Learning with Few: A Comparative Study of Multilingual Text Anomaly Detection

Fabio Masaracchia Maia[1], **Anna Helena Reali Costa**[1]

[1]Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil

fabio.masaracchia@gmail.com, anna.reali@usp.br

Abstract. *Detecting anomalies in textual data is a critical task in domains such as content moderation, fraud detection, and risk monitoring. However, this remains challenging due to the semantic complexity of language and the scarcity of labeled anomalies in real-world scenarios. This paper presents a comprehensive benchmark study that integrates multiple perspectives: representation strategies, learning paradigms, and linguistic diversity. We evaluate unsupervised and semi-supervised models—including deep learning approaches—across datasets in both Portuguese and English. Additionally, we assess the impact of sentence embeddings, comparing multilingual encoders with language-specific models. Our findings show that representation choice and limited supervision strongly influence model performance in few-shot settings.*

1. Introduction

Anomaly detection aims to identify patterns in data that deviate from expected behavior and plays a critical role in applications such as fraud detection, hate speech moderation, and quality assurance. While this task has been extensively studied in structured domains, its application to unstructured text presents unique challenges due to linguistic variability, lack of structure, and high dimensionality [Chandola et al. 2009, Pang et al. 2021].

In recent years, deep learning has driven progress in anomaly detection, particularly through unsupervised and semi-supervised methods capable of capturing complex data representations. However, many models were initially developed and tested on structured or visual datasets [Pang et al. 2021], such as DeepSAD and Deep SVDD on image datasets like MNIST and CIFAR-10 [Ruff et al. 2018, Ruff et al. 2020], or DevNet [Pang et al. 2019] on tabular and high-dimensional vectorized data.

In the textual domain, research often repurposes English-language classification datasets—such as 20 Newsgroups, IMDB, AG News, and Reuters-21578—for anomaly detection tasks, despite their original design for supervised learning [Ruff et al. 2019b, Manolache et al. 2021, Xu et al. 2023]. This has led to a lack of multilingual benchmarks, especially for low-resource languages like Portuguese. A recent study [Maia and Costa 2024] addressed this gap using a semi-supervised approach on two Portuguese datasets. Expanding on this line of research, the present study benchmarks a broader range of models, supervision levels, and embedding strategies across both Portuguese and English corpora. The results extend previous insights, offering a more comprehensive perspective on multilingual anomaly detection.

A central challenge in anomaly detection is the scarcity of labeled anomalies. While unsupervised methods are traditionally favored for their ability to detect rare patterns without annotation, there is growing interest in semi-supervised strategies that use a small number of labeled outliers to improve performance [Pang et al. 2021, Xu et al. 2023]. These approaches are particularly relevant in real-world scenarios where annotation is expensive or infeasible, aligning naturally with few-shot learning paradigms.

This work provides a systematic benchmark of text-based anomaly detection under realistic constraints. We evaluate both unsupervised and semi-supervised models across diverse textual domains and linguistic contexts, comparing multilingual and language-specific embeddings on datasets that differ in complexity, structure, and language (Portuguese and English). By analyzing model behavior across different levels of supervision, we identify effective combinations of architecture, representation, and linguistic setting. This deep dive into learning with limited labels contributes to our understanding of how to improve anomaly detection in text, particularly under multilingual and low-resource conditions.

To guide our investigation, we structure the study around the following research questions:

- **RQ1:** Does supervision improve anomaly detection performance?
- **RQ2:** How does representation type affect performance?
- **RQ3:** Which models are most robust across tasks and datasets?
- **RQ4:** How much supervision is needed to outperform unsupervised methods?

2. Related Work

Anomaly detection has been widely studied in machine learning, with classical approaches such as One-Class SVM (OC-SVM) [Schölkopf et al. 2001], Isolation Forest (IForest) [Liu et al. 2008], and Local Outlier Factor (LOF) [Breunig et al. 2000] relying on geometric or density-based heuristics to identify deviations. These models typically operate in unsupervised or shallow semi-supervised settings and have served as the foundation for many traditional applications.

More recently, deep learning has enabled more expressive and scalable anomaly detection methods. Deep SVDD [Ruff et al. 2018] and its semi-supervised extension DeepSAD [Ruff et al. 2020] aim to learn compact latent representations of normal data. DevNet [Pang et al. 2019] introduces a deviation loss to explicitly separate labeled anomalies from normal instances, showing strong performance under limited supervision. Hybrid methods such as XGBOD [Zhao and Hryniewicki 2018] combine unsupervised outlier scores with a supervised XGBoost classifier, while REPEN [Liang et al. 2018] learns embeddings optimized for random distance-based detectors. Additionally, deep generative models like AutoEncoder [Hinton and Salakhutdinov 2006] and Variational AutoEncoder (VAE) [Kingma and Welling 2013] are frequently used as unsupervised baselines, alongside classical algorithms like OC-SVM [Schölkopf et al. 2001], IForest [Liu et al. 2008], and SVDD [Tax and Duin 2004].

Despite recent advances, many of these models were initially developed for structured or visual domains, with benchmarks commonly based on tabular datasets such as Arrhythmia and Shuttle [Pang et al. 2019], or image datasets like MNIST and CIFAR-10 [Ruff et al. 2019b, Ruff et al. 2020]. Their generalizability to unstructured text data remains limited [Pang et al. 2021, Xu et al. 2023], highlighting the need for domain-specific adaptations in natural language settings.

To address this, recent work has explored models tailored to textual data. CVDD [Ruff et al. 2019a] applies multi-head self-attention to pre-trained word embeddings for context-aware, interpretable anomaly detection. DATE [Munoz-Galeano et al. 2021] adopts a self-supervised approach by corrupting tokens and training a discriminator to detect perturbations. These approaches better exploit linguistic structure but are still primarily evaluated on English datasets such as 20 Newsgroups [Lang 1995] and AG News [Zhang et al. 2015]. Surveys confirm that most anomaly detection research in text focuses on English corpora, with limited attention to multilingual or low-resource settings [Pang et al. 2021, Boutalbi et al. 2023]. Xu et al. [Xu et al. 2023], for instance, conducted a broad comparison involving 22 algorithms across 17 corpora, yet lacked a systematic evaluation of deep models for Portuguese, reducing the generalizability of their conclusions.

In this context, evaluating embedding strategies becomes essential. Prior work has emphasized the importance of representation learning in anomaly detection [Pang et al. 2021], but few studies have systematically compared multilingual and language-specific embeddings in this setting. Multilingual models such as DistilUSE [Reimers and Gurevych 2019] and XLM-R [Conneau et al. 2020] are commonly used due to their broad cross-lingual coverage, while Portuguese-specific models like BERTimbau [Souza et al. 2020] and Serafim [Gomes et al. 2024] offer embeddings tailored to linguistic nuances of the language. However, their comparative effectiveness across tasks such as sentiment analysis(SA), hate speech detection(HS), and topic classification(TC)—particularly under semi-supervised anomaly detection scenarios—remains underexplored.

3. Methodology

We address the task of anomaly detection in textual data under two distinct learning scenarios. Let $X = \{x_1, x_2, \ldots, x_N\}$ represent the training dataset with N instances. Each $x_i \in X$ corresponds to a textual document.

- **Unsupervised setting:** All instances in X are treated as unlabeled during training. The objective is to detect anomalies without any label information.
- **Semi-supervised setting:** A small subset $A \subset X$ containing K labeled anomalies is available for training, with $K \ll N$. The remaining $(N - K)$ instances are treated as inliers during model training.

Each instance $x_i \in X$ is transformed into a dense vector $z_i \in \mathbb{R}^d$ using a fixed or pre-trained encoder, forming the embedding set $Z = \{z_1, z_2, \ldots, z_N\}$. The model then learns a scoring function $\phi : \mathbb{R}^d \to \mathbb{R}$ that assigns an anomaly score to each instance, with higher scores indicating a greater likelihood of being anomalous.

The central challenge lies in distinguishing true anomalies from semantically novel but legitimate samples, particularly in low-supervision scenarios where models must generalize from very few labeled outliers. This is especially difficult in the unsupervised case, where no anomaly examples are provided and decisions rely solely on distributional assumptions. We compare shallow and neural models under both unsupervised and weakly-supervised conditions, focusing on how different architectures handle the scarcity of labeled data. Fully supervised approaches are not included in this study, as they require extensive labeled datasets and shift the task toward standard classification. Such methods are less representative of realistic anomaly detection scenarios, where anomalies are rare, diverse, and often unlabeled [Chandola et al. 2009, Pang et al. 2021].

3.1. Models and Training Strategies

To enable a comprehensive benchmark, we evaluate a wide range of models across both paradigms:

- **Unsupervised models:** One-Class SVM (OCSVM), Isolation Forest (IForest), Local Outlier Factor (LOF), Histogram-based Outlier Score (HBOS), Kernel Density Estimation (KDE), AutoEncoder (AE), Variational AutoEncoder (VAE), and DeepSVDD. These models rely on statistical, geometric, density-based, or reconstruction-based criteria and operate without any labeled data.
- **Semi-supervised models:** DeepSAD [Ruff et al. 2020], DevNet [Pang et al. 2019], XGBOD [Zhao and Hryniewicki 2018], and a feed-forward MLP trained using a small set of labeled anomalies. While our primary focus is on neural approaches, we include XGBOD as a shallow semi-supervised baseline that combines unsupervised outlier scores with a boosted tree classifier.

From a modeling perspective, our benchmark spans a diverse set of anomaly detection strategies commonly found in the literature. Table 1 summarizes the evaluated models, indicating their supervision type, architectural depth, and modeling approach. This categorization allows us to explore the interplay between supervision, model complexity, and anomaly detection performance.

Our selection includes classical one-class classifiers (e.g., OCSVM, DeepSVDD, DeepSAD), reconstruction-based models (e.g., AE, VAE), and ensemble methods (e.g., XGBOD). Predictive neural networks such as DevNet and MLP are also evaluated, along with clustering-based (e.g., LOF), probabilistic (e.g., KDE), and proximity-based (e.g., HBOS) techniques. This taxonomy is guided by prior surveys on anomaly detection in both general and textual domains [Pang et al. 2021, Xu et al. 2023]. All models are trained using a consistent evaluation pipeline. In the semi-supervised setting, we simulate label scarcity by randomly sampling a small number of labeled anomalies. This ensures comparability across methods and aligns with the few-shot assumptions of real-world anomaly detection.

3.2. Embedding Strategies

Text instances are transformed into fixed-size embeddings using a set of pre-trained Transformer-based models. These encoders capture either multilingual or Portuguese-specific semantics and are used to generate dense representations $z_i \in \mathbb{R}^d$ for downstream anomaly detection models. We use the SentenceTransformer framework [Reimers and Gurevych 2019] when available, and implement custom encoding logic for models not natively supported (e.g., BERTimbau variants), applying either [CLS] token extraction or mean pooling over contextualized token embeddings.

Table 2 summarizes the embedding models employed in this study. Multilingual models—such as DistilUSE [Reimers and Gurevych 2019] and XLM-R [Conneau et al. 2020]—and Portuguese-specific alternatives like BERTimbau [Souza et al. 2020] and Serafim [Gomes et al. 2024] enable a contrast between

Table 1. Summary of anomaly detection models evaluated in this benchmark.

Model	Supervision	Architecture	Modeling Strategy
OCSVM	Unsupervised	Shallow	One-class classification
IForest	Unsupervised	Shallow	Ensemble / Tree-based
LOF	Unsupervised	Shallow	Clustering / Density
HBOS	Unsupervised	Shallow	Proximity-based
KDE	Unsupervised	Shallow	Probabilistic
AE	Unsupervised	Deep	Reconstruction-based
VAE	Unsupervised	Deep	Generative / Reconstruction-based
DeepSVDD	Unsupervised	Deep	One-class classification
DeepSAD	Semi-supervised	Deep	One-class (with entropy loss)
DevNet	Semi-supervised	Deep	Predictive / Deviation loss
XGBOD	Semi-supervised	Mixed	Ensemble (unsupervised + XGBoost)
MLP	Semi-supervised	Deep	Predictive

Table 2. Pre-trained sentence encoders used for embedding text instances.

Short Name	Model Name	Type	Language	Dim
distiluse-v1	distiluse-base-multilingual-cased-v1	Sentence-BERT	Multilingual	512
distiluse-v2	distiluse-base-multilingual-cased-v2	Sentence-BERT	Multilingual	512
XLM-RoBERTa	xlm-roberta-large	Transformer	Multilingual	1024
BERT-base-PT	bert-base-portuguese-cased	Transformer	Portuguese	768
BERT-large-PT	bert-large-portuguese-cased	Transformer	Portuguese	1024
Serafim	serafim-100m-portuguese-pt	Transformer	Portuguese	768

general-purpose and language-tuned representations. These models offer a diverse range of representational scopes and dimensionalities, enabling a comparative analysis between general-purpose multilingual encoders and models tuned specifically for Portuguese.

3.3. Datasets and Evaluation

We benchmark models on six datasets spanning both English and Portuguese corpora. These datasets cover diverse tasks, including hate speech detection, sentiment analysis, and topic classification. To ensure comparability and computational feasibility, we limit each dataset to a maximum of 60,000 samples using stratified sampling when applicable.

For consistency across experiments, we adopt a common anomaly designation strategy: the most frequent class in each dataset is treated as the inlier (normal), while all other classes are considered anomalies. This approach simulates a realistic scenario in which one dominant class is well-represented, and deviations from it are rare or unexpected. To align with anomaly detection assumptions and ensure fair comparison across encoders and models, all datasets are subsampled to maintain a fixed contamination rate of 5% . This design reflects the low-prevalence nature of anomalies while enabling controlled evaluation across diverse configurations.

The selected datasets include: WikiNews (pt) [Garcia et al. 2024], 20 Newsgroups [Lang 1995], Portuguese Tweets [Portuguese Tweets Dataset 2018], TweetEval [Rosenthal et al. 2017], TOLD-Br [Leite et al. 2020], and Hate Speech Tweets [Sharma 2018]. Table 3 provides an overview of each dataset, including task type, language, number of samples used in our experiments, and the class labels designated as anomalous and normal.

Table 3. Datasets used in the benchmark. All datasets were standardized to 5% anomaly rate during training by subsampling the anomaly class. The "Size" column shows the number of samples before contamination adjustment.

Dataset	Source	Task	Lang	Size	Anomaly Tag(s)	Orig. Anom.%	Normal Tag(s)
WikiNews	Wikimedia Dump	TC	pt	10,581	All but "Política"	44.91%	"Política"
20 Newsgroups	SetFit / Hugging-Face	TC	en	18,846	All but "comp.graphics"	95.76%	"comp.graphics"
Portuguese Tweets	Kaggle	SA	pt	60,000	Positivo	33.48%	Negativo
TweetEval	HuggingFace	SA	en	59,899	Negative	18.99%	Neutral / Positive
TOLD-Br	Twitter/Reddit	HD	pt	21,000	Hate	44.07%	Non-hate
Hate Speech Tweets	HuggingFace	HD	en	31,962	Hate	7.01%	Non-hate

4. Results and Discussion

All experiments were conducted using a unified benchmarking pipeline to ensure consistency across models and datasets. We evaluate both unsupervised and semi-supervised scenarios. In the semi-supervised setup, anomaly labels are provided for all anomalous training samples. All remaining training samples are treated as inliers. This simulates a realistic few-shot scenario in which only a small fraction of data is explicitly labeled as anomalous. Reported results aggregate performance across different combinations of models, datasets, and embeddings, reflecting variability from these factors.

Experiments were executed in a Python 3 environment with a single NVIDIA Tesla T4 GPU (16 GB memory), running CUDA 12.4. All models were implemented using a combination of Scikit-learn [Pedregosa et al. 2011], PyOD [Zhao et al. 2019], and DeepOD [Xu 2022], with custom wrappers used for unified evaluation and embedding integration.[1]

4.1. RQ1: Does Supervision Improve Anomaly Detection Performance?

Figure 1 presents the distribution of AUC scores across all datasets under both unsupervised and semi-supervised settings. The results reveal a consistent pattern: the inclusion of minimal supervision (5% labeled anomalies) leads to substantial performance improvements. Across all datasets, semi-supervised models achieve higher median AUC scores with significantly reduced variance.

This effect is especially evident in semantically complex tasks like sentiment analysis and hate speech detection, where unsupervised models often fail to capture ambiguity, subtle phrasing, irony, or culturally specific hostility. For example, in TweetEval (en) and

[1] The source code is available at https://github.com/FabioMMaia/Multilingual-Text-Anomaly-Detection.

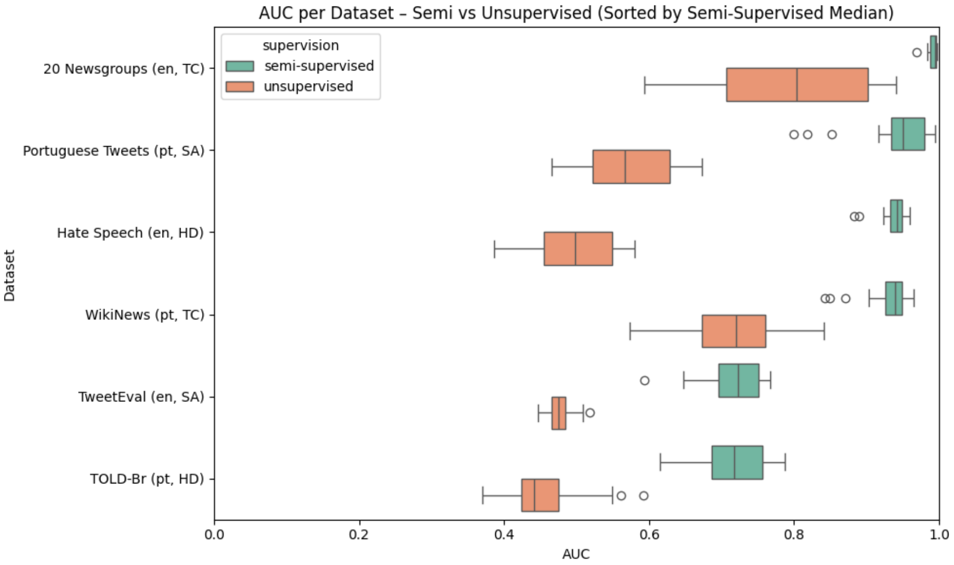

Figure 1. AUC per dataset comparing semi-supervised and unsupervised settings. Datasets are annotated with language and task type, and sorted by median AUC in the semi-supervised setting.

Portuguese Tweets (pt), distinguishing sentiment nuances is challenging without labels. Similarly, hate speech detection tasks such as TOLD-Br and Hate Speech Tweets benefit greatly from even a small number of labeled examples, which helps models establish clearer decision boundaries. By contrast, topic classification datasets like 20 Newsgroups (en) and WikiNews (pt) contain more distinct lexical signals—e.g., "graphics" or "politics"—allowing unsupervised models to perform reasonably well. Still, semi-supervised models outperform them, though with smaller gains.

These results challenge the common belief that unsupervised methods suffice due to their generality and low annotation cost. While this may apply to structured data, our findings show that textual anomaly detection benefits substantially from limited supervision, especially in linguistically rich contexts. We also observe cross-lingual differences: English datasets exhibit higher and more stable AUCs, while Portuguese datasets show greater variance and lower baselines—possibly due to label noise, fewer pre-trained encoders, or increased linguistic diversity. The Wilcoxon signed-rank test [Wilcoxon 1945], applied to paired AUC scores from multiple model-encoder combinations on each dataset, confirms that the performance improvements with semi-supervised models are statistically significant across all datasets evaluated ($p < 0.001$).

4.2. RQ2: How does representation type affect performance?

Figure 2 presents the performance distribution of multilingual and Portuguese-specific encoders across the three Portuguese datasets (TOLD-Br, WikiNews, and Portuguese Tweets), analyzed separately under unsupervised and semi-supervised conditions. In the semi-supervised setting, both encoder types achieved competitive AUCs, with models like XLM-RoBERTa (multilingual) and BERT-large-PT (Portuguese-specific) showing relatively stronger results. However, performance variability was evident across encoders, with Serafim, for instance, presenting lower median AUCs compared to the other Portuguese-specific models. In the unsupervised setting, AUC distributions were broadly dispersed across all encoders, with substantial variability within and across encoder types. Although Portuguese-specific encoders showed slightly higher median AUCs in some cases, the performance distributions of multilingual and Portuguese-specific models largely overlapped.

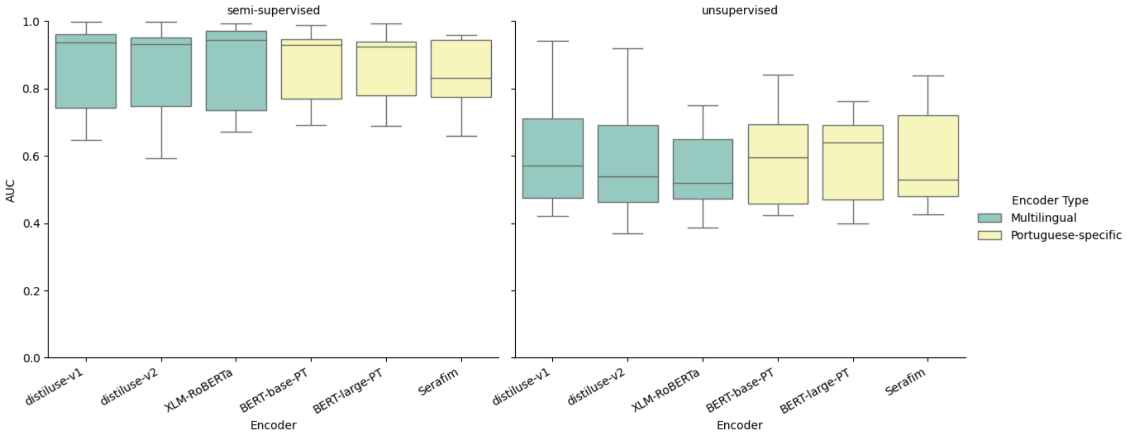

Figure 2. AUC for multilingual and Portuguese-specific encoders under unsupervised and semi-supervised settings (Portuguese datasets only). Encoder names are abbreviated for readability, as listed in Table 2.

To assess whether encoder type influenced anomaly detection performance, we used the Mann–Whitney U test [Mann and Whitney 1947] to compare AUC distributions between multilingual and Portuguese-specific encoders across the Portuguese datasets. The test indicated no statistically significant difference in either supervision setting (semi-supervised: $p = 0.77$; unsupervised: $p = 0.39$), suggesting that performance variations are more dependent on dataset-specific factors and model configurations than on encoder language scope

4.3. RQ3: Which Models Are Most Robust Across Tasks and Datasets?

To evaluate model robustness, we analyze the distribution of AUC scores per model across all datasets and task types under both supervision regimes (Figure 3). Models such as DevNet and XGBOD consistently deliver strong and stable performance, particularly in semi-supervised settings. Surprisingly, a simple feed-forward MLP also proves competitive, often outperforming more complex unsupervised architectures—highlighting the importance of high-quality representations even in lightweight models.

However, performance varies by task. In topic classification datasets such as *20 Newsgroups* and *WikiNews*, unsupervised methods perform reasonably well. These tasks involve lexical and thematic shifts that align well with density- and distance-based heuristics. In contrast, sentiment analysis and hate speech detection present subtler, context-dependent anomalies that challenge unsupervised models and benefit more from supervision. In such semantically rich settings, even minimal labeled data enables semi-supervised models to calibrate more discriminative boundaries. This effect is especially apparent in DevNet, XGBOD, and MLP, which frequently exceed 0.90 AUC in semi-supervised mode. These models leverage few-shot signals effectively, suggesting that sparse supervision can dramatically enhance generalization in high-dimensional text spaces. Conversely, models such as DeepSVDD, OCSVM, and LOF show limited effectiveness—particularly in unsupervised configurations. Reconstruction-based methods like AutoEncoder (AE) and Variational AutoEncoder (VAE) achieve intermediate results but generally lag behind their semi-supervised counterparts.

Overall, our findings suggest that model robustness in textual anomaly detection depends on both architectural flexibility and task complexity. The best-performing models pair adaptable learning objectives with strong embeddings, but task semantics remain a critical factor: unsupervised methods are more viable in structured, topical datasets, while semantically nuanced tasks demand supervision to achieve competitive performance.

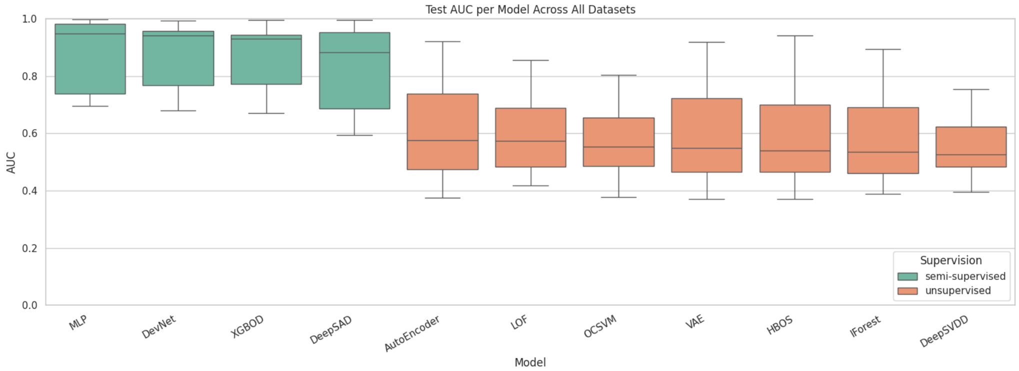

Figure 3. **Test AUC per model across all datasets under unsupervised and semi-supervised settings. Robust models exhibit high median performance and low variance across diverse tasks.**

4.4. RQ4: How Much Supervision Is Needed to Outperform Unsupervised Methods?

To assess how much supervision is necessary to surpass unsupervised baselines, we focused on the `DevNet` model—previously identified as one of the most effective semi-supervised approaches—and applied it across three Portuguese datasets. We progressively increased the number of labeled anomalies, starting from 10 examples and incrementally reaching up to 5% of the training set. In each round, a small batch of additional anomalies was randomly selected and incorporated into the labeled set. This process was repeated across five random seeds to ensure robustness.

Figure 4 displays the AUC progression for each dataset as a function of labeled anomaly percentage. Results show that even small amounts of supervision can lead to substantial performance improvements—especially in the early rounds. However, the supervision-efficiency curve varies by task: performance trends differ by task. In *Portuguese Tweets* (sentiment analysis), near-optimal AUC is achieved with less than 1% labeled anomalies, and variance stabilizes quickly. In *WikiNews* (topic classification), performance improves steadily but saturates earlier and more modestly, suggesting that some tasks generalize well with minimal labels. In contrast, *TOLD-Br* (hate speech detection) exhibits a more gradual and noisy curve, with no clear saturation by 5%, indicating that semantically complex or ambiguous tasks may require more supervision.

In addition to mean AUC, we observe a clear difference in variance reduction across tasks. For instance, Portuguese Tweets (sentiment analysis) exhibits low variability even with a small number of labels, suggesting consistent patterns are quickly learned. In contrast, TOLD-Br (hate speech detection) shows high variance throughout, underscoring the semantic ambiguity and domain-specific nuances that make the task more labelhungry. These findings suggest that, in few-shot anomaly detection scenarios, the amount of supervision required to unlock strong performance depends heavily on task complexity. Nevertheless, even randomly selected anomalies are often sufficient to yield meaningful gains, reinforcing the value of lightweight semi-supervised pipelines in practical settings.

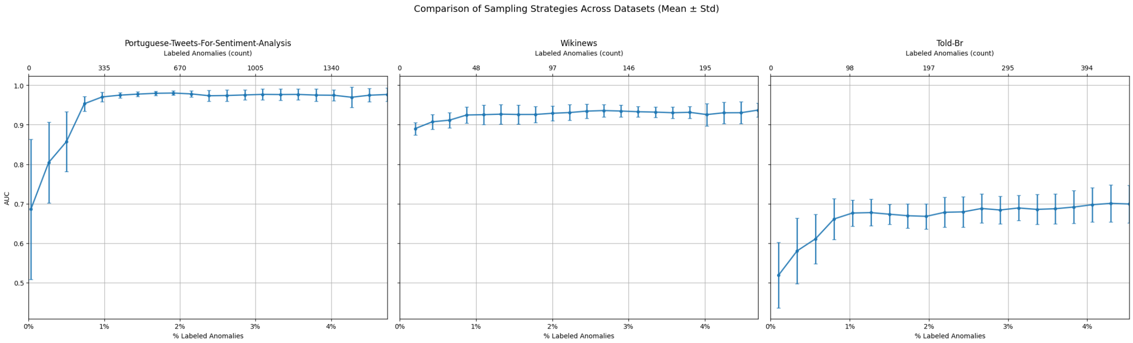

Figure 4. **AUC as a function of labeled anomaly proportion (bottom x-axis) and raw count (top x-axis) for three Portuguese datasets using `DevNet` with random sampling. Each line shows the mean ± standard deviation across five runs. Saturation is reached at different supervision levels depending on task complexity.**

5. Conclusion and Future Work

This study introduced a comprehensive benchmark for anomaly detection in textual data, emphasizing low-supervision scenarios across both Portuguese and English corpora. We compared unsupervised and semi-supervised approaches over six datasets and multiple embedding strategies, including multilingual and language-specific encoders. Our experiments show that even minimal supervision—labeling 1–5% of anomalies—can yield substantial improvements, especially in semantically complex tasks like sentiment analysis and hate speech detection.

Semi-supervised models such as DevNet, XGBOD, and DeepSAD consistently outperformed unsupervised baselines. Multilingual encoders demonstrated greater robustness when labeled data was scarce, while language-specific models proved highly effective when more labels were available. Importantly, we observed that performance often saturated after only a few annotation rounds, reinforcing the value of few-shot learning for scalable anomaly detection.

Looking ahead, we see several promising directions: (i) leveraging large language models (LLMs) to generate weak labels, assist with difficult cases, and enable real-time adaptation to new data; (ii) expanding beyond sentence-level analysis to entities, phrases, and documents to uncover finer-grained anomalies; (iii) extending this benchmark to additional languages and domains, with particular attention to low-resource settings where labeled anomalies are rare but domain knowledge is crucial; and (iv) exploring how different representation learning objectives—such as contrastive or attention-based pretraining—affect anomaly detection quality.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors would like to thank the National Council for Scientific and Technological Development (CNPq, grant #312360/2023-1).

References

Boutalbi, K., Loukil, F., Verjus, H., Telisson, D., and Salamatian, K. (2023). Machine learning for text anomaly detection: A systematic review. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1319–1324.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA. Association for Computing Machinery.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):71–97.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Garcia, K., Shiguihara, P., and Berton, L. (2024). Breaking news: Unveiling a new dataset for portuguese news classification and comparative analysis of approaches. *PLOS ONE*, 19(1):1–15.

Gomes, L., Branco, A., Silva, J. a., Rodrigues, J. a., and Santos, R. (2024). Open sentence embeddings for portuguese with the serafim pt* encoders family. In *Progress in Artificial Intelligence: 23rd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana Do Castelo, Portugal, September 3–6, 2024, Proceedings, Part III*, page 267–279, Berlin, Heidelberg. Springer-Verlag.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313:504–7.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lang, K. (1995). Newsweeder: Learning to filter netnews. `http://www.cs.cmu.edu/~kenlang`. CMU, unpublished manuscript.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR*, abs/2010.04543.

Liang, Y., Zhao, Y., Hu, Y., Li, Z., Liu, W., Akoglu, L., and Ding, B. (2018). Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 735–744. ACM.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Maia, F. and Costa, A. H. (2024). Anomaly detection in text data: A semi-supervised approach applied to the portuguese domain. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 288–293, Porto Alegre, RS, Brasil. SBC.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Manolache, A., Brad, F., and Burceanu, E. (2021). DATE: Detecting anomalies in text via self-supervision of transformers. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 267–277, Online. Association for Computational Linguistics.

Munoz-Galeano, N., Borchmann, L., and Reimers, N. (2021). Date: Detecting anomalies in text via self-supervision of transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 282–293. ACL.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2).

Pang, G., Shen, C., and van den Hengel, A. (2019). Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 353–362, New York, NY, USA. Association for Computing Machinery.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Portuguese Tweets Dataset (2018). Portuguese tweets for sentiment analysis. `https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis`. Accessed: 2025-09-03.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Ruff, L., Kauffmann, J., Vandermeulen, R., Montavon, G., Samek, W., Kloft, M., and Müller, K.-R. (2019a). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071. ACL.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., and Kloft, M. (2020). Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. (2019b). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.

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.

Sharma, R. (2018). Twitter sentiment analysis for hate speech detection. `https://huggingface.co/datasets/tweets-hate-speech-detection/tweets_hate_speech_detection`. Accessed: 2025-09-03.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Tax, D. M. and Duin, R. P. (2004). Support vector data description. *Machine Learning*, 54(1):45–66.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Xu, H. (2022). Deepod: A benchmarking framework for deep outlier detection. `https://github.com/xuhongzuo/DeepOD`.

Xu, Y., Gábor, K., Milleret, J., and Segond, F. (2023). Comparative analysis of anomaly detection algorithms in text data. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1234–1245, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural*

Information Processing Systems - Volume 1, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Zhao, Y. and Hryniewicki, M. K. (2018). Xgbod: Improving supervised outlier detection with unsupervised representation learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7.