REATO: Outlier Ensemble for Textual Data with Robust Subspace Recovery Autoencoders

Anonymous ACL submission

Abstract

Outlier detection is a recurring challenge in machine learning, actively researched across various domains including computer vision, time series analysis, and high-dimensional data. Recently, the interest in textual outlier detection and textual anomaly detection has blossomed, bringing forth unique challenges. Unfortunately, existing approaches often overlook a critical consideration: the specific type of textual outlier they aim to detect. We found that the experimental protocol of the literature does not identify different kind of textual outliers. To solve this issue, we present a novel approach of textual outlier detection using robust ensemble autoencoders that succeed to retrieve difficult anomalies. To enhance the robustness of our autoencoders, we introduce a novel robust subspace recovery loss function that takes into account the locality in the latent space. Our ensemble learning strategy involves randomly connected autoencoders. Additionnaly, we address the issue of limited corpus availability by preparing two types of outliers: independent and contextual. An intriguing aspect of our work is the distinction between these two outlier types, which we formalize and demonstrate to be fundamentally different to handle within a corpus. Notably, our approach not only delivers competitive results when compared to existing methods but also excels in handling contextual outliers.

1 Introduction

Outlier Detection (OD) (Hawkins, 1980; Hodge and Austin, 2004; Zhang, 2013; Aggarwal, 2017) is the task that aims to estimate whether an observation is normal or not. Depending of the data or of the case study, this task is similar to the task of Anomaly Detection (AD) (Chandola et al., 2009; Ruff et al., 2021). Several works have been devoted to study and categorise the different characteristics of an outlier, and there exists three principal kinds of outliers: independent, contextual and collective.

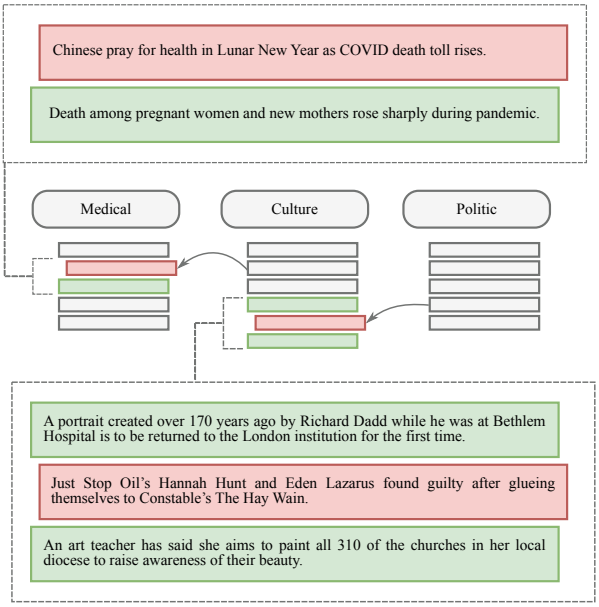


Figure 1: Presentation of the studied problem with three documents topics: medical, culture and politic. Under each topic we represent a textual document with colored rectangles. Gray and green are inliers and red ones are outliers. The detailed documents are the abstract of the news articles taken from sources like Reuters, New York times, BBC, ... The first scenario is the apparition of a culture-related document in a medical feed, and the second scenario is a political document in the culture feed.

The type of outlier is important because the lack of dedicated dataset often leads to prepare an artificial contamination. With the emergence of new machine learning methods and the availability of many datasets and corpora, outlier detection can be addressed through many approaches. In most cases, the models for this task are based on one-class classification (OCC) (Khan and Madden, 2014; Ruff et al., 2018; Sohn et al., 2021).

Performing outlier detection on textual data is less common than many other types of data (image, time series and medical) but it comes with several useful applications that helps discerning

wrong web content, hateful message, spam or also errors in news feed. The difficulty to reproduce experimental protocols and results from the literature is one of the reason of the unpopularity of the task with text. Indeed, there is a great difference between tackling independent outliers and contextual outliers (Mahapatra et al., 2012; Fouché et al., 2020) using semantic in text. For the former, the classifier needs to differentiate two kinds of documents that come from unrelated topics (sports and computer) but the for the latter, one topic is contaminated with another "sibling" topic. The Figure 1 describes such scenario. Most of the recent works are contaminating corpora without addressing the problem of which kind of anomaly/outlier is added (Manevitz and Yousef, 2001; Kannan et al., 2017; Ruff et al., 2019; Lai et al., 2020).

Recent advances in word embedding with language models like GloVe (Pennington et al., 2014), Fast-Text (Bojanowski et al., 2017), BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) have shown promising characteristics for outlier detection. Only few methods of the literature propose their usage (Manolache et al., 2021; Ruff et al., 2019). Other methods like One-Class Support Vector Machine (OCSVM) (Schölkopf et al., 2001) and Textual Outlier using Nonnegative Matrix Factorization (TONMF) (Kannan et al., 2017) rely on tf-idf. On the other hand, recent methods are not using outlier ensemble methods (Aggarwal and Sathe; Zhao et al., 2019a; Zimek et al., 2014) for performing outlier detection with text data. Additionnaly, AutoEncoders (AE) have been used for anomaly/outlier detection with high-dimensional data (Chen et al., 2017; Kieu et al., 2019) and are also successful with other kind of data (An and Cho, 2015; Chen et al., 2018; Lai et al., 2020; Zhou and Paffenroth, 2017), but the risk of using autoencoders with language models leads to the apparition of degenerate solution in the learning step. Robust properties are needed in such scenario.

We propose in this work a novel outlier ensemble method that performs outlier detection on text using word embedding and a Robust Subspace Recovery (RSR) (Lerman and Maunu, 2018; Rahmani and Atia, 2017) layer. The autoencoder use the RSR layer for mapping the normal distribution in a subspace where outliers are at the edge (Lai et al., 2020). Our method, called Robust subspace recovery Autoencoder ensemble for Text Outlier (REATO), build a RSRAE ensemble whose are randomly connected. RATO can also be seen as an ensemble of several subspace that aims to find normal data with different *manifold*. In short, such learning method are making the hypothesis that the distribution is highly contaminated and the inliers (normal data) lie in a low-dimensional subspace. The performance of REATO are experimented against other state of the art methods on a total of eight corpora. We are proposing a definition of two different outliers that can be applied on available corpora and REATO outperforms the literature with more robust results.

2 REATO: Robust subspace recovery ensemble autoencoder for text outliers

This section presents our approach, REATO, and the description of its properties. While robust subspace recovery autoencoders have successfully tackle anomaly detection with text, they lack locality and geometry awareness for mitigating manifold collapse in transformer-based language models. For this reason we introduce Robust subspace recovery Ensemble Autoencoder for Text Outliers (REATO) which integrates locality in the latent representation through locally linear embedding technique.

The section is structured with a presentation of the randomly connected autoencoders, followed by a presentation of RSR loss. We then introduce the locally linear embedding loss term of REATO before presenting its ensemble method. Finally, we present the representation of text.

2.1 Randomly Connected One-Class Autoencoder

Instead of using fully connected autoencoders, we propose to use randomly connected autoencoders. In the case of RSRAE, it is a novel approach and allow us to build ensemble autoencoders with different base detectors.

Let X be a dataset of N instances such as $\mathrm{X} = \{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{N}\}$. Each instance has D dimension which correspond to its attributes: $\mathrm{x_i} = \{x_1, ..., x_D\}$. An Autoencoder is a neural networks in which the encoder $\mathcal{E}$ maps an instance $\mathrm{x_i}$ in a latent representation noted $\mathrm{z}_i = \mathcal{E}(\mathrm{x}_i) \in \mathbb{R}^e$ of dimension e. The RSR layer is a linear transformation $\mathbf{A} \in \mathbb{R}^{d \times e}$ that reduces the dimension to d. We note $\hat{\mathrm{z}}_\mathrm{i}$ the representation of $\mathrm{z_i}$ through the RSR layer, such as $\hat{\mathrm{z}}_\mathrm{i} = \mathbf{A}\mathrm{z_i} \in \mathbb{R}^d$ The decoder $\mathcal{D}$ maps $\hat{\mathrm{z}}_\mathrm{i}$ to $\hat{\mathrm{x}}_\mathrm{i}$ in the original space D. The matrix

$\mathbf{A}$ and the parameters of $\mathcal{E}$ and $\mathcal{D}$ are obtained with the minimization of a loss function.

Similarly to Chen et al. (2017) we introduce autoencoders with random connection such as we increase the variance of our model. In the autoencoders ensemble each autoencoder has a random probability of having several of its connections to be cut. Thus, we setup the probability disconnection with a random rate between $[0.2, 0.5]$.

2.2 Robust Subspace Recovery Layer

We present the RSR AutoEncoder that aims to robustly and nonlinearly reduce the dimension of the original data (Lerman and Maunu, 2018). The RSR layer maps the inliers around their original locations and the outliers far from their original locations. The loss function minimizes the sum of the autoencoder loss function noted L_{AE} with the RSR loss function noted L_{RSR}.

$$L^p_{AE}(\mathcal{E}, \mathbf{A}, \mathcal{D}) = \sum_{i=1}^{N} ||\mathrm{x_i} - \hat{\mathrm{x}}_\mathrm{i}||_2^p \qquad (1)$$

which is the $l_{2,p} - norm$ based loss function for $p > 0$.

For performing the subspace recovery, we denote two terms that have different roles in the minimization process. The first term enforces the RSR layer to be robust (PCA estimation) and the second enforces the projection to be orthogonal:

$$\begin{aligned} L^q_{RSR}(\mathbf{A}) =& \lambda_1 \sum_{i=1}^{N} ||\mathrm{z_i} - \mathbf{A}^\mathrm{T}\hat{\mathrm{z}}_\mathrm{i}||_2^q \\ &+ \lambda_2 \sum_{i=1}^{N} ||\mathbf{A}\mathbf{A}^\top - \mathbf{I}_d||_f^q \end{aligned} \qquad (2)$$

with $\mathbf{A}^\top$ the transpose of $\mathbf{A}$, $\mathbf{I}_d$ the $d \times d$ matrix and $||\cdot||_f$ the Frobenius norm. λ_1 and λ_2 are hyperparameters and $q = 1$ is corresponding to the optimal $l_{p,q}$ norm (Maunu et al., 2019). If we simplify Equation 2 we have:

$$\begin{aligned} L_{RSRAE}(\mathcal{E}, \mathbf{A}, \mathcal{D}) =& \lambda_1 L^1_{AE}(\mathcal{E}, \mathbf{A}, \mathcal{D}) \\ &+ \lambda_2 L^1_{RSR}(\mathbf{A}) \end{aligned} \qquad (3)$$

2.3 Locally linear embedding term

Locally Linear Embedding (LLE) (Roweis and Saul, 2000; Chen and Liu, 2011) is a popular nonlinear dimensionality reduction technique that aims to preserve the local geometry of the data in a lower-dimensional subspace. It is based on the assumption that data points in a local neighborhood can be linearly represented by their neighboring data points. The LLE term in the loss function encourages the autoencoder to learn representations that preserve the relationships between data points in their local neighborhoods. By doing so, it helps to project the Euclidean distance with its neighbors in the learned subspace. Based on Equation 2, the reconstruction loss function of RSRAE enforces robustness with L^1_{AE} and the orthogonality with L^1_{RSR}. Because the learned representation of the encoder is compressed in a e dimension space, the locality of the subspace is not handled.

For tackling this problem, we propose to introduce a third term to L_{RSRAE} based on locally linear embedding. Given a set of data points $\{\mathrm{x_i}\}_{i=1}^N$ in the input space, the goal of LLE is to find a lower-dimensional representation $\{\mathrm{z_i}\}_{i=1}^N$ in the output space (the subspace learned by the autoencoder) such that the local relationships between data points are preserved. We note:

$$L_{LLE}(\mathbf{A}) = \sum_{i=1}^{N} \left\| \mathrm{x_i} - \sum_{j \in \mathcal{N}_i} w_{ij} \mathrm{x_j} \right\|_2^2 \qquad (4)$$

where $\mathcal{N}_i$ represents the set of indices of the k-nearest neighbors of $\mathrm{x_i}$ (excluding $\mathrm{x_i}$ itself) and w_{ij} are the weights assigned to the neighboring data point $\mathrm{x_j}$ in the linear reconstruction of $\mathrm{x_i}$. The weights w_{ij} can be computed using the least squares method to minimize the reconstruction error: $\min_{\mathbf{w_i}} \left\| \mathrm{x_i} - \sum_{j \in \mathcal{N}_i} w_{ij} \mathrm{x_j} \right\|_2^2$ subject to the constraint $\sum_{j \in \mathcal{N}_i} w_{ij} = 1$.

The LLE term encourages the autoencoder to find a representation for each data point as a linear combination of its k-nearest neighbors in the input space. By minimizing the LLE term in the loss function, the autoencoder learns to preserve the local linear relationships, which ultimately helps to project the Euclidean distance with its neighbors in the learned subspace. The reconstruction errors of LLE is measured by the cost function:

$$L_{LLE}(\mathbf{A}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} w_{ij} \|\mathbf{A}\mathrm{x_i} - \mathbf{A}\mathrm{x_j}\|_2^2 \qquad (5)$$

The weight w_j assigned to the neighbor $\mathrm{x_{ij}}$ in the local linear reconstruction of $\mathrm{x_i}$ are determined based on the distance between data points and their neighbors. The inclusion of the LLE term in the loss function encourages the autoencoder to preserve the local geometric structure of the data in the learned subspace.

Algorithm 1 CTO: Contextual Contamination for Topic-level Outliers

Require: Inlier topic ζ, corpus X, split size l, contamination rate ν
Ensure: $0 < l \leq N$
  $c \leftarrow l\nu$
  $i \leftarrow 0$
  Initialize empty matrix Z
  $\mathcal{O} \leftarrow \{\mathrm{x}_j \times \mathrm{y}_j \in \mathrm{X} \times \mathrm{Y} | \forall j \in [0, N], \mathrm{y}_j \neq \zeta\}$
  ▷ Outlier Matrix
  $\mathrm{X}_\zeta \leftarrow \{\mathrm{X} \backslash \mathcal{O}\}$ ▷ Inlier Matrix
  while $|\mathrm{Z}| < c$ **do**
    if Parent$(\mathrm{y}_i) \neq$ Parent(ζ) **then**
      Append$(\mathrm{x}_i, \mathrm{y}_i)$ to Z
    end if
    $i \leftarrow i + 1$
  end while
  Fill Z with X_ζ until $|\mathrm{Z}| = l$
  return Shuffle(Z)

Finally, the REATO cost function is measured as follows:

$$L_{REATO}(\mathcal{E}, \mathbf{A}, \mathcal{D}) = L_{RSRAE}(\mathcal{E}, \mathbf{A}, \mathcal{D}) + \lambda_3 L_{LLE}(\mathbf{A}) \quad (6)$$

The parameter λ_3 controls the influence of the LLE term on the overall loss. Because it controls the influence of locality of the manifold, the term is preferred to be low for avoiding collapsing results.

2.4 Ensemble Learning

The main idea behind ensemble methods is that a combination of several models, also called *base detectors*, and their outputs is more robust than usage of a single model. Such robustness can be observed against the bias-variance tradeoff and also for tackling the issue of overfitting. Although the possibility to combine multiple base detectors is intuitive, the design of such approaches needs special attention regarding normalization of outputs. In REATO, we use the RSR reconstruction error of each autoencoders and then we normalise each base detector scores through the standard deviation of one unit. We then take the median value for each observation.

2.5 Text Representation

In our REATO approach, we use RoBERTA (Liu et al., 2019) for text representation instead of GloVe, FastText or TFIDF. Ruff et al. (2019); Manolache et al. (2021) have recorded their results on these language model, in addition of BERT, but with meticulous observation of the results of RoBERTA is a top performing representation. The REATO model is not based on the self-attention mechanism, such as for Ruff et al. (2019); Manolache et al. (2021), and we propose to use the implementation of Reimers and Gurevych (2019). Similarly to Ruff et al. (2019), we do not find substantial difference between glove and BERT performances.

Dataset	Task	Hierarchy
20 Newsgroups	Classification	Yes
DBpedia 14	Classification	Yes
Reuters-21578	Classification	No
Web of Science	Classification	Yes
Enron	Spam	No
SMS Spam	Spam	No
IMDB	Sentiment	No
SST2	Sentiment	No

Table 1: Presentation of datasets from the literature on outlier detection and the inherent tasks. We describe these corpora by indicating the existence of a topic hierarchy in the labels of the original corpus.

3 Different types of outlier

A large number of contributions have settled anomaly/outlier taxonomies (Hawkins, 1980; Hodge and Austin, 2004; Zhang, 2013; Aggarwal, 2017; Ruff et al., 2019) several types of outliers have been proposed in the literature: Point outlier, Conditional/Contextual outlier and Collective/Group outlier. A similar taxonomy can be applied to textual data. Consequently, various types of outliers frequently coexist within the documents of a given corpus. The definition of a topic can be assimilated to the subject matter that a document addresses. Depending on the document type, there may be multiple subtopics within a broader category (e.g., a sports topic that encompasses football and tennis). Thus, accounting for this hierarchical structure introduces a form of contextual outlier. These contextual outliers may appear unremarkable in isolation but are considered outliers when associated with a small subset of the corpus.

Collective outliers pose challenges in terms of formalization due to the contextual nature of textual data. To illustrate, consider a legal document mentioning a football player, which would be anoma-

lous if incorrectly appearing in a sports-related corpus. Point outliers represent observations that lack any meaningful relationship with other topics. Specifically, outlier topics and inlier topics have different hierarchical parents within the category structure. Let a labelled document of a corpus $(\mathrm{x}, y) \in \mathrm{X} \times \mathcal{Y}$ and ζ be the inlier category, and its corresponding subset $\mathrm{X}_\zeta \subseteq \mathrm{X}$. We define $\mathcal{O}$ the subset of all outliers such as $\mathcal{O} \subset \mathrm{X}$. We have:

$$\mathcal{O} = \mathrm{X} \backslash \mathrm{X}_\zeta \tag{7}$$

Regarding $\mathcal{O}$, we can make the distinction with two different constraints. We note $P(y)$ the direct parent of y in a given hierarchy. First, an observation $\mathrm{x_i}$ is considered to be an outlier if its parent topic is different of inlier parent topics such as:

$$\mathcal{O}_p(\zeta) = \{\mathrm{P}(\zeta) \neq \mathrm{P}(\mathrm{y}) | (\mathrm{o}, \mathrm{y}) \in \mathcal{O} \times \mathrm{Y}\} \tag{8}$$

The second constraint is corresponding to documents that do not lie in X_ζ but share the same parent topic as ζ. These observations are identified as another kind of outlier: contextual outliers. We write:

$$\mathcal{O}_c(\zeta) = \{\mathrm{P}(\zeta) = \mathrm{P}(\mathrm{y}) | (\mathrm{o}, \mathrm{y}) \in \mathcal{O} \times \mathrm{Y}, \mathcal{O} \backslash \mathcal{O}_p\} \tag{9}$$

4 Experiments

In this section we present conducted experiments on both independent outliers and contextual outliers. We present corpora and how CTO can be applied. A comparison of the model scores is proposed, highlighting the robustness of REATO.

4.1 Data

Although there are dedicated datasets for outlier detection, such as ODDS or UCI, they mainly provide multidimensional data, time series and computer vision data. Applications such as spam detection and text classification have a rich set of corpora available. Recent work (Lai et al., 2020; Ruff et al., 2019; Kannan et al., 2017; Mahapatra et al., 2012) uses classification datasets such as Reuters-21578[1] and 20 Newsgroups[2] with a dedicated preparation in order to compare their approaches.

We use the corpus presented in the section and for each available category, we apply the preparation of independent outliers and contextual outliers with CTO (Algorithm 1). For reasons of fairness between each method and each dataset, we set the size of the preparation subset to 350 and the results are averaged over 10 of runs. The data is pre-processed by removing lower case and stop words. The train part of each corpus is used for training and evaluation. The tfidf model is applied to the whole corpus and only tokens that appear at least three times are kept in the vocabulary. In a first step, we set $\nu = 0.10$.

20 Newsgroups For 20 Newsgroups we separate the subtopics into seven main topics: computer, forsale, motors, politics, religion, science, sports. We do not count the topic forsale for contextual outliers because it has no subtopics.

Reuters-21578 The Reuters-21578 corpus contains documents associated with several topics. We delete all these documents to keep only those associated with a single topic. We reorganise the topics in order to obtain a hierarchy, based on the work of Toutanova et al. (2001). Thus, four parent themes are created: commodities, finance, metals and energy. We apply GenTO to the eight topics that have the greatest number of training documents.

DBpedia 14 For DBpedia 14 we create the topic hierarchy based on the ontology provided[3] and has six parent topics.

Web Of Science Web of Science is often used as a reference for hierarchical classification and provides three levels of topic hierarchy. The third level topics are distributed among the corresponding first level parents. Thus, seven parent topics are present and for child topics that are associated with more than one parent, we keep the largest set of children and delete the others.

4.2 Setup

We use CTO for preparing contextual contamination on each candidate inliers possible with $\nu = 0.1$ (CTO1) and a split size of 350. All results are performed on AUROC and AUPRC reference works from the previous Section. We integrate results of one-class autoencoder and we also benchmark results on a randomly connected autoencoder ensemble (RAE) (Chen et al., 2017). The architecture

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

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

[3] mappings.dbpedia.org/server/ontology/classes/

Model	Newsgroups	Reuters	WOS	DBpedia 14	Enron	SMS Spam	IMDB	SST2
OCSVM	0.948	0.917	0.981	0.993	0.723	0.693	0.539	0.575
PCC	**0.952**	0.938	0.982	0.992	0.724	0.685	0.542	0.576
OC-AE	0.697	0.732	0.856	0.837	0.592	0.514	0.517	0.499
RSRAE	0.949	0.940	0.982	0.994	0.731	0.704	0.540	0.577
REA	0.884	0.704	0.935	0.918	0.636	0.553	0.665	0.614
REATO	0.949	**0.953**	**0.989**	**0.991**	**0.749**	**0.898**	**0.704**	**0.627**

Table 2: Results of state of the art models for independent outliers with the contamination rate $\nu = 0.10$. Area under ROC (AUROC) is the evaluation metric. The experimental study is performed on Distill RoBERTA. Each result is performed on test split prepared through Algorithm 1.

Model	Newsgroups		Reuters		WOS		DBpedia 14	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
OCSVM	0.282	0.750	0.491	0.811	0.599	0.889	0.759	0.945
PCC	0.314	0.776	0.518	0.828	0.613	0.897	0.771	**0.954**
OC-AE	0.191	0.623	0.246	0.604	0.249	0.680	0.348	0.735
RSRAE	0.309	0.779	0.506	0.821	0.621	0.900	0.762	0.936
REA	0.194	0.623	0.278	0.615	0.448	0.810	0.368	0.747
REATO	**0.362**	**0.793**	**0.538**	**0.880**	**0.687**	**0.921**	**0.840**	0.951

Table 3: Results of state of the art models for contextual outliers with contamination rate $\nu = 0.10$. Average precision (AUPRC) and Area under ROC (AUROC) are evaluation metric.The experimental study is performed on Distill RoBERTA. Each result is performed on test split prepared through Algorithm 1.

is similar to Chen et al. (2017) and the autoencoders are following their settings. The same goes for our approach REATO that follows the setup of Lai et al. (2020). We also keep the number of runs for each corpus to 10.

For REATO and RAE we setup similarly than with the autoencoder and we setup the number of base predictors to 50. We provide the code of our approach[4] using the PyOD base implementation(Zhao et al., 2019b). Additionally, we also set hyperparameters $\lambda_1 = 0.1$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.05$. For avoiding manifold collapse problem and degenerates solutions, we advise that $\lambda_3 < \lambda_1$. On the other hand, we set the epoch number to 50 and random connection probability between $[0.2, 0.5]$.

We propose to compare our approach against OCSVM(Schölkopf et al., 2001), PCC(Shyu et al., 2003), a simple one-class autoencoder (OC-AE) and RSRAE(Lai et al., 2020). For OCSVM, PCC and OC-AE we use the implementation from PyOD(Zhao et al., 2019b) and for RSRAE we use their implementation. We rigorously follow the guidelines provided by Lai et al. (2020).

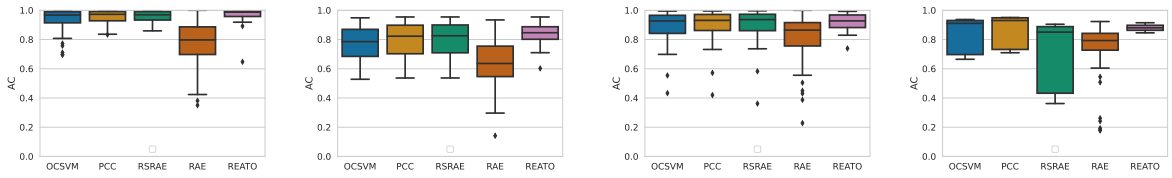
Alternatively, we propose to use a variant of Algorithm 1 considering $P(y) = P(\zeta)$ for independent contamination. Also we propose to benchmark our results on corpora presented in Table 1.

[4] anonymous

4.3 Results

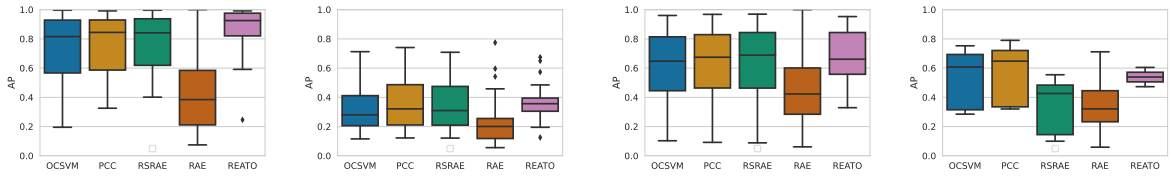
We propose to present our results on three principal points: independent contamination, contextual contamination and robustness of model scores. Table 2 displays results on CTO1 (contamination of 0.1) and independent contamination. We can see that REATO is out performing all models except on 20 Newsgroups. While the results are similar, our approach is notably standing over the others on SMS Spam and IMDB corpora. Thus, our ensemble method presents success for semantic related, spam and sentiment corpora.

Table 3 displays the experimental results conducted with our approach REATO. We observe that our approach is outperforming others model with AUROC metric and AUPRC metric. We can see that usage of REATO allow to mitigate unstable decision of the original RSRAE. We can also see significant difference of performance with Web of Science corpus and Reuters-21578. PCC is the only approach that succeeds to beat our approach against AUROC metric of DBpedia 14. Additionally, we can observe that the original one-class autoencoder highly benefit from randomly connection and en-



(a) DBpedia 14 (b) 20 Newsgroups (c) Web of Science (d) Reuters-21578

Figure 2: Results of our experimental study with $\nu = 0.1$, split size of 350 and number of base detector of 25. The performance metric is AUROC (AC) and the text representation is RoBERTA.



(a) DBpedia 14 (b) 20 Newsgroups (c) Web of Science (d) Reuters-21578

Figure 3: Boxplots of results of our experimental study with $\nu = 0.1$, split size of 350 and number of base detector of 25. The performance metric is AUPRC (AP) and the text representation is RoBERTA.

semble technique, as it close the gap with other models.

While our performances are competitive, the principal purpose of tackling outlier detection with ensemble methods is to mitigate the bias-variance tradeoff. We propose to compare the model results with boxplots, similarly to the previous chapter. The main objective of our contribution is to robust outlier scores for contextual outliers with text. The Figure 2 and the Figure 3 displays an outperforming results from our approach. We can see that the variance of our model is noticeable as the box variance are always smaller than its competitors. Also, the min and max possible scores are close from the median scores, concluding to see that our approach is more efficient, more robust and can handle well language model like RoBERTA.

5 Conclusion

In this work we have introduced REATO, an ensemble approach with RSR a autoencoders, otpimzed through LLE for tackling contextual outlier in text. One perspective is to study the integration of attention head for mitigating the black box problem of our model. It is common, recently, to display text with their corresponding temperature, thanks to recent language model based on transformers. The representation of text is a key concept that we want to investigate in the near future. Our approach has proven state of the art results and a great robustness against two kinds of outliers and with a small amount of available documents. Furthermore, we have displayed that reference contributions have not put sufficient effort to the contamination process in their protocol. One promising perspective is to propose an unsupervised approach for generating different kinds of outliers. Also, our work mainly focuses on the semantic structure of text but syntax is also a promising direction.

6 Ethical statement

Our approach considers contamination of corpora with a robust approach. Our research can leads to generation of malicious contamination in news feeds and procution of fake news. With such risks, our approach can also provide a response to the detection of malicious contamination. As stated in the results section, our approach succeeds to determine spam and tone-like documents.

References

Charu C. Aggarwal. 2017. Outlier Detection in Categorical, Text, and Mixed Attribute Data. In *Outlier Analysis*, pages 249–272. Springer International Publishing, Cham.

Charu C. Aggarwal and Saket Sathe. Theoretical foundations and algorithms for outlier ensembles. 17(1):24–47.

Jinwon An and Sungzoon Cho. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Jing Chen and Yang Liu. 2011. Locally linear embedding: a survey. *Artificial Intelligence Review*, 36:29–48.

Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. 2017. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 90–98. SIAM.

Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. 2018. Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*, pages 1–5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Fouché, Yu Meng, Fang Guo, Honglei Zhuang, Klemens Böhm, and Jiawei Han. 2020. Mining text outliers in document directories. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 152–161. IEEE.

D. M Hawkins. 1980. *Identification of Outliers*. Springer Netherlands, Dordrecht.

Victoria Hodge and Jim Austin. 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126.

Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. 2017. Outlier Detection for Text Data. *SDM International Conference on Data Mining*, 17:489–497.

Shehroz S. Khan and Michael G. Madden. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374.

Tung Kieu, Bin Yang, Chenjuan Guo, and Christian S. Jensen. 2019. Outlier detection for time series with recurrent autoencoder ensembles. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2725–2732. International Joint Conferences on Artificial Intelligence Organization.

Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. 2020. Robust Subspace Recovery Layer for Unsupervised Anomaly Detection. *ICLR International Conference on Learning Representations*.

Gilad Lerman and Tyler Maunu. 2018. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Amogh Mahapatra, Nisheeth Srivastava, and Jaideep Srivastava. 2012. Contextual Anomaly Detection in Text Data. *Algorithms*, 5(4):469–489.

Larry M. Manevitz and Malik Yousef. 2001. One-Class SVMs for Document Classification. *Journal of Machine Learning Research*, 2:139–154.

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.

Tyler Maunu, Teng Zhang, and Gilad Lerman. 2019. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Mostafa Rahmani and George K. Atia. 2017. Randomized robust subspace recovery and outlier detection for high dimensional data matrices. *IEEE Transactions on Signal Processing*, 65(6):1580–1594.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire 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*, 109(5):756–795.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*, 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.

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.

Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. 2021. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, volume 9.

Kristina Toutanova, Francine Chen, Kris Popat, and Thomas Hofmann. 2001. Text classification in a hierarchical mixture model for small training sets. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 105–113.

Ji Zhang. 2013. Advancements of Outlier Detection: A Survey. *ICST Transactions on Scalable Information Systems*, 13(1).

Yue Zhao, Zain Nasrullah, Maciej K Hryniewicki, and Zheng Li. 2019a. Lscp: Locally selective combination in parallel outlier ensembles. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 585–593. SIAM.

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

Chong Zhou and Randy C. Paffenroth. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 665–674, New York, NY, USA. Association for Computing Machinery.

Arthur Zimek, Ricardo JGB Campello, and Jörg Sander. 2014. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter*, 15(1):11–22.

A Example Appendix

This is an appendix.