Using Locally Learnt Word Representations for better Textual Anomaly Detection

Alicia Breidenstein[1,2] and Matthieu Labeau[1]

[1]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
[2]Caisse des Dépôts et Consignations, Paris, France
{alicia.breidenstein, matthieu.labeau}@telecom-paris.fr

Abstract

The literature on general purpose textual Anomaly Detection is quite sparse, as most textual anomaly detection methods are implemented as out of domain detection in the context of pre-established classification tasks. Notably, in a field where pre-trained representations and models are of common use, the impact of the pre-training data on a task that lacks supervision has not been studied. In this paper, we use the simple setting of k-classes out anomaly detection and search for the best pairing of representation and classifier. We show that well-chosen embeddings allow a simple anomaly detection baseline such as OC-SVM to achieve similar results and even outperform deep state-of-the-art models.

1 Introduction

Anomaly Detection (AD) consists in detecting observations that deviate from *normality*: what is normal is defined by available data and assumed to be bounded (Ruff et al., 2021), while anomalies (which can be called *outliers*, or *novelty* depending on the application) are outside this bound. The most obvious hurdle with AD is that it is usually not possible to characterize anomalies: models are mostly not designed to target a specific type of outlier, and the assumptions made on data are rarely stated. In this context, supervision usually comes from *normal* data. However, most NLP models employ pre-trained representations: the impact that this kind of prior knowledge may have on AD is difficult to appreciate, and overlooked.

A first attempt to characterize outliers in natural language data was made by Arora et al. (2021), classifying them as coming from either *background* shifts (coming from a shift in domain) or *semantic* shifts (coming from a shift in content), and bringing insights on which detection method might better work on each. Arora et al. (2021) showed that background shifts are well detected by language models, which are able to estimate the density of normal data; there is furthermore an abundant literature on adapting pre-trained language model to new normal data (Ramponi and Plank, 2020).

We hence focus on semantic shifts, which are shown to be well detected by calibration methods. However, this assumes access to a classification model trained on relevant categories; we however prefer to not assume access to any labels, and adopt a simple but convenient way of evaluating AD: repurposing classification datasets by declaring one class to be normal and the others as anomalies, in what is called *k-classes-out*. In this setting, existing approaches are fewer. Some are inspired by topic modeling: they learn topic models optimized to reconstruct normal data well, aiming to detect anomalies by failing to accurately reconstruct them. For example, CVDD (Ruff et al., 2019) learns a limited number of topic-centroid vectors by applying attention upon pre-trained word-embeddings. A second direction is to train deep self-supervised models to recognize anomalies that are simulated, for example through random perturbation of data, as for DATE (Manolache et al., 2021). While both these models were previously compared on common datasets, CVDD uses pre-trained representations and DATE is only trained on the data available for the AD task.

In this paper, our goal is to investigate the impact of the pre-training data on anomaly detection performance in the k-classes-out setting; we experiment with static and contextual representations, off-the-shelf or obtained strictly on the AD training data, on three datasets. Our results show that the most simple configuration - a simple non-neural classification model, when equipped with textual representations obtained from the AD training data, can beat state-of-the-art models on our AD task.

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 82–91
June 20, 2024 ©2024 Association for Computational Linguistics

2 Background

2.1 Preliminaries

Anomaly score: To classify a data point $x \in \mathcal{X}$ as an anomaly, we compute an anomaly score $s : \mathcal{X} \to \mathbb{R}$ indicating its *degree of anomalousness* $s(x)$ (Ruff et al., 2021); then, a threshold δ is used as cutoff. However, we will here use measures that evaluate the performance of AD models using only s and independently of the choice of δ[1].

Data: We consider a training set of documents $\mathcal{D}_{train} = \{x_i\}_{i=1}^n$ for our task, which is part of a larger dataset: $\mathcal{D}_{train} \subset \mathcal{D}$. A document $x = (w_1, w_2, \dots, w_l)$ is a sequence of $l \in \mathbb{N}$ words from a vocabulary $\mathcal{V}$. We will use different vector representations $\mathbf{x}$ of x depending on the method.

Pre-training word embeddings: In Ruff et al. (2019), CVDD is tested with embeddings $\mathbf{W} \in \mathbb{R}^{d \times |\mathcal{V}|}$ pre-trained with *FastText* and *GloVe*. However, those where trained on an external dataset, which might be very different than $\mathcal{D}_{train}$: hence, we propose to experiment with representations pre-trained on $\mathcal{D}_{train}$ and $\mathcal{D}$. We choose to use FastText, as the better performing static word representation algorithm. However, to avoid training prediction-based representations on datasets that are too small, we also use a traditional alternative in NLP, the *PPMI*(Church and Hanks, 1990) matrix, which we reduce to the appropriate dimension d using the SVD. As DATE is based on ELECTRA (Clark et al., 2020), we also experiment with representations obtained through its off-the-shelf version, and through one pre-trained on $\mathcal{D}$.

2.2 Anomaly Detection methods

We present in this section the necessary background information about the two models we experiment with, CVDD and DATE, as well as the chosen baseline. We follow Ruff et al. (2019) and use OC-SVM, a one-class classification-based AD model[2].

CVDD: CVDD scores a document by computing an average anomaly score over r *topics*. It takes as input word representations $\mathbf{X} = (\mathbf{w}_{w_j})_{j=1}^l \in \mathbb{R}^{d \times l}$. It learns jointly two components: (1) a multi-head self-attention mechanism, which computes sets of attention scores over the l input word embeddings for each of the r attention heads, grouped in $\mathbf{A} \in \mathbb{R}^{l \times r}$, allowing to aggregate them into r representations $\mathbf{M} = \mathbf{XA} \in \mathbb{R}^{d \times r}$, and (2) a set of r topic vectors $\mathbf{C} = (\mathbf{c}_k)_{k=1}^r \in \mathbb{R}^{d \times r}$ whose cosine distances with the corresponding training data representations $d(\mathbf{c}_k, \mathbf{m}_k)$ are minimized through the training objective. The anomaly score is, for a new document x_{test}, computed as follows:

$$s_{CVDD}(x_{test}) = \sum_{k=1}^{r} d(\mathbf{c}_k, \mathbf{X}_{test}\mathbf{a}_k)$$

DATE: DATE masks some of the tokens of the document, uses a generator to replace them, and learns through a transformer model D based on ELECTRA to detect the tokens which were modified, via a binary classification task called *Replaced Token Detection* (RTD). Motivated by computational efficiency, the authors propose to use as score the probability of each token *not being modified*:

$$s_{DATE}(x_{test}) = \frac{1}{l} \sum_{j=1}^{l} P_{RTD}(m_j = 0 | x_{test}, D)$$

where m_j is a boolean indicating if the token w_j has been modified in the input to the model D. The model is trained to maximize the log-likelihood of this distribution on perturbed data. It is trained jointly using the *Replaced Mask Detection* (RMD) objective, which aims at predicting which masking pattern is used, and with the Masked Language Modeling (MLM) objective. DATE jointly learns its own contextual word representations, and is given the document x as input. It takes decisions at the token-level, which is made possible by using contextual representations. Note that the score s_{DATE} will give a high value to inliers examples, and should be reversed for comparison.

OC-SVM: We define our OC-SVM model following the baseline of CVDD: it uses the Scikit-learn (Pedregosa et al., 2011) implementation, based on the model described by Schölkopf et al. (2001). It takes as input the aggregate[3] $\mathbf{x}^{aggr} = \frac{1}{l}\sum_{j=1}^{l} \mathbf{w}_{w_j} \in \mathbb{R}^d$ and aims at separating all the training data points *from the origin* in the feature space $\mathcal{F}_k$. This space is defined as the reproducing

[1] Selecting this threshold is a difficult problem in itself, with values selected by validation not generalizing well (Khosla and Gangadharaiah, 2022).

[2] We also experimented on TONMF (Kannan et al., 2017) and their baseline LSA as well, but the results of these baselines were worse than the ones we obtain with CVDD, DATE and OC-SVM.

[3] Contrarily to Ruff et al. (2019), we don't present results using tf-idf to weight word embeddings, as we did not find it to produce competitive results.

kernel Hilbert space (RKHS) associated to the chosen positive semi-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and corresponding feature map $\phi_k : \mathbb{R}^d \to \mathcal{F}_k$. Separating data from the origin is done looking for a hyper-plane $\boldsymbol{\omega} \in \mathcal{F}_k$ maximizing a margin:

$$\min_{\boldsymbol{\omega},\rho,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{\omega}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i$$
$$s.t \quad \rho - \langle \phi_k(\mathbf{x}_i^{aggr}), \boldsymbol{\omega} \rangle \leq \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

where the margin to the origin is given by $\frac{\rho}{\|\boldsymbol{\omega}\|}$, and the $\boldsymbol{\xi}$ are the slack variables. The decision function should be positive for most training data points $\mathbf{x}_i^{aggr}$. Here, ν does not control the smoothness of the margin, but the *fraction of the data which the model will be allowed to consider as outliers*. Finally, the scoring function is simply minus the value given by f:

$$s_{OCSVM}(x_{test}) = \langle \phi_k(\mathbf{x}_{test}^{aggr}), \boldsymbol{\omega} \rangle - \rho$$

3 Experimental setting

We evaluate the performance of these models quantitatively on several datasets: after exploring the impact that the pre-training data used for word representations has on anomaly detection with OC-SVM and CVDD, we compare all models. [4]

3.1 Datasets

Following Manolache et al. (2021), we first compare the different methods on two publicly available textual datasets containing news articles for classification purposes: *20 Newsgroups*[5] and *AG News*[6]. The third dataset, *RNCP*[7], for *Répertoire National des Certifications Professionelles*, was built from a public official french repository with training certifications. The relevant statistics for the datasets are given in Table 5. For all datasets, we follow the pre-processing from Ruff et al. (2019).

3.2 Experimental details

Most of our experimental choices are made following Ruff et al. (2019). We mainly extend their experimental framework by looking at supplementary representations for the OC-SVM and CVDD models, trying to compare these approaches more fairly with respect to the data available to the model. Unless mentioned, for each model, we chose hyperparameters following the reference paper.

Evaluation with k-classes-out: Noting $\mathcal{C}$ the set of classes of the dataset $\mathcal{D}$, for each $c \in \mathcal{C}$ we have a train and test sets $\mathcal{D}^c_{train}$ and $\mathcal{D}^c_{test}$. In order to adapt the datasets to AD, one class c_{normal} is picked, while the others are considered to be anomalous. In our experimental setting, which we call *semi-supervised*, we consider that the normal class has been properly labeled, and the model is trained with exactly $\mathcal{D}^{c_{normal}}_{train}$. It is then evaluated on $\mathcal{D}_{test} = \bigcup_{c\in\mathcal{C}} \mathcal{D}^c_{test}$, where only elements of $\mathcal{D}^{c_{normal}}_{test}$ are to be recognized as inliers by the model. Experiments are repeated with taking every $c \in \mathcal{C}$ as c_{normal}. We present similar experiments in an *unsupervised* setting, where anomalies are present in the training data, in Appendix B.2.

Evaluation metrics: We use the Area Under Receiver Operating Curve (AUROC, or AUC) which is widely employed in the AD literature. It allows to measure the performance of a binary classifier by computing the area under the ROC curve, obtained by plotting the true positive rate against the false positive rate: hence, it covers the range of possible thresholds δ between normality and anomalies over the possible outputs of the anomaly score $s(x)$.

Experimenting with pre-trained representations: Following Section 2.1, we propose to experiment with various sets of representations for OC-SVM and CVDD: first, the FastText representations for English (and French, for RNCP) trained on Wikipedia and Common Crawl[8], which we note FT$_{Large}$. Then, we train our own embeddings with FastText on $\mathcal{D}$, and $\mathcal{D}^{c_{normal}}_{train}$, noting them respectively FT$_{\mathcal{D}}$ and FT$_{\mathcal{C}}$. Similarly, we note the representations obtained by reducing the dimension of a PPMI matrix[9] PPMI$_{\mathcal{D}}$ and PPMI$_{\mathcal{C}}$. For these pre-trained representations, we use $d = 300$. Lastly, we experimented with the ELECTRA model available on Huggingface[10] and one we trained on $\mathcal{D}$; as well as those obtained through the corresponding DATE model. As **none of the contextual representations gave competitive results**, we only display

[4] The code is available at https://github.com/abreidenstein/TextualAD

[5] http://qwone.com/~jason/20Newsgroups/

[6] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

[7] https://www.data.gouv.fr/en/datasets/repertoire-national-des-certifications-professionnelles-et-repertoire-specifique/

[8] https://fasttext.cc/docs/en/crawl-vectors.html

[9] We follow here Turney (2012, Section 3.6) and don't use the eigenvalues when reducing the dimension.

[10] https://huggingface.co/google/electra-small-discriminator

20Ng	OC-SVM			CVDD
	Linear	Poly	RBF	Best $r \in [1, 10]$
FT$_{\mathcal{D}}$	**81.4** ± 0.1	**76.3** ± 0.1	58.4 ± 0.1	55.8 ± 0.3 ($r = 10$)
FT$_{\mathcal{C}}$	69.9 ± 0.2	69.7 ± 0.1	35.6 ± 0.1	50.0 ± 0.3 ($r = 5$)
FT$_{Large}$	66.0 ± 0.2	65.9 ± 0.2	66.4 ± 0.1	68.0 ± 0.1 ($r = 3$)
PPMI$_{\mathcal{D}}$	59.4 ± 0.2	59.1 ± 1.6	**75.1** ± 0.1	**70.4** ± 1.6 ($r = 2$)
PPMI$_{\mathcal{C}}$	74.5 ± 0.1	74.6 ± 0.1	40.6 ± 0.2	55.7 ± 0.2 ($r = 2$)

Table 1: AUCs of AD experiments over 20Ng, with OC-SVM with Linear, Poly and RBF kernels, and CVDD.

AGNews	OC-SVM			CVDD
	Linear	Poly	RBF	Best $r \in [1, 10]$
FT$_{\mathcal{D}}$	**89.8** ± 0.01	**87.7** ± 0.03	72.6 ± 0.1	86.5 ± 0.5 ($r = 1$)
FT$_{\mathcal{C}}$	79.6 ± 0.1	87.3 ± 0.1	20.8 ± 0.1	62.8 ± 0.6 ($r = 1$)
FT$_{Large}$	82.2 ± 0.1	82.0 ± 0.1	79.1 ± 0.1	**87.2** ± 0.7 ($r = 2$)
PPMI$_{\mathcal{D}}$	61.2 ± 0.1	60.6 ± 0.1	**89.4** ± 0.01	83.9 ± 0.2 ($r = 2$)
PPMI$_{\mathcal{C}}$	79.5 ± 0.1	79.8 ± 0.1	29.9 ± 0.1	58.7 ± 0.9 ($r = 5$)

Table 2: AUCs of AD experiments over AG News, with OC-SVM with Linear, Polynomial and RBF kernels, and CVDD.

RNCP	OC-SVM			CVDD
	Linear	Poly	RBF	Best $r \in [1, 15]$
FT$_{\mathcal{D}}$	**63.7** ± 0.05	**61.5** ± 0.04	**57.8** ± 0.05	**58.3** ± 0.4 ($r = 8$)
FT$_{\mathcal{C}}$	60.6 ± 0.04	60.8 ± 0.04	41.3 ± 0.1	52.2 ± 0.3 ($r = 10$)
FT$_{Large}$	56.2 ± 0.1	56.2 ± 0.2	55.0 ± 0.04	56.6 ± 0.3 ($r = 12$)
PPMI$_{\mathcal{D}}$	58.4 ± 0.04	58.6 ± 0.03	57.2 ± 0.1	56.9 ± 0.2 ($r = 2$)
PPMI$_{\mathcal{C}}$	57.4 ± 0.1	58.8 ± 0.1	49.0 ± 0.04	52.2 ± 0.1 ($r = 1$)

Table 3: AUCs of AD experiments over RNCP, with OC-SVM with Linear, Poly and RBF kernels, and CVDD.

the corresponding results in Appendix B.3.

4 Results

Choosing word representations: The results for CVDD and OC-SVM[11] obtained with the remaining static representations are presented in Table 1, 2 and 3 for two of the datasets. FT$_{\mathcal{D}}$ representations show consistently better performances than FT$_{Large}$, and the best overall, especially when used with an OC-SVM with a linear kernel. With class-based representations, the results of OC-SVM models seem to vary following the size of the dataset: the larger it is, the closer the results get to those of dataset-based representation. In particular, FT$_{\mathcal{C}}$ representations give great results on AG News with a polynomial kernel, as reported in Table 2. We hence postulate that the poorer performance of FT$_{\mathcal{C}}$ representations is linked to a lack of training data. With linear and polynomial kernels, PPMI$_{\mathcal{C}}$ give good results and largely beats PPMI$_{\mathcal{D}}$ on 20 Newsgroups and AG News, opposite to what we see with FT representations. We assume here that statistics obtained only on class data are more representative, and hence work better with simpler kernels. We discuss the poor performance of class-based representations with the RBF kernel in Appendix A.2.

	AGNews	20Ng	RNCP
OC-SVM + FT$_{Large}$	82.2 ± 0.1	66.0 ± 0.2	56.2 ± 0.1
OC-SVM + *ours*	**89.8** ± 0.01	**81.4** ± 0.1	**63.7** ± 0.05
CVDD + FT$_{Large}$	87.2 ± 0.7	68.0 ± 0.1	56.6 ± 0.3
CVDD + *ours*	86.5 ± 0.5	70.4 ± 1.6	58.3 ± 0.4
DATE	*88.5* ± 0.2	*70.9* ± 0.4	*59.2* ± 0.1

Table 4: AUCs and standard deviations of AD experiments over all datasets, with all models. For OC-SVM and CVDD, we show the best results across hyperparameters with FT$_{Large}$, and across our own word representations, for which we took FT$_{\mathcal{D}}$ representations except for CVDD with 20 Newsgroups, where PPMI$_{\mathcal{D}}$ provide better results.

Overall comparison: Table 4 presents the best results obtained for each model, with comparison to DATE; additionally, for OC-SVM and CVDD, we present results for our representations (noted *ours*) and external representations separately. OC-SVM outperforms CVDD on all datasets. It reaches better results than DATE, especially on 20 Newsgroups and RNCP, although being far simpler. For all the models, the AUC values on the RNCP dataset are lower, which can be due to the shortness of the documents in this dataset, making the AD task more challenging.

On the performance of OC-SVM: our results show that, with appropriate representations, a simple OC-SVM model outmatches complex models such as CVDD and DATE. We hypothesize that, in our setting especially, AD approaches based on one-class classification are at an advantage; but the objective with which DATE is trained may lead the model away from what is needed in the k-classes-out setting, as it learns to detect random replacements. Here, the simplicity of an OC-SVM is a strength, though it has the disadvantage of not providing any density score nor possible word-level interpretation, contrarily to CVDD (through the attention mechanism) and DATE.

On the performance of dataset-based representations: our results show the clear superiority of representations pre-trained on the same data that will be used on the AD task. While dataset-based representations will generally not be available at

[11] The scikit-learn implementation of OC-SVM is deterministic. Variations in our results come from the composition of document representations from word embeddings; we suppose this is due to how padding is handled in the implementation of (Ruff et al., 2019).

training time, we argue that an OC-SVM model with class-based representations and a polynomial kernel should provide results that are very competitive with state-of-the-art models; the choice of representation pre-training method should depend on the quantity of training data available. Our results with contextual representations are in line with previous results from Ruff et al. (2019).

5 Conclusion

In this paper, we implement a fair comparison between existing textual anomaly detection methods in a k-classes-out setting and show that training the models on only the data available for the AD task can lead to better results. This allows methods regarded as baselines, such as OC-SVM models, to achieve impressive results, challenging state-of-the-art models based on deep neural architectures, with only the data available at hand. We intend to extend this line of work towards more challenging textual AD tasks. We also believe our results are indicative of the potential of model adaptation methods for semantic anomaly detection, which is a direction that has only been seldom explored (Xu et al., 2021). In the future, we also intend to extend our investigation to larger, more recent language models for obtaining representations.

6 References

Mira Ait-Saada and Mohamed Nadif. 2023. Unsupervised anomaly detection in multi-topic short-text corpora. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1392–1403, Dubrovnik, Croatia. Association for Computational Linguistics.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. 2022. Adbench: Anomaly detection benchmark.

Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. 2017. *Outlier Detection for Text Data*.

Sopan Khosla and Rashmi Gangadharaiah. 2022. Evaluating the practical utility of confidence-score based techniques for unsupervised open-world classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 18–23, Dublin, Ireland. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Nianzu Ma, Alexander Politowicz, Sahisnu Mazumder, Jiahua Chen, Bing Liu, Eric Robertson, and Scott Grigsby. 2021. Semantic novelty detection in natural language descriptions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 866–882, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Larry M. Manevitz and Malik Yousef. 2002. One-class svms for document classification. *J. Mach. Learn. Res.*, 2.

Andrei Manolache, Florin Brad, and Elena Burceanu. 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 267–277, Online. Association for Computational Linguistics.

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

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. 2021. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*.

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.

Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. 2001. Estimating support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. 44:533–585.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1052–1061, Online. Association for Computational Linguistics.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 167–174, Online. Association for Computational Linguistics.

A Datasets and hyperparameters

A.1 Dataset description

Textual AD datasets and evaluation: To the best of our knowledge, only a handful of AD-specific textual datasets have been released: among them, CLINC150 (Larson et al., 2019), an intent classification[12] dataset comprising OOD examples, and the recent NSD2 (Ma et al., 2021) proposing anomalies that are created as fine-grained semantic modifications. As our objective is to get a clearer view of the performance of existing models, we choose to stay in the simple but popular setting of k-classes-out: we should note that this effectively restricts our study to the detection of what Arora et al. (2021) call semantic shifts. Many classification datasets have been used this way, a few of them being part of the recently released AD benchmark *ADBench* (Table B1: Han et al., 2022). Among those, we choose to re-use 20 Newsgroups and AG News, which DATE was applied to (Manolache et al., 2021). Following Ait-Saada and Nadif (2023), we diversify our experiments with a difficult classification dataset based on the French repository of training certifications, containing short texts (certification titles) with little lexical overlap within classes.

20 Newsgroups: This dataset is composed of newsgroups posts from 20 topics split between a training and a testing set. We reproduce the setup of Ruff et al. (2019); Manolache et al. (2021) and group the articles into 6 top-level categories.

AG News: This topic classification dataset was built by choosing the 4 largest classes from the original AG dataset and contains news articles collected from numerous news sources, and also includes an *train/test* split.

[12] Intent classification has attracted a large part of the efforts dedicated to textual AD, including a dedicated comparative framework (Zhang et al., 2021).

	$\lvert\mathcal{D}_{train}\rvert$			**General statistics on $\mathcal{D}$**			
Dataset	Smallest	Largest	Median	$\lvert\mathcal{D}_{train}\rvert/\lvert\mathcal{D}_{test}\rvert$ Ratio	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{V}\rvert$	Median(l)
AG News	30000	30000	30000	30000/1900	4	61230	24
20 newsgroups	577	2857	1916	0.6/0.4	6	76807	44
RNCP	927	14413	2957	0.75/0.25	16	4116	7

Table 5: Description of the datasets through key statistics.

RNCP: This dataset contains French training certification contents provided by the public organisation France Compétences. Following Ait-Saada and Nadif (2023), we build it into a classification dataset by taking as textual input the "Intitulé" (title) field, and using the ROME code of each certification (which are linked to thematic topics) for determining the class. However, their split into train and test sets was not made available: hence, while we keep the same $75\%/25\%$ ratio, we chose to work with an updated (and thus larger) version of the dataset.

A.2 Hyperparameter tuning

Hyperparameters and computation of results: Following Ruff et al. (2019), all presented values are obtained by averaging results over 5 runs. For OC-SVM, we present results over the best $\nu \in [0.05, 0.1, 0.2, 0.5]$. The best value of ν is then kept for experiments in section 4 and B.2. For CVDD, we only present the best results obtained over the number of attention heads r. Similarly, the best r are re-used in section section 4 and B.2. All our results are micro-averaged over all classes in the dataset, meaning that we average the values obtained for each model trained on $\mathcal{D}^{c}_{train}, \forall c \in \mathcal{C}$, weighted with $\lvert\mathcal{D}^{c}_{train}\rvert$. The standard deviation values presented are obtained using these averages over 5 different runs.

Choice of r for CVDD: Following (Ruff et al., 2019), we experiment with a large array of values for the number of context vectors r in CVDD. In our results, the best value seems to depend on both the dataset and the representation used, and needs to be tuned according to these two factors. The AUC variations given r on 20 Newsgroups for the 5 representations are presented in Figure 1. The best AUC values for FT$_{Large}$ and PPMI$_{\mathcal{D}}$ are obtained with $r = 3$ and $r = 2$ respectively. On the whole, the best values of r in Tables 2, 1, 3 show that more complex datasets lead CVDD to need more context vectors. Indeed, while the classes of AG News are thematically consistent, those of 20 Newsgroups aggregate several lower-level themes, and the documents in the RNCP classes are also quite diverse (Ait-Saada and Nadif, 2023).

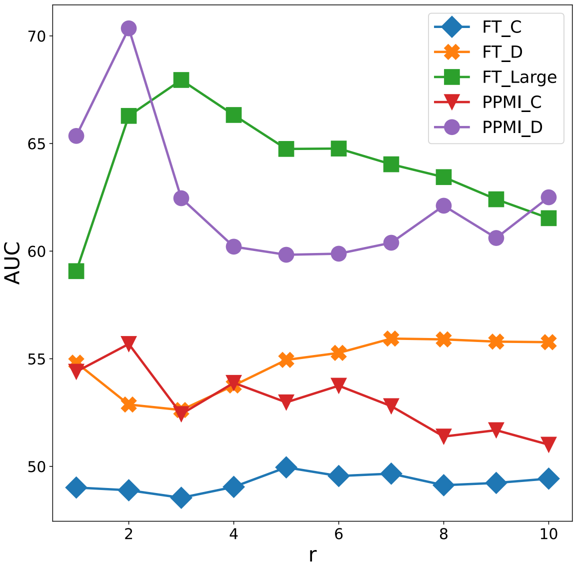

Figure 1: AUCs over 20 Newsgroups for CVDD models trained with our 5 pre-trained word embeddings (FT$_{\mathcal{C}}$, FT$_{\mathcal{D}}$, FT$_{Large}$, PPMI$_{\mathcal{C}}$, PPMI$_{\mathcal{D}}$), depending on the number of attention heads r.

A.3 Choice of OC-SVM Kernel

The RBF kernel is usually the default kernel for OC-SVMs (Manevitz and Yousef, 2002). However, a linear kernel provides here the best results. We can infer that the geometry of the FastText representations is well adapted to our AD task, and that using a more complex kernel makes the model prone to overfitting. In particular, following Ruff et al. (2019), we applied our hyperparameter search to ν only, whereas the γ hyperparameter of the RBF kernel is set automatically through a method proper to scikit-learn, inversely proportional to the variance of the training data. We hypothesize that the surprising counter performance of RBF kernel on class-based representations could be linked to this way of choosing γ. It may also be caused by examples very representative of normal data lying close to the origin in the feature space and being selected in the portion of training data ν allowed to be labelled by the OC-SVM as anomalies during training.

B Additional results

B.1 Evaluation metrics - AUPR

We also compare the performances of the different models using the Area Under Precision Recall curve (AUPR), which is less prevalent: it allows to measure the performance on imbalanced datasets, which is important in AD where the proportion of anomalies can be very low, although their detection matters the most. While this is not the case in our k-classes-out setting, we use this measure for complementary analysis.

Table 6 is an extended version of Table 4, which also includes values of AUPR-i and AUPR-o metrics, which are the AUPR values computed respectively for the inlier and the outlier classes. For this measure, the performances of a random classifier correspond to the number of positive examples divided by the size of the testing set. Thus, the results vary from one dataset to another, not only depending on the performances of the model, but also with the numbers of classes and their sizes. Overall the AUPR scores follow the same trend as the AUC score, except for CVDD which gets a better performance than the other models on the AUPR-i on 20 Newsgroups.

B.2 Unsupervised setting

Setting description: In our *unsupervised* setting, randomly selected documents of other classes are added to the normal class at a specified *contamination* rate r_{cont}. This corresponds to real-case scenarios where the data has not been properly labelled and may be contaminated with anomalies. More formally, this corresponds to using:

$$\mathcal{D}_{cont}^{c_{normal}} = \mathcal{D}_{train}^{c_{normal}} \cup \mathcal{D}_{anom}^{c_{normal}}$$

where $\mathcal{D}_{anom}^{c_{normal}}$ contains examples sampled from $\mathcal{D}_{train} \setminus \mathcal{D}_{train}^{c_{normal}}$ and r_{cont} is the proportion of these samples in $\mathcal{D}_{cont}^{c_{normal}}$. We experiment with several values of r_{cont} to evaluate the models robustness to anomalies in training data, and evaluate with the same $\mathcal{D}_{test}$. For fair comparison, we use the same contaminated datasets $\mathcal{D}_{cont}^{c_{normal}}$ for each model. Again, experiments are repeated by picking every $c \in \mathcal{C}$ to be c_{normal}.

Results: Figure 2 presents the results for several contamination rates r_{cont} corresponding to the proportion of anomalies added to the training set of the normal class $\mathcal{D}_{train}^{c_{normal}}$, for the three datasets. Unsurprisingly, the more the contamination rate rises, the lower the results get. We can notice that on 20 Newsgroups OC-SVM with a linear kernel seems less robust to anomalies in training data than the other methods. However, it still gives the best results. Overall, no particular trend stands out. While results obtained on RNCP decrease less with contamination, they are very unsatisfactory, for all models. We take note that specifically-designed methods based on a priori assumptions on the dataset reach better results (Ait-Saada and Nadif, 2023).

We should note that the results we obtain are, in some settings, notably worse than the ones presented in Manolache et al. (2021), especially on the dataset 20 Newsgroups, although we re-used the implementation provided by the authors and tried our best to reproduce their results following the paper. The discrepancy is particularly high for the OC-SVM and DATE models, while CVDD stays stable.

B.3 OC-SVM with DATE and Electra representations

To better understand the impact of local representations on AD, we experimented using the contextual representations from DATE with an OC-SVM model. These representations are learnt locally on each class of the dataset. To get a document-level representation, we used the $[CLS]$ token. We also experimented using Electra representations learnt locally without the additional RMD task present in DATE. Figure 7 presents the results on the different datasets.

On 20Ng and AGNews, combining DATE representations with OC-SVM shows worse performances than the ones obtained by DATE (with DATE representations) or OC-SVM (with FastText representations) in Table 4. On the RNCP Dataset however, using OC-SVM with DATE representations gets the best results. We hypothesize that the shortness of RNCP documents leads smaller models such as FastText to have more difficulties to extract the relevant information in the representations. However, AD methods specifically designed for short text documents such as the one presented by Ait-Saada and Nadif (2023) still provide the best results.

	AGNews			**20Ng**			**RNCP**		
	AUC	AUPR-i	AUPR-o	AUC	AUPR-i	AUPR-o	AUC	AUPR-i	AUPR-o
OC-SVM + FT$_{Large}$	82.2 ± 0.1	68.1 ± 0.2	90.8 ± 0.04	66.0 ± 0.2	34.8 ± 0.2	86.8 ± 0.1	56.2 ± 0.1	12.6 ± 0.1	91.7 ± 0.04
OC-SVM + *ours*	**89.8** ± 0.01	**75.7** ± 0.1	**96.0** ± 0.01	**81.4** ± 0.1	*44.3* ± 0.2	**94.4** ± 0.1	**63.7** ± 0.05	**14.5** ± 0.03	**93.2** ± 0.01
CVDD + FT$_{Large}$	87.2 ± 0.7	71.6 ± 0.8	94.4 ± 0.4	68.0 ± 0.1	42.5 ± 0.2	86.6 ± 0.03	56.6 ± 0.3	12.8 ± 0.2	91.5 ± 0.1
CVDD + *ours*	86.5 ± 0.5	70.3 ± 1.1	94.2 ± 0.2	70.4 ± 1.6	**45.3** ± 1.8	88.2 ± 0.4	58.3 ± 0.4	12.8 ± 0.2	91.8 ± 0.1
DATE	*88.5* ± 0.2	*73.7* ± 0.6	*95.2* ± 0.1	*70.9* ± 0.4	41.8 ± 0.5	*89.8* ± 0.1	*59.2* ± 0.1	*13.1* ± 0.1	*92.6* ± 0.04

Table 6: AUCs of AD experiments over all datasets, with all models. For OC-SVM and CVDD, we show the best results across hyperparameters with FT$_{Large}$, and across our own word representations.

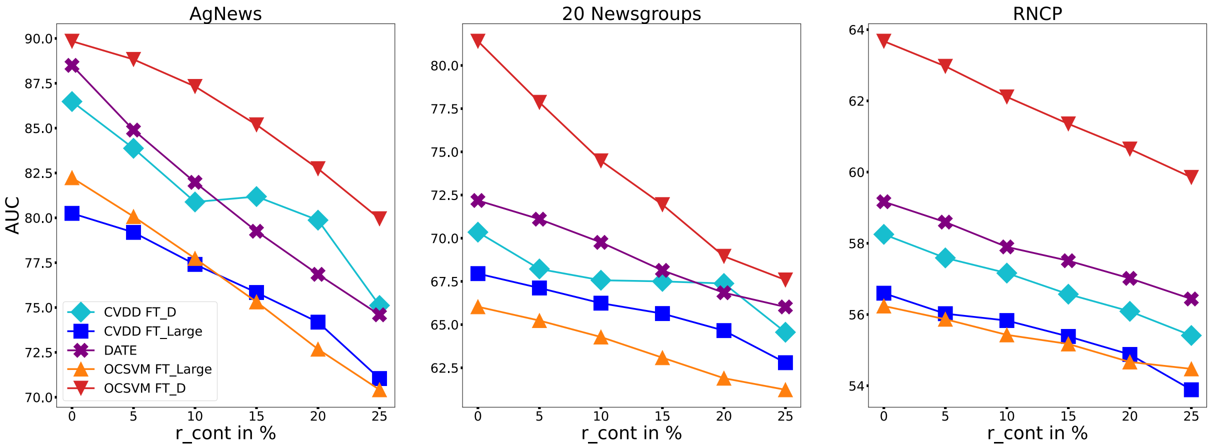

Figure 2: AUCs of AD experiments over AG News, 20 Newsgroups and RNCP, with 5 of the models shown in Table 6, for a contamination rate r_{cont} varying from 0 to 25%.

		AGNews	**20Ng**	**RNCP**
DATE	OC-SVM Linear	73.1	63.3	65.5
	OC-SVM Poly	73.3	63.2	67.7
	OC-SVM Rbf	73.2	63.8	66.7
Electra	OC-SVM Linear	41.5	61.4	59.2
	OC-SVM Poly	40.9	61.3	59.3
	OC-SVM Rbf	40.4	61.4	59.3

Table 7: AUCs of AD experiments over all datasets, with OC-SVM using representations from DATE and Electra learnt on each class of the dataset.

Using locally trained Electra representations combined with OC-SVM gets worse results than using DATE representations. This underlines the contribution of the RMD task introduced by Manolache et al. (2021) for AD. We also experimented on OC-SVM with pre-trained Electra embeddings, but got notably worse results than the ones presented in Table 7. We recall that Ruff et al. (2019) also experimented with BERT representations but found the results to be lacking and did not display them.

B.4 Results detailed by class

20 Ng	**OC-SVM**		**CVDD**		**DATE**
Class	linear - FT$_{\mathcal{D}}$	linear - FT$_{Large}$	FT$_{\mathcal{D}}$ ($r = 2$)	FT$_{Large}$ ($r = 3$)	DATE
$0 - comp$	87.4	78.9	78.0	73.7	87.7
$1 - misc$	86.5	63.8	65.1	74.0	54.8
$2 - pol$	82.7	58.6	76.3	71.4	61.3
$3 - rec$	82.4	63.4	69.1	60.5	66.9
$4 - rel$	83.1	67.2	75.8	77.9	71.0
$5 - sci$	69.4	57.2	55.8	58.2	64.5

Table 8: AUCs of AD experiments over the different classes of 20 Newsgroups dataset, with all models. For OC-SVM and CVDD, we show the best results across hyperparameters with FT$_{Large}$, and across our own word representations.

AG News	**OC-SVM**		**CVDD**		**DATE**
Class	linear - FT$_{\mathcal{D}}$	linear - FT$_{Large}$	FT$_{\mathcal{D}}$ ($r = 2$)	FT$_{Large}$ ($r = 3$)	DATE
$0 - business$	85.2	77.8	83.9	87.9	88.7
$1 - science$	86.3	74.8	80.7	83.4	82.6
$2 - sports$	95.7	92.1	94.7	95.7	94.5
$3 - world$	92.1	83.3	86.5	81.8	88.2

Table 9: AUCs of AD experiments over the different classes of AG News dataset, with all models. For OC-SVM and CVDD, we show the best results across hyperparameters with FT$_{Large}$, and across our own word representations.

RNCP	**OC-SVM**		**CVDD**		**DATE**
Class	linear - FT$_{\mathcal{D}}$	linear - FT$_{Large}$	FT$_{\mathcal{D}}$ ($r = 8$)	FT$_{Large}$ ($r = 12$)	DATE
$1 - environnement$	52.7 6	50.2	53.5	52.8	55.1
$2 - defense$	73.7	51.9	66.2	59.4	38.7
$3 - patrimoine$	63.1	47.5	59.3	55.4	63.5
$4 - economie$	58.7	56.6	53.0	53.8	55.6
$5 - recherche$	65.7	58.4	65.7	65.1	66.5
$6 - nautisme$	57.1	50.9	55.7	54.1	57.7
$7 - aronautique$	68.7	63.5	63.7	62.7	66.4
$8 - scurit$	72.3	65.4	72.1	74.3	57.2
$9 - multimdia$	71.7	62.1	57.6	56.0	60.2
$10 - humanitaire$	61.3	51.8	56.8	54.6	58.3
$11 - nuclaire$	69.4	63.2	62.7	61.2	63.1
$12 - enfance$	81.4	55.5	67.5	56.3	61.7
$13 - saisonnier$	76.7	51.5	70.6	54.8	44.0
$14 - assistance$	65.5	41.5	49.7	38.7	50.1
$15 - sport$	68.1	51.2	56.9	48.3	58.2
$16 - ingnierie$	67.8	62.7	62.2	63.3	65.6

Table 10: AUCs of AD experiments over the different classes of RNCP dataset, with all models. For OC-SVM and CVDD, we show the best results across hyperparameters with FT$_{Large}$, and across our own word representations.