Anomaly Detection in Text Data: A Semi-Supervised Approach Applied to the Portuguese Domain

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. *Anomaly detection, driven by advancements in machine learning and deep learning, has gained significant importance across various fields. However, its application to unstructured textual data, particularly in Portuguese, remains underexplored. In textual analysis, these techniques are crucial for detecting deviations within text collections. This paper investigates state-of-the-art methods for anomaly detection in Portuguese text corpora and introduces a new, flexible loss function designed to enhance detection across different contamination levels. By evaluating these methods on benchmark datasets, specifically in the contexts of hate speech detection and sentiment analysis, we address existing challenges and contribute to the development of more effective anomaly detection techniques for Portuguese text data.*

1. Introduction

Anomaly detection refers to the identification of patterns in data that deviate from expected norms [Chandola et al. 2009]. Anomalies, often termed outliers or exceptions, are distinct from the majority of observations that define the "normal" pattern. While anomaly detection techniques have been extensively applied to structured data, such as continuous and categorical variables [Boutalbi et al. 2023], less attention has been given to unstructured data like text — the focus of this work.

Anomaly detection began with statistical methods in the late 19th century [Edgeworth 1887] and has since expanded, with deep learning broadening its scope to unstructured domains like images and text [Chandola et al. 2009, Pimentel et al. 2014]. However, its application to textual anomaly detection remains limited [Pang et al. 2019]. Detecting anomalies in text data is particularly challenging due to the variety of linguistic levels involved, such as spelling, syntax, and semantics [Xu et al. 2023b]. Leveraging deep learning's ability to model complex patterns has led to significant advancements in the field.

Traditionally treated as an unsupervised task due to the absence of ground truth labels, anomaly detection has employed techniques like autoencoders and GANs [Pang et al. 2019]. Popular approaches in this domain include DeepSVDD [Ruff et al. 2018] and Deep Isolation Forest [Xu et al. 2023a], both focusing on modeling normal data and identifying deviations as anomalies. However, recent semi-supervised neural approaches, such as DevNet [Pang et al. 2019] and DeepSAD [Ruff et al. 2020], have shown improved detection accuracy by integrating limited labeled anomalies into the training process [Xu et al. 2023b], bridging the gap between unsupervised and supervised learning. Despite recent advancements, there remains a significant gap in comprehensive research focused on anomaly detection in Portuguese text corpora.

In this paper, we extend neural network-based anomaly detection techniques to handle these complexities in Portuguese text data. Furthermore, we propose a change in the loss function in order to establish a compromise between the samples that correspond to anomalies in relation to the others. Experiments show that this approach is quite promising. To effectively address the unique challenges of representing textual data, we employ two pre-trained BERT-based models, checking the strengths and weaknesses of each representation in the different tasks.

2. Methodology

2.1. Problem Definition

Given a dataset $X = \{x_1, x_2, \ldots, x_{N+K}\}$, where $U = \{x_1, x_2, \ldots, x_N\}$ is unlabeled data and $K = \{x_{N+1}, \ldots, x_{N+K}\}$ represents labeled anomalies ($K \ll N$), the goal is to train a model to identify these rare anomalies. This task is challenging due to the imbalance between the large unlabeled set and the small labeled anomaly set. The process involves two key steps:

1. **Embedding Transformation:** Data X is transformed into embeddings $Z = \{z_1, z_2, \ldots, z_{N+K}\}$, with each z_i being a vector in $\mathbb{R}^d$.
2. **Scoring Function:** A neural network learns a scoring function $\phi : Z \to \mathbb{R}$ to ensure that $\phi(z_i) > \phi(z_j)$ when z_i is an anomaly and z_j is normal, minimizing the use of labeled examples.

We adopted the DevNet model due to its demonstrated good performance obtained in studies considering textual domain [Xu et al. 2023b] along with its ability to effectively manage high-dimensional spaces, such as embeddings. Additionally, the model's interpretable loss function, based on a straightforward Z-score strategy, provides valuable insights that can be later used to assess text identified as anomalies.

2.2. DevNet for Anomalous Text

The DevNet algorithm [Pang et al. 2019] introduces a semi-supervised approach that learns an interpretable outlier scoring function, $\phi(z; \Theta)$, using a Z-score deviation loss. While the original formulation is based on raw data points x, we denote the embeddings as z to reflect the transformed data representations. This approach assumes a prior normal distribution over anomaly scores, modeled with l random objects $r_i \in \mathbb{R}$ sampled from a standard normal distribution $\mathcal{N}(\mu_R, \sigma_R)$:

$$\text{dev}(z) = \frac{\phi(z; \Theta) - \mu_R}{\sigma_R}, \tag{1}$$

where μ_R and σ_R are the mean and standard deviation of anomaly scores within the reference distribution. This deviation is then incorporated into a contrastive loss function to enhance the distinction between anomalous and normal samples, where y indicates anomaly status, and a ensures a minimum separation between classes [Pang et al. 2019].

$$L(\phi(z; \Theta), \mu_R, \sigma_R) = (1 - y)|\text{dev}(z)| + y \cdot \max(0, a - \text{dev}(z)), \tag{2}$$

2.3. Proposal

To provide flexibility, the parameter $\eta \in [0, 1]$ is introduced in the DevNet loss function given in Eq. 2, controlling the balance between regular and anomalous samples,

$$L(\phi(x; \Theta), \mu_R, \sigma_R) = (1 - \eta)(1 - y)|\text{dev}(x)| + \eta \cdot y \cdot \max(0, a - \text{dev}(x)), \qquad (3)$$

The η parameter adjusts the model's emphasis on anomalies, allowing adaptation to varying levels of contamination (i.e., percentage of labeled anomalies) and data availability. We investigate different proportions of labeled anomalies to assess the robustness of the solution in various scenarios, aiming to determine the minimum amount of labeled data needed for good performance. Additionally, we employ two distinct text representation strategies: the monolingual BERTimbau model [Souza et al. 2020], specifically designed for processing Portuguese text, and the multilingual Sentence-BERT (SBERT) [Reimers and Gurevych 2019], which generates sentence-level embeddings across multiple languages, including Portuguese. Our customized DevNet implementation, named η-DevNet, was evaluated against its original version using both representation strategies.

3. Experiments and Results

3.1. Experiments

To evaluate the performance of different representation methods and loss functions, we first tested η values ranging from 0.5 to 1 using the BERTimbau embedding strategy, where $\eta = 0.5$ corresponds to the original DevNet formulation. The other parameters were adopted from the DevNet reference: $a = 5$, $l = 5000$, $\mu_R = 0$, and $\sigma_R = 1$. Contamination levels were adjusted by introducing between 5 and 1000 anomalies across the experiments, with anomalies randomly selected. After identifying the optimal η value, we applied it in subsequent experiments to compare both embedding strategies across different datasets. The mean ROC-AUC values were calculated over 10 experimental runs for each scenario.

3.2. Dataset

We evaluate our approach using two Brazilian datasets. The first, Told-Br [Leite et al. 2020], contains 21,000 labeled instances of tweets tagged with hate speech, categorized into themes such as homophobia, racism, and misogyny, with hate speech serving as the anomaly class. The second dataset, UTLC-Movies [Sousa et al. 2019], comprises over one million movie reviews. From this dataset, we sampled 40,000 reviews for sentiment analysis, where negative sentiment is treated as the anomaly class.

3.3. Results

Figure 1, shows performance and stability improvements as η is adjusted, with $\eta = 0.7$ yielding optimal performance. This value was subsequently used for further analysis. The results shown in Table 1 outline these results for both tasks, demonstrating that in most cases, the adapted loss function led to performance improvements. Reaching a reasonable level of accuracy requires a minimum threshold of labeled examples, which varies with task complexity. In our experiments, sentiment analysis needed only 0.87% of labeled anomalies to achieve a ROC-AUC of 0.85, while hate speech detection required 2.59% to

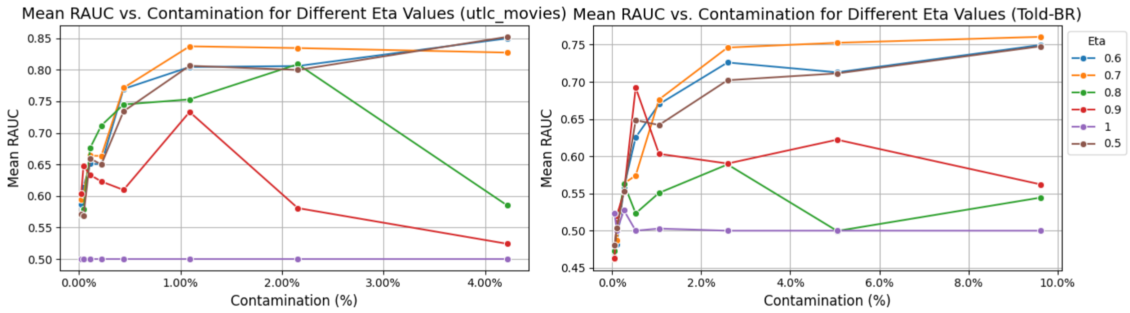

Figure 1. η **Comparison across different** η **values with varying amounts of labeled anomalies using BERTimbau embedding strategy.**

reach a ROC-AUC of 0.73. This discrepancy likely arises from the greater complexity of hate speech detection, which involves subtle linguistic nuances and diverse expressions. Additionally, pre-trained models may not fully capture slang and politically specific contexts, which are common in hate speech but may be underrepresented during training.

Table 1. Comparison of ROC-AUC values across different scenarios and contamination levels for UTLC-Movies and Told-BR datasets when $\eta = 0.7$**, where % refers to the contamination level.**

	UTLC-Movies					Told-BR				
Nb. Outliers	**%**	**Bη**	**BD**	**Mη**	**MD**	**%**	**Bη**	**BD**	**Mη**	**MD**
5	0.02	**0.58**	0.57	0.52	0.52	0.05	0.46	0.48	**0.50**	**0.50**
10	0.04	**0.62**	0.58	0.54	0.57	0.11	0.48	0.50	**0.53**	**0.53**
25	0.09	0.66	0.65	0.60	**0.68**	0.27	0.57	0.57	0.54	**0.60**
50	0.18	**0.70**	0.69	0.64	0.66	0.53	0.58	**0.63**	0.56	**0.60**
100	0.35	**0.78**	0.71	0.71	0.70	1.05	**0.66**	0.63	0.61	0.60
250	0.87	**0.82**	0.76	0.75	0.69	2.59	**0.73**	0.68	0.56	0.52
500	1.73	**0.83**	**0.83**	0.81	0.79	5.05	**0.75**	0.73	0.50	0.51
1000	3.41	**0.85**	0.84	0.81	0.75	9.62	**0.76**	**0.76**	0.50	0.52

Acronyms: BERTimbau η-loss (Bη), BERTimbau Devnet loss (BD), multilingual SBERT η-loss (Mη), multilingual SBERT Devnet loss (MD).

Our results show that the BERTimbau representation [Souza et al. 2020] consistently outperformed the multilingual model across tasks. This advantage can be traced to BERTimbau's specialization in Portuguese, allowing it to capture more intricate linguistic nuances, such as idiomatic expressions and regional variations.

4. Conclusion and Future Work

This study shows that BERTimbau, tailored for Portuguese, consistently outperforms multilingual models in anomaly detection, with the customized loss function providing notable improvements. These results highlight the potential of semi-supervised methods for tasks like harmful content detection and sentiment analysis in Portuguese contexts with limited labeled data.

Future work may expand this approach to related tasks such as topic modeling, fake news detection, and fraud detection. Although some labeling effort is still required for good performance, the small amount of labeled data needed makes this approach feasible in resource-constrained scenarios. Furthermore, the promising advances in Large Language Models (LLMs) could not only serve as valuable tools for benchmarking but also automate anomaly tagging, reducing manual effort and enhancing adaptability and scalability across various real-world applications.

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.

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

Edgeworth, F. Y. (1887). Xli. on discordant observations. *Philosophical Magazine Series 1*, 23:364–375.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

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.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

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.

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., Müller, K.-R., and Kloft, M. (2020). Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*.

Sousa, R. F. d., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Symposium in Information and Human Language Technology - STIL*. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Xu, H., Pang, G., Wang, Y., and Wang, Y. (2023a). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604.

Xu, Y., Gabor, K., Milleret, J., and Segond, F. (2023b). Comparative analysis of anomaly detection algorithms in text data. pages 1234–1245.