Comparative Analysis of Anomaly Detection Algorithms in Text Data

Yizhou Xu[1,2], **Kata Gábor**[1], **Jérôme Milleret**[2], **Frédérique Segond**[1,3]

[1]ERTIM, INaLCO, 2 Rue de Lille, 75007 Paris, France
[2]ChapsVision, 4 rue du Port Aux Vins, 92150 Suresnes, France
[3]Inria, 860 Rue Saint Priest, 34095 Montpellier, France
{yxu, jmilleret}@chapsvision.com
kata.gabor@inalco.fr, frederique.segond@inria.fr

Abstract

Text anomaly detection (TAD) is a crucial task that aims to identify texts that deviate significantly from the norm within a corpus. Despite its importance in various domains, TAD remains relatively underexplored in natural language processing. This article presents a systematic evaluation of 22 TAD algorithms on 17 corpora using multiple text representations, including monolingual and multilingual SBERT. The performance of the algorithms is compared based on three criteria: degree of supervision, theoretical basis, and architecture used. The results demonstrate that semi-supervised methods utilizing weak labels outperform both unsupervised methods and semi-supervised methods using only negative samples for training. Additionally, we explore the application of TAD techniques in hate speech detection. The results provide valuable insights for future TAD research and guide the selection of suitable algorithms for detecting text anomalies in different contexts.

1 Introduction

Anomaly detection is a fundamental process in data analysis, aiming to identify inconsistent data points that deviate significantly from expected behaviors or established norms within a dataset. Such anomalies can emerge from various factors, including human errors, malicious behaviors, unusual events, or unexpected changes. Effective anomaly detection can facilitate swift problem recognition, proactive measures for error correction, and future problem prevention. It enhances data quality, aids in risk identification, and empowers decision-making across diverse domains, with its utility extending to various data types such as tabular data, graphs, time series, texts, images, and videos.

In the context of text data, the anomalies refer to specific texts or textual fragments that deviate significantly from established norms, which can be determined based on the overall text or corpus, regular language usage, or common sense. These anomalies may manifest at various linguistic levels, such as orthographic (spelling), lexical (word usage), syntactic (sentence structure), semantic (meaning), and discourse (overall context) levels (Wang et al., 2014; Saranya et al., 2014; Wahl, 2021; Sufi and Alsulami, 2021). Detecting anomalies in text data holds vital importance in applications like language development assessment, plagiarism detection, quality control in data processing, and identifying abnormal language usage in cybersecurity (Cichosz, 2020; Szoplák and Andrejková, 2021).

In this article, we concentrate on Text Anomaly Detection (TAD) at the semantic and discourse levels, where norms are established on a corpus scale. It is important to note that the definition of anomalies can be further refined and may slightly differ according to specific contexts. For instance, in the realm of competitive intelligence, anomalies often relate to abnormal themes or topics, while in the field of online reputation monitoring, they typically pertain to negative sentiments.

Despite the broad utility of TAD and the potential benefits it offers, TAD has not been as extensively explored as other topics within Nature Language Processing (NLP). While previous research works have approached the field of TAD, they have often been limited either by the scope of algorithms considered or by the range of textual representations evaluated (Barrett et al., 2019; Pantin et al., 2022). Unlike these studies, this paper aims to provide a comprehensive overview of TAD by evaluating a wide array of algorithms on several corpora across different languages, making our approach distinctive in its breadth and depth.

Our primary objective is to provide a systematic evaluation of 22 TAD algorithms applied to 17 corpora across three languages. We assess these algo-

Proceedings of Recent Advances in Natural Language Processing, pages 1234–1245
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_131

rithms' performance in detecting textual anomalies and examine the use of various text representations, including monolingual and multilingual Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Additionally, we investigate the potential application of TAD techniques in detecting hate speech, aiming to gain insights into the effectiveness of TAD in this specific domain.

2 Related Work

Text Anomaly Detection (TAD) stands as a comparatively less explored intersection of Data Mining (DM) and Natural Language Processing (NLP). While extensive research has been conducted in DM dedicated to anomaly detection, scant attention has been given to the application of these techniques to text data. On the other side of the spectrum, NLP, despite significant progress in text understanding and generation, exhibits a noticeable deficiency in research focusing on the detection of anomalous text. Consequently, dedicated algorithms for text data anomaly detection are rare, and corpora specific to this task are either completely inaccessible or simply nonexistent.

Anomaly Detection Algorithms In the DM field, a wealth of systematic analyses and evaluations of anomaly detection algorithms have been carried out (Markou and Singh, 2003a,b; Chandola et al., 2009; Pimentel et al., 2014; Aggarwal, 2015, 2017; Chalapathy and Chawla, 2019; Pang et al., 2021). However, these studies have largely overlooked the performance of these algorithms on text data. In contrast, within the NLP realm, research efforts were largely channeled towards adapting techniques proposed for other domains, such as image and video data, to handle text data. However, these studies often adopted a narrow focus, examining a particular algorithm and contrasting it against a limited set of others (Drozdyuk and Eke, 2017; Ruff et al., 2019; Jafari, 2022). This resulted in a fragmented and insufficiently broad approach that fell short of providing an all-encompassing assessment of various anomaly detection methods' performance on text data. To address this deficiency, recent efforts have been made by researchers. Yap et al. (2020) proposed an algorithm based on Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), contrasting its performance against state-of-the-art methods like CVDD (Ruff et al., 2019). Barrett et al. (2019) undertook a comparative study of six algorithms using three corpora and four representation strategies, namely TF-IDF, One-hot, Bag of Words, and PCA. In a significant systematic endeavor, Pantin et al. (2022) compared ten algorithms on two corpora using a novel anomaly generator, GenTO.

Data The scarcity or complete absence of human-annotated anomaly detection corpora presents a considerable challenge in anomaly detection. The rarity of anomalies and the subjectivity involved in defining and annotating them contribute to this scarcity. To counteract this problem, three main strategies have been proposed in the literature. The first strategy involves leveraging artificially generated data to construct a corpus, with a particular focus on creating anomaly samples (Christophe et al., 2019). The second strategy combines diverse text sources to create a corpus, drawing "normal" examples from one source and anomalies from another (Dasigi and Hovy, 2014). The third and most common approach involves adapting existing corpora originally created for different tasks for anomaly detection. In this approach, researchers often repurpose corpora that are initially designed for tasks such as topic or sentiment classification. Datasets commonly used in this context include Reuters (Barrett et al., 2019; Yap, 2020; Han et al., 2022; Pantin et al., 2022), AGNews (Zeng et al., 2022; Han et al., 2022), 20NewsGroups (Barrett et al., 2019; Hu et al., 2021; Pantin et al., 2022), and IMDB (de la Torre-Abaitua et al., 2021; Han et al., 2022).

Text Representation Techniques Finally, in TAD, as with many other text classification tasks, the choice of text representation techniques is critical. While traditional encoding strategies such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are prevalent (Barrett et al., 2019; Pantin et al., 2022), the text embeddings generated by pre-trained models, like Sentence-BERT (SBERT)(Reimers and Gurevych, 2019), remain relatively underutilized in this field. This limited adoption of contextual embeddings presents promising area for exploration and potential improvement in TAD methodology. The systematic evaluation of TAD algorithms on various text representation techniques could yield significant insights and drive advancements in this field.

3 Comparative Analysis of Text Anomaly Detection Algorithms

3.1 Corpus Assembly

Dataset Selection Due to the absence of a dedicated corpus to Text Anomaly Detection (TAD), we have repurposed various datasets that were originally designed for different NLP tasks. In this study, we utilized a collection of 14 datasets, each primarily designed to address either binary or multi-class text classification challenges, covering a wide range of application scenarios (refer to Table 1). Our selection includes datasets employed for topic or thematic classification (TC) and sentiment analysis (SA), which are common in the literature, and those used for hate speech detection (HD). The inclusion of the latter is intended to explore the potential of considering hate speech and offensive language as forms of textual anomalies. In contrast to many previous studies that have solely focused on English, our datasets encompass texts in three different languages: English, French, and Chinese. The data we used were collected from a variety of sources, including news agencies (such as ABC News and Reuters), forums (like Stormfront), social media platforms (Twitter, Weibo, and others), and various websites (Amazon, IMDB, etc.).

Dataset Adaptation We curated 17 different corpora for TAD based on the datasets mentioned above (see Table 1). In order to adapt these datasets to TAD, we employed the following strategies:

1. For TC data, we selected pairs of topics/themes, designating one as the "normal" class and the other as the "anomalous" class. If the available number of documents for a topic/theme was insufficient to form a class, we combined two or more topics/themes into one class.

2. For SA data, if labels were in the form of sentiment polarity, we labeled the "positive" class as "normal" and the "negative" class as "anomalous". If the data was annotated on a 5-point evaluation scale, we classified texts with 1 or 2 points as "anomalous" and those with 4 or 5 points as "normal".

3. For HD data, in the case of binary classification, we designated the "positive (hateful/offensive)" class as "anomalous" and the "negative" class as "normal". For multi-class classification, we grouped different types of hate speech into an "anomalous" class and non-hateful texts into a "normal" class.

4. To ensure comparability across datasets, we uniformly set the anomaly ratio to 10%. This decision aligns with common practice in the field, where a 10% anomaly ratio is frequently used (Pantin et al., 2022). Moreover, it is consistent with the default contamination rate usually adopted in anomaly detection tools (Buitinck et al., 2013; Zhao et al., 2019).

5. We created the corpora by conducting stratified random sampling from the datasets, respecting the predefined anomaly ratio.

3.2 Text Representation

The texts in the corpora are transformed into vectors using two distinct strategies: TF-IDF (Term Frequency-Inverse Document Frequency) and SBERT (Sentence-BERT) (Reimers and Gurevych, 2019). The selection of these techniques formed an essential step in preparing the data for the subsequent application of anomaly detection algorithms.

The TF-IDF technique was employed to generate vectors where the weighting of each term was determined by its frequency within a document but inversely proportional to its frequency across the entire corpus (represented by the training subset in our case).

Simultaneously, we utilized Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to generate embeddings for our text data. SBERT is a modification of the pre-trained BERT network that allows for the computation of semantically meaningful sentence embeddings. To account for linguistic variations across our multilingual dataset, we employed a selection of pre-trained monolingual SBERT models specific to each language under consideration: all-mpnet-base-v2 (en), all-MiniLM-L6-v2 (en), all-distilroberta-v1 (en), sentence-camembert-large (fr), sentence-camembert-base (fr), text2vec-base-chinese (zh), sbert-base-chinese-nli (zh), and sbert-chinese-dtm-domain-v1-distill (zh). To further diversify our text representation and explore potential generalizability across languages, we also incorporated multilingual SBERT models into our study. These models, such as distiluse-base-multilingual-cased-v1, paraphrase-multilingual-mpnet-base-v2, and paraphrase-multilingual-MiniLM-L12-v2, were chosen based on their

Corpus	Dataset	Citation	Source	Task	Lang	Size	AnormalTag	NormalTag
TDT2	Topic Detection and Track	Cieri et al. 1999	Press	TC	en	1000	topic 6/10/51	topic 1
20NG	20 Newsgroups		Press	TC	en	2000	politics.guns	sport
AGNews	AG News Topic Classification Dataset	Zhang et al. 2015	Press	TC	en	35000	Sci/Tech	Business
Reuters	Reuters-21578 Text Categorization Collection Dataset	Lewis 1997	Press	TC	en	4000	cpi/interest	earn
Amazon-en	Multilingual Amazon Reviews Corpus	Keung et al. 2020	Amazon	SA	en	8000	4/5 star	1/2 star
Amazon-fr			Amazon	SA	fr	10000	4/5 star	1/2 star
Amazon-zh			Amazon	SA	zh	25000	4/5 star	1/2 star
IMDB	Large Movie Review Dataset	Maas et al. 2011	IMDB	SA	en	25000	negative	positive
Yelp	Large Yelp Review Dataset	Zhang et al. 2015	Yelp	SA	en	10000	negative	positive
HTPO-Trump	Hate Towards the Political Opponent	Grimminger and Klinger 2021	Twitter	SA	en	1000	Against	Favor
HTPO-HOF			Twitter	HD	en	2500	Hateful	Non-Hateful
Stormfront	Hate Speech Dataset from a White Supremacy Forum	de Gibert et al. 2018	Forum	HD	en	10000	hate	nonHate
OLID	Offensive Language Identification Dataset	Zampieri et al. 2019	Twitter	HD	en	10000	OFF	NOT
COLD	Complex Offensive Language Dataset	Palmer et al. 2020	Twitter	HD	en	700	offensive/hateful	nonNone
COLDataset	Chinese Offensive Language Detection	Deng et al. 2022	Zhihu/Weibo	HD	zh	21000	1	0
SWSR	Sina Weibo Sexism Review	Jiang et al. 2021	Weibo	HD	zh	6000	1	0
MLMA-fr	MultiLingual Multi-Aspect hate speech	Ousidhoum et al. 2019	Twitter	HD	fr	900	offensive/hateful	normal

Table 1: Overview of the Datasets Utilized for Corpus Construction. The table provides details about each corpus, including the corpus ID, the original dataset name along with its citation, the source of the texts, the original task for which the dataset was created (TC: Topic Classification, SA: Sentiment Analysis, HD: Hate Speech Detection), the size of the corpus, and the tags used to denote anomalies and normal data.

demonstrated performance in processing a variety of languages, aligning well with the linguistic diversity present within our corpora.

3.3 Algorithm Comparison

In this study, we conducted an investigation of 22 distinct algorithms (refer to Table 2) on 17 different corpora. Considering the diverse taxonomy of approaches proposed in the literature (Chandola et al., 2009; Pimentel et al., 2014; Aggarwal, 2017), we opted to classify the algorithms from three unique angles: the utilization of neural networks, the degree of supervision, and the underlying theory driving the method. This approach not only allowed us to compare individual algorithmic performances, but also facilitated a comparison of categories of algorithms against each other.

Neural Networks Based on their architecture, the algorithms can be divided into two distinct types: deep algorithms that harness neural networks, and shallow algorithms that do not employ them (Han et al., 2022).

Supervision Based on the degree of supervision, or the extent to which they rely on labels, we can distinguish three categories of algorithms: supervised, semi-supervised, and unsupervised. Given the rarity of anomalies, procuring sufficient labels for abnormal (or positive) data often poses a significant challenge. Hence, within the domain of TAD, our primary focus is on the latter two types of algorithms: semi-supervised and unsupervised algorithms.

- **Semi-supervised algorithms** make use of partially labeled data for training. Certain anomaly detection techniques, such as OCSVM and LOF, assume that only normal (negative) instances are available during the training phase, leading them to be also known as "novelty detection" algorithms. In contrast, other algorithms leverage labeled and unlabeled data, utilizing the labeled data, which includes information about both normal and abnormal instances, to guide the learning process. By learning from the labeled data, these algorithms seek to predict anomalies in the unlabeled data, thereby detecting instances that deviate from normal behavior. Recently proposed algorithms like XGBOD (Zhao and Hryniewicki, 2018) and DevNet(Pang et al., 2019) demonstrate the ability to exploit weak labels, which could be limited or noisy. These algorithms are designed to perform effectively even when the available labels for abnormal instances are neither exhaustive nor accurate.

- **Unsupervised algorithms** do not rely on labeled data during the training process. The training set consists of both normal and abnormal instances, resulting in a dataset considered to be contaminated with outliers. These methods aim to identify anomalies in a dataset by exclusively analyzing the characteristics and patterns present in the unlabeled data. Unsupervised methods are grounded in the concept that anomalies significantly diverge from the expected behavior of the majority of the data points.

Underlying Theory Anomaly detection algorithms assess the abnormality or deviation of each

Algo. ID	**Name**	**Citation**	**Supervision**	**Theory**	**Architecture**
ABOD	Angle-based Outlier Detector	Kriegel et al. 2008	Unsup.	Proximity	Shallow
ALAD	Adversarially Learned Anomaly Detection	Zenati et al. 2018	Unsup.	Reconstruction	Deep
AnoGAN	Anomaly Detection with Generative Adversarial Networks	Schlegl et al. 2017	Unsup.	Reconstruction	Deep
AutoEncoder	Auto Encoder		Unsup.	Reconstruction	Deep
CBLOF	Clustering Based Local Outlier Factor	He et al. 2003	Unsup.	Proximity	Shallow
COF	Connectivity-Based Outlier Factor	Tang et al. 2002	Unsup.	Proximity	Shallow
COPOD	Copula Based Outlier Detector	Li et al. 2020	Unsup.	Probabilistic	Shallow
DeepSAD	Deep Semi-supervised Anomaly Detection	Ruff et al. 2020	Semi	Reconstruction	Deep
DeepSVDD	Deep One-Class Classifier with AutoEncoder	Ruff et al. 2018	Unsup.	Domain	Deep
DevNET	Deviation Networks	Pang et al. 2019	Semi	Reconstruction	Deep
ECOD	Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions	Li et al. 2022	Unsup.	Probabilistic	Shallow
GMM	Gaussian Mixture Model		Unsup.	Probabilistic	Shallow
HBOS	Histogram-based Outlier Detection	Goldstein and Dengel 2012	Unsup.	Probabilistic	Shallow
IForest	Isolation Forest	Liu et al. 2008	Unsup.	Ensemble	Shallow
KNN	k-Nearest Neighbors Detector	Ramaswamy et al. 2000	Unsup.	Proximity	Shallow
LOF	Local Outlier Factor	Breunig et al. 2000	Semi	Proximity	Shallow
KDE	Outlier Detection with Kernel Density Functions	Latecki et al. 2007	Unsup.	Probabilistic	Shallow
OCSVM	One Class Support Vector Machine	Schölkopf et al. 2001	Semi	Domain	Shallow
PCA	Principal Component Analysis	Shyu et al. 2003	Unsup.	Reconstruction	Shallow
PReNet	Pairwise Relation prediction-based ordinal regression Network	Pang et al. 2020	Semi	Ensemble	Deep
VAE	Variational Autoencoder	Kingma and Welling 2013	Unsup.	Reconstruction	Deep
XGBOD	Extreme Gradient Boosting Outlier Detection	Zhao and Hryniewicki 2018	Semi	Ensemble	Shallow

Table 2: Overview of Investigated Anomaly Detection Algorithms: Algorithm ID, Full Algorithm Name, Original Paper Citation, Degree of Supervision, Underlying Theory, and Model Architecture (Deep/Shallow)

data point by calculating an anomaly score. This score is then contrasted against a predefined threshold set for the entire dataset. Anomaly detection algorithms can be categorized into five groups based on the underlying theory driving the algorithm and methodology used to calculate the anomaly score.

- **Probabilistic or statistical algorithms** function by estimating the generative probability density function of the data. They model the probability distribution of the data using probability and statistical tools, such as Gaussian distribution or logistic regression. Data points that yield a low probability of conforming to the distribution model are considered as potential anomalies.

- **Proximity-based algorithms** identify a data point as an anomaly if it is surrounded by a sparsely populated or dissimilar neighborhood. The anomaly score is calculated based on the degree of deviation or isolation of a data point from its immediate neighbors. Based on their definition of proximity, these techniques are further classified into three subcategories: cluster-based algorithms, density-based algorithms, and distance-based algorithms.

- **Domain-based algorithms** utilize training data to define a domain that encapsulates the normal class. The model created in this process describes the boundary or region of the normal class and determines whether a data point belongs to this class based on its position relative to the boundary. The anomaly score is typically derived from the distance or proximity of a data point to the boundary of the designated normal region (Pimentel et al., 2014).

- **Reconstruction-based algorithms** aim to compress the data into a space of lower dimensionality and subsequently reconstruct the original data from this condensed representation. The reconstruction error, defined as the difference between the original and the reconstructed data, is used to compute the anomaly score. The principle is straightforward: the greater the reconstruction error, the higher the likelihood of the data point being anomalous (Pimentel et al., 2014).

- **Ensemble algorithms** combine the outputs from multiple base algorithms or detectors to create a unified, more robust output (Aggarwal, 2017). These algorithms leverage the diversity of individual detectors and strive to enhance the overall performance by aggregating their results. Common ensemble techniques include voting, averaging, stacking, and boosting, among others.

4 Experiments

4.1 Experimental Settings

Evaluation A multitude of metrics are traditionally employed to gauge the effectiveness of anomaly detection algorithms. These include Precision, Recall, F-score, ROC AUC (Area Under the Receiver Operating Characteristic Curve), PR

AUC (Precision-Recall Area Under the Curve), and MCC (Matthews Correlation Coefficient) (Manevitz and Yousef, 2001; Dasigi and Hovy, 2014; Ruff et al., 2019; Todd et al., 2020; Pantin et al., 2022; Barrett et al., 2019). In this study, we have chosen to focus on ROC AUC, the most prevalent metric within the domain of anomaly detection. In this context, the ROC (Receiver Operating Characteristic) curve plots the true positive rate ($sensitivity$) against the false positive rate ($1 - specificity$) over a range of threshold settings. The ROC AUC score, a numerical value between 0 and 1, offers an indicative measure of the classification capability of the model. A score of 0.5 corresponds to a random classifier, while a score of 1 signifies a perfect classifier. A model's capacity to distinguish between normal and anomalous instances is typically associated with a higher ROC AUC score.

Data Partitioning and Independent Trials To ensure the robustness of our experimental findings, we employed a 10-fold cross-validation methodology for data partitioning. In each fold, 90% of the data was reserved for training and the remaining 10% for testing purposes. Stratified sampling ensured a consistent anomaly ratio across both the training and test sets within each fold. Anomaly detection models were individually trained on the data from each fold and subsequently evaluated against the corresponding test set. The average ROC AUC score, calculated over all 10 folds, served as the aggregate measure of the model's ability to accurately differentiate between normal and anomalous instances.

Hyperparameters It is common practice to run an algorithm multiple times to select the parameters that optimize the ROC AUC. However, this approach is not suitable for anomaly detection as it inadvertently introduces a form of supervision by using knowledge of the anomaly labels to select parameters (Aggarwal, 2017). To ensure a fair comparison, it is essential to adhere to an unsupervised approach. Therefore, in this work, we strictly employ the default hyperparameter settings as provided in the original papers of all the algorithms.

Implementation The experiments were conducted using three Python libraries: scikit-learn (Buitinck et al., 2013), PyOD (Zhao et al., 2019), and DeepOD (Xu, Hongzuo).

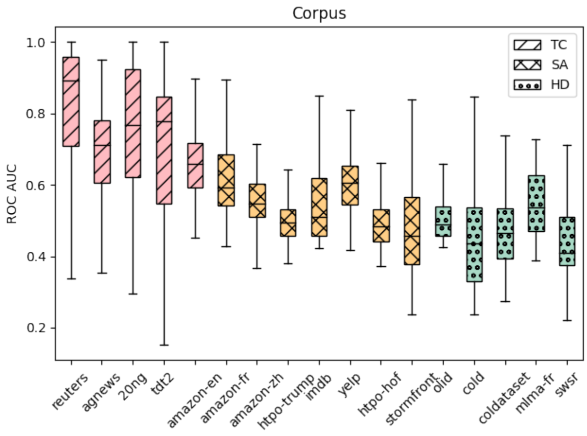


Figure 1: Performance (avg. ROC AUC) comparison of anomaly detection algorithms across 17 corpora grouped by original tasks: Topic Classification (TC), Sentiment Analysis (SA), and Hate Speech Detection (HD)

4.2 Results and Discussion

Corpus Figure 1 illustrates the performance of the 22 algorithms tested across 17 diverse corpora, which are divided into 3 categories: corpora for topic classification (TC), sentiment analysis (SA), and hate speech (HD). Notably, the TC corpora achieve the highest scores, with a median ROC AUC of 0.768. In contrast, the HD corpora, incorporated into TAD testing for the first time, exhibit a median ROC AUC of 0.474. This suggests a performance level below random chance, indicating that the TAD algorithms have room for improvement when it comes to effectively identifying hate speech. It's important, however, to bear in mind that the TC corpora mainly comprise press texts, while the HD corpora are largely made up of noisy social media texts. Further experiments are necessary to evaluate and mitigate the potential impact of textual noise on algorithm performance.

Text Representation Figure 2 presents the performance of algorithms categorized based on the representations used: TF-IDF model, monolingual SBERT models, and multilingual SBERT models. The TF-IDF model shows fairly stable results, albeit with a noticeably lower upper limit compared to SBERT models. Among the monolingual SBERT models, the Chinese models exhibit weaker performance, indicated by a median ROC AUC of 0.464, which is significantly beneath the level of random chance. This could be due to the specific concentration of Chinese corpora on hate

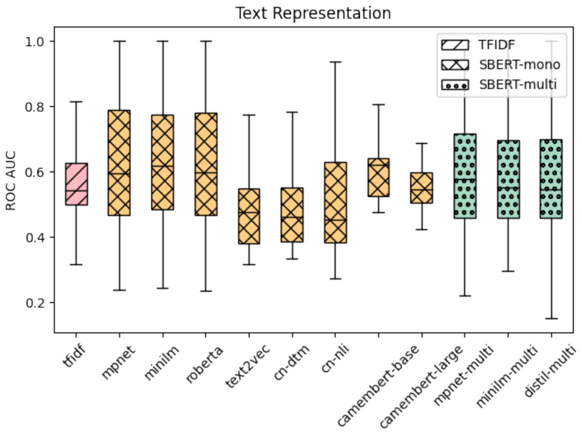


Figure 2: Performance (avg. ROC AUC) comparison of anomaly detection algorithms using different text representation strategies: TF-IDF, Monolingual SBERT, and Multilingual SBERT

speech detection, which may not align well with the TAD task. When excluding the Chinese models, the monolingual SBERT models perform slightly better than the multilingual ones, even though the difference is not substantial.

Algorithms Figures 3 to 5 depict the performance of the 22 selected algorithms evaluated across 17 different corpora using three types of representations. The algorithms are grouped from three perspectives:

- In terms of **degree of supervision**, the unsupervised approaches register a mean ROC AUC score of 0.539. This relatively lower score indicates that these methods may have struggled to effectively detect anomalies in the text data without any labeled information or prior knowledge. Semi-supervised methods, particularly OCSVM and LOF, which utilize only negative samples for training, perform slightly better with a mean ROC AUC score of 0.581. Nevertheless, semi-supervised methods that employ weak labels show a markedly improved performance, demonstrating a mean ROC AUC score of 0.721. This improvement hints at the significant role weak labels can play in enhancing anomaly detection performance.

- In terms of **underlying theory** for anomaly scores, proximity-based, probabilistic-based, and domain-based approaches exhibit relatively lower mean ROC AUC scores (0.538, 0.550, and 0.541, respectively), indicating their limitations in accurately identifying anomalies based on proximity or probabilistic reasoning. In contrast, reconstruction-based methods show a stronger performance with a mean ROC AUC score of 0.613. However, the most promising results are obtained by the ensemble methods, which achieve the highest mean ROC AUC score (0.825). These methods, leveraging the combination of multiple anomaly detection techniques or models, demonstrate superior performance in identifying anomalies in text data. Notably, the best-performing methods overall are the reconstruction-based and ensemble methods when utilizing weak labels within a semi-supervised learning context.

- In terms of **model architecture**, deep models utilizing neural networks achieve a mean ROC AUC of 0.621, demonstrating their relatively higher efficiency in detecting anomalies in text data compared to shallow models. Excluding XGBOD, the shallow models exhibit a lower mean ROC AUC of 0.549, suggesting their limited effectiveness. However, XGBOD, a shallow model employing extreme gradient boosting, stands out with an exceptional mean ROC AUC of 0.862, surpassing both deep and other shallow models. These findings highlight the advantage of deep neural networks in text data anomaly detection. Nevertheless, XGBOD defies expectations as a shallow model by delivering outstanding performance. Consequently, model architecture selection demands careful consideration, as both deep models and well-optimized shallow models, like XGBOD, can yield effective anomaly detection outcomes in text data.

5 Conclusion

In summary, this paper provides a comprehensive evaluation of 22 anomaly detection algorithms applied to 17 corpora derived from datasets associated with three distinct tasks. The evaluation considers three types of text representations: TF-IDF, monolingual SBERT, and multilingual SBERT models. The findings shed light on several key insights regarding the performance and limitations of these algorithms.

The analysis reveals variations in algorithm performance across different corpora categories. The

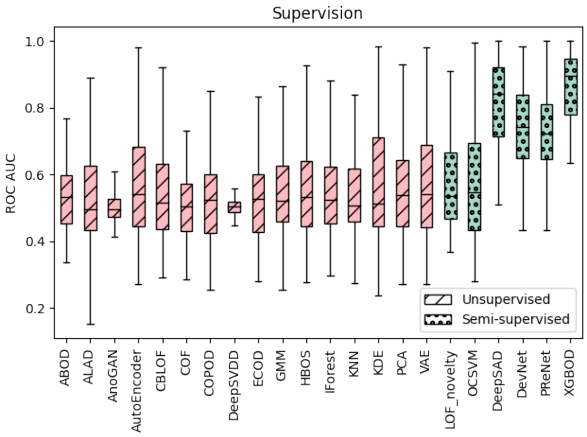


Figure 3: Performance (avg. ROC AUC) comparison of anomaly detection algorithms based on supervision level: Semi-supervised and Unsupervised

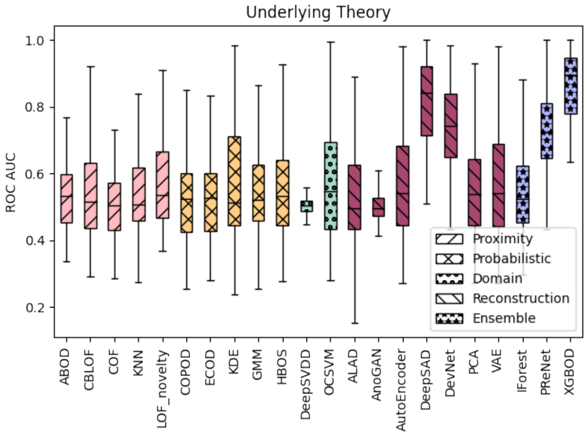


Figure 4: Performance (avg. ROC AUC) comparison of anomaly detection algorithms based on underlying theory for anomaly scores: Proximity-based, Probabilistic-based, Domain-based, Reconstruction-based, and Ensemble methods

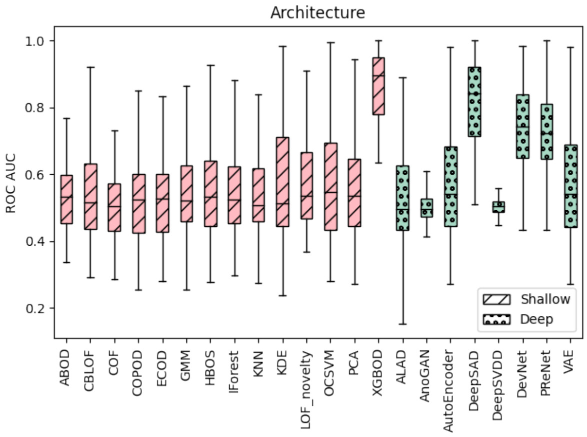


Figure 5: Performance (avg. ROC AUC) comparison of anomaly detection algorithms based on model architecture: Deep Models (with neural networks) and Shallow Models

corpora designed for topic classification exhibit the highest scores, indicating their suitability for anomaly detection tasks. In contrast, the hate speech corpora pose considerable challenges, with algorithms underperforming possibly due to the noisy social media text they contain. Addressing the impact of textual noise on algorithm performance becomes a crucial area for future research. Furthermore, the evaluation of different text representations demonstrates that the TF-IDF model shows stable performance but with a lower upper limit compared to SBERT models. Excluding the Chinese models, monolingual SBERT models outperformed the multilingual ones, emphasizing the importance of language-specific representations for anomaly detection. From the perspectives of degree of supervision, underlying theory for anomaly scores, and model architecture, the study offers a detailed comparative analysis of the algorithms. The findings highlight the superior performance of reconstruction-based and ensemble methods in a semi-supervised setting, and the advantage of deep models over shallow models, except for XGBOD.

Looking ahead, several potential avenues of investigation could further enrich the field of text anomaly detection. Firstly, the exploration of supervised algorithms could provide an opportunity to bolster anomaly detection performance, especially in contexts where labeled data is available. Secondly, the incorporation of advanced technologies, such as language models like ChatGPT, opens up novel possibilities for innovative anomaly detection methodologies that can adapt to evolving data landscapes. Another promising direction lies in the creation of specialized datasets explicitly designed for anomaly detection tasks. Such datasets could allow for the refining and optimization of current detection algorithms while enabling the development of new, more effective methods. Lastly, delving deeper into the study of different types of text anomalies could provide a more nuanced understanding of their unique characteristics and the detection strategies that work best for each.

References

Charu C. Aggarwal. 2015. *Data Mining*. Springer International Publishing, Cham.

Charu C. Aggarwal. 2017. *Outlier Analysis*. Springer International Publishing, Cham.

Leslie Barrett, Sidney Fletcher, and Robert Kingan.

2019. Textual Outlier Detection and Anomalies in Financial Reporting. In *2nd KDD Workshop on Anomaly Detection in Finance*, page 6.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD workshop: Languages for data mining and machine learning*, pages 108–122.

Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. *arXiv:1901.03407 [cs, stat]*. ArXiv: 1901.03407.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58.

Clément Christophe, Julien Velcin, Jairo Cugliari, Philippe Suignard, and Manel Boumghar. 2019. How to detect novelty in textual data streams? A comparative study of existing methods. *arXiv:1909.05099 [cs, stat]*. ArXiv: 1909.05099.

Paweł Cichosz. 2020. Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation. *Natural Language Engineering*, 26(5):551–578.

Chris Cieri, David Graff, Mark Liberman, Nii Martey, Stephanie Strassel, and others. 1999. The TDT-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News workshop*, pages 57–60.

Pradeep Dasigi and Eduard Hovy. 2014. Modeling Newswire Events using Neural Networks for Anomaly Detection. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 9.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A Benchmark for Chinese Offensive Language Detection. ArXiv:2201.06025 [cs].

Andriy Drozdyuk and Norbert Eke. 2017. Anomaly detection with Generative Adversarial Networks and text patches. page 13.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. ArXiv:1809.04444 [cs].

Markus Goldstein and Andreas Dengel. 2012. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63. Publisher: Citeseer.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*. ArXiv: 1406.2661.

Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. page 10.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. 2022. ADBench: Anomaly Detection Benchmark. Number: arXiv:2206.09426 arXiv:2206.09426 [cs].

Zengyou He, Xiaofei Xu, and Shengchun Deng. 2003. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650.

Chenlong Hu, Yukun Feng, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2021. One-class Text Classification with Multi-modal Deep Support Vector Data Description. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, volume Main Volume, pages 3378–3390. Association for Computational Linguistics.

Amir Jafari. 2022. A Deep Learning Anomaly Detection Method in Textual Data. ArXiv:2211.13900 [cs].

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2021. SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection. ArXiv:2108.03070 [cs].

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The Multilingual Amazon Reviews Corpus. *arXiv:2010.02573 [cs]*. ArXiv: 2010.02573.

Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. ArXiv:1312.6114 [cs, stat].

Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. 2008. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452, Las Vegas Nevada USA. ACM.

Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. 2007. Outlier Detection with Kernel Density Functions. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 4571, pages 61–75. Springer Berlin Heidelberg, Berlin, Heidelberg. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

David D. Lewis. 1997. Reuters-21578 Text Categorization Collection Data Set.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. 2020. COPOD: Copula-Based Outlier Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123, Sorrento, Italy. IEEE.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. 2022. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1. ArXiv:2201.00382 [cs, stat].

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, Pisa, Italy. IEEE.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, page 9.

Larry M Manevitz and Malik Yousef. 2001. One-Class SVMs for Document Classification. page 16.

Markos Markou and Sameer Singh. 2003a. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

Markos Markou and Sameer Singh. 2003b. Novelty detection: a review—part 2:. *Signal Processing*, 83(12):2499–2521.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. *arXiv:1908.11049 [cs]*. ArXiv: 1908.11049.

Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. 2020. COLD: Annotation scheme and evaluation data set for complex offensive language in English. *Journal for Language Technology and Computational Linguistics*, 34(1):1–28.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):1–38. ArXiv: 2007.02500.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep Anomaly Detection with Deviation Networks. ArXiv:1911.08623 [cs, stat].

Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. 2020. Deep Weakly-supervised Anomaly Detection. ArXiv:1910.13601 [cs, stat].

Jeremie Pantin, Marie-Jeanne Lesot, and Christophe Marsala. 2022. Analyse de données aberrantes pour le texte: Taxonomie et étude expérimentale. *TextMine'22*.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing*, 99:215–249.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient algorithms for mining outliers from large data sets.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*. ArXiv: 1908.10084.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. 2020. Deep Semi-Supervised Anomaly Detection. *arXiv:1906.02694 [cs, stat]*. ArXiv: 1906.02694.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. 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, Florence, Italy. Association for Computational Linguistics.

S Saranya, R Rajeshkumar, and S Shanthi. 2014. A survey on anomaly detection for discovering emerging topics. *International Journal of Computer Science and Mobile Computing*, 310(10):895–902.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. *arXiv:1703.05921 [cs]*. ArXiv: 1703.05921.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. 2003. A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.

Fahim K. Sufi and Musleh Alsulami. 2021. Automated Multidimensional Analysis of Global Events With Entity Detection, Sentiment Analysis and Anomaly Detection. *IEEE Access*, 9:152449–152460.

Zoltán Szoplák and Gabriela Andrejková. 2021. Anomaly detection in text documents using HTM networks. In *ITAT*, pages 20–28.

Jian Tang, Zhixiang Chen, Ada Wai-chee Fu, and David W. Cheung. 2002. Enhancing effectiveness of outlier detections for low density patterns. In *Advances in knowledge discovery and data mining*, pages 535–548, Berlin, Heidelberg. Springer Berlin Heidelberg.

Graham Todd, Catalin Voss, and Jenny Hong. 2020. Unsupervised Anomaly Detection in Parole Hearings using Language Models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 66–71, Online. Association for Computational Linguistics.

Gonzalo de la Torre-Abaitua, Luis Fernando Lago-Fernández, and David Arroyo. 2021. A Compression-Based Method for Detecting Anomalies in Textual Data. *Entropy*, 23(5):618.

Mathias Wahl. 2021. Detecting Hate Speech in Norwegian Texts Using BERT Semi-Supervised Anomaly Detection. Master's thesis, Norwegian University of Science and Technology.

Zhaoxia Wang, Victor Joo, Chuan Tong, Xin Xin, and Hoong Chor Chin. 2014. Anomaly Detection through Enhanced Sentiment Analysis on Social Media Data. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pages 917–922, Singapore, Singapore. IEEE.

Xu, Hongzuo. DeepOD: Python deep Outlier/Anomaly detection. Tex.version: 0.2.

Tec Yan Yap. 2020. Text Anomaly Detection with ARAE-AnoGAN. *Honors Projects*, 22.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Houssam Zenati, Manon Romain, Chuan Sheng Foo, Bruno Lecouat, and Vijay Ramaseshan Chandrasekhar. 2018. Adversarially Learned Anomaly Detection. ArXiv:1812.02288 [cs, stat].

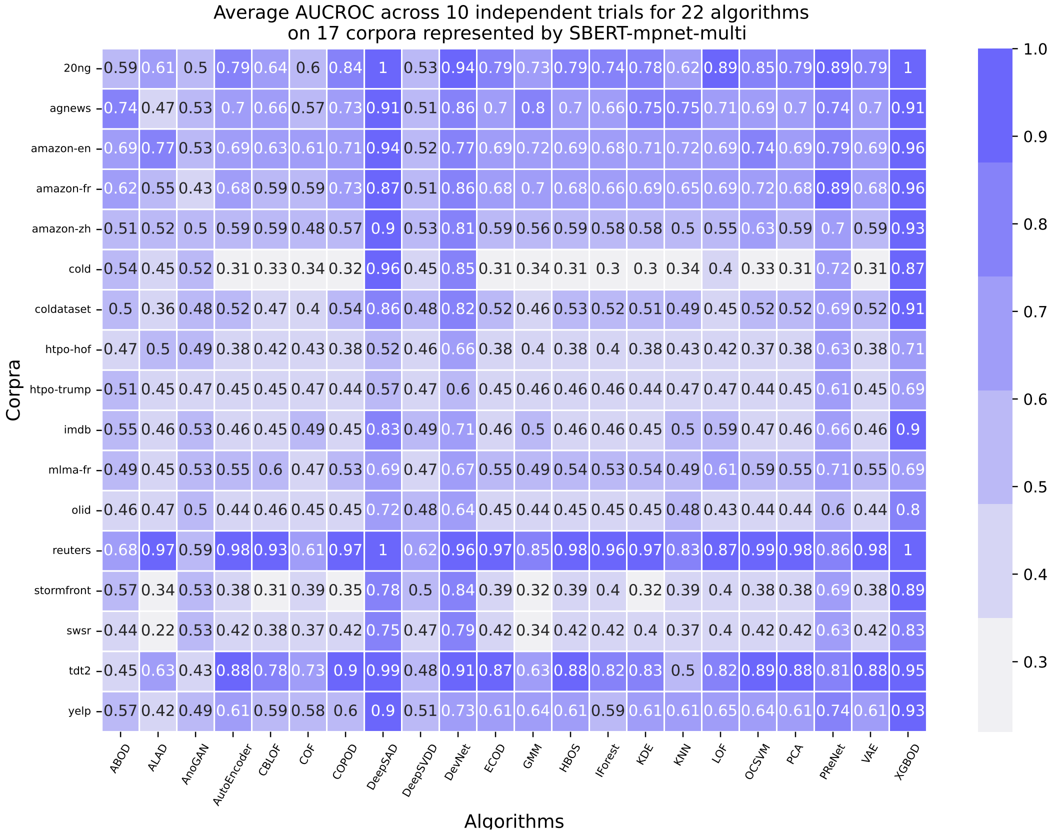
Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly Supervised Text Classification using Supervision Signals from a Language Model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2295–2305, Seattle, United States. Association for Computational Linguistics.

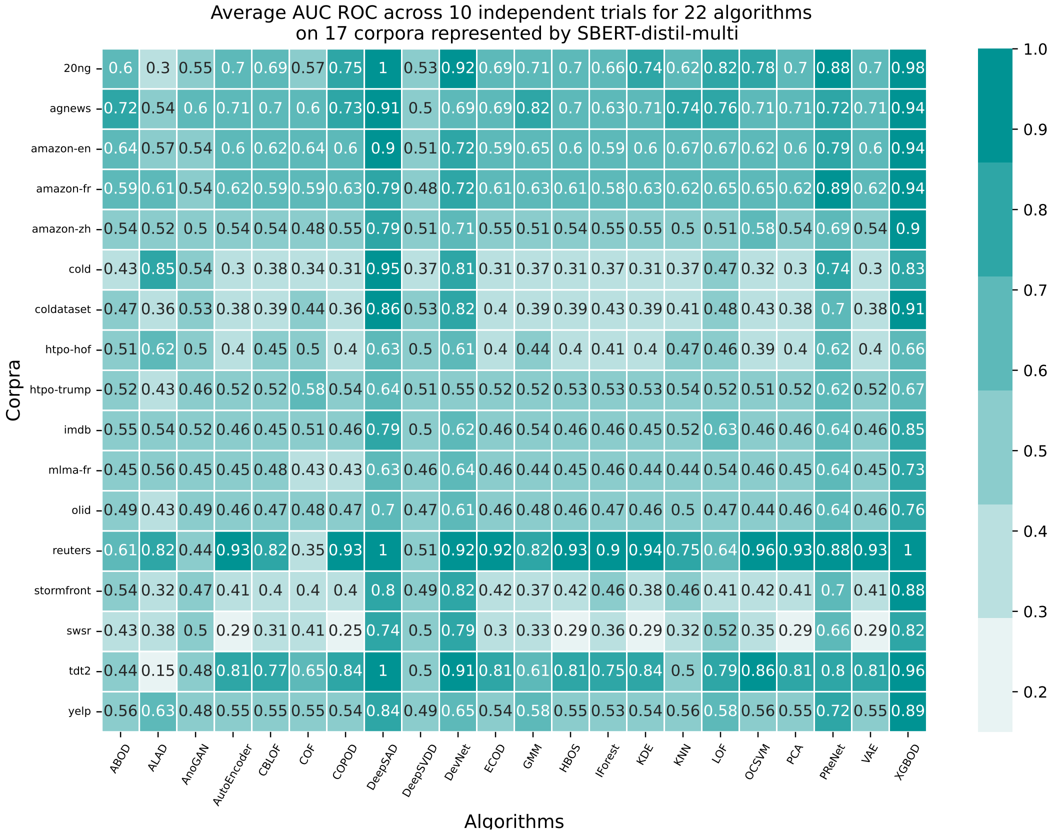
Xiang Zhang, Zhao Junbao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. *Advances in neural information processing systems*, 28. Number: arXiv:1502.01710 arXiv:1502.01710 [cs].

Yue Zhao and Maciej K. Hryniewicki. 2018. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rio de Janeiro. IEEE.

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

A Partial Results of the Experiments


Average AUCROC across 10 independent trials for 22 algorithms
on 17 corpora represented by SBERT-mpnet-multi
Corpra
Algorithms
ABOD ALAD AnoGAN AutoEncoder CBLOF COF COPOD DeepSAD DeepSVDD DevNet ECOD GMM HBOS IForest KDE KNN LOF OCSVM PCA PReNet VAE XGBOD
20ng 0.59 0.61 0.5 0.79 0.64 0.6 0.84 1 0.53 0.94 0.79 0.73 0.79 0.74 0.78 0.62 0.89 0.85 0.79 0.89 0.79 1
agnews 0.74 0.47 0.53 0.7 0.66 0.57 0.73 0.91 0.51 0.86 0.7 0.8 0.7 0.66 0.75 0.75 0.71 0.69 0.7 0.74 0.7 0.91
amazon-en 0.69 0.77 0.53 0.69 0.63 0.61 0.71 0.94 0.52 0.77 0.69 0.72 0.69 0.68 0.71 0.72 0.69 0.74 0.69 0.79 0.69 0.96
amazon-fr 0.62 0.55 0.43 0.68 0.59 0.59 0.73 0.87 0.51 0.86 0.68 0.7 0.68 0.66 0.69 0.65 0.69 0.72 0.68 0.89 0.68 0.96
amazon-zh 0.51 0.52 0.5 0.59 0.59 0.48 0.57 0.9 0.53 0.81 0.59 0.56 0.59 0.58 0.58 0.5 0.55 0.63 0.59 0.7 0.59 0.93
cold 0.54 0.45 0.52 0.31 0.33 0.34 0.32 0.96 0.45 0.85 0.31 0.34 0.31 0.3 0.3 0.34 0.4 0.33 0.31 0.72 0.31 0.87
coldataset 0.5 0.36 0.48 0.52 0.47 0.4 0.54 0.86 0.48 0.82 0.52 0.46 0.53 0.52 0.51 0.49 0.45 0.52 0.52 0.69 0.52 0.91
htpo-hof 0.47 0.5 0.49 0.38 0.42 0.43 0.38 0.52 0.46 0.66 0.38 0.4 0.38 0.4 0.38 0.43 0.42 0.37 0.38 0.63 0.38 0.71
htpo-trump 0.51 0.45 0.47 0.45 0.45 0.47 0.44 0.57 0.47 0.6 0.45 0.46 0.46 0.46 0.44 0.47 0.47 0.44 0.45 0.61 0.45 0.69
imdb 0.55 0.46 0.53 0.46 0.45 0.49 0.45 0.83 0.49 0.71 0.46 0.5 0.46 0.46 0.45 0.5 0.59 0.47 0.46 0.66 0.46 0.9
mlma-fr 0.49 0.45 0.53 0.55 0.6 0.47 0.53 0.69 0.47 0.67 0.55 0.49 0.54 0.53 0.54 0.49 0.61 0.59 0.55 0.71 0.55 0.69
olid 0.46 0.47 0.5 0.44 0.46 0.45 0.45 0.72 0.48 0.64 0.45 0.44 0.45 0.45 0.45 0.48 0.43 0.44 0.44 0.6 0.44 0.8
reuters 0.68 0.97 0.59 0.98 0.93 0.61 0.97 1 0.62 0.96 0.97 0.85 0.98 0.96 0.97 0.83 0.87 0.99 0.98 0.86 0.98 1
stormfront 0.57 0.34 0.53 0.38 0.31 0.39 0.35 0.78 0.5 0.84 0.39 0.32 0.39 0.4 0.32 0.39 0.4 0.38 0.38 0.69 0.38 0.89
swsr 0.44 0.22 0.53 0.42 0.38 0.37 0.42 0.75 0.47 0.79 0.42 0.34 0.42 0.42 0.4 0.37 0.4 0.42 0.42 0.63 0.42 0.83
tdt2 0.45 0.63 0.43 0.88 0.78 0.73 0.9 0.99 0.48 0.91 0.87 0.63 0.88 0.82 0.83 0.5 0.82 0.89 0.88 0.81 0.88 0.95
yelp 0.57 0.42 0.49 0.61 0.59 0.58 0.6 0.9 0.51 0.73 0.61 0.64 0.61 0.59 0.61 0.61 0.65 0.64 0.61 0.74 0.61 0.93


Average AUC ROC across 10 independent trials for 22 algorithms
on 17 corpora represented by SBERT-distil-multi
Corpra
Algorithms
ABOD ALAD AnoGAN AutoEncoder CBLOF COF COPOD DeepSAD DeepSVDD DevNet ECOD GMM HBOS IForest KDE KNN LOF OCSVM PCA PReNet VAE XGBOD
20ng 0.6 0.3 0.55 0.7 0.69 0.57 0.75 1 0.53 0.92 0.69 0.71 0.7 0.66 0.74 0.62 0.82 0.78 0.7 0.88 0.7 0.98
agnews 0.72 0.54 0.6 0.71 0.7 0.6 0.73 0.91 0.5 0.69 0.69 0.82 0.7 0.63 0.71 0.74 0.76 0.71 0.71 0.72 0.71 0.94
amazon-en 0.64 0.57 0.54 0.6 0.62 0.64 0.6 0.9 0.51 0.72 0.59 0.65 0.6 0.59 0.6 0.67 0.67 0.62 0.6 0.79 0.6 0.94
amazon-fr 0.59 0.61 0.54 0.62 0.59 0.59 0.63 0.79 0.48 0.72 0.61 0.63 0.61 0.58 0.63 0.62 0.65 0.65 0.62 0.89 0.62 0.94
amazon-zh 0.54 0.52 0.5 0.54 0.54 0.48 0.55 0.79 0.51 0.71 0.55 0.51 0.54 0.55 0.55 0.5 0.51 0.58 0.54 0.69 0.54 0.9
cold 0.43 0.85 0.54 0.3 0.38 0.34 0.31 0.95 0.37 0.81 0.31 0.37 0.31 0.37 0.31 0.37 0.47 0.32 0.3 0.74 0.3 0.83
coldataset 0.47 0.36 0.53 0.38 0.39 0.44 0.36 0.86 0.53 0.82 0.4 0.39 0.39 0.43 0.39 0.41 0.48 0.43 0.38 0.7 0.38 0.91
htpo-hof 0.51 0.62 0.5 0.4 0.45 0.5 0.4 0.63 0.5 0.61 0.4 0.44 0.4 0.41 0.4 0.47 0.46 0.39 0.4 0.62 0.4 0.66
htpo-trump 0.52 0.43 0.46 0.52 0.52 0.58 0.54 0.64 0.51 0.55 0.52 0.52 0.53 0.53 0.53 0.54 0.52 0.51 0.52 0.62 0.52 0.67
imdb 0.55 0.54 0.52 0.46 0.45 0.51 0.46 0.79 0.5 0.62 0.46 0.54 0.46 0.46 0.45 0.52 0.63 0.46 0.46 0.64 0.46 0.85
mlma-fr 0.45 0.56 0.45 0.45 0.48 0.43 0.43 0.63 0.46 0.64 0.46 0.44 0.45 0.46 0.44 0.44 0.54 0.46 0.45 0.64 0.45 0.73
olid 0.49 0.43 0.49 0.46 0.47 0.48 0.47 0.7 0.47 0.61 0.46 0.48 0.46 0.47 0.46 0.5 0.47 0.44 0.46 0.64 0.46 0.76
reuters 0.61 0.82 0.44 0.93 0.82 0.35 0.93 1 0.51 0.92 0.92 0.82 0.93 0.9 0.94 0.75 0.64 0.96 0.93 0.88 0.93 1
stormfront 0.54 0.32 0.47 0.41 0.4 0.4 0.4 0.8 0.49 0.82 0.42 0.37 0.42 0.46 0.38 0.46 0.41 0.42 0.41 0.7 0.41 0.88
swsr 0.43 0.38 0.5 0.29 0.31 0.41 0.25 0.74 0.5 0.79 0.3 0.33 0.29 0.36 0.29 0.32 0.52 0.35 0.29 0.66 0.29 0.82
tdt2 0.44 0.15 0.48 0.81 0.77 0.65 0.84 1 0.5 0.91 0.81 0.61 0.81 0.75 0.84 0.5 0.79 0.86 0.81 0.8 0.81 0.96
yelp 0.56 0.63 0.48 0.55 0.55 0.55 0.54 0.84 0.49 0.65 0.54 0.58 0.55 0.53 0.54 0.56 0.58 0.56 0.55 0.72 0.55 0.89