domains like legal or clinical tasks also remains to be seen. Second, ClaD relies on the ability of the target class manifold to be modeled as a multivariate normal distribution. While our experiments suggest that training can nudge embeddings closer to a normal manifold, this assumption may not hold universally: certain representations or highly imbalanced corpora may exhibit multimodal or heavy-tailed distributions that deviate substantially from normality, potentially affecting performance. Third, although ClaD's fast convergence leads to significant computational savings compared to multi-epoch fine-tuning, maintaining a dynamically updated covariance matrix in the Mahalanobis distance computation can be memory-intensive for very large datasets. Further advances in this line of research will likely require more memory-efficient approximations or low-rank updates.

These limitations may be addressed by exploring alternative distributional assumptions (*e.g.*, Gaussian mixtures) to accommodate more complex embedding spaces, conducting broader evaluations across languages and task domains, and developing lightweight variants of Mahalanobis-based training to reduce the memory overhead. They have the potential to further enhance ClaD's versatility and impact in real-world applications.

Ethics Statement

This work adheres to ethical standards in NLP research by ensuring transparency, reproducibility, and fairness in our experiments. Our study does not involve human subjects or sensitive data, and all datasets used are publicly available with appropriate licenses. While our findings highlight efficiency gaps in large-scale language models, we acknowledge that their broader societal impacts, including biases and potential misuse, require further investigation. We encourage responsible deployment of our proposed methods and emphasize the need for continued ethical scrutiny in model development and evaluation.

References

Theodore W. Anderson and Donald Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212.

Angeliki Athanasiadou. 2024. On the margins of figurative thought and language. *Lingua*, 299:103655.

Richard Bellman. 1957. *Dynamic Programming*. Princeton University Press.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Julia Birke and Anoop Sarkar. 2006. A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Saugata Bose and Guoxin Su. 2022. Deep one-class hate speech detection model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7040–7048.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. *arXiv*, 2205.12404v3.

Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. *arXiv*, 1901.03407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: from identification, interpretation, generation to application. *Artif. Intell. Rev.*, 56(2):1829–1895.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A Report on the 2020 Sarcasm Detection Shared Task.

Hunter Gregory, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin. 2020. A transformer approach to contextual sarcasm detection in twitter. In *Proceedings of the second workshop on figurative language processing*, pages 270–275.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv*, 2501.12948.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT With Disentangled Attention. In *International Conference on Learning Representations*.

Norbert Henze and Bernd Zirkler. 1990. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-theory and Methods*, 19:3595–3617.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321 – 377.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Kristina Kasparian. 2013. Hemispheric differences in figurative language processing: Contributions of neuroimaging methods and challenges in reconciling current empirical findings. *Journal of Neurolinguistics*, 26(1):1–21.

Roger Kiska. 2012. Hate Speech: A Comparison between the European Court of Human Rights and the United States Supreme Court Jurisprudence . *Regent University Law Review*, 25:107.

Amardeep Kumar and Vivek Anand. 2020. Transformers on sarcasm detection with context. In *Proceedings of the second workshop on figurative language processing*, pages 88–92.

Huiyuan Lai and Malvina Nissim. 2024. A Survey on Automatic Generation of Figurative Language: From Rule-based Systems to Large Language Models. *ACM Comput. Surv.*, 56(10).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations (ICLR)*.

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *proceedings of the second workshop on figurative language processing*, pages 264–269.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, et al. 2024. DeepSeek-V3 Technical Report. Technical report, DeepSeek-AI.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2010. On Detecting Clustered Anomalies Using SCiForest. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–290, Berlin, Heidelberg. Springer Berlin Heidelberg.

Prasanta Chandra Mahalanobis. 1936. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

Rishabh Misra and Prahal Arora. 2019. Sarcasm Detection using Hybrid Neural Network. *arXiv*, 1908.07414.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Zineb Noumir, Paul Honeine, and Cédue Richard. 2012. On simple one-class classification methods. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2022–2026.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Karl Pearson. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559 – 572.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3980–3990. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Michel Rosenfeld. 2002. Hate Speech in Constitutional Jurisprudence: A Comparative Analysis. *Cardozo Law Review*, 24:1523.

Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4390–4399. PMLR.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A Report on the FigLang 2022 Shared Task on Understanding Figurative Language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. 2018. Adversarially learned one-class classifier for novelty detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3379–3388. Computer Vision Foundation / IEEE Computer Society.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8990–9005. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579 – 2605.

Dimitrios Ververidis and Constantine Kotropoulos. 2008. Gaussian mixture modeling by exploiting the mahalanobis distance. *IEEE transactions on signal processing*, 56(7):2797–2811.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing Metaphor Detection by Gloss-based Interpretations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1971–1981, Online. Association for Computational Linguistics.

Mingyu Wan, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *CoRR*, abs/2312.01044.

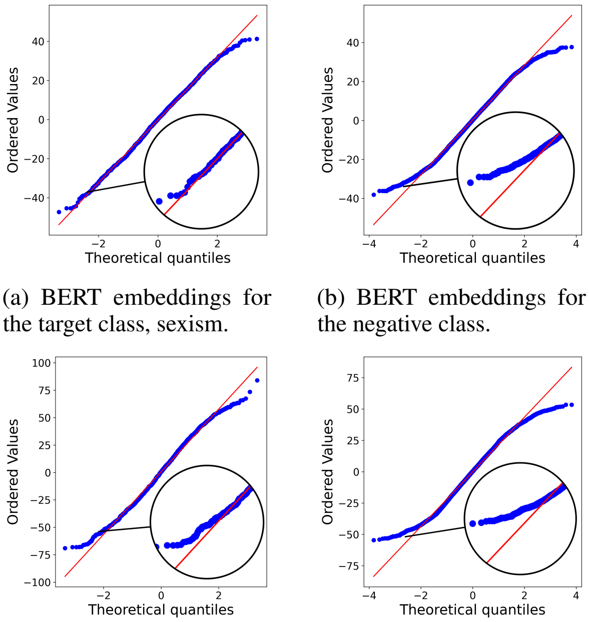
Samuel S Wilks. 1962. Mathematical statistics. a wiley publication in mathematical statistics john wiley & sons. *Inc., New York-London*, pages 0173–45805.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv*, 2205.01068.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7158–7166.



(a) BERT embeddings for the target class, sexism.

(b) BERT embeddings for the negative class.

(c) SimCSE embeddings for the target class, sexism.

(d) SimCSE embeddings for the negative class.

Figure 6: Q-Q (quantile-quantile) plots to assess the goodness-of-fit of target and non-target classes in the CMSB corpus for sexism detection. Shown here are the first t-SNE dimensions of pretrained BERT-base (*a* and *b*) and SimCSE (*c* and *d*) embeddings.

A A Visual Approach to Goodness-of-Fit in Target Class Distributions

Our aim is to unveil the distributional properties shared across diverse datasets and tasks, where identifying the minority target class amidst a spectrum of heterogeneous linguistic expressions (with no common pragmatic language function) is paramount. It is well known, however, that there are challenges to such analyses regardless of the manifold structure. As the number of dimensions increase, the volume of space grows exponentially and tests based on density estimation or empirical distance measures can struggle to maintain accuracy due to the increased sparsity and spread of data points.[12] Statistical tests also suffer from reduced power in higher dimensions. Moreover, estimating the covariance matrix becomes problematic in higher dimensions (for instance, due to ill-conditioned or singular matrices).

To mitigate these problems, we reduce the number of dimensions using t-SNE (van der Maaten and Hinton, 2008).[13] The Quantile-Quantile (Q-Q) plots of the first latent dimension are shown in Figure 6, for visual assessment of adherence to a normal distribution. For the sake of brevity, we present only the Q-Q plots for sexism detection using two language models, BERT (Devlin et al., 2019) and SimCSE (Gao et al., 2021), as similar patterns were observed for the other tasks.

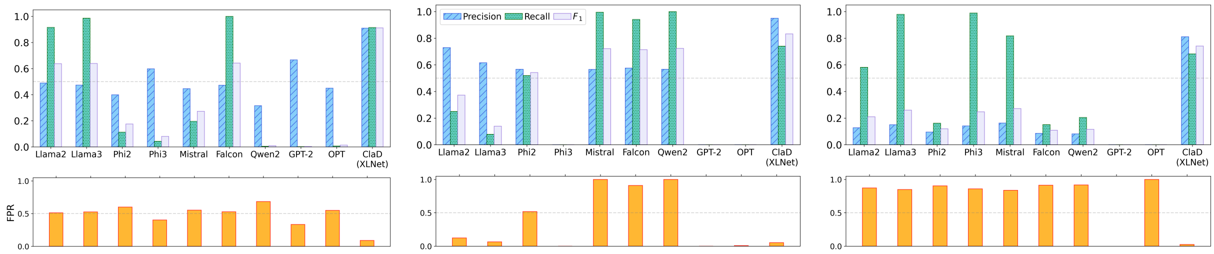
B Configuration

Baseline Models: Four baseline models were fine-tuned – SimCSE, ALBERT, XLNet, and DistilBERT – on three datasets using a consistent set of configurations. The models were trained with a per-device batch size of 16 for both training (up to 5 epochs) and evaluation (at the end of each epoch). The learning rate is set to 1×10^{-5} for all models, with 50 warm-up steps and a weight decay of 0.01. For tokenization, inputs were padded and truncated to a maximum sequence length of 512. We used the Adam optimizer for parameter updates.

Few-shot and Low-resource Experiments: We conducted fine-tuning experiments for the classification models under both few-shot and low-resource scenarios. In the few-shot setting, we used 5 training samples, while in the low-resource setting, we used 100 training samples. Training was conducted up to 10 epochs. We selected a variety of mainstream LLMs: Falcon, GPT-2, Llama2, Llama3, Mistral-7B, OPT, Phi2, Phi3, and Qwen2. These models are loaded via the `AutoModelForSequenceClassification` module provided by HuggingFace Transformers. To train the models with limited computational resources, we employed 4-bit quantization (e.g., `nf4`) in conjunction with Low-Rank Adaptation (LoRA) (Hu et al., 2022) for efficient parameter tuning. Specifically, we set the LoRA rank (`r`) to 16 and the LoRA scaling factor (`lora_alpha`) to 8 (32 for GPT-2), with a dropout rate (`lora_dropout`) of 0.05 (0.1 for GPT-2). For optimization, HuggingFace `Trainer` was used with its default settings, with cross-entropy loss adopted for binary classification. The learning rate was set to 5×10^{-5}, the weight decay set to 0.01, and the batch sizes were 24 for training and 6 for validation (per device).

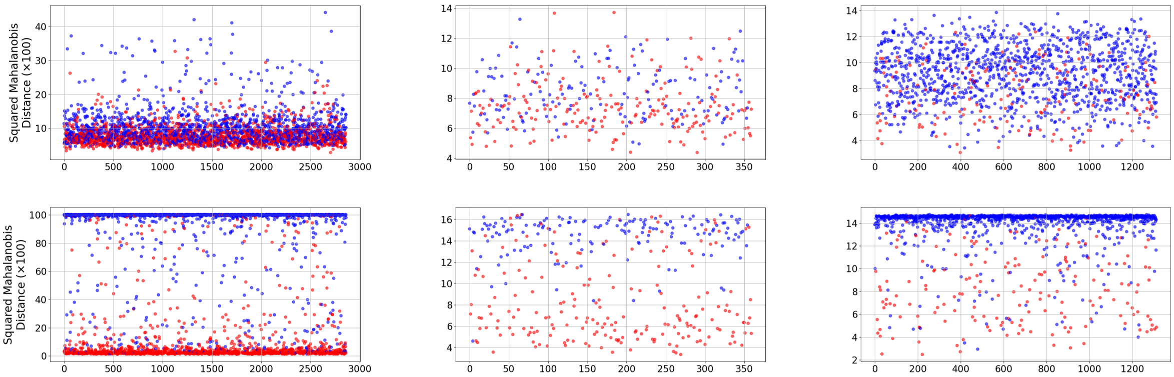
[12] The "curse of dimensionality" strikes again, as Bellman (1957) presciently described this exponential increase in problem complexity with growing number of dimensions.

[13] We also experiment with dimensionality reduction by means of studying anisotropy (Ethayarajh, 2019) and dominant dimensions as defined by Timkey and van Schijndel (2021), as well as with principal component analysis (Pearson, 1901; Hotelling, 1936). In each case, the results of the statistical tests of manifold structure are nearly identical. So, for conciseness, we omit the details of these other approaches.



(a) Figurative language: sarcasm (b) Figurative language: metaphors (c) Deviant language: sexism

Figure 7: Comparison of zero-shot evaluation of a suite of nine large language models (left to right): Llama2, Llama3, Phi2, Phi3, Mistral-7B, Falcon, Qwen2, GPT-2, and OPT, against ClaD's single-epoch training (rightmost).



(a) Sarcasm detection (*SH*) (b) Metaphor detection (*TroFi*)) (c) Sexism detection (*CMSB*)

Figure 8: Scatter plots of squared Mahalanobis distance for three test datasets before and after training. Red and blue indicate the target and non-target classes, respectively.

Zero-shot Experiments: The Transformer library's `AutoModelForSequenceClassification` API is used. This API automatically adds a linear classification head on top of the model's pooled output, enabling it to be handled as logits for classification tasks. Specifically, for binary classification, the output consists of 2D logits, representing the likelihood of the input belonging to either class. This approach focuses on *classification*, as opposed to letting a LLM *generate* its answer (e.g., `1` or `0`) in response. To guide the classification, the following prompt is prepended to every input:

```
Please identify if the following text is an
example of <task-word>.  Reply with 1 if it
exhibits <task-word>, and 0 otherwise:
<input sentence>
```

where the placeholder `<task-word>` is replaced by the specific task of interest (i.e., sarcasm, metaphor, sexism), and `<input sentence>` is the text being classified. This enables the model to classify the input sentence based on the particular task while maintaining the flexibility to adapt to various tasks by simply changing the task-word.

Class Distillation Experiments: We used a sliding window mechanism (window size: 100 × batch size for small datasets, and expanded to 500 × batch size for large datasets, update frequency: batch size) to efficiently update the Mahalanobis distance parameters (mean and covariance matrix) during training by incrementally processing batch data and computing statistics using the latest model-generated embeddings, dynamically adapting the parameters while maintaining computational efficiency. We use batch sizes of 16 for the sarcasm detection and metaphor tasks, and 40 for the sexism detection task.

C Zero-shot classification

Converting LLM logits to probabilities via softmax, we observe that ClaD's single-epoch training substantially outperforms zero-shot predictions (shown in Figure 7. While this is perhaps unsurprising, the magnitude of improvement is notable: F_1 score improvements range from +27.0% (vs. Falcon) to +90.9% (vs. GPT-2) on *SH*, +10.8% (vs. Qwen2) to +83.2% (vs. OPT) on *TroFi*, and +46.9% (vs. Mistral) to +74.1% (vs. OPT) on *CMSB*. In sar-

	Acc	Pr	FPR	F_1
Sarcasm detection on *Sarcasm Headlines*				
Mahalanobis β-decision	0.580	0.557	0.100	0.547
One-class SVM	0.494	0.465	0.494	0.473
Isolation Forest	0.472	0.472	0.998	0.641
Autoencoder	0.530	0.517	0.046	0.099
Metaphor detection on *TroFi*				
Mahalanobis β-decision	0.678	0.680	0.500	0.741
One-class SVM	0.614	0.602	0.814	0.734
Isolation Forest	0.614	0.610	0.737	0.721
Autoencoder	0.578	0.575	0.942	0.724
Sexism detection on *Call Me Sexist But*				
Mahalanobis β-decision	0.134	0.134	1.000	0.236
One-class SVM	0.602	0.141	0.364	0.206
Isolation Forest	0.135	0.133	0.996	0.234
Autoencoder	0.827	0.121	0.051	0.066

Table 4: Comparing Mahalanobis β-decision (Algorithm 1) and standard outlier detection methods (One-class SVM, Isolation Forest, and Autoencoder), demonstrating that the former achieves higher F_1 scores and lower false positive rates (FPR) across most tasks.

casm detection, LLMs perform near-randomly (accuracies: 0.47-0.53, AUC ≈ 0.5), with models either aggressively over-predicting (Falcon, Llama2, Llama3: recall ≥ 0.91, but very low precision) or being overly conservative (GPT-2, Phi3: extremely low recall with moderate precision). Metaphor detection shows modest improvements (accuracies: 0.43-0.57, max. AUC: 0.56), though extreme behaviors persist: Falcon, Qwen2 and Mistral favor recall over precision, while Llama2 do the opposite. Sexism detection reveals poor adaptation to class imbalance, with extremely high FPR ($\geq 83\%$) across all LLMs. Some models also exhibit task-specific inconsistencies, such as GPT-2 alternating between conservative and aggressive predictions in sarcasm and metaphor detection, respectively.

D Comparing Anomaly Detection Methods

Additionally, Table 4 shows that treating deviant or figurative language merely as "out of the ordinary" is insufficient. We compare our β-decision algorithm against traditional anomaly detection methods like isolation forest (IF) (Liu et al., 2008, 2010), autoencoders (Chalapathy and Chawla, 2019), and one-class SVM (Noumir et al., 2012) – all on the same pretrained model – and further demonstrate the necessity of an inference algorithm based on an understanding of the target class manifold. Our novel decision algorithm (Algorithm 1) achieves superior F_1 scores across all tasks, with one exception: IS performs better on sarcasm detection. However, this happens at the expense of significantly higher FPR and lower target-class precision.

E Mahalanobis Contrast and Separability

Training with our Mahalanobis mean loss function, $\mathcal{L}_{\text{MAH},\mu}$ (Equation 2) has a significant impact on classification, evident in the results presented in our ablation experiments subsection 5.3. Here, we add visualizations of class separability, using the squared Mahalanobis distance (Equation 3). Figure 8 presents scatter plots of squared Mahalanobis distance for the three test datasets before and after training, where the target and non-target class instances are shown in red and blue, respectively. It is clearly demonstrated that training with our Mahalanobis loss function leads to a distinct increase in class separability.