

DEEP ANOMALY DETECTION IN TEXT



Andrei-Marian Manolache

coordinated by
Conf. Dr. Bogdan Alexe

Faculty of Mathematics and Computer Science
University of Bucharest

Andrei-Marian Manolache: *Deep Anomaly Detection in Text* MSc thesis, June 2021.

 andrei_mano@outlook.com

This MSc thesis was written with L^AT_EX using ArsClassica, a reworking of the ClassicThesis style designed by André Miede, inspired to the masterpiece *The Elements of Typographic Style* by Robert Bringhurst. This template was provided by Florin-Radu Gogianu.

ABSTRACT

Deep anomaly detection methods have become increasingly popular in recent years, with methods like Stacked Autoencoders , Variational Autoencoders and Generative Adversarial Networks greatly improving the state-of-the-art. Other methods rely on augmenting classical models (such as the One-Class Support Vector Machine), by learning an appropriate kernel function using Neural Networks. Recent developments in representation learning by self-supervision are proving to be very beneficial in the context of anomaly detection. Inspired by the advancements in anomaly detection using self-supervised learning in the field of computer vision, this thesis aims at developing a method for detecting anomalies by exploiting “pretext tasks” tailored for text corpora. This approach greatly improves the state-of-the-art on two datasets - 20Newsgroups and AG News, for both semi-supervised and unsupervised anomaly detection, thus proving the potential for self-supervised anomaly detectors in the field of natural language processing.

Anomaly Detection (AD)
Autoencoder (AE)
Generative Adversarial Network (GAN)
One-Class Support Vector Machine (OC-SVM)
Neural Network (NN)
Self-Supervised Learning (SSL)
Computer Vision (CV)
Natural Language Processing (NLP)

The truth is rarely pure and never simple. — Oscar Wilde

ACKNOWLEDGEMENTS

This thesis is based on the research performed by the author over the last years at Bitdefender's Theoretical Research laboratory. I am extremely grateful for the fruitful discussions I've had with my laboratory colleagues and for the guidance from various professors at the University of Bucharest during my Master's studies.

In particular, I would like to thank to my thesis advisor, professor Bogdan Alexe, for the provided supervision, counseling and for the excellent courses held at the Faculty of Mathematics and Computer Sciences.

The thesis is based on the paper ([Manolache et al., 2021](#)), therefore I would like to thank my co-authors Florin Brad and Elena Burceanu for their contributions, support, and inexhaustible patience.

I would also like to thank to professor Marius Popescu for his insights and advises on the problem of authorship detection, and to Sergiu Nisioi for the suggestions regarding further developments.

Parts of this work have been supported by UEFISCDI, under Project PN-III-P2-2.1-PTE-2019-0532.

CONTENTS

1	PROLOGUE	1
1.1	Introduction	1
1.2	Contribution	2
1.3	Outline	2
2	REPRESENTATION LEARNING	3
2.1	Self-Supervised Learning	4
2.2	Word Embeddings	6
2.3	Transformers	8
2.3.1	BERT	10
2.3.2	ELECTRA	11
3	ANOMALY DETECTION	13
3.1	Classical Models	14
3.1.1	One-Class Support Vector Machine	14
3.1.2	Isolation Forest	16
3.2	Deep Models	17
3.2.1	E ³ Outlier	17
3.2.2	Context Vector Data Description	20
4	DATE	23
4.1	Pretext tasks	23
4.1.1	DATE Architecture	24
4.1.2	Anomaly Detection score	26
4.2	Experimental analysis	27
4.2.1	Experimental setup	27
4.2.2	Ablation studies	28
4.2.3	Comparison with other AD methods	30
4.2.4	Unsupervised AD	31
4.2.5	Qualitative results	32
5	CONCLUSION AND FURTHER DEVELOPMENTS	33
	BIBLIOGRAPHY	35
	Appendices	49
A	HYPERPARAMETERS AND IMPLEMENTATION	51
A.1	Disjoint patterns analysis	51
A.2	Model implementation	52
B	OTHER RESULTS	53
B.1	More qualitative and quantitative results on AG News	53
B.2	Authorship detection	54
B.3	Music genre detection	55
B.4	Native language detection	57

1

PROLOGUE

1.1 INTRODUCTION

ANOMALY DETECTION is the task of detecting data points that are deviating from the expected data distribution (inliers). Such a data point is said to be an *outlier*, *abnormality*, or *deviant*. Hawkins (D. M. Hawkins, 1980) notably defined an anomaly as follows:

inlier, anomaly, outlier, abnormality, deviant

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

The early work on anomaly detection was historically produced in the statistics community, but the rapid growth of available data banks and the development of computer hardware and software solutions has led to increasing amounts of interest from Data Scientists and the broader Computer Science community.

The interest of researchers in outliers span several decades (Aggarwal and Reddy, 2014; Chandola et al., 2009; Knorr and Ng, 1997), some of the first formal mentions about "discordant observations" going back to the 19th century (Edgeworth, 1887). Efforts for developing Anomaly Detection methods have proven very fruitful, with applications in credit card fraud detection (Dorronsoro et al., 1997), network monitoring (B. Radford et al., 2018; Sharafaldin et al., 2018), intrusion detection systems (Banerjee et al., 2017), time series (Braei and Wagner, 2020), medical imaging (Wei et al., 2018), and manufacturing (Kammerer et al., 2019).

discordant

Most anomaly detection systems are usually producing an *anomaly score* for a given data sample. Such a score is a very general form of output and can be used to asset the "abnormality degree" of the sample (see Fig. 1). Such a scoring method is desirable since an anomaly highly depends on the intended use-case and the tolerance to false-positives that the system should have. For example, we would not want an automatic fraud detector used in a banking system to flag unusual (but otherwise benign) transactions as being fraudulent. Having a granular scoring scheme enables us to possibly forward such a transaction to further evaluation by some human in the loop.

anomaly score



Figure 1: The anomaly spectrum. Noisy data can be regarded as "weakly" anomalous. (figure adapted from: Aggarwal and Reddy, 2014)

1.2 CONTRIBUTION

Transformer, DATE

Replaced Mask Detection (RMD), Replaced Token Detection (RTD)

Pseudo Label (PL)

In this thesis, we present various architectures for anomaly detection in text. Particularly, we introduce a novel deep learning method based on the Transformer architecture, called DATE ([Manolache et al., 2021](#)). To our knowledge, DATE is the first end-to-end deep anomaly detection method on text that is using self-supervision to produce an anomaly score. Our contributions are summarized below:

- We introduce a sequence-level self-supervision task called *Replaced Mask Detection* that, when used in conjunction with the Replaced Token Detection task ([Clark, Luong, Q. V. Le, et al., 2020](#)), improves the anomaly detection capabilities of our model and stabilizes the training. RMD is used to distinguish between text that is corrupted in different ways.
- We formulate an efficient Pseudo Label score for text anomalies. We remove the need of doing inference on text corrupted in multiple ways ([S. Wang et al., 2019](#)) by looking at the individual tokens probabilities. The token-level PL score significantly speeds up inference times and makes the results interpretable.
- We outperform existing state-of-the-art semi-supervised text AD models on two datasets: 20Newsgroups (+4.7% AUROC) and AG News (+6.9% AUROC). More so, when doing experiments in the unsupervised AD settings, DATE surpasses all other methods trained with 0% contamination, even at 10% train data contamination rate.
- We prove that anomaly detectors in text can be used for authorship detection by training the models on the individual authors' text corpus and treating other texts as anomalies.

1.3 OUTLINE

The rest of this thesis is organized as follows:

*representation learning
self-supervised learning
word embedding*

Transformer

pretext tasks

Chapter 2 introduces notions about representation learning , with an emphasis on deep self-supervised learning for text. We briefly discuss about models that are able to produce neural word embeddings , then we focus on some neural architectures that can be used to deal with sequential data. We study two particular neural language models based on the Transformer ([Vaswani et al., 2017](#)) architecture.

In Chapter 3 we introduce the problem of anomaly detection, which is the main concern of this thesis. We present some classical models and the recent advancements that are using deep neural networks.

We finally introduce our solution for detecting anomalies in text using deep learning in Chapter 4. We describe how we exploit our “pretext task” to do training and inference, then we go in detail about our experimental setup, quantitative and qualitative results, and a rigorous ablation study. We also showcase an application for anomaly detection in text, in the form of authorship detection in the supplementary material B.2.

2

REPRESENTATION LEARNING

NEURAL NETWORKS have recently seen great popularity, mainly due to the fact that AI systems based on neural networks are able to learn from massive amounts of data, given enough compute power. The development of GPU acceleration techniques (Chellapilla et al., 2006) and dedicated hardware accelerators (Jouppi et al., 2017) has enabled researchers and practitioners to train neural networks that can achieve human-level or even super-human performance on various tasks (Badia et al., 2020; Cireşan et al., 2012). The popularity of supervised deep learning approaches has escalated after the 2012 edition of the *ImageNet Large Scale Visual Recognition Challenge* (Russakovsky et al., 2015), where Alex Krizhevsky has obtained a top-5 error of 15.3% with AlexNet, a convolutional neural network (Fukushima, 1980; LeCun, B. Boser, J. Denker, et al., 1990; LeCun, B. Boser, J. S. Denker, et al., 1989; Waibel et al., 1989). This approach brought an improvement of 41.4% over the next best solution (Krizhevsky et al., 2012), and was the start of a complete shift in interest from manually constructed features to learned features. The AlexNet architecture is shown in Fig. 2.

This renewed interest in deep neural networks has led to many different architectures which greatly improved the state-of-the-art in computer vision (He et al., 2016; Simonyan and Zisserman, 2014; Szegedy et al., 2015), natural language processing (Cho et al., 2014b; Hochreiter and Schmidhuber, 1997; Mikolov, K. Chen, et al., 2013; Vaswani et al., 2017), and deep reinforcement learning (Berner et al., 2019; Mnih et al., 2013; Silver et al., 2016; Vinyals et al., 2019).

Most of the architectures used in computer vision were trained via supervised learning, where models are instructed to solve particular tasks. The idea of transfer learning (Caruana et al., 1995) by using some of the learned weights and extracting representations or fine-tuning the model quickly re-emerged and is still widely used for training when limited amounts of data are provided (Zhuang et al., 2020). In natural language processing, since text is unstructured and varies in length, it was already very desirable to

Convolutional Neural Network (CNN)

*supervised learning
transfer learning
representations, fine-tuning*

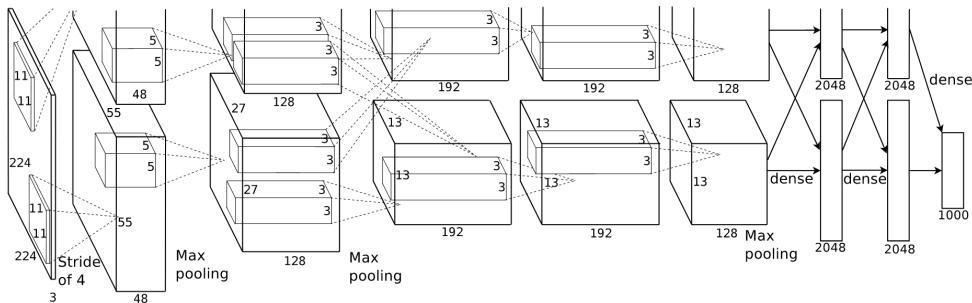


Figure 2: The AlexNet CNN architecture. (figure from: Krizhevsky et al., 2012)

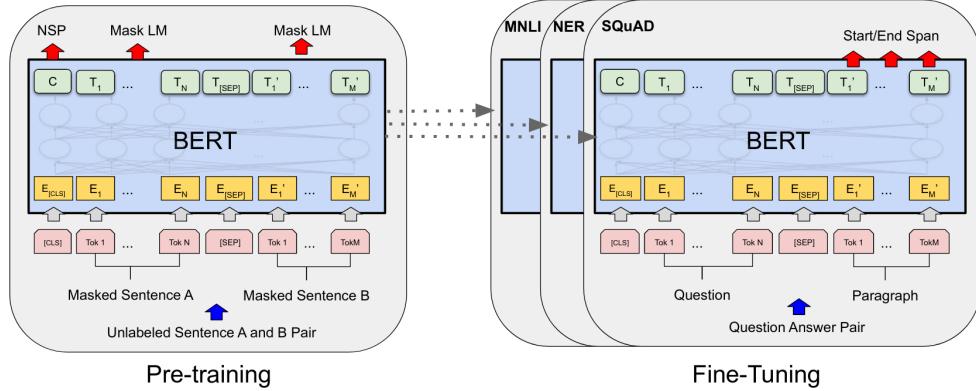


Figure 3: The training and fine-tuning procedures for the BERT Transformer-based model. More details about Transformer models will be provided in section 2.3. The BERT architecture will be detailed in subsection 2.3.1. (figure from: Devlin et al., 2018)

learn general word representations which could be used in conjunction with specialized models. This has led to the development of many techniques for representation learning in NLP (Bojanowski et al., 2017; Mikolov, K. Chen, et al., 2013; Pennington et al., 2014) culminating with deep language models (Peters et al., 2018) based on the Transformer architecture (Brown et al., 2020; Clark, Luong, Q. V. Le, et al., 2020; Devlin et al., 2018; Vaswani et al., 2017), which are usually trained via self-supervision.

deep language models

self-supervision

2.1 SELF-SUPERVISED LEARNING

As previously mentioned, much of the recent progress in the field of machine learning is due to models trained to solve particular tasks in a supervised fashion. These models require large collections of (usually) expensive labeled data. One way of alleviating the requisite of annotated data is by transferring knowledge from models trained on immense labeled data banks. This approach assumes that these kinds of data banks exist, which can lead to domain-specific solutions or, at best, sub-optimal results. For instance, the knowledge transfer from a 2D computer vision dataset to a 3D task can be quite limited. Similarly, we can't expect that the weights of a model trained on pictures of cats and dogs to be sufficient for classifying medical images. Moreover, the task of labeling everything in the world is impossible. These limitations can make us wonder if there's a more efficient way of learning general representations. One way of tackling this problem is self-supervised learning.

Self-supervised learning is a form of unsupervised learning where the data provides the supervision (Zisserman, 2018). Generally, the self-supervised approach is done by corrupting a part of the input and making the model reconstruct some of the lost information. In this way, the model should be able to learn the general structure of the data in an unsupervised way, thus enabling us to use the learned representations for other downstream tasks via transfer learning. The self-supervision task is sometimes referred to as

downstream task

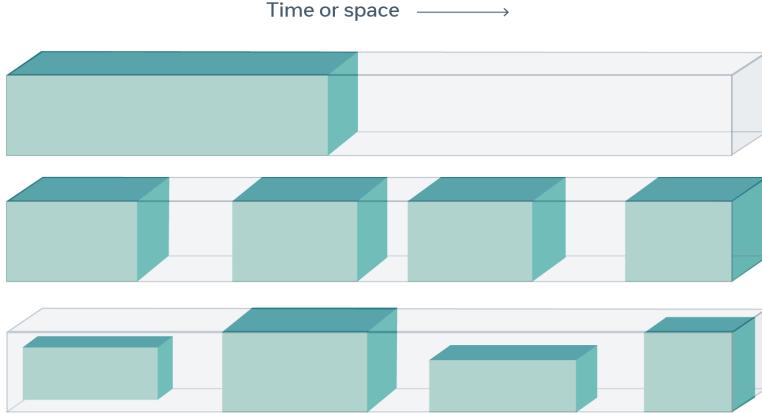


Figure 4: Example of general pretext tasks for self-supervised learning. In natural language processing we could try to predict the next token, or some masked tokens, from a sequence of tokens. In computer vision we can build a model that reconstructs the missing parts of a corrupted image. (figure from: LeCun, 2021)

a “*pretext task*”, and the loss function is usually called “*proxy loss*”. By using self-supervised learning, the model obtains supervisory signals from the structure of the data, therefore the reliance on labeled datasets is eliminated.

The field of natural language processing is the earliest success story of SSL. By using huge amounts of unlabeled text, deep language models trained using self-supervision resulted in which many consider being “*NLP’s ImageNet moment*”. These models are usually given an input phrase that has some of its parts masked, and then they are assigned to complete the phrase. For example, when given the text “*The students are (blank) mathematics at the faculty of (blank)*”, the model should be able to determine that the most likely action for the students to do in this context is to learn and that the most common faculty where one learns mathematics is the faculty of mathematics, thus one viable (but not unique) completion would be “*The students are learning mathematics at the faculty of mathematics*”. As a result, these masked language models learn to represent the contexts and the significance of the words in different contexts (LeCun, 2021).

proxy loss

Masked Language Model (MLM)

The development of SSL models is not limited to the field of NLP. Recent models trained with self-supervision tasks in the field of computer vision have obtained state-of-the-art results when fine-tuned to solve classification tasks (T. Chen, Kornblith, Norouzi, et al., 2020; Grill et al., 2020). Furthermore, some models trained via self-supervision and fine-tuned on fractions of the dataset are able to surpass their fully-supervised counterparts, even when using just 10% labeled data (T. Chen, Kornblith, Swersky, et al., 2020).

But still, CV has not benefited from SSL to the same extent as NLP, in part because in vision it’s more difficult to represent uncertainty due to the domain in which we need to make predictions being virtually infinite. We are able to fix a finite vocabulary and get a probability distribution over the entire vocabulary for a masked word, but if given a picture of a book on

contrastive learning

a table, it's practically impossible to get a distribution over all the possible objects that may be on that table. As a result, SSL in vision is usually done by predicting various transformations (such as rotations or translations) of the input (S. Wang et al., 2019), agreeing upon various views of the same object (Grill et al., 2020), or by using *contrastive learning* (T. Chen, Kornblith, Norouzi, et al., 2020).

In the next sections, we'll focus on self-supervised learning for natural language processing and we'll discuss some of the more common models and approaches.

2.2 WORD EMBEDDINGS

Word embeddings are dense vectorial representations of words. A key property of these embeddings is that the words that have similar semantic meaning also tend to have similar encodings.

The idea of obtaining dense vectors from words using neural networks is not new (Bengio et al., 2003), but the *Collobert-Weston model* (Collobert and Weston, 2008) was one of the first successful attempts at learning general-purpose word embeddings for multi-task learning. The authors describe a convolutional neural network which is leveraging unlabeled data using a language model that minimizes the following ranking-type loss:

$$\sum_{s \in S} \sum_{w \in D} \max(0, 1 - f(s) + f(s^w)) \quad (1)$$

where S is the set of context windows of the text, D is the dictionary of words, and $f(\cdot)$ is the neural network without the softmax¹ layer, and s^w is a context window where the middle word has been replaced by a random word w . In this way, the language model is trained to solve a binary classification task: whether the word in the middle of the input is related to the context or not.

This approach has led to improvements when sharing weights and jointly training the LM along with supervised tasks such as *part-of-speech tagging*, *named entity recognition*, *chunking*, and *semantic role labeling*. Furthermore, the model produced high-quality word embeddings as a side-effect (Table 1).

FRANCE 454	JESUS 1973	XBOX 6909	REDDISH 11724	SCRATCHED 29869
SPAIN	CHRIST	PLAYSTATION	YELLOWISH	SMASHED
ITALY	GOD	DREAMCAST	GREENISH	RIPPED
RUSSIA	RESURRECTION	PSNUMBER	BROWNISH	BRUSHED
POLAND	PRAYER	SNES	BLUISH	HURLED
ENGLAND	YAHWEH	WII	CREAMY	GRABBED
DENMARK	JOSEPHUS	NES	WHITISH	TOSSED
GERMANY	MOSES	NINTENDO	BLACKISH	SQUEEZED
PORTUGAL	SIN	GAMECUBE	SILVERY	BLASTED

Table 1: Qualitative result regarding the word embeddings produced by the Collobert-Weston model. The queried words are followed by their index in the dictionary and their nearest 8 neighbors. (table from: Collobert and Weston, 2008)

¹ The softmax function,
 $\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$
 $\forall i \in \{1, \dots, K\}$ and
 $z = (z_1, \dots, z_K)$ is an activation function used to normalize the output of a network to a probability distribution over the predicted classes.

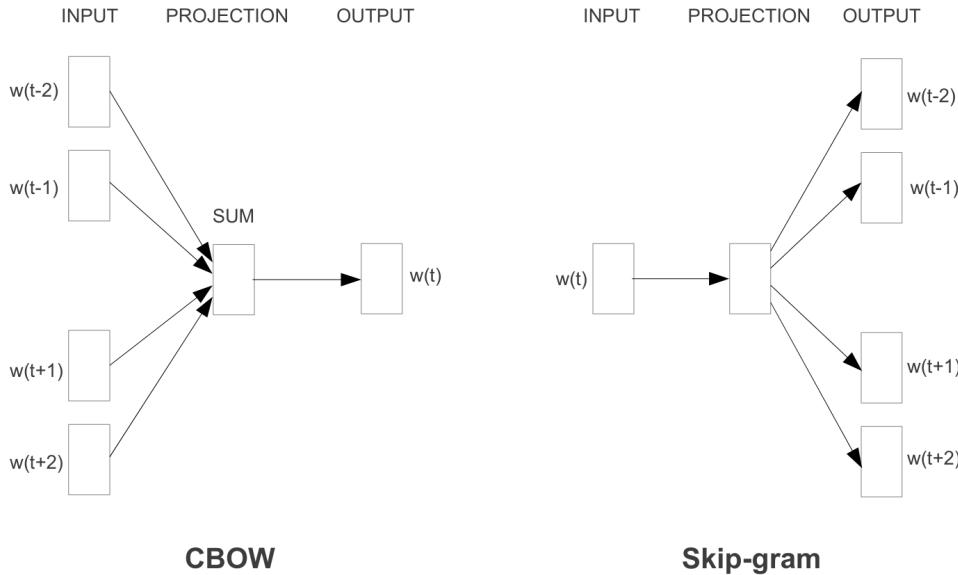


Figure 5: The CBOW and Skip-Gram models. CBOW is predicting the word with respect to the context, while Skip-Gram is predicting the context with respect to the word. (figure from: Mikolov, Q. V. Le, et al., 2013)

The next great leap for word embeddings was brought by a class of techniques called *Word2Vec* (Mikolov, K. Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Mikolov, Q. V. Le, et al., 2013). This time, the models are specifically trained to produce word embeddings, the quality of the embeddings being measured with a word similarity task and compared to the previous state-of-the-art. We'll briefly examine two of the models from the Word2Vec family: the *Skip-Gram* model and the *Continuous Bag of Words* (Fig. 5).

Word2Vec

Skip-Gram, Continuous Bag of Words (CBOW)

Negative sampling (NEG)

The Skip-gram model is using the Negative sampling objective to predict the context of a word, on the other hand, CBOW's objective is to predict the word in the middle of the context. One key advantage of these models is that they require relatively low compute power, thus they can be trained on large corpora. The authors note that Skip-gram delivers better performance on small monolingual datasets, while CBOW is faster and benefits from large amounts of data.

Given a middle token w_t and a training context of k tokens, the Skip-gram model tries to compute the probability of the neighbouring tokens as being placed correctly. Formally, Skip-gram maximizes the following expression:

$$\frac{1}{T} \sum_{t=1}^T \left[\sum_{j=-k}^k \log P(w_{t+j}|w_t) \right], \quad (2)$$

$P(w_i|w_j)$ is defined as:

$$P(w_i|w_j) = \frac{e^{u_w^\top v_{w_j}}}{\sum_{\xi=1}^{|D|} e^{u_\xi^\top v_{w_j}}}, \quad (3)$$

where $|D|$ is the cardinality of the dictionary, u_w is the associated input representation, and v_w is the associated output representation.

The procedure for CBOW is analogous, except for the fact that we're now trying to maximize the log probability $\log P(w_t|w_{t-j:t+j})$.

An interesting property of the learned vector representations is that linear operations seem to be able to capture semantic meaning. For example, the operation $\text{vector}(\text{"King"}) - \text{vector}(\text{"Man"}) + \text{vector}(\text{"Woman"})$ yields a vector close to $\text{vector}(\text{"Queen"})$ (Mikolov, Yih, et al., 2013), and $\text{vector}(\text{"France"}) - \text{vector}(\text{"Paris"})$ is close to $\text{vector}(\text{"Italy"}) - \text{vector}(\text{"Rome"})$ (Mikolov, Q. V. Le, et al., 2013).

In the following years, other popular methods based on the Skip-Gram and CBOW methods were developed, the most notable being *GloVe* (Pennington et al., 2014) and *fastText* (Bojanowski et al., 2017). *GloVe* builds on Word2Vec by using global word statistics of the corpus during training, thus obtaining state-of-the-art performance on the word analogy dataset and producing vector spaces with meaningful sub-structure. The *fastText* method extends Word2Vec by using n-grams of characters instead of words. This approach helps to capture the significance of shorter words, suffixes, prefixes, and rare words.

2.3 TRANSFORMERS

Word embedding models are great tools for obtaining token-level semantic vectors, but they offer little contextual information. It can be argued that an abstract representation of a phrase can be obtained by applying an operation (such as addition) to all the extracted vectors (Mikolov, Sutskever, et al., 2013), but this is not a direct *inductive bias* of the model, but a property in the resulting vector space. Also, by doing something like summing all of the vectors in order to obtain a sequence representation, we lose information such as the order of the tokens. For tasks where sequentiality is important, as is the case in machine translation, this kind of information loss is not acceptable. A more expressive *contextual embedding* is desirable. This contextual embedding should be able to represent an entire sequence of words or the contextual meaning of a word in a sequence, with respect to the other words and their order.

The modelling of sequential data was usually done using 1-dimensional Convolutional Neural Networks or *Recurrent Neural Networks* and the gated variants of the latter, such as the *Long Short-Term Memory* (Hochreiter and Schmidhuber, 1997) and *Gated Recurrent Units* (Cho et al., 2014a) models. The problem with CNNs is that they have a limited context window, and some locality information is lost by applying operations such as pooling. RNNs, on the other hand, are able to encode a longer succession of tokens, but are notoriously hard to parallelize and can become very memory-expensive when training on long examples. Various tricks (Kuchalev and Ginsburg, 2017; Shazeer et al., 2017) can help alleviate these issues, but the fundamental architectural constraints remain.

A model that is able to circumvent these limitations is the Transformer (Vaswani et al., 2017). Transformer networks are relinquishing recurrence in favour of the *self-attention mechanism* to draw global dependencies between

Global Vectors (GloVe)
fastText

n-gram

inductive bias

contextual embedding

Recurrent Neural Network (RNN)

Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)

Unit (GRU)

self-attention

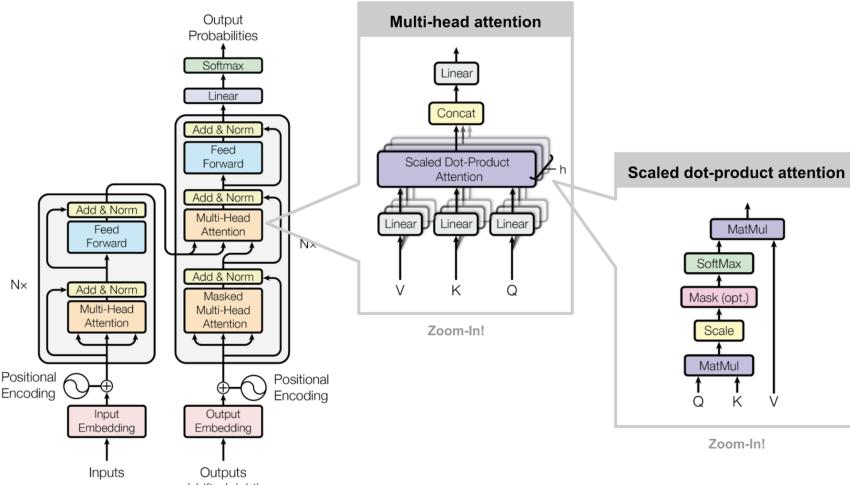


Figure 6: The original Transformer encoder-decoder architecture used for machine translation. The encoder layer has two sub-layers: a multi-head self-attention layer, and a fully-connected layer. One decoder layer contains two multi-head self-attention layers and a fully-connected layer. (figure from: Vaswani et al., 2017; Weng, 2018)

the input and the output. This approach enables much more parallelization, leading to much larger scale language models.

The Transformer has an encoder-decoder structure, therefore it is able to map a sequence $x = (x_1, \dots, x_n)$ of symbols to a continuous representation $z = (z_1, \dots, z_n)$ which can be used in an *auto-regressive* way to compute an output sequence of symbols $y = (y_1, \dots, y_n)$. This encoder-decoder structure is constructed using just stacked self-attention and fully-connected layers, for both the encoder and the decoder, as can be seen in Fig. 6.

auto-regressive

The key component in the Transformer architecture is the multi-head self-attention mechanism. Self-attention is used to calculate a weighted average of the sequence representations, which are encoded into a *key-value-query* pair. The values of the (K, V, Q) pair are obtained by linearly projecting the word embeddings using a dot product between them and some learned weights (linear projection). A dot product between the query and the keys is then computed. The final weighted average is obtained by performing another dot product with the values matrix:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where d_k is the dimension of the queries and keys. The scaling factor $\frac{1}{\sqrt{d_k}}$ is used to prevent having large magnitude dot products, which could “push the softmax functions into regions with extremely small gradients”².

key, value, query (K, V, Q)

This operation is done in parallel using multiple *heads*, each head containing parameters for different linear projection matrices, thus allowing the model to process the information from different linear subspaces, resulting in more expressiveness. The final output is the following:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^O \quad (5)$$

² From Vaswani et al., 2017: “To illustrate why the dot products get large, assume that the components of q and k are independent random variables with mean 0 and variance 1. Then their dot product, $d \times k = \sum_{i=1}^{d_k} q_i k_i$, has mean 0 and variance d_k ”

where $\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$ and $X \in \mathbb{R}^{\text{batch} \times \text{tokens} \times \text{dim}}$ is the input, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are the learnable linear projection matrices, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ is another learnable linear projection matrix.

The output from the self-attention block can be then forwarded to a fully-connected network that's using a *ReLU* activation function:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2. \quad (6)$$

The original Transformer model obtained state-of-the-art performances on multiple machine translation benchmarks while requiring fewer FLOPs to train than its counterparts. Models based on the Transformer architecture became very popular in natural language processing, especially when trained using self-supervised language modelling tasks, with the intention of being later fine-tuned for downstream tasks. Recently, Transformers have also been successfully used in computer vision (Caron et al., 2021; M. Chen et al., 2020; Dosovitskiy et al., 2020), and many various augmentations to the architecture have been proposed (Beltagy, Peters, et al., 2020; Kitaev et al., 2020; Xiong et al., 2021).

The reader should be aware that recent work in computer vision (Tolstikhin et al., 2021) suggests that self-attention is not needed and that one can design entirely fully-connected networks with similar performance to that of Transformers. However, this does not imply that self-attention and the Transformer are redundant, but merely that it's useful to also explore other inductive biases (Wolpert and Macready, 1997).

2.3.1 BERT

As mentioned in the previous subsection, a prevalent technique is to train large language models based on the Transformer and then use the language model for a specific (usually supervised) downstream task, by either extracting the representations or by fine-tuning the model on the new task.

Bidirectional Encoder Representations from Transformers (BERT)

One of the more successful models for accomplishing this is *BERT*. The authors argue that a key difference between BERT and previous techniques (Peters et al., 2018; A. Radford and Narasimhan, 2018) is that the former “restrict the power of the pre-trained representations” mainly because they are unidirectional language models, therefore limiting the attention to look only at the previous tokens during the pre-training phase, which could lead to sub-optimal results for sentence-level tasks.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Table 2: BERT scores on the GLUE test set. (table from: Devlin et al., 2018)

BERT proposes two different tasks for pre-training: the masked language modelling task, which is inspired by the Cloze task (Taylor, 1953), and the *next sentence prediction* task. The MLM task consists of randomly masking parts of the input and predicting the masked input with respect to the unmasked tokens. This allows for the self-attention mechanism to attend to the entire context during training, unlike the classical unidirectional language modeling task. The NSP task is a binary classification task for predicting whether the second sentence in a pair is a continuation of the first one. NSP is justified by the fact that it could improve the performance on various downstream tasks, such as question answering. However, further empirical research shows that the NSP task does not significantly improve downstream performance (Y. Liu et al., 2019). An illustration of the training and fine-tuning procedures for BERT can be consulted at Fig. 3.

*next sentence prediction
(NSP)*

The architecture of BERT consists of a multi-layer bidirectional Transformer encoder and the original model comes in two configurations: BERT_{BASE} and BERT_{LARGE}. The first contains 12 layers with a hidden size of 768 and 12 attention heads, with a total of 110M parameters, while the former contains 24 layers with a hidden size of 1024 and 16 attention heads, totaling in 340M parameters. BERT is able to improve the state-of-the-art on 11 NLP tasks, including GLUE (Table 2) (A. Wang et al., 2018), SQuAD (Rajpurkar et al., 2016), and SWAG (Zellers et al., 2018).

Many improvements and adaptations have been proposed over the original BERT. Some address other languages (de Vries et al., 2019; Dumitrescu et al., 2020; Martin et al., 2020), domain-specific solutions (Beltagy, Lo, et al., 2019; Lee et al., 2019), or more efficient pre-training (Clark, Luong, Q. V. Le, et al., 2020; Lan et al., 2019). We'll next present one such BERT-based model which leverages the idea of pre-training using two separate Transformers: a discriminator and a generator.

2.3.2 ELECTRA

ELECTRA (Clark, Luong, Q. V. Le, et al., 2020) is a method of pre-training by using a BERT-like generator and discriminator. The generator is trained with the MLM objective, as in BERT, and generates plausible alternatives for the masked tokens. The generated sequence is then fed to the discriminator which classifies every token as being *original* or *replaced* - this task is called replaced token detection (see Fig. 7, Fig. 15). The network is trained using the following loss:

$$\mathcal{L}_{\text{ELECTRA}} = \mathcal{L}_{\text{MLM}} + \lambda \mathcal{L}_{\text{RTD}} \quad (7)$$

where \mathcal{L}_{MLM} is the masked language modeling loss:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E} \left[\sum_{i \in m} -\log \mathbb{P}_G(x_i | \hat{x}(m); \theta_G) \right], \quad (8)$$

\mathcal{L}_{RTD} is the replaced token detection loss:

$$\mathcal{L}_{\text{RTD}} = \mathbb{E} \left[\sum_{i=1}^T -\log \mathbb{P}_D(m_i | \tilde{x}(m); \theta_D) \right], \quad (9)$$

Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)

Model	Train / Infer FLOPs	Speedup	Params	Train Time + Hardware	GLUE
ELMo	3.3e18 / 2.6e10	19x / 1.2x	96M	14d on 3 GTX 1080 GPUs	71.2
GPT	4.0e19 / 3.0e10	1.6x / 0.97x	117M	25d on 8 P6000 GPUs	78.8
BERT-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	75.1
BERT-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	82.2
ELECTRA-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	79.9
50% trained	7.1e17 / 3.7e9	90x / 8x	14M	2d on 1 V100 GPU	79.0
25% trained	3.6e17 / 3.7e9	181x / 8x	14M	1d on 1 V100 GPU	77.7
12.5% trained	1.8e17 / 3.7e9	361x / 8x	14M	12h on 1 V100 GPU	76.0
6.25% trained	8.9e16 / 3.7e9	722x / 8x	14M	6h on 1 V100 GPU	74.1
ELECTRA-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	85.1

Table 3: ELECTRA results on the GLUE dataset. (table from: [Clark, Luong, Q. V. Le, et al., 2020](#))

θ_G , θ_D are the generators' and discriminators' parameters, the set m is the set of indices where the masking is applied, $\hat{x}(m)$ is the sequence without the replaced tokens at positions m , $\tilde{x}(m)$ is the sequence with the tokens replaced at positions m , λ is a loss weighting hyperparameter, \mathbb{P}_G is the vocabulary probability distribution for each token, and \mathbb{P}_D is the probability that a token was replaced or not.

The generator is smaller than the discriminator, containing about $\frac{1}{4}$ of the discriminator parameters. The final model doesn't have shared parameters, except for the embedding layers. The reader should be aware of the fact that, even if the objective is similar to the one of a generative adversarial network, there are some key differences. First, the gradients from the discriminator are not propagating to the generator, mainly because there's a non-differentiable sampling phase when generating the corrupted input. Second, both the generator and the discriminator are trained using maximum likelihood, the method is thus distinct from the usual "mini-max game" played by the GAN discriminator and generator.

During training, 15% of the tokens are masked. After training, the generator is discarded and only the discriminator is used for downstream tasks. Overall, this method greatly speeds up the training time. Furthermore, the fine-tuned discriminator is obtaining similar or better performance on all of the tasks (see Table 3).

A follow-up paper from the authors of ELECTRA explores the theoretical soundness of the training procedure ([Clark, Luong, Q. Le, et al., 2020](#)). As a result, an *energy-based* Cloze model called Electric is developed, and it is argued that ELECTRA is a special case of Electric. This work is outside the scope of this thesis, but the author highly recommends that the interested reader should study both ELECTRA and Electric.

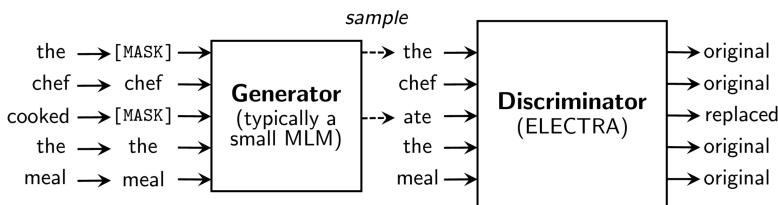


Figure 7: ELECTRA training scheme. (figure from: [Clark, Luong, Q. V. Le, et al., 2020](#))

3 ANOMALY DETECTION

MANY TECHNIQUES were proposed to address the outlier detection problem. If labels are available, we can treat the anomaly detection problem as an imbalanced supervised classification task and use common classification algorithms. However, data imbalance must be addressed. This is usually done by weighting the data or by adaptive resampling methods, such as SMOTE (Bowyer et al., 2011). The scenarios that interest us are the semi-supervised one, where we have access to clean data, and the unsupervised one, where we assume that our training set contains a percentage of outliers.

Some of the classic methods are based on *density estimation* (Breunig et al., 2000; F. T. Liu et al., 2008; Rousseeuw and Driessen, 1999), one-class learning (L. Manevitz and Yousef, 2001; L. M. Manevitz and Yousef, 2002; Tax and Duin, 2004), Bayesian networks (Adams and MacKay, 2007; Cheeseman and Stutz, 1996), ensemble methods (Aggarwal, 2013), or approaches based on dictionary learning (Baltoiu et al., 2020; Irofti and Baltoiu, 2020).

density estimation, one-class learning, bayesian network, ensemble method, dictionary learning

Methods based on deep neural networks have become popular for AD due to their superior performance and the availability of large amounts of data. These methods usually involve learning the inlier class features using autoencoders (J. Chen et al., 2017; S. Hawkins et al., 2002; Sakurada and Yairi, 2014) or generative adversarial networks (Deecke et al., 2019; H. Wang et al., 2019) and produce normality scores using the reconstruction error.

Recently, deep AD methods based on self-supervision have emerged in the field of computer vision, obtaining better quantitative and qualitative results. These approaches usually rely on training deep neural networks to distinguish between various transformations applied to the input images (Golan and El-Yaniv, 2018; S. Wang et al., 2019) and an anomaly score is computed by aggregating model predictions across multiple transformed input images. Such self-supervised approaches have also recently become common when dealing with video data (Georgescu et al., 2020). In the text domain, deep methods are more scarce and usually rely on pretrained word embeddings (Hu et al., 2021; Ruff, Zemlyanskiy, et al., 2019).

In this chapter, we'll first describe two powerful classical baselines: the Isolation Forest (F. T. Liu et al., 2008) and the One-Class SVM (L. M. Manevitz and Yousef, 2002), and then we'll present two recent deep approaches - the *E³ Outlier* (S. Wang et al., 2019) framework, which enables the training of end-to-end self-supervised anomaly detectors by predicting transformations applied to the original images, and the *Context Vector Data Description* method, which is using pretrained word embeddings and self-attention to produce “*context vectors*”, which can be leveraged to produce anomaly scores for text sequences based on the *cosine similarity* between the embeddings and the context vectors.

E³ Outlier

Context Vector Data Description (CVDD)

cosine similarity

3.1 CLASSICAL MODELS

3.1.1 One-Class Support Vector Machine

Support Vector Machines (B. E. Boser et al., 1992) are a class of models used to do binary classification. They achieve this by projecting the samples in a high-dimensional (or infinite dimensional) space, where the data is *linearly separable*, and choosing the decision boundary which has the *maximum margin*. The concept of margin is defined as being the smallest distance between each sample and the decision boundary. Intuitively, this can be thought of as finding the hyperplane that separates the data into two classes and has the largest distance to all the nearest samples in the training data.

The One-Class SVM is a one-class classification technique based on the SVM. The strategy is to project the data into another space and then separate the data from the origin which has the maximum margin. By doing so, we can attribute the outlier class to the samples which are closer to the origin of the new space. Formally, let \mathcal{X} be the initial feature space, and $\Phi : \mathcal{X} \rightarrow \mathcal{F}$ be a transformation that maps points from \mathcal{X} to a space \mathcal{F} which is equipped with a dot product that can be evaluated with a kernel function $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ which has the form:

$$k(x, y) = (\Phi(x) \cdot \Phi(y)). \quad (10)$$

To separate the data from the origin, we need to minimize the following expression:

$$\mathcal{L} = \frac{1}{2} \|W\|^2 + \frac{1}{vN} \sum_{i=1}^N \max\{b - W \cdot \Phi(x_i), 0\} - b, \quad (11)$$

where W is the coefficient vector, N is the number of samples in the training set, $\frac{1}{2}\|W\|^2$ is a regularization term, b is the bias, $\max\{b - W \cdot \Phi(X), 0\}$ is the so-called *slack penalty*, and v is the term that regulates the trade-off between *false positives* and *false negatives*. The term v can be also be thought as being a “prior probability that a data point in the training set is an outlier” (Aggarwal, 2015).

A direct solution for the optimization problem is difficult to obtain due to the usage of $\Phi(\cdot)$, the projection function. Moreover, $\Phi(\cdot)$ is often not used directly when training SVMs, the transformation being implicitly done via the kernel function $k(\cdot, \cdot)$. One solution to this is by solving a *dual representation* of the maximum margin problem, and using the *kernel trick* within the dual formulation.

The dual problem is a constrained optimization problem and is outside the scope of this thesis. The reader should refer to Chapter 7.1 of Christopher Bishop’s excellent book (Bishop, 2006) for a more elaborate discussion regarding the optimization of SVMs. For the particular case of optimizing the OC-SVM, one can also refer to Chapter 3.4 of Charu C. Aggarwal’s book (Aggarwal, 2015) on outlier detection.

*linearly separable
decision boundary, margin*

slack penalty

false positive, false negative

dual problem, kernel trick

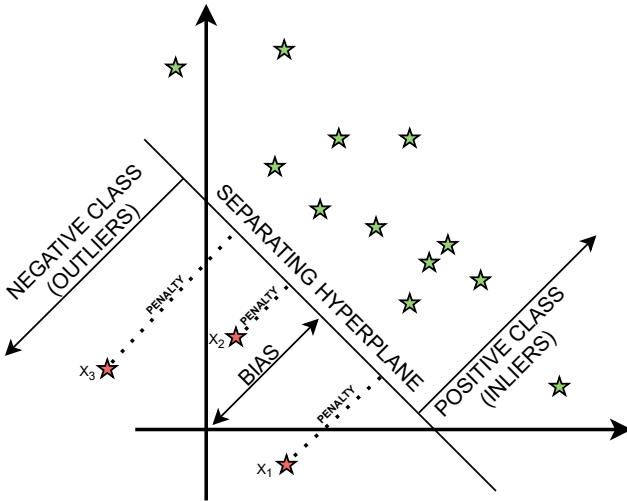


Figure 8: A solution to the OC-SVM optimization problem. Points on the left side of the hyperplane are classified as outliers, while points on the right side are inliers. Outliers are penalized with penalty = $b - W \cdot \Phi(X_i)$. Inlier points are not penalized. (figure adapted from: Aggarwal, 2015)

After training, an anomaly score can be obtained by computing:

$$\text{Score}(x) = \sum_{i=1}^N \alpha_i \cdot k(x, x_i) - b, \quad (12)$$

where α_i is a Lagrangian parameter corresponding to the slack variables from the dual form.

A model that is closely related to the OC-SVM is the *Support Vector Data Description* (Tax and Duin, 2004). The SVDD is an SVM that finds a hypersphere of radius R in which the inliers are enclosed (as shown in Fig. 9). This differs from the standard OC-SVM approach where a linear separator from the origin is found. The SVDD can be seen as a special case of the OC-SVM when using the *Radial Basis Function* kernel, which embeds the points on an unit sphere. Nonetheless, the two methods usually produce different solutions, and empirical evidence (Emmott et al., 2013) suggests that the SVDD has a slight advantage over the OC-SVM using the RBF kernel.

Support Vector Data Description (SVDD)

Radial Basis Function (RBF)
 $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

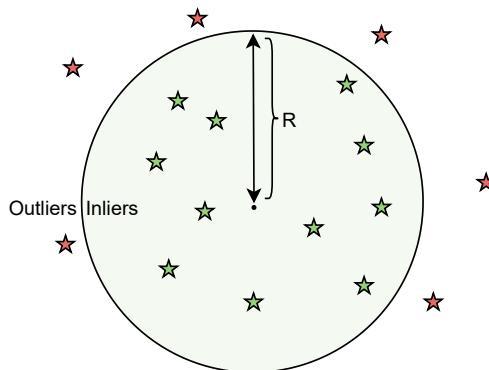


Figure 9: A SVDD solution. Inliers are inside the hypersphere of radius R , while outliers are outside of it.

3.1.2 Isolation Forest

Isolation Forest (iForest) Instead of learning how the data distribution of inliers looks like, the *Isolation Forest* (F. T. Liu et al., 2008) is identifying anomalous entries by isolating them into nodes and calculating an anomaly score based on the distance of the nodes from the root.

Isolation Tree (iTree)

Decision Tree (DT)

The Isolation Forest is built using an ensemble of *Isolation Trees*, which are an unsupervised formulation of *Decision Trees*. The iForest assumes that anomalies are a minority in the dataset which have attributes that greatly differ from the inlier class. iTrees are exploiting these properties of anomalies by recursively partitioning randomly selected attributes of the data with random cuts that are parallel to the axis. This has the effect of isolating outliers in shallow tree branches due to the points being located in sparse neighborhoods and thus making them more susceptible to isolation. The final decision is made by averaging the distances from the leaves to the root of the tree in different trees of the iForest.

The recursive partitioning done by an iTree is performed in the following way: given a sample of n d-dimensional data points $X = \{x_1, \dots, x_n\}$, we consider X as being the root node. We now divide X by randomly picking an attribute i and a random cutoff point $c \in [i_{\min}, i_{\max}]$, where i_{\min} and i_{\max} are the minimum and maximum values for attribute i . This gives us two subsets $X_{\leq} = \{x \in X | x^{(i)} \leq c\}$, and $X_{>} = \{x \in X | x^{(i)} > c\}$. The two subsets represent the leaves of the initial node. We repeat this procedure recursively for every new node until either: (i) the tree reaches a height limit that was set priorly, (ii) the new subsets have exactly one element, or (iii) the new subsets contain the same value. A toy example of this procedure is shown in Figure 10.

³ the Isolation Tree is a Binary Search Tree

The average path length of an unsuccessful search in a Binary Search Tree³ is $c(n) = 2H(n-1) - (2(n-1)/2)$, where n is the number of external nodes, and $H(n) \approx \ln(n) + e$ is the harmonic number. The anomaly score for data point x is defined as:

$$\text{Score}(x, n) = 2^{-\frac{\frac{1}{n_t} \sum_{i=1}^{n_t} h(x)}{c(n)}}, \quad (13)$$

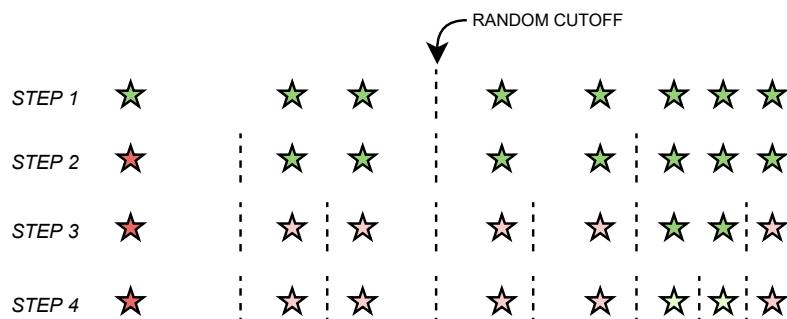


Figure 10: Isolation Tree partitioning process. The first data point is isolated at step 2 (higher anomaly score). All the points are isolated at the end of the recursive partitioning.

where n_t is the number of iTrees in the Isolation Forest and $h(x)$ is the path length of a point measured by the number of edges that need to be traversed from the root to the point's corresponding node in the iTree.

The authors suggest that the following decisions can be made depending on the value of Score:

- if $s(x, n)$ is close to 1, then x is an anomaly.
- if $s(x, n)$ is less than 0.5, then x can be regarded as an inlier.
- if $\forall x \in X$ we have $s(x, n) \approx 0.5$, then we are unable to tell if X contains anomalies.

3.2 DEEP MODELS

3.2.1 E³ Outlier

Some of the first artificial neural networks designed to tackle the anomaly detection problem were the “*replicator networks*” (S. Hawkins et al., 2002). The replicator network is an autoencoder trained to reconstruct the input data. The reconstruction error is then used to produce an anomaly score. Many recent deep AD methods rely on various autoencoder architectures (Xia et al., 2015; Zhou and Paffenroth, 2017), but these kinds of networks suffer from the fact that autoencoders are generally not good at representation learning when compared to discriminative DNN architectures, such as ResNet (He et al., 2016). This begs the question if discriminative networks could somehow be trained in an end-to-end fashion to do Outlier Detection.

The E³ (Effective End-to-End) Outlier (S. Wang et al., 2019) is a neural network training and unsupervised anomaly detection framework for computer vision. Outlier detection is achieved by training the models via self-supervision⁴ with the task of discriminating between various transformations or combinations of transformations that are applied to the input data. An anomaly score is then calculated based on the model’s predictions on multiple transformations or by using extracted representations in conjunction with other AD methods (such as the iForest).

⁴ In the original paper, the terms used are *surrogate supervision* and *surrogate supervision based discriminative network* (SSD).

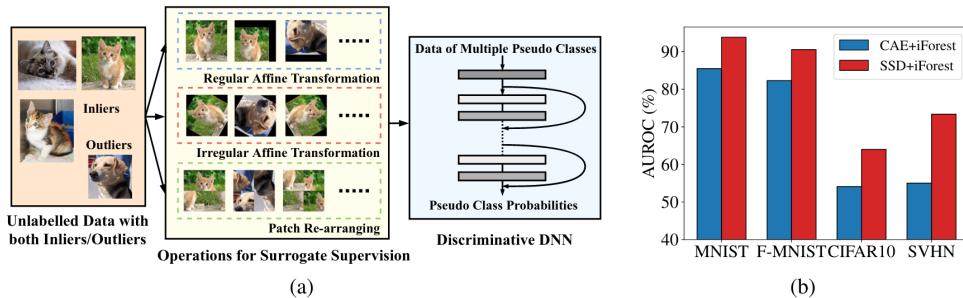


Figure 11: a) E³ Outlier self-supervision training schema. b) A performance comparison between representations extracted from the self-supervised discriminative network (SSD) and a convolutional autoencoder (CAE). (figure from: S. Wang et al., 2019)

$${}^5 \text{CE}(\hat{y}, y) = -\frac{1}{K} \sum_{i=1}^K y_i \cdot \log(\hat{y}_i)$$

The proposed self-supervision task can be defined in the following way: construct a set $\mathcal{O} = \{O(\cdot|y)\}_{y=1}^k$ of K transformations that can be applied to the input images and label them with a label y . We can generate a new “pseudo-labeled” data point $x^{(y)}$ from an entry x by applying the transformation $O(\cdot|y) \in \mathcal{O}$. Now we can train a SSD model which has a K -dimensional softmax layer in a self-supervised fashion by feeding it $x^{(y)}$, calculating the loss using *cross entropy*⁵, backpropagating, and updating the neural network weights, as we would do with any supervised neural network. Please note that this technique requires no manually labeled data and that the “pseudo-label” is created at the moment in which we pick the transformation which we want to apply. We are not using any prior knowledge about the dataset itself.

The following sets of transformations are used:

1. Rotations: the images are rotated clock-wisely by a certain degree.
2. Flip: the images are flipped or not.
3. Shifting: some pixels of the images are shifted along the x or y axis.
4. Jigsaw: the images are partitioned into several equally-sized patches and then the patches are permuted.

Based on these sets, we can define the “regular affine transformations” subset \mathcal{O}_{RA} , the “irregular affine transformations” subset \mathcal{O}_{IA} , and the “patch re-arranging set” \mathcal{O}_{PR} . These subsets can contain combinations of transformations, which have their own class. The final operation set used is $\mathcal{O}_{RA} \cup \mathcal{O}_{IA} \cup \mathcal{O}_{PR}$ and consists of 111 transformations.

After the learning process, the learned representations can be used to train other anomaly detectors (as suggested in figure 11b). Another method of obtaining an anomaly score is by directly utilizing the output from the softmax layer. Indeed, this approach produces better results when compared to the former.

To extract an anomaly score for an image x , we apply to every transformation $O(\cdot|y)$ in from the set \mathcal{O} , therefore obtaining a set of K transformed images $\{x^{(1)}, \dots, x^{(K)}\}$ and their pseudo-labels. The transformed images are then fed through the network, producing a set of K logits $\{P(x^{(y)}|\Theta)\}$, where Θ are the parameters of the network. Now we can apply one of the proposed scores:

1. Pseudo Label (PL): an anomaly score is obtained by averaging the probabilities of the **correct class** from the obtained logits, i.e.:

$$\text{PL}(x) = \frac{1}{K} \sum_{y=1}^K P^{(y)}(x^{(y)}|\Theta). \quad (14)$$

2. Maximum Probability (MP): A issue with the PL score is that the classes obtained by applying the transformations could not be sufficiently separable. For example when we flip the number “8” we’re still getting the number “8”. This can make our models unable to detect the correct transformation and can hurt our anomaly score. One way

to solve this issue is to take the **maximum probability from the logits** for every transformation:

$$\text{MP}(x) = \frac{1}{K} \sum_{y=1}^K \max_t \mathbb{P}^{(t)}(x^{(y)}|\Theta). \quad (15)$$

3. Negative Entropy (NE): The previous two scores are using information from a single classes' probability, ignoring the other ($K - 1$) classes. One assumption that we can make is that, when given an anomaly, the neural network will be unable to decide what transformation was applied, thus producing an uninformative probability distribution that has high entropy, as seen in Fig. 12. Under this assumption, we can use the entropy to produce an anomaly score in the following way:

$$\text{NE}(x) = -\frac{1}{K} \sum_{y=1}^K H(\mathbb{P}(x^{(y)}|\Theta)) = \frac{1}{K} \sum_{y=1}^K \sum_{t=1}^K \mathbb{P}^{(t)}(x^{(y)}|\Theta) \log(\mathbb{P}^{(t)}(x^{(y)}|\Theta)). \quad (16)$$

In practice, the Negative Entropy score gives the best results.

The authors are training multiple discriminative DNN models such as ResNet, Wide ResNet (Zagoruyko and Komodakis, 2016), and DenseNet (Huang et al., 2017), but the method can be applied to any neural network that can be trained on image data. The datasets used are MNIST (LeCun and Cortes, 2010), Fashion MNIST (Xiao et al., 2017), CIFAR10, SVHN (Netzer et al., 2011), and CIFAR100 (Krizhevsky, 2012). The semi-supervised outlier detection setup is the following: given a dataset, consider that one class with one semantic concept (e.g. “car”, “human”) is the inlier class and use it as training data. For outliers, sample a percentage of the images from every other class. For the fully unsupervised scenario, simply add a percentage of outliers in the training dataset.

The models trained in the E³ Outlier manner are greatly outperforming the competition, usually improving the AUROC and AUPR scores by 5% – 30% in both semi-supervised and unsupervised scenarios, as can be seen in Table 4 and Fig. 13.

Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPR)

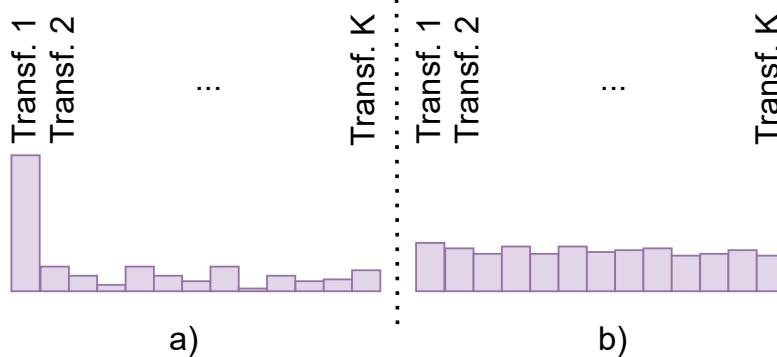


Figure 12: a) Low entropy “spikey” distribution. b) High entropy “flat” distribution. A high entropy gives a higher anomaly score.

Table 4: AUROC/AUPR-inlier/AUPR-outlier (%) scores for unsupervised outlier detection. Bold represents the best scores. ρ is the training set contamination rate. CAE: Convolutional autoencoder. CAE-IF: CAE+iForest. DRAE: Discriminative reconstruction based autoencoder. RDAE: Robust deep autoencoder. DAGMM: Deep autoencoding gaussian mixture model (table from: S. Wang et al., 2019)

Dataset	ρ	CAE	CAE-IF	DRAE	RDAE	DAGMM	SSD-IF	E^3 Outlier
MNIST	10%	68.0/92.0/32.9	85.5/97.8/49.0	66.9/93.0/30.5	71.8/93.1/35.8	64.0/92.9/26.6	93.8/99.2/ 68.7	94.1/99.3/67.5
	20%	64.0/82.7/40.7	81.5/93.6/57.2	67.2/86.6/42.5	67.0/84.2/43.2	65.9/86.4/41.3	90.5/97.3/71.0	91.3/97.6/72.3
F-MNIST	10%	70.3/94.3/29.3	82.3/97.2/40.3	67.1/93.9/25.5	75.3/95.8/31.7	64.0/92.7/30.3	90.6/98.5/68.6	93.3/99.0/75.9
	20%	64.4/85.3/36.8	77.8/92.2/49.0	65.7/86.9/36.6	70.9/89.2/41.4	66.0/86.7/43.5	87.6/95.6/71.4	91.2/97.1/78.9
CIFAR10	10%	55.9/91.0/14.4	54.1/90.2/13.7	56.0/90.7/14.7	55.4/90.7/14.0	56.1/91.3/15.6	64.0/93.5/18.3	83.5/97.5/43.4
	20%	54.7/81.6/25.5	53.8/80.7/25.3	55.6/81.7/26.8	54.2/81.0/25.7	54.7/81.8/26.3	60.2/85.0/28.3	79.3/93.1/52.7
SVHN	10%	51.2/90.3/10.6	55.0/91.4/11.9	51.0/90.3/10.5	52.1/90.6/10.8	50.0/90.0/19.3	73.4/95.9/22.0	86.0/98.0/36.7
	20%	50.7/80.2/20.7	54.0/82.0/22.4	50.6/80.4/20.5	51.8/80.9/21.1	50.0/79.9/29.6	69.2/89.5/33.7	81.0/93.4/47.0
CIFAR100	10%	55.2/91.0/14.5	54.5/90.7/13.8	55.6/90.9/15.0	55.8/90.9/15.0	54.9/91.1/14.2	55.6/91.5/13.0	79.2/96.8/33.3
	20%	54.4/81.7/25.6	53.5/80.9/25.1	55.5/81.8/27.0	54.9/81.5/26.5	53.8/81.5/24.7	54.3/82.1/23.4	77.0/92.4/46.5

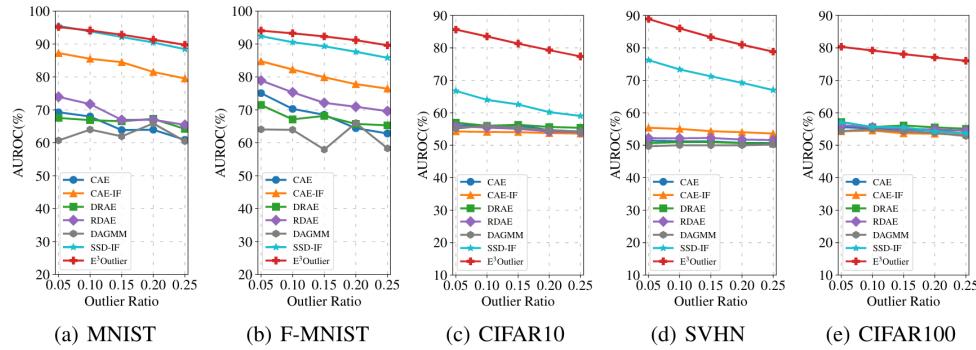


Figure 13: Unsupervised outlier detection comparison for various rates of contamination. (figure from: S. Wang et al., 2019)

3.2.2 Context Vector Data Description

While many anomaly detection solutions using deep neural networks were specifically designed for computer vision, only a few have been proposed for natural language processing. One method for detecting anomalies in text is the Context Vector Data Description (Ruff, Zemlyanskiy, et al., 2019), which leverages pretrained word embeddings in order to construct “*context vectors*” using the self-attention mechanism. These context vectors can capture semantic contexts and can be used to perform contextual anomaly detection with respect to the themes and concepts which are encoded in the learned context vectors.

The CVDD method is using a slightly different multi-head self-attention mechanism than the one presented in (Vaswani et al., 2017). For a given sentence $S = (w_1, \dots, w_l)$, we can extract a matrix $H = (h_1, \dots, h_l) \in \mathbb{R}^{p \times l}$ containing the corresponding word embeddings, extracted with some universal word embedding models (e.g. Word2Vec, FastText, GloVe), or by some language model (e.g. BERT, ELECTRA). Now, by using the multi-head self-attention mechanism, we can obtain a single fixed-size representation from our matrix H . First, we compute the attention matrix A in the following way:

$$A = \text{softmax}(\tanh(H^T W_1) W_2), \quad (17)$$

where $W_1 \in \mathbb{R}^{p \times d_a}$ and $W_2 \in \mathbb{R}^{d_a \times r}$ are learnable weight matrices, and d_a is the model dimension. The *tanh*⁶ activation function is applied element-wise while the softmax activation is applied column-wise. By doing this we obtain an attention matrix $A = (a_1, \dots, a_r)$ containing r attention heads. Now, a sentence embedding matrix $M = (m_1, \dots, m_r) \in \mathbb{R}^{p \times r}$ can be calculated using the self-attention matrix:

$$M = H A. \quad (18)$$

The context vectors can now be constructed as follows: let r be the number of attention heads and $C = (c_1, \dots, c_r) \in \mathbb{R}^{p \times r}$ a matrix containing r vectors. We say that C is the context matrix and c_1, \dots, c_r are the context vectors. The Context Vector Data Description loss is defined in the following way:

$$\min_{C, W_1, W_2} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^r \sigma_k(H^{(i)}) d(c_k, m_k^{(i)}) + \|C^T C - I\|_F^2, \quad (19)$$

where $d(c_k, m_k^{(i)})$ is the cosine distance between the context vector c_k and the data representation $m_k^{(i)}$, I is the $r \times r$ identity matrix, $\|\cdot\|_F$ is the Frobenius norm, and $\sigma_1(H), \dots, \sigma_r(H)$ are the input-dependent weights which are passed through a *parametrized softmax function*⁷. Note that there are as many context vectors as attention heads, thus constraining the network to

$$^6 \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

```

greatest show ever mad full stop greatest show ever mad full stop greatest show ever mad full stop
greatest show ever mad full stop greatest show ever mad full stop greatest show ever mad full stop
lived let tell idea heck bear walk never heard whole years really funny beginning went hill quickly
ten minutes people spewing gallons pink vomit recurring scenes enormous piles dog excrement need one say
john made two one man shows rama freaks neither one shown dvd john john put dvd john people see need see john case anyones keeping watchful eye
suspenseful subtle much much disturbing

```

(a) Overall most anomalous reviews in IMDB test set according to CVDD.

c_1	c_2	c_3
great	excellent	awful
good	superb	downright
well	wonderful	stupid
nice	best	irrepting
terrific	beautiful	pathetic

(b) Top words per context by self-attention weights from the most similar sentences per CVDD context.

<small>terrible performance many key place steve sandbois jacqueline basett rebekah johnson superb story reels movie emotional yet light hearted movie everyone shared loves great mixed company crowd nice production cheap budget well organized keeps interest uses some newer ideas flashbacks one point keeps viewer edge seat matter walk life viewer buy one viewpoint film worth watching whole lot rated film lol independent films seem bit overrated days well hollywood films matter movie fantastic really great bad reach huge audience superno really love alicia determination really makes look upon life gift set privileged education aside movie really proves good artist tell good story making difference everyone adored it didn't really learn something roger ebert know one person been year certainly one trutiful</small>	<small>tiny appealing waste space friend tried watch film conclusion switch minutes end count films switched end one hand script direction leaden deeply uninspiring surprised found script pile cast scripts example irritating</small>	<small>supernatural plot black white complex movie character塑造力チャーチャー complex moto charlie chan series subplots unexpected twists appearances number movie recognize immediately none ever made status wonderfully unpredictable minutes would love see boxed dvd series films</small>
<small>awful majority exclusively made video low budget fight flicks invariably stink worse raunchy old socks particularly dismal amateurish budget chicago set bargain basement nasty necrophiliac nutcase loose bloodbath serves depressing affirmation borderline fact bearded disheveled long haired beat flower shirt wearing wild eyed psycho hippie fruitcake embarks standard random</small>	<small>plotting errors unattractive middle aged man called lester meets rich unattractive middle aged woman via lonely hearts ads murders money needs feed growing addition whole plot really happens along way attempt intriguing lester starts get phone calls mysterious stranger taunts knowing</small>	<small>supernatural plot black white complex movie character塑造力チャーチャー complex moto charlie chan series subplots unexpected twists appearances number movie recognize immediately none ever made status wonderfully unpredictable minutes would love see boxed dvd series films</small>
<small>majority exclusively made video low budget fight flicks invariably stink worse raunchy old socks particularly dismal amateurish budget chicago set bargain basement nasty necrophiliac nutcase loose bloodbath serves depressing affirmation borderline fact bearded disheveled long haired beat flower shirt wearing wild eyed psycho hippie fruitcake embarks standard random</small>	<small>enjoy hack something clearly changed storylines seem much stronger plot may still tax surreal characters developed depth surreal plot forgiven attribute fine acting every show come charging starting gate winner need time pick speed glad kept watching program really hope lasts</small>	<small>supernatural plot black white complex movie character塑造力チャーチャー complex moto charlie chan series subplots unexpected twists appearances number movie recognize immediately none ever made status wonderfully unpredictable minutes would love see boxed dvd series films</small>
<small>majority exclusively made video low budget fight flicks invariably stink worse raunchy old socks particularly dismal amateurish budget chicago set bargain basement nasty necrophiliac nutcase loose bloodbath serves depressing affirmation borderline fact bearded disheveled long haired beat flower shirt wearing wild eyed psycho hippie fruitcake embarks standard random</small>	<small>enjoy hack something clearly changed storylines seem much stronger plot may still tax surreal characters developed depth surreal plot forgiven attribute fine acting every show come charging starting gate winner need time pick speed glad kept watching program really hope lasts</small>	<small>supernatural plot black white complex movie character塑造力チャーチャー complex moto charlie chan series subplots unexpected twists appearances number movie recognize immediately none ever made status wonderfully unpredictable minutes would love see boxed dvd series films</small>

(c) Most normal reviews in IMDB test set for CVDD contexts c_1 , c_2 , and c_3 with highlighted self-attention weights.

Figure 14: Qualitative results on the IMDB Movie Reviews dataset. a) The most anomalous entries in the test set. b) Words that are clustering around the context vectors - the first and second context vectors seem to model positive and negative sentiments, while the third one models cinematic language. c) Top most normal entries in the test set. Most abnormal words are in a more intense red. (figure from: [Ruff, Zemlyanskiy, et al., 2019](#))

Normal Class	GloVe							fastText						
	OC-SVM			CVDD				OC-SVM			CVDD			
	mean	tf-idf	max	r=3	r=5	r=10	c*	mean	tf-idf	max	r=3	r=5	r=10	c*
Reuters														
earn	91.1	88.6	77.1	94.0	92.8	91.8	97.6	87.8	82.4	74.9	95.3	92.7	93.9	94.5
acq	93.1	77.0	81.4	90.2	88.7	91.5	95.6	91.8	74.1	80.2	91.0	90.3	92.7	92.4
crude	92.4	90.3	91.2	89.6	92.5	95.5	89.4	93.3	90.2	84.7	90.9	94.1	97.3	85.0
trade	99.0	96.8	93.7	98.3	98.2	99.2	97.9	97.6	95.0	92.1	97.9	98.1	99.3	97.7
money-fx	88.6	81.2	73.6	82.5	76.7	82.8	99.7	80.5	82.6	73.8	82.6	79.8	82.5	99.5
interest	97.4	93.5	84.2	92.3	91.7	97.7	98.4	91.6	88.7	82.8	93.3	92.1	95.9	97.4
ship	91.2	93.1	86.5	97.6	96.9	95.6	99.7	90.0	90.6	85.0	96.9	94.7	96.1	99.7
20 News														
comp	82.0	81.2	54.5	70.9	66.4	63.3	86.6	77.5	78.0	65.5	74.0	68.2	64.2	88.2
rec	73.2	75.6	56.2	50.8	52.8	53.3	68.9	66.0	70.0	51.9	60.6	58.5	54.1	85.1
sci	60.6	64.1	53.0	56.7	56.8	55.7	61.0	61.0	64.2	57.0	58.2	57.6	55.9	64.4
misc	61.8	63.1	54.1	75.1	70.2	68.6	83.8	62.3	62.1	55.7	75.7	70.3	68.0	83.9
pol	72.5	75.5	64.9	62.9	65.3	65.1	75.4	73.7	76.1	68.1	71.5	66.4	67.1	82.8
rel	78.2	79.2	68.4	76.3	72.9	70.7	87.3	77.8	78.9	73.9	78.1	73.2	69.5	89.3

Table 5: Semi-supervised outlier detection scores (AUROC %) on the Reuters and 20 Newsgroups datasets. The c* context vector is hand-picked and produces the best scores, thus making it dependent on the ground truth. (table from: Ruff, Zemlyanskiy, et al., 2019)

learn attention weights which are able to extract the most common themes and concepts from the given corpora. The $\|C^T C - I\|_F^2$ regularization term is also used to promote diverse context vectors, by encouraging orthogonality. The network is trained using gradient descent, with respect to the weights W_1, W_2 , and the context matrix C .

After training, an anomaly detection score can be calculated for a sample S with an embedding H by averaging the contextual anomaly scores:

$$\text{Score}(H) = \frac{1}{r} \sum_{k=1}^r d(c_k, m_k(H)). \quad (20)$$

Quantitative and qualitative experiments are performed on 20Newsgroups (Lang and Rennie, 1999), Reuters-21578 (Lewis, 1999) (See Table 5 and Table 6), and IMDB Movie Reviews (Maas et al., 2011) datasets (See Fig. 14). The official train/test split is used. One class is considered as being inliers, while the rest are considered outliers.

The word embeddings used are GloVe and fastText. On every dataset, the text is stripped of punctuation, numbers, redundant whitespaces, and stopwords. The text is also converted to lowercase and the only words that are considered are the words with more than three characters.

computer			politics			religion		
c_1	$c_2 (c^*)$	c_3	c_1	c_2	$c_3 (c^*)$	c_1	$c_2 (c^*)$	c_3
get	windows	use	kill	think	government	example	god	one
help	software	using	killed	know	peace	particular	christ	first
thanks	disk	used	escape	say	arab	specific	christians	two
appreciated	dos	uses	away	really	political	certain	faith	three
got	unix	possible	back	thing	occupation	analysis	jesus	also
know	computer	system	shoot	anyone	forces	rather	christianity	later
way	hardware	need	shot	guess	support	therefore	bible	time
try	desktop	allow	crying	something	movement	consistent	scripture	last
tried	macintosh	could	killing	understand	leaders	often	religion	year
take	cpu	application	fight	sure	parties	context	worship	four

Table 6: Top words per context on the 20 Newsgroups comp, pol, and rel subsets with 3 context vectors. (table from: Ruff, Zemlyanskiy, et al., 2019)

4 | DATE

AS ANNOUNCED, in this chapter we shall introduce our solution to the problem of detecting anomalies in text using deep learning. Our method is called DATE (which is short for Detecting Anomalies in Text using ELECTRA) ([Manolache et al., 2021](#)). Our solution is an end-to-end AD method relying on self-supervision in order to produce an anomaly score. We will first concern ourselves with the RMD and RTD self-supervision tasks which were also described in the Prologue chapter of the thesis.

4.1 PRETEXT TASKS

The discriminative task called RMD has the purpose of creating training data by altering a given text using one out of K given operations and it also asks to ascertain which one of the operations is the correct one given the transformed text. The modification of the text is obtained by following the next steps:

1. *masking* some of the input words using a predefined mask pattern;
2. replacing the *masked* words with alternative ones⁸.

⁸ For example *frog* with *animal*

We shall now proceed to describe our method more formally. To this end, let $\mathbf{m} \in \{0, 1\}^T$ be a mask pattern corresponding to the text input $\mathbf{x} = [x_1, \dots, x_T]$. For training, we generate and fix K mask patterns $\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(K)}$ by randomly sampling a constant number of 1's out of the T available ones. Unlike ELECTRA which masks random tokens spontaneously, our model first samples a mask model from the K predefined ones. We then proceed to mask our input with \mathbf{m} and obtain the vector $\hat{\mathbf{x}}(\mathbf{m}) := [\hat{x}_1, \dots, \hat{x}_T]$, where

$$\hat{x}_i = \begin{cases} x_i, & m_i = 0 \\ [\text{MASK}], & m_i = 1 \end{cases}$$

or simply $\hat{x}_i = x_i(1 - m_i) + [\text{MASK}]m_i$. For instance, given an input $\mathbf{x} = [\text{they, were, ready, to, go}]$ and a mask pattern $\mathbf{m} = [0, 0, 1, 0, 1]$, the resulting masked input will be $\hat{\mathbf{x}}(\mathbf{m}) = [\text{they, were, [MASK], to, [MASK]}]$.

Each of the masked tokens could then be replaced by a word token. A naive way to do this is by sampling uniformly from the vocabulary. However, for more plausible alternatives, masked tokens can be sampled from a Masked Language Model generator such as BERT, which outputs a proba-

bility distribution \mathbb{P}_G over the vocabulary for each corrupted token. We will denote the plausibly corrupted text by $\tilde{x}(\mathbf{m}) = [\tilde{x}_1, \dots, \tilde{x}_T]$, where

$$\tilde{x}_i = \begin{cases} x_i, & \mathbf{m}_i = 0 \\ w_i \sim \mathbb{P}_G(x_i|\hat{x}(\mathbf{m}); \theta_G), & \mathbf{m}_i = 1 \end{cases}$$

For instance, for the masked input from before, a plausibly corrupted input could be $\tilde{x}(\mathbf{m}) = [\text{they}, \text{were}, \text{prepared}, \text{to}, \text{depart}]$.

RTD is a binary sequence tagging task, where some tokens in the input are corrupted with plausible alternatives, which is similar to how RMD works. The key difference is that the discriminator must then predict for *each token* if it's the *original* token or a *replaced* one. Distinctly from RTD, which is a *token-level* discriminative task, RMD is a *sequence-level* one. This means that the model distinguishes between a fixed number of predefined transformations applied to the input. As such, RMD can be seen as the text counterpart task for the self-supervised classification of geometric alterations applied to images, as described in Subsec. 3.2.1. While RTD predictions could be used to sequentially predict an entire mask pattern, they can lead to masks that are not part of the predefined K patterns⁹. The RMD constraint overcomes this behaviour. We thus train DATE to solve both tasks simultaneously, which increases the AD performance compared to solving one task only, as shown in Subsec. 4.2.2. Furthermore, we observed empirically that this approach also improves training stability.

⁹ Suppose that we have the input text “*the cat is purring*” and the masking pattern set $\mathcal{M} = \{[0, 1, 0, 1], [1, 0, 0, 1]\}$. In this situation, we could

use the RTD head to produce both available masking patterns, but it could also produce the pattern $[1, 1, 1, 1]$, which does not exist in \mathcal{M} .

4.1.1 DATE Architecture

We propose simultaneously solving RMD and RTD by jointly training a generator, \mathbf{G} , and a discriminator, \mathbf{D} .

\mathbf{G} is a MLM used to *replace* the masked tokens with plausible alternatives. We also consider a setup with a *random generator*, in which we sample tokens uniformly from the vocabulary. Indeed, this approach seems to work better for our AD setup than a parameterized generator, as can be seen in the ablation study at Subsec. 4.2.2. We suspect that this is happening because the anomalies are a lot more different than the inliers, thus making the plausible text generated by a parameterized generator look much like the inlier class, ultimately confusing the discriminator.

\mathbf{D} is a deep neural network with two prediction heads used to distinguish between *corrupted* and original tokens (RTD) and to predict which mask pattern was applied to the corrupted input (RMD). At test time, \mathbf{G} is discarded, and \mathbf{D} 's probabilities are used to compute an anomaly score.

Both \mathbf{G} and \mathbf{D} models are based on a BERT encoder, which consists of several stacked Transformer blocks. The BERT encoder transforms an input token sequence $\mathbf{x} = [x_1, x_2, \dots, x_T]$ into a sequence of contextualized word embeddings $\mathbf{h}(\mathbf{x}) = [h_1, h_2, \dots, h_T]$.

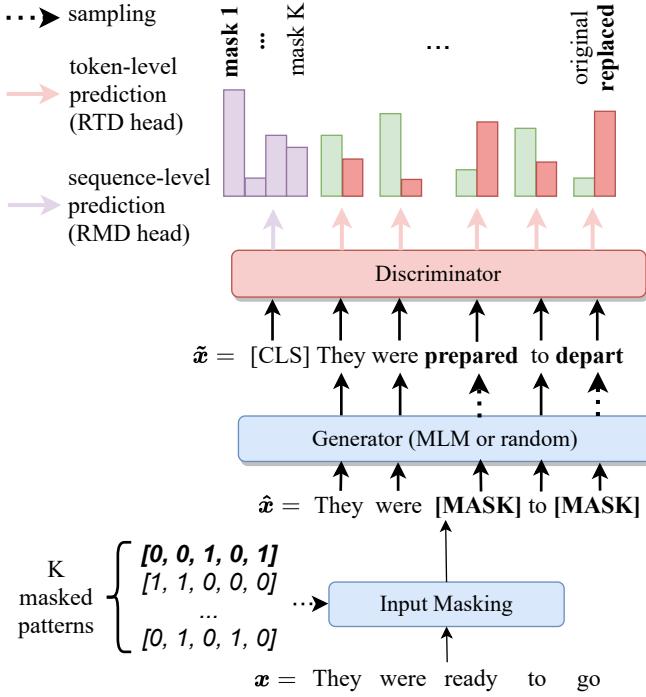


Figure 15: DATE Training. Firstly, the input sequence is masked using a sampled masked pattern and a generator fills in new tokens in place of the masked ones. Secondly, the discriminator receives supervision signals from two tasks: RMD (which mask pattern was applied to the input sequence) and RTD (the per-token status: original or replaced). (figure from: [Manolache et al., 2021](#))

When parameterized, \mathbf{G} is a BERT encoder with a linear layer on top that outputs the probability distribution P_G for each token. The generator is trained using the MLM loss:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E} \left[\sum_{\substack{i=1; \\ s.t. m_i=1}}^T -\log P_G(x_i | \tilde{\mathbf{x}}(\mathbf{m}); \theta_G) \right] \quad (21)$$

\mathbf{D} is a BERT encoder with two prediction heads applied over the contextualized word representations:

1. **RMD head.** This head outputs a vector of logits for all mask patterns $\mathbf{o} = [o_1, \dots, o_K]$. We use the contextualized hidden vector $h_{[\text{CLS}]}$ (corresponding to the $[\text{CLS}]$ special token at the beginning of the input) for computing the mask logits \mathbf{o} and P_M , the probability of each mask pattern:

$$P_M(\mathbf{m} = \mathbf{m}^{(k)} | \tilde{\mathbf{x}}(\mathbf{m}^{(k)}); \theta_D) = \frac{e^{o_k}}{\sum_{i=1}^K e^{o_i}} \quad (22)$$

2. **RTD head.** This head outputs scores for the two classes (*original* and *replaced*) for each token x_1, x_2, \dots, x_T , by using the contextualized hidden vectors h_1, h_2, \dots, h_T .

We train the DATE network in a maximum-likelihood fashion using the $\mathcal{L}_{\text{DATE}}$ loss:

$$\min_{\theta_D, \theta_G} \sum_{x \in \mathcal{X}} \mathcal{L}_{\text{DATE}}(\theta_D, \theta_G; x) \quad (23)$$

The loss contains both the token-level losses in ELECTRA, as well as the sequence-level mask detection loss \mathcal{L}_{RMD} :

$$\mathcal{L}_{\text{DATE}}(\theta_D, \theta_G; x) = \mu \mathcal{L}_{\text{RMD}}(\theta_D; x) + \mathcal{L}_{\text{MLM}}(\theta_G; x) + \lambda \mathcal{L}_{\text{RTD}}(\theta_D; x), \quad (24)$$

where the discriminator losses are:

$$\mathcal{L}_{\text{RMD}} = \mathbb{E} \left[-\log \mathbb{P}_M(m | \tilde{x}(m); \theta_D) \right], \quad (25)$$

$$\mathcal{L}_{\text{RTD}} = \mathbb{E} \left[\sum_{\substack{i=1; \\ x_i \neq [\text{CLS}]}}^T -\log \mathbb{P}_D(m_i | \tilde{x}(m); \theta_D) \right], \quad (26)$$

where \mathbb{P}_D is the probability distribution that a token was replaced or not.

The ELECTRA loss enables \mathbf{D} to learn good feature representations for language understanding. Our RMD loss puts the representation in a larger sequence-level context. After pre-training, \mathbf{G} is discarded and \mathbf{D} can be used as a general-purpose text encoder for downstream tasks. Output probabilities from \mathbf{D} are further used to compute an anomaly score for new examples.

4.1.2 Anomaly Detection score

We adapt the Pseudo Label (PL) based score from the E³Outlier framework in a novel and efficient way. In its general form, the PL score aggregates responses corresponding to multiple transformations of x . This

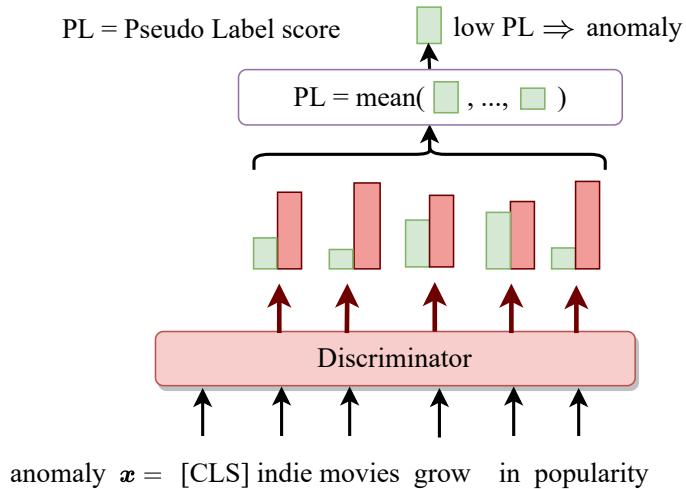


Figure 16: DATE Testing. The input text sequence is fed to the discriminator, resulting in token-level probabilities for the normal class, which are further aggregated into an anomaly score, as detailed in Sec. 4.1.2. For deciding whether a sample is either normal or abnormal, we aggregate over all of its tokens. (figure from: [Manolache et al., 2021](#))

approach requires K input transformations over an input x and K forward passes through a discriminator. It then takes the probability of the ground truth transformation and averages it over all K transformations.

To compute PL for our RMD task, we take x to be our input text and the K mask patterns as the possible transformations. We corrupt x with mask $\mathbf{m}^{(i)}$ and feed the resulted text to the discriminator. We take the probability of the i^{th} mask from the RMD head. We repeat this process K times and average over the probabilities of the correct mask pattern. This formulation requires K feed-forward steps through the DATE network, which slows down inference. We propose a more computationally efficient approach next.

Instead of aggregating *sequence-level* responses from multiple transformations over the input, we can aggregate *token-level* responses from a single model over the input to compute an anomaly score. More specifically, we can discard the generator and feed the original input text to the discriminator directly. We then use the probability of each token being *original* (not *corrupted*) and then average over all the tokens in the sequence:

$$\text{PL}_{\text{RTD}}(x) = \frac{1}{T} \sum_{i=1}^T P_D(m_i = 0 | \tilde{x}(\mathbf{m}^{(0)}); \theta_D), \quad (27)$$

where $\mathbf{m}^{(0)} = [0, 0, \dots, 0]$ effectively leaves the input unchanged. As can be seen in Fig. 16, the RTD head will be less certain in predicting the *original* class for *outliers* (having a probability distribution unseen at training time), which will lead to lower PL scores for *outliers* and higher PL scores for *inliers*. We use PL at testing time, when the entire input is either normal or abnormal. Our method also speeds up inference, since we only do one feed-forward pass through the discriminator instead of k passes. Moreover, having a per token anomaly score helps us better understand and visualize the behavior of our model, as shown in Fig. 18b.

4.2 EXPERIMENTAL ANALYSIS

In this section, we detail the empirical validation of our method by presenting: the semi-supervised and unsupervised experimental setup, a comprehensive ablation study on DATE, and the comparison with state-of-the-art on the semi-supervised and unsupervised AD tasks. DATE does not use any form of pre-training or knowledge transfer (from other datasets or tasks), learning all the embeddings from scratch. Using pre-training would introduce unwanted prior knowledge about the outliers, making our model considering them known (normal).

4.2.1 Experimental setup

We describe next the Anomaly Detection setup, the datasets and the implementation details of our model. The code is publicly available ¹⁰.

We use a semi-supervised setting in Sec. 4.2.2-4.2.3 and an unsupervised one in Sec. 4.2.4. In the semi-supervised case, we successively treat one class as normal (*inliers*) and all the other classes as abnormal (*outliers*). In

¹⁰ <https://github.com/bit-ml/date>

the unsupervised AD setting, we add a fraction of outliers to the inliers training set, thus contaminating it. We compute the Area Under the Receiver Operating Curve (AUROC) for comparing our method with the previous state-of-the-art. For a better understanding of our model’s performance in an unbalanced dataset, we report the Area Under the Precision-Recall curve (AUPR) for inliers and outliers per split in the supplementary material B.1.

We test our solution using two text classification datasets, after stripping headers and other metadata. For the first dataset, 20Newsgroups, we keep the exact setup, splits, and preprocessing (lowercase, removal of: punctuation, number, stop words, and short words) as in (Ruff, Zemlyanskiy, et al., 2019), ensuring a fair comparison with previous text anomaly detection methods. As for the second dataset, we use a significantly larger one, AG News, better suited for deep learning methods.

¹¹ <http://qwone.com/~jason/20Newsgroups/>

1. **20Newsgroups**¹¹: We only take the articles from six top-level classes: *computer, recreation, science, miscellaneous, politics, religion*, like in (Ruff, Zemlyanskiy, et al., 2019). This dataset is relatively small, but a classic for NLP tasks (for each class, there are between 577-2856 samples for training and 382-1909 for validation).
2. **AG News** (Zhang et al., 2015): This topic classification corpus was gathered from multiple news sources, for over more than one year ¹². It contains four topics, each class with 30000 samples for training and 1900 for validation.

For training the DATE network we follow the pipeline in Fig. 15. In addition to the parameterized generator, we also consider a *random generator*, in which we replace the masked tokens with samples from a uniform distribution over the vocabulary.

The *discriminator* is composed of four Transformer layers, with *two prediction heads* on top (for RMD and RTD tasks). We provide more details about the model in the supplementary material A.2. We train the networks with AdamW with amsgrad (Loshchilov and Hutter, 2019), $1e^{-5}$ learning rate, using sequences of maximum length 128 for AG News, and 498 for 20Newsgroups. We use $K = 50$ predefined masks, covering 50% of the input for AG News and $K = 25$, covering 25% for 20Newsgroups. The training converges on average after 5000 update steps and the inference time is 0.005 sec/sample in PyTorch (Paszke et al., 2017), on a single GTX Titan X.

4.2.2 Ablation studies

To better understand the impact of different components in our model and making the best decisions towards a higher performance, we perform an extensive set of experiments (see Tab. 7). Note that we successively treat each AG News split as inlier and report the mean and standard deviations over the four splits. The results show that our model is robust to domain shifts.

A. Anomaly score. We explore three anomaly scores introduced in the E³Outlier framework (S. Wang et al., 2019) on semi-supervised and unsupervised AD tasks in Computer Vision: Maximum Probability (MP), Negative

Abl.	Method	Variation	AUROC(%)
	CVDD	best	83.1 ± 4.4
	OCSVM	best	84.0 ± 5.0
	ELECTRA	adapted for AD	84.6 ± 4.5
	DATE (Ours)		90.0 ± 4.2
<hr/>			
A.	Anomaly score	MP	72.4 ± 3.7
		NE	73.1 ± 3.9
B.	Generator	small	89.3 ± 4.2
		large	89.8 ± 4.4
C.	Loss function	RTD only	89.4 ± 4.4
		RMD only	85.9 ± 4.1
D.	Masking patterns	5 masks	87.5 ± 4.5
		10 masks	89.2 ± 4.3
		25 masks	89.8 ± 4.3
		100 masks	89.8 ± 4.3
E.	Mask percent	15%	89.5 ± 4.1
		25%	89.5 ± 4.1

Table 7: Ablation study. We show results for the competition and report ablation experiments which are only one change away from our best DATE configuration: **A.** PL_{RTD}; **B.** Rand **C.** RTD + RMD; **D.** 50 masks; **E.** 50%. For the ELECTRA line, we use: A. PL_{RTD}; B. Rand; C. RTD only; D. Unlimited; E. 15%. A. The Anomaly Score used over classification probabilities shows that PL_{RTD} (used in DATE) is the best in predicting anomalies, meaning that our self-supervised classification task is well defined, with few ambiguous samples; B. A learned Generator does not justify its training cost; C. RMD Loss proved to be complementary with RTD Loss, their combination (in DATE) increasing the score and stabilizes the training; D+E. (table from: [Manolache et al., 2021](#))

Entropy (NE), and our modified Pseudo Label (PL_{RTD}). These scores are computed using the softmax probabilities from the final classification layer of the discriminator. PL is an ideal score if the self-supervised task manages to build and learn well separated classes. The way we formulate our mask prediction task enables a very good class separation, as theoretically proved in detail in the supplementary material [A.1](#). Therefore, PL_{RTD} proves to be significantly better in detecting the anomalies compared with MP and NE metrics, which try to compensate for ambiguous samples.

B. Generator performance. We tested the importance of having a learned generator, by using a one-layer Transformer with hidden size 16 (small) or 64 (large). The *random generator* proved to be better than both parameterized generators.

	Inlier class	iForest best	OC-SVM best	CVDD best	mSVDD best	DATE (Ours)
20 News	comp	66.1	78.0	74.0	84.4	92.1
	rec	59.4	70.0	60.6	77.7	83.4
	sci	57.8	64.2	58.2	67.3	69.7
	misc	62.4	62.1	75.7	76.2	86.0
	pol	65.3	76.1	71.5	79.2	81.9
	rel	71.4	78.9	78.1	83.6	86.1
AG News	business	79.6	79.9	84.0	-	90.0
	sci	76.9	80.7	79.0	-	84.0
	sports	84.7	92.4	89.9	-	95.9
	world	73.2	83.2	79.6	-	90.1

Table 8: Semi-supervised performance (AUROC%). We test on the 20Newsgroups and AG News datasets, by comparing DATE against several strong baselines and state-of-the-art solutions (with multiple variations, choosing the best score per split as detailed in Sec. 4.2.3): iForest, OC-SVM, CVDD, and mSVDD. We largely outperform competitors with an average improvement of 4.7% on 20Newsgroups and 6.9% on AG News compared with the next best solution. Note that DATE uses the same set of hyper-parameters per dataset. (table adapted from: [Manolache et al., 2021](#))

C. Loss function. For the final loss, we combined RTD (which sanctions the prediction per token) with our RMD (which enforces the detection of the mask applied on the entire sequence). We also train our model with RTD or RMD only, obtaining weaker results. This proves that combining losses with supervisions at different scales (locally: token-level and globally: sequence-level) improves AD performance. Moreover, when using only the RTD loss, the training can be very unstable (AUROC score peaks in the early stages, followed by a steep decrease). With the combined loss, the AUROC is only stationary or increases with time.

D. Masking patterns. The mask patterns are the root of our task formulation, hiding a part of the input tokens and asking the discriminator to classify them. As experimentally shown, having more mask patterns is better, encouraging increased expressiveness in the embeddings. Too many masks on the other hand can make the task too difficult for the discriminator and our ablation shows that having more masks does not add any benefit after a point. We validate the percentage of masked tokens in **E. Mask percent** ablation.

4.2.3 Comparison with other AD methods

We compare our method against classical AD baselines like Isolation Forest ([F. T. Liu et al., 2008](#)) and existing state-of-the-art One Class SVM ([L. M. Manevitz and Yousef, 2002](#)), CVDD ([Ruff, Zemlyanskiy, et al., 2019](#)), and

mSVDD (Hu et al., 2021) We outperform all previously reported performances on all 20Newsgroups splits by a large margin: 13.5% over the best reported CVDD and 11.7% over the best OCSVM, and 4.7% over the best mSVDD, as shown in Tab. 8. In contrast, DATE uses the same set of hyperparameters for a dataset, for all splits. For a proper comparison, we keep the same experimental setup as the one introduced in (Ruff, Zemlyanskiy, et al., 2019).

ISOLATION FOREST. We apply it over fastText or GloVe embeddings, varying the number of estimators (64, 100, 128, 256), and choosing the best model per split. In the unsupervised AD setup, we manually set the percent of outliers in the train set.

OCSVM. We use the One-Class SVM model implemented in the CVDD work. For each split, we choose the best configuration (fastText vs Glove, rbf vs linear kernel, $\nu \in [0.05, 0.1, 0.2, 0.5]$).

CVDD. For each split, we chose the best column out of all reported context sizes (r). The scores reported using the c^* context vector depends on the ground truth and it only reveals “*the potential of contextual anomaly detection*”, as the authors mention.

MSVDD. This model (Hu et al., 2021) is the current state-of-the-art solution for AD on text. For each split, we chose the best column out of all reported configurations, excluding those which use negative supervision.

4.2.4 Unsupervised AD

We further analyse how our algorithm works in a fully unsupervised scenario, namely when the training set contains some anomalous samples (which

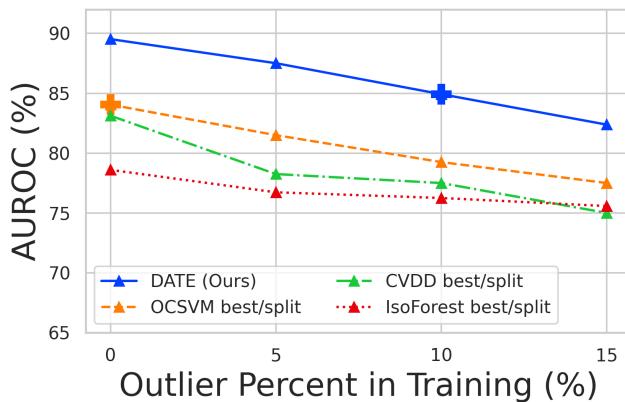
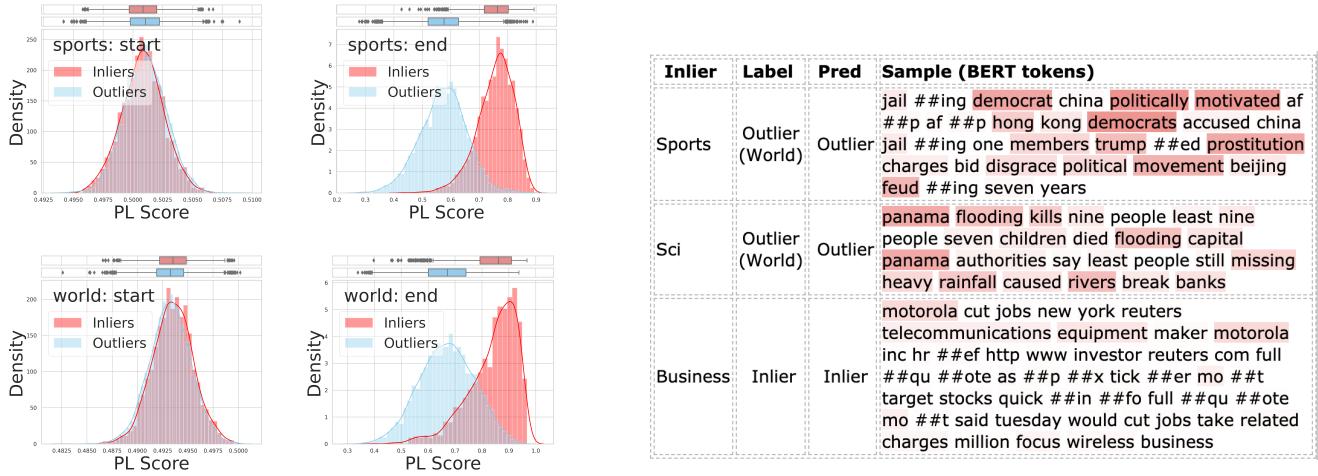


Figure 17: Unsupervised AD. We test the performance of our method when training on impure data, which contains anomalies in various percentages: 0%-15%. The performance slowly decreases when we increase the anomaly percentage, but even at 10% contamination, it is still better than state-of-the-art results on self-supervised anomaly detection in text (Ruff, Zemlyanskiy, et al., 2019), which trains on 0% anomalous data, proving the robustness of our method. Experiments were done on all AG News splits. (figure from: Manolache et al., 2021)



(a) Normalized histogram for anomaly score. We see how the anomaly score (PL) distribution varies among inliers and outliers, from the beginning of the training (1st column) to the end (2nd column), where the two become well separated, with relatively low interference between classes. Note that a better separation is correlated with high performance (1st line split has 95.9% AUROC, while the 2nd has only 90.1%).

(b) Qualitative examples. Lower scores are shown in a more intense red, and point to anomalies. In the 1st example, words from politics are flagged as anomalous for sports. In the 2nd one, words describing natural events are outliers for technology. In the 3rd row, while few words have higher anomaly potential for the business domain, most of them are appropriate.

Figure 18: Qualitative results. a) Evolution of the anomaly scores on the sports and sci test sets. b) Anomalous words for three different scenarios. (figures from: [Manolache et al., 2021](#))

we treat as normal ones). By definition, the quantity of anomalous events in the training set is significantly lower than the normal ones. In this experiment, we show how our algorithm performance is influenced by the percentage of anomalies in training data. Our method proves to be extremely robust, surpassing state-of-the-art, which is a semi-supervised solution, trained over a clean dataset (with 0% anomalies), even at 10% contamination, with +0.9% in AUROC (see Fig. 17). By achieving outstanding performance in the unsupervised setting, we make unsupervised AD in text competitive against other semi-supervised methods. The reported scores are the mean over all AG News splits. We compare against the same methods presented in Sec. 4.2.3.

4.2.5 Qualitative results

We show in Fig. 18b how DATE performs in identifying anomalies in several examples. Each token is colored based on its PL score.

We also show how our anomaly score (PL) is distributed among normal vs abnormal samples. For visualization, we chose two splits from AG News and report the scores from the beginning of the training to the end. We see in Fig. 18a that, even though at the beginning, the outliers' distribution of scores fully overlaps with the inliers, at the end of training the two are well separated, proving the effectiveness of our method.

5

CONCLUSION AND FURTHER DEVELOPMENTS

We propose DATE, a model for tackling Anomaly Detection in Text, and formulate an innovative self-supervised task, based on masking parts of the initial input and predicting which mask pattern was used. After masking, a generator reconstructs the initially masked tokens and the discriminator predicts which mask was used. We optimize a loss composed of both token and sequence-level parts, taking advantage of powerful supervision, coming from two independent pathways, which stabilizes learning and improves AD performance. For computing the anomaly score, we alleviate the burden of aggregating predictions from multiple transformations by introducing an efficient variant of the Pseudo Label score, which is applied per token, only on the original input. We show that this score separates very well the abnormal entries from normal ones, leading DATE to outperform state-of-the-art results on all AD splits from 20Newsgroups and AG News datasets, by a large margin, both in the semi-supervised and unsupervised AD settings.

We would like to further investigate and analyse other self-supervision tasks for AD in text. In its current form, DATE is randomly masking tokens from the input sequence. It would be interesting to see how masking specific parts of speech would affect the anomaly detection capabilities. Another intriguing approach to self-supervision is by contrastive learning ([T. Chen, Kornblith, Norouzi, et al., 2020](#)). One way in which we could adapt DATE to contrastive learning is to apply different augmentations (by applying different masking patterns) and minimize the cosine distance between the representations which come from the same text corrupted in different ways while maximizing the cosine distance between different texts. An alternative way of generating anomaly scores is by using energy-based models. Recently, EBMs have become increasingly popular. An EBM can theoretically be used as an anomaly detector by directly evaluating the energy function ([Ruff, Kauffmann, et al., 2020](#)) to produce anomaly scores. Models such as ELECTRA can be formulated as energy-based models ([Clark, Luong, Q. Le, et al., 2020](#)), this raises the question if they can also be used to detect anomalies. Finally, DATE demonstrates capabilities for authorship detection and can detect stylistic differences between texts written by different authors (see supplementary material [B.2](#)). We would like to further evaluate deep anomaly detectors for such tasks.

BIBLIOGRAPHY

- Adams, R. P. and D. J. C. MacKay
2007 *Bayesian Online Changepoint Detection*, arXiv: [0710.3742 \[stat.ML\]](#).
- Aggarwal, C. C.
2015 “Outlier analysis”, in *Data mining*, Springer, pp. 237-263.
2013 “Outlier Ensembles: Position Paper”, *SIGKDD Explor. Newsl.*, 14, 2 (Apr. 2013), pp. 49-58, ISSN: 1931-0145, DOI: [10.1145/2481244.2481252](#), <https://doi.org/10.1145/2481244.2481252>.
- Aggarwal, C. C. and C. K. Reddy
2014 (eds.), *Data Clustering: Algorithms and Applications*, CRC Press, ISBN: 978-1-46-655821-2, <http://www.crcpress.com/product/isbn/9781466558212>.
- Badia, A. P., B. Piot, S. Kapiturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, and C. Blundell
2020 “Agent57: Outperforming the Atari Human Benchmark”, in *Proceedings of the 37th International Conference on Machine Learning*, ed. by H. D. III and A. Singh, Proceedings of Machine Learning Research, PMLR, vol. 119, pp. 507-517, <http://proceedings.mlr.press/v119/badia20a.html>.
- Baltoiu, A., A. Patrascu, and P. Irofti
2020 “Graph Anomaly Detection Using Dictionary Learning”, *IFAC-PapersOnLine*, 53 (Jan. 2020), pp. 3551-3558, DOI: [10.1016/j.ifacol.2020.12.1731](#).
- Banoth, L., M. Teja, M. Saicharan, and N. Chandra
2017 “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, *International Journal of Research*, 4, pp. 406-412.
- Beltagy, I., M. E. Peters, and A. Cohan
2020 “Longformer: The Long-Document Transformer”, *arXiv:2004.05150*.
- Beltagy, I., K. Lo, and A. Cohan
2019 “SciBERT: A Pretrained Language Model for Scientific Text”, 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)*, Association for Computational Linguistics, Hong Kong, China, pp. 3615-3620, DOI: [10.18653/v1/D19-1371](#), <https://www.aclweb.org/anthology/D19-1371>.
- Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin
2003 “A Neural Probabilistic Language Model”, *J. Mach. Learn. Res.*, 3, null (Mar. 2003), pp. 1137-1155, ISSN: 1532-4435.

- Berner, C., G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. W. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang
 2019 "Dota 2 with Large Scale Deep Reinforcement Learning", *ArXiv*, abs/1912.06680.
- Bevendorff, J., B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel Pardo, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, and E. Zangerle
 2020 "Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection", in pp. 372-383, ISBN: 978-3-030-58218-0, DOI: [10.1007/978-3-030-58219-7_25](https://doi.org/10.1007/978-3-030-58219-7_25).
- Bishop, C. M.
 2006 *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, ISBN: 0387310738.
- Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov
 2017 "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, 5, pp. 135-146, DOI: [10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051), <https://www.aclweb.org/anthology/Q17-1010>.
- Boser, B. E., I. M. Guyon, and V. N. Vapnik
 1992 "A Training Algorithm for Optimal Margin Classifiers", in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, pp. 144-152, ISBN: 089791497X, DOI: [10.1145/130385.130401](https://doi.org/10.1145/130385.130401), <https://doi.org/10.1145/130385.130401>.
- Bowyer, K. W., N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer
 2011 "SMOTE: Synthetic Minority Over-sampling Technique", *CoRR*, abs/1106.1813, arXiv: [1106.1813](https://arxiv.org/abs/1106.1813), [http://arxiv.org/abs/1106.1813](https://arxiv.org/abs/1106.1813).
- Braei, M. and S. Wagner
 2020 "Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art" (Apr. 2020).
- Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander
 2000 "LOF: Identifying Density-Based Local Outliers", in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, Association for Computing Machinery, Dallas, Texas, USA, pp. 93-104, ISBN: 1581132174, DOI: [10.1145/342009.335388](https://doi.org/10.1145/342009.335388), <https://doi.org/10.1145/342009.335388>.

- Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei
- 2020 "Language Models are Few-Shot Learners", in *Advances in Neural Information Processing Systems*, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Curran Associates, Inc., vol. 33, pp. 1877-1901, <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>.
- Caron, M., H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin
- 2021 "Emerging Properties in Self-Supervised Vision Transformers", *arXiv preprint arXiv:2104.14294*.
- Caruana, R., D. L. Silver, J. Baxter, T. M. Mitchell, L. Y. Pratt, and S. Thrun
- 1995 *Learning to learn: knowledge consolidation and transfer in inductive systems.*
- Chandola, V., A. Banerjee, and V. Kumar
- 2009 "Anomaly Detection: A Survey", *ACM Comput. Surv.*, 41, 3 (July 2009), ISSN: 0360-0300, DOI: [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882), <https://doi.org/10.1145/1541880.1541882>.
- Cheeseman, P. and J. Stutz
- 1996 "Bayesian Classification (AutoClass): Theory and Results", in *Advances in Knowledge Discovery and Data Mining*, American Association for Artificial Intelligence, USA, pp. 153-180, ISBN: 0262560976.
- Chellapilla, K., S. Puri, and P. Simard
- 2006 "High Performance Convolutional Neural Networks for Document Processing" (Oct. 2006).
- Chen, J., S. Sathe, C. C. Aggarwal, and D. S. Turaga
- 2017 "Outlier Detection with Autoencoder Ensembles", in *SDM*.
- Chen, M., A. Radford, R. Child, J. Wu, H. Jun, P. Dhariwal, D. Luan, and I. Sutskever
- 2020 "Generative Pretraining from Pixels".
- Chen, T., S. Kornblith, M. Norouzi, and G. Hinton
- 2020 "A Simple Framework for Contrastive Learning of Visual Representations", in *Proceedings of the 37th International Conference on Machine Learning*, ed. by H. D. III and A. Singh, Proceedings of Machine Learning Research, PMLR, vol. 119, pp. 1597-1607, <http://proceedings.mlr.press/v119/chen20j.html>.

- Chen, T., S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton
 2020 “Big Self-Supervised Models are Strong Semi-Supervised Learners”, in *Advances in Neural Information Processing Systems*, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Curran Associates, Inc., vol. 33, pp. 22243-22255, <https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf>.
- Cho, K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio
 2014a “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1724-1734, DOI: [10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179), <https://www.aclweb.org/anthology/D14-1179>.
- Cho, K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio
 2014b “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1724-1734, DOI: [10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179), <https://www.aclweb.org/anthology/D14-1179>.
- Cireşan, D., U. Meier, J. Masci, and J. Schmidhuber
 2012 “Multi-Column Deep Neural Network for Traffic Sign Classification”, *Neural networks : the official journal of the International Neural Network Society*, 32 (Feb. 2012), pp. 333-8, DOI: [10.1016/j.neunet.2012.02.023](https://doi.org/10.1016/j.neunet.2012.02.023).
- Clark, K., M.-T. Luong, Q. Le, and C. D. Manning
 2020 “Pre-Training Transformers as Energy-Based Cloze Models”, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 285-294, DOI: [10.18653/v1/2020.emnlp-main.20](https://doi.org/10.18653/v1/2020.emnlp-main.20), <https://www.aclweb.org/anthology/2020.emnlp-main.20>.
- Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning
 2020 “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”, *ArXiv*, abs/2003.10555.
- Collobert, R. and J. Weston
 2008 “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning”, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, Association for Computing Machinery, Helsinki, Finland, pp. 160-167, ISBN: 9781605582054, DOI: [10.1145/1390156.1390177](https://doi.org/10.1145/1390156.1390177), <https://doi.org/10.1145/1390156.1390177>.

- Deecke, L., R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft
 2019 "Image Anomaly Detection with Generative Adversarial Networks", in *Machine Learning and Knowledge Discovery in Databases*, ed. by M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Springer International Publishing, Cham, pp. 3-17, ISBN: 978-3-030-10925-7.
- Deng, L.
 2012 "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]", *IEEE Signal Processing Magazine*, 29, pp. 141-142.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova
 2018 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *CoRR*, abs/1810.04805, arXiv: [1810.04805](https://arxiv.org/abs/1810.04805), <http://arxiv.org/abs/1810.04805>.
- De Vries, W., A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim
 2019 "BERTje: A Dutch BERT Model", *CoRR*, abs/1912.09582, arXiv: [1912.09582](https://arxiv.org/abs/1912.09582), <http://arxiv.org/abs/1912.09582>.
- Dorronsoro, J., F. Ginel, C. Sanchez, and C. Cruz
 1997 "Neural fraud detection in credit card operations", *IEEE transactions on neural networks*, 84, pp. 827-34.
- Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby
 2020 "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *CoRR*, abs/2010.11929, arXiv: [2010.11929](https://arxiv.org/abs/2010.11929), <https://arxiv.org/abs/2010.11929>.
- Dumitrescu, S. D., A.-M. Avram, and S. Pyysalo
 2020 "The birth of Romanian BERT", *arXiv preprint arXiv:2009.08712*.
- Edgeworth, F. Y.
 1887 "XLI. On discordant observations", *Philosophical Magazine Series 1*, 23, pp. 364-375.
- Emmott, A. F., S. Das, T. Dietterich, A. Fern, and W.-K. Wong
 2013 "Systematic Construction of Anomaly Detection Benchmarks from Real Data", in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, ODD '13, Association for Computing Machinery, Chicago, Illinois, pp. 16-21, ISBN: 9781450323352, DOI: [10.1145/2500853.2500858](https://doi.org/10.1145/2500853.2500858), <https://doi.org/10.1145/2500853.2500858>.
- Fukushima, K.
 1980 "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", *Biological Cybernetics*, 36, pp. 193-202.

- Georgescu, M.-I., A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah
 2020 “Anomaly Detection in Video via Self-Supervised and Multi-Task Learning”, *CoRR*, abs/2011.07491, arXiv: 2011.07491, <https://arxiv.org/abs/2011.07491>.
- Golan, I. and R. El-Yaniv
 2018 “Deep Anomaly Detection Using Geometric Transformations”, *CoRR*, abs/1805.10917, arXiv: 1805.10917, <http://arxiv.org/abs/1805.10917>.
- Grill, J.-B., F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko
 2020 “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning”, *ArXiv*, abs/2006.07733.
- Hawkins, D. M.
 1980 *Identification of outliers*, Springer, vol. 11.
- Hawkins, S., H. He, G. J. Williams, and R. A. Baxter
 2002 “Outlier Detection Using Replicator Neural Networks”, in *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2000, Springer-Verlag, Berlin, Heidelberg, pp. 170-180, ISBN: 3540441239.
- He, K., X. Zhang, S. Ren, and J. Sun
 2016 “Deep Residual Learning for Image Recognition”, in pp. 770-778, DOI: 10.1109/CVPR.2016.90.
- Hochreiter, S. and J. Schmidhuber
 1997 “Long Short-Term Memory”, *Neural Comput.*, 9, 8 (Nov. 1997), pp. 1735-1780, ISSN: 0899-7667, DOI: 10.1162/neco.1997.9.8.1735, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hu, C., Y. Feng, H. Kamigaito, H. Takamura, and M. 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: Main Volume*, Association for Computational Linguistics, Online, pp. 3378-3390, <https://www.aclweb.org/anthology/2021.eacl-main.296>.
- Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger
 2017 “Densely Connected Convolutional Networks”, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261-2269, DOI: 10.1109/CVPR.2017.243.
- Irofti, P. and A. Baltoiu
 2020 “Unsupervised Dictionary Learning for Anomaly Detection”, *CoRR*, abs/2003.00293, arXiv: 2003.00293, <https://arxiv.org/abs/2003.00293>.

- Jouppi, N. P., C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-I. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leahy, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon
- 2017 "In-Datacenter Performance Analysis of a Tensor Processing Unit", in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ISCA '17, Association for Computing Machinery, Toronto, ON, Canada, pp. 1-12, ISBN: 9781450348928, DOI: [10.1145/3079856.3080246](https://doi.org/10.1145/3079856.3080246), <https://doi.org/10.1145/3079856.3080246>.
- Kammerer, K., B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert
- 2019 "Anomaly Detections for Manufacturing Systems Based on Sensor Data—Insights into Two Challenging Real-World Production Settings", *Sensors (Basel, Switzerland)*, 19.
- Kitaev, N., L. Kaiser, and A. Levskaya
- 2020 "Reformer: The Efficient Transformer", *CoRR*, abs/2001.04451, arXiv: [2001.04451](https://arxiv.org/abs/2001.04451), <https://arxiv.org/abs/2001.04451>.
- Knorr, E. M. and R. T. Ng
- 1997 "A Unified Notion of Outliers: Properties and Computation", in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, KDD'97, AAAI Press, Newport Beach, CA, pp. 219-222.
- Krizhevsky, A.
- 2012 "Learning Multiple Layers of Features from Tiny Images", *University of Toronto* (May 2012).
- Krizhevsky, A., I. Sutskever, and G. E. Hinton
- 2012 "ImageNet Classification with Deep Convolutional Neural Networks", in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, Curran Associates Inc., Lake Tahoe, Nevada, pp. 1097-1105.
- Kuchaiev, O. and B. Ginsburg
- 2017 "Factorization tricks for LSTM networks" (Mar. 2017).
- Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut
- 2019 "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", *CoRR*, abs/1909.11942, arXiv: [1909.11942](https://arxiv.org/abs/1909.11942), [http://arxiv.org/abs/1909.11942](https://arxiv.org/abs/1909.11942).

- Lang, K. and J. Rennie
 1999 “20 Newsgroups Dataset”, ed. by empty, <http://people.csail.mit.edu/jrennie/20Newsgroups/>.
- LeCun, Y.
 2021 *Self-supervised learning: The dark matter of intelligence*, <https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/>.
- LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel
 1990 “Handwritten Digit Recognition with a Back-Propagation Network”, in *Advances in Neural Information Processing Systems*, ed. by D. Touretzky, Morgan-Kaufmann, vol. 2, <https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fc54-Paper.pdf>.
- LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel
 1989 “Backpropagation Applied to Handwritten Zip Code Recognition”, *Neural Comput.*, 1, 4 (Dec. 1989), pp. 541–551, ISSN: 0899-7667, DOI: [10.1162/neco.1989.1.4.541](https://doi.org/10.1162/neco.1989.1.4.541), <https://doi.org/10.1162/neco.1989.1.4.541>.
- LeCun, Y. and C. Cortes
 2010 “MNIST handwritten digit database”, <http://yann.lecun.com/exdb/mnist/>.
- Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang
 2019 “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, *Bioinformatics* (Sept. 2019), ISSN: 1367-4803, DOI: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682), <https://doi.org/10.1093/bioinformatics/btz682>.
- Lewis, D.
 1999 *Reuters-21578 Dataset*, <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>.
- Liu, F. T., K. M. Ting, and Z.-H. Zhou
 2008 “Isolation Forest”, in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM ’08, IEEE Computer Society, USA, pp. 413–422, ISBN: 9780769535029, DOI: [10.1109/ICDM.2008.17](https://doi.org/10.1109/ICDM.2008.17), <https://doi.org/10.1109/ICDM.2008.17>.
- Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov
 2019 “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, *CoRR*, abs/1907.11692, arXiv: [1907.11692](https://arxiv.org/abs/1907.11692), [http://arxiv.org/abs/1907.11692](https://arxiv.org/abs/1907.11692).
- Loshchilov, I. and F. Hutter
 2019 “Decoupled Weight Decay Regularization”, in *ICLR 2019*.

- Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts
 2011 "Learning Word Vectors for Sentiment Analysis", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 142-150, <http://www.aclweb.org/anthology/P11-1015>.
- Manevitz, L. and M. Yousef
 2001 "One-Class SVMs for Document Classification", *J. Mach. Learn. Res.*, 2, pp. 139-154.
- Manevitz, L. M. and M. Yousef
 2002 "One-Class SVMs for Document Classification", *J. Mach. Learn. Res.*, 2 (Mar. 2002), pp. 139-154, ISSN: 1532-4435.
- Manolache, A., F. Brad, and E. 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*, Association for Computational Linguistics, Online, pp. 267-277, <https://www.aclweb.org/anthology/2021.naacl-main.25>.
- Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot
 2020 "CamemBERT: a Tasty French Language Model", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 7203-7219, DOI: [10.18653/v1/2020.acl-main.645](https://doi.org/10.18653/v1/2020.acl-main.645), <https://www.aclweb.org/anthology/2020.acl-main.645>.
- Mikolov, T., K. Chen, G. S. Corrado, and J. Dean
 2013 *Efficient Estimation of Word Representations in Vector Space*, <http://arxiv.org/abs/1301.3781>.
- Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean
 2013 "Distributed Representations of Words and Phrases and Their Compositionality", in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, Curran Associates Inc., Lake Tahoe, Nevada, pp. 3111-3119.
- Mikolov, T., W.-t. Yih, and G. Zweig
 2013 "Linguistic Regularities in Continuous Space Word Representations", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, pp. 746-751, <https://www.aclweb.org/anthology/N13-1090>.
- Mikolov, T., Q. V. Le, and I. Sutskever
 2013 "Exploiting Similarities among Languages for Machine Translation", *CoRR*, abs/1309.4168, arXiv: [1309.4168](https://arxiv.org/abs/1309.4168), <http://arxiv.org/abs/1309.4168>.

- Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller
 2013 “Playing Atari with Deep Reinforcement Learning” (Dec. 2013).
- Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng
 2011 “Reading Digits in Natural Images with Unsupervised Feature Learning”, in
- Nisioi, S., E. Rabinovich, L. P. Dinu, and S. Wintner
 2016 “A Corpus of Native, Non-native and Translated Texts”, in *LREC*.
- Paszke, A., S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer
 2017 *Automatic differentiation in PyTorch*.
- Pennington, J., R. Socher, and C. D. Manning
 2014 “Glove: Global Vectors for Word Representation.”, in *EMNLP*, vol. 14, pp. 1532–1543.
- Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer
 2018 *Deep contextualized word representations*, cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready, <http://arxiv.org/abs/1802.05365>.
- Radford, A. and K. Narasimhan
 2018 “Improving Language Understanding by Generative Pre-Training”, in
- Radford, B., L. Apolonio, A. Trias, and J. Simpson
 2018 “Network Traffic Anomaly Detection Using Recurrent Neural Networks” (Mar. 2018).
- Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang
 2016 “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, pp. 2383–2392, DOI: [10.18653/v1/D16-1264](https://doi.org/10.18653/v1/D16-1264), <https://www.aclweb.org/anthology/D16-1264>.
- Rousseeuw, P. and K. Driessen
 1999 “A Fast Algorithm for the Minimum Covariance Determinant Estimator”, *Technometrics*, 41 (Aug. 1999), pp. 212–223, DOI: [10.1080/00401706.1999.10485670](https://doi.org/10.1080/00401706.1999.10485670).
- Ruff, L., J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller
 2020 “A Unifying Review of Deep and Shallow Anomaly Detection”, *CoRR*, abs/2009.11732, arXiv: [2009.11732](https://arxiv.org/abs/2009.11732), <https://arxiv.org/abs/2009.11732>.

- Ruff, L., Y. Zemlyanskiy, R. Vandermeulen, T. Schnake, and M. 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*, Association for Computational Linguistics, Florence, Italy, pp. 4061-4071, DOI: [10.18653/v1/P19-1398](https://doi.org/10.18653/v1/P19-1398), <https://www.aclweb.org/anthology/P19-1398>.
- Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei
- 2015 "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision (IJCV)*, 115, 3, pp. 211-252, DOI: [10.1007/s11263-015-0816-y](https://doi.org/10.1007/s11263-015-0816-y).
- Sakurada, M. and T. Yairi
- 2014 "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction", in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, Association for Computing Machinery, Gold Coast, Australia QLD, Australia, pp. 4-11, ISBN: 9781450331593, DOI: [10.1145/2689746.2689747](https://doi.org/10.1145/2689746.2689747), <https://doi.org/10.1145/2689746.2689747>.
- Sharafaldin, I., A. Habibi Lashkari, and A. Ghorbani
- 2018 "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", in pp. 108-116, DOI: [10.5220/0006639801080116](https://doi.org/10.5220/0006639801080116).
- Shazeer, N., A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean
- 2017 "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", in <https://openreview.net/pdf?id=B1ckMDqlg>.
- Silver, D., A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis
- 2016 "Mastering the game of Go with deep neural networks and tree search", *Nature*, 529 (Jan. 2016), pp. 484-489, DOI: [10.1038/nature16961](https://doi.org/10.1038/nature16961).
- Simonyan, K. and A. Zisserman
- 2014 "Very Deep Convolutional Networks for Large-Scale Image Recognition", *CoRR*, abs/1409.1556, <http://arxiv.org/abs/1409.1556>.
- Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich
- 2015 "Going deeper with convolutions", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, DOI: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).

- Tax, D. M. J. and R. P. W. Duin
 2004 "Support Vector Data Description", *Mach. Learn.*, 54, 1 (Jan. 2004), pp. 45-66, ISSN: 0885-6125, DOI: [10.1023/B:MACH.0000008084.60811.49](https://doi.org/10.1023/B:MACH.0000008084.60811.49).
- Taylor, W. L.
 1953 ""Cloze Procedure": A New Tool for Measuring Readability", *Journalism Quarterly*, 30, 4, pp. 415-433, DOI: [10.1177/107769905303000401](https://doi.org/10.1177/107769905303000401), eprint: <https://doi.org/10.1177/107769905303000401>, <https://doi.org/10.1177/107769905303000401>.
- Tolstikhin, I., N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy
 2021 *MLP-Mixer: An all-MLP Architecture for Vision*, arXiv: [2105.01601](https://arxiv.org/abs/2105.01601) [cs.CV].
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin
 2017 "Attention Is All You Need", *CoRR*, abs/1706.03762, arXiv: [1706.03762](https://arxiv.org/abs/1706.03762), <http://arxiv.org/abs/1706.03762>.
- Vinyals, O., I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, A. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver
 2019 "Grandmaster level in StarCraft II using multi-agent reinforcement learning", *Nature*, pp. 1-5.
- Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang
 1989 "Phoneme recognition using time-delay neural networks", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37, 3, pp. 328-339, DOI: [10.1109/29.21701](https://doi.org/10.1109/29.21701).
- Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman
 2018 "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", in *BlackboxNLP@EMNLP*.
- Wang, H., M. Li, F. Ma, S.-L. Huang, and L. Zhang
 2019 "Unsupervised Anomaly Detection via Generative Adversarial Networks: Poster Abstract", in *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, IPSN '19, Association for Computing Machinery, Montreal, Quebec, Canada, pp. 313-314, ISBN: 9781450362849, DOI: [10.1145/3302506.3312605](https://doi.org/10.1145/3302506.3312605), <https://doi.org/10.1145/3302506.3312605>.
- Wang, S., Y. Zeng, X. Liu, E. Zhu, J. Yin, C. Xu, and M. Kloft
 2019 "Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discriminative Network", in *NeurIPS*.

- Wei, Q., Y. Ren, R. Hou, B. Shi, J. Y. Lo, and L. Carin
 2018 "Anomaly detection for medical images based on a one-class classification", in *Medical Imaging 2018: Computer-Aided Diagnosis*, ed. by N. Petrick and K. Mori, International Society for Optics and Photonics, SPIE, vol. 10575, pp. 375-380, doi: [10.1117/12.2293408](https://doi.org/10.1117/12.2293408), <https://doi.org/10.1117/12.2293408>.
- Weng, L.
 2018 *Attention? Attention!*, <https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>.
- Wolpert, D. and W. Macready
 1997 "No free lunch theorems for optimization", *IEEE Transactions on Evolutionary Computation*, 1, 1, pp. 67-82, doi: [10.1109/4235.585893](https://doi.org/10.1109/4235.585893).
- Xia, Y., X. Cao, F. Wen, G. Hua, and J. Sun
 2015 "Learning Discriminative Reconstructions for Unsupervised Outlier Removal", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Xiao, H., K. Rasul, and R. Vollgraf
 2017 "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms", *CoRR*, abs/1708.07747, arXiv: [1708.07747](https://arxiv.org/abs/1708.07747), <http://arxiv.org/abs/1708.07747>.
- Xiong, Y., Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh
 2021 *Nyströmlformer: A Nyström-Based Algorithm for Approximating Self-Attention*, arXiv: [2102.03902 \[cs.CL\]](https://arxiv.org/abs/2102.03902).
- Zagoruyko, S. and N. Komodakis
 2016 "Wide Residual Networks", *CoRR*, abs/1605.07146, arXiv: [1605.07146](https://arxiv.org/abs/1605.07146), <http://arxiv.org/abs/1605.07146>.
- Zellers, R., Y. Bisk, R. Schwartz, and Y. Choi
 2018 "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 93-104, doi: [10.18653/v1/D18-1009](https://www.aclweb.org/anthology/D18-1009), <https://www.aclweb.org/anthology/D18-1009>.
- Zhang, X., J. J. Zhao, and Y. LeCun
 2015 "Character-level Convolutional Networks for Text Classification", in *NeurIPS*.
- Zhou, C. and R. 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, Association for Computing Machinery, Halifax, NS, Canada, pp. 665-674, ISBN: 9781450348874, doi: [10.1145/3097983.3098052](https://doi.org/10.1145/3097983.3098052), <https://doi.org/10.1145/3097983.3098052>.

- Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He
2020 "A Comprehensive Survey on Transfer Learning", *Proceedings of the IEEE*, PP (July 2020), pp. 1-34, DOI: [10.1109/JPROC.2020.3004555](https://doi.org/10.1109/JPROC.2020.3004555).
- Zisserman, A.
2018 *Self-Supervised Learning*, <https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf>.

Appendices

A | HYPERPARAMETERS AND IMPLEMENTATION

A.1 DISJOINT PATTERNS ANALYSIS

We start from two observations regarding the performance of DATE, our Anomaly Detection algorithm. First, a discriminative task performs better if the classes are well separated (Deng, 2012) and there is a low probability for confusions. Second, the PL score for anomalies achieves the best performance when the probability distribution for its input is clearly separated. Intuitively, for three classes, $\text{PL}([0.9, 0.05, 0.05])$ is better than $\text{PL}([0.5, 0.3, 0.2])$ because it allows PL to give either near 1 score if the class is correct, either near 0 score if it is not, avoiding the zone in the middle where we depend on a well chosen threshold.

Since the separation between the mask patterns greatly influences our final performance, we next analyze our AD task from the mask pattern generation point of view. Ideally, we want to have a sense of how disjoint our randomly sampled patterns are and make an informed choice for the pattern generation hyper-parameters.

First, we start by computing an upper bound for the probability of having two patterns with at least p common masked points. We have $\binom{S}{M}$ patterns, where S is the sequence length and M is the number of masked tokens. We fix the first p positions that we want to mask in any pattern. Considering those fixed masks, the probability of having a sequence with M masked tokens, with p tokens in the first positions is r :

$$r = \frac{\binom{S-p}{M-p}}{\binom{S}{M}}. \quad (28)$$

Next, the probability that two sequences mask the first p tokens is r^2 . But we can choose those two positions in a $\binom{S}{p}$ ways. So the probability that any two sequences have at least p common masked tokens is lower than UB_2 :

$$\text{UB}_2 = \binom{S}{p} r^2 \quad (29)$$

Next, out of our generated patterns, we sample N masks, so the probability becomes less than the upper bound UB_N :

$$\begin{aligned} \text{UB}_N &= \binom{N}{2} \text{UB}_2 = \binom{N}{2} \binom{S}{p} r^2 \\ &= \binom{N}{2} \binom{S}{p} \left(\frac{\binom{S-p}{M-p}}{\binom{S}{M}} \right)^2. \end{aligned} \quad (30)$$

In our experiments, the sequence length is $S = 128$ and we chose the number of masked tokens to be between 15% and 50% (M between 19 and

64). We consider that two patterns are disjoint when they have less than p masked tokens in common, for N sampled patterns.

The probability that any two patterns collide (have more than p masked tokens in common) is very low. We compute several values for its upper bound: $UB_{N=100,p=12} = 5e-4$, $UB_{N=100,p=15} = 1e-9$, $UB_{N=10,p=15} = 1e-11$, $UB_{N=10,p=13} = 1e-7$.

In conclusion, for our specific setup, the probability for two masks to largely overlap (large p compared with S) is extremely small, ensuring us a good performance in the discriminator. We take advantage of this property of our pretext task by combining the discriminator output probabilities with the PL score.

A.2 MODEL IMPLEMENTATION

We add next more details on the implementation of the modules: from the ablation experiments in Tab. 7, **Generator (small)**: 1 Transformer layer, with 4 self-attention heads, token and positional embeddings of size 128, hidden layer of size 16, feedforward layer of sizes 1024 and 16; **Generator (large)**: 1 Transformer layer, with 4 self-attention heads, token and positional embeddings of size 128, hidden layer of size 64, feedforward layer of sizes 1024 and 64; As empirical experiments showed us, we choose a *random Generator* (samples were drawn from a uniform distribution over the vocabulary) in our final model. **Discriminator**: 4 Transformer layers, each with 4 self-attention heads, hidden layers of size 256, feedforward layers of sizes of 1024 and 256, 128-dimensional token and positional embeddings, which are *tied* with the generator. For other unspecified hyper-parameters we use the ones in ELECTRA-Small model. **Prediction Heads**: both heads have 2 linear layers separated by a non-linearity, ending in a classification. **Loss weights**: We set the RTD λ weight to 50 as in (Clark, Luong, Q. V. Le, et al., 2020), and the RMD μ weight to 100.

B

OTHER RESULTS

B.1 MORE QUALITATIVE AND QUANTITATIVE RESULTS ON AG NEWS

In Fig. 19 we show more qualitative results, trained on different inliers. To encourage further more detailed comparisons, we report the AUPR metric on AG News for inliers and outliers (see Tab. 9). When all the other metrics are almost saturated, we notice that AUPR-in better captures the performance on a certain split.

Inlier	Label	Pred	Sample (BERT tokens)
Sports	Outlier (Business)	Outlier	airways watch bankruptcy court judge stephen mitchell hear arguments today asking rec ##ons ##ider four month percent pay cut imposed many union ##ized workers last month
Sci	Outlier (World)	Outlier	bush defend ##s decision invade iraq president bush went skeptical hall world leaders tuesday mount vigorous defense war iraq telling united nations iraqi people
Business	Outlier (Sci)	Outlier	nokia nec test new multimedia sub ##sy ##ste ##m two high tech communications players completed first phase series tests show next generation data communications infrastructure works
World	Outlier (Sports)	Outlier	man plan athens four years ago sydney gymnast ##s gone medal free olympics first time years federation president bob cola ##ross ##i sitting table explaining turn ##around already begun women moved sixth fourth world one year men sixth fifth
Sports	Inlier	Inlier	dolphins steelers play sunday night reuters reuters miami dolphins pittsburgh steelers play scheduled game sunday night [SEP]
Sci	Inlier	Inlier	microsoft launch new search engine software giant microsoft corp decided release beta versions updated ms ##n search engine earlier today company hopes compete leading search engines google yahoo [SEP]

Figure 19: More qualitative examples.

Subset	business	sci	sports	world
AUPR-in	74.8	62.4	88.8	81.9
AUPR-out	96.1	93.5	98.5	95.5

Table 9: We report AUPR metric for AG News splits, on inliers and outliers since this is a more relevant metric for unbalanced classes (which is the case for all splits in text AD, as explained in Anomalies setup).

B.2 AUTHORSHIP DETECTION

The PAN 2020 (Bevendorff et al., 2020) data is collected from the FanFiction website¹³ and represents an excerpt of one of the four shared tasks on authorship analysis presented at PAN 2020, namely, the second one. The data is aimed at cross-domain authorship verification, with the purpose of understanding the associations between authors and their texts in a context agnostic to domain-specific vocabulary.

We've sampled the top 19 authors from our curated dataset, each of them having written 30 articles. We've then picked a subset of 10 authors from the initial 19 for training and testing (due to time and resource constraints). We're treating the Authorship Verification task as a semi-supervised Anomaly Detection problem in the following way - Let A_1, \dots, A_{19} be the texts written by the authors, and $\mathcal{A} = \{A_1, \dots, A_{19}\}$ be the set of all texts, then given an $A_k \in \mathcal{A}$, we set A_k to be the set of normal texts, and $\mathcal{A} \setminus A_k$ to be the set of anomalous texts.

We are trying to profile authors using DATE. We are using the same hyperparameters as in the “AG News” dataset experiments. The only preprocessing that was done to the data is making it lower-cased. We’re reporting scores at 1500 steps into the training process. The results of the experiments are in Table 10.

Figure 20 shows how DATE is able to detect stylistic changes when correctly profiling an author. The model considers that certain punctuation and

train/val	Usr1	Usr2	Usr3	Usr4	Usr5	Usr6	Usr7	Usr8	Usr9	Usr10
25/5 (AUROC)	100	100	98.2	99.6	100	100	96.9	100	98.4	100
25/5 (AUPR-in)	100	100	87.2	93.8	100	100	65.7	100	84.1	100
25/5 (AUPR-out)	100	100	99.9	99.9	100	100	99.8	100	99.9	100
15/15 (AUROC)	96.7	98.5	97.3	99.3	98.2	100	95.2	100	93.8	98.6
15/15 (AUPR-in)	68.2	83.6	74.6	89.9	85.0	100	57.5	100	61.1	89.9
15/15 (AUPR-out)	99.8	64.4	99.9	99.9	99.9	100	99.7	100	99.6	99.9
5/25 (AUROC)	82.7	89.1	81.7	88.6	83.2	97.4	79.4	82.3	67.7	84.4
5/25 (AUPR-in)	42.3	64.4	39.1	52.8	56.1	83.9	34.6	65.9	22.0	42.0
5/25 (AUPR-out)	98.8	89.1	98.7	99.3	98.7	99.9	98.3	97.8	97.3	99.0
1/29 (AUROC)	72.7	82.0	73.2	87.2	82.2	95.4	83.2	80.3	53.1	65.1
1/29 (AUPR-in)	20.2	33.1	15.3	43.6	41.2	66.8	29.6	42.0	12.3	17.8
1/29 (AUPR-out)	98.1	98.8	97.7	99.2	98.7	99.7	98.9	97.9	95.5	97.2

Table 10: Semi-supervised AD Performance (AUROC%, AUPR-in%, AUPR-out%) on the PAN2020 dataset (custom setup) using DATE: The train/val ratio is the number of samples for training per author, and number of outlier articles per author when testing respectively. No hyperparameter tuning was done. The AUPR-out scores are very high because AUPR is sensible to imbalanced data.

strength . " " and **we** don " t have that time . " luke realized . " the other fi ##end ##s will know **we** " re serious **now** . they " ll be ready for us , and the longer **we** wait , the more time **we** give them to prepare . " " exactly . . . but you " ve risen so far in so little time already , remember . " earth polished his glasses , smiling again . " i think you **can** do it . . . and so do my brother and sisters , though they remain un ##rise ##n . **we** " re counting on you . " " then **we** should get to crescent lake , and fast . " the fighter decided . " more than ever , how **we** need to con ##ven ##e with the sage ##s . . . and there " s another fi ##end there too , isn " t there ? " ka ##ry , fi ##end of fire . " earth confirmed . " **now** then . . . shall i send you back to the surface , or was there anything else ? " " i thought there was something , but i **can** " t remember . . . "

for something useful and tell us which way to go ? " the cavern fork ##ed in a t - shape to the left and right . " uh . . . i don " t know . " drake admitted . " the only ones who have ever been in this place and made it out alive are the wizards , and they don " t exactly give us maps to be nice guys . " " you " re useless . " vlad sighed . " somebody pick one , then . " " right . " luke shrugged and walked ahead , then paused . " uh - oh . company . " two crawl were advancing , pale and gray colored , and with them was a monstrous green wolf , eyes glowing red and fangs dripping . " i " ll take the dog ! " yelling , the fighter engaged it while the other spread out . " watch out , its bite is poisonous ! " drake warned him as he and mary began stabbing and hacking at one crawl . " and these ones stu ##n , so be careful about them too ! " " these would be the stronger types , then . " vlad noted calmly , freezing the other crawl then blasting it apart with a lightning bolt . " still just as ugly . " mary smashed the crawl flat , and drake stabbed it a few more times to make

(a) *Usr9* correctly identified as *Usr9*. Interestingly, the network identifies the use of stop-words as being anomalous.

comforting things | i stared then out at nothing really | just being truly afraid now | i glanced back at the window once then hugged my sister and anne close | and could already feel the tears that were forming in my eyes now | the long walk that seemed to take forever | marie could only keep her gaze forward | the echoing of her heels hitting against the stone pavement | she felt like screaming | like shivering | like crying but she didn " | she walked like a cold block of stone | which did not feel nor spoke | her hands folded together in front of her | as she stood behind her sister | heading through the crowd | to get to the front | she wore a long dark orange dress | and her hair was tied up today | not lying past her shoulders | it seemed so cold to her | all of this |

here and watch | it wasn " | | | this wasn " | supposed to happen | it wasn " | fair | but marie knew things weren " | fair really | not all the time | and that people had expected this | she felt so uneasy here | and she just wanted to leave | anne had wanted so much | she had gotten so many things | and now it was being taken away | it wasn " | the fair thing to do as said | it wasn " | how it was supposed to happen | and this was supposed to be like a fairy - tale | he was supposed to love her | be a husband to her | and take care of her | but henry seemed to have other ideas | | henry the eighth didn " | want to have anne as his wife anymore | he preferred jane seymour now | it wasn " | that marie didn " | like jane | it was the fact of losing a friend that made her feel like this | she lost so much | and it wasn " | stopping | it wouldn " | stop | how could he do this ? how could he let his wife be treated this way ? she tried so hard to hold the tears back | to keep calm for anne | but when she saw anne walk up the steps | she could hear her own gasp then | " i don " | think i can | | | cara

(b) Network trained on *Usr9*. The network correctly classifies the texts as not being written by *Usr9*. Notice the fact that the network considers that the use of punctuation is dubious.

Figure 20: DATE Qualitative results on the PAN2020 dataset. Lower PL scores are shown in a more intense red and contribute more to a higher anomaly probability.

word usage is dubious. When scoring an inlier entry, DATE seems to think that some stopword usage is anomalous, this could indicate a change of stylistic approach when an author is writing texts in different fandoms.

B.3 MUSIC GENRE DETECTION

The Song Lyrics dataset¹⁴ is composed of four sources and consists of over 290,000 multilingual song lyrics and their respective genres: Rock, Metal, Pop, Hip-Hop, Electronic, R&B, Country, Folk and Indie. The initial data was forwarded from the 2018 Textract Hackathon¹⁵. This was enhanced with data from three kaggle datasets: *150K Lyrics Labeled with Spotify Valence, dataset lyrics musics* and *AZLyrics song lyrics*.

Apart from the original 2018 Textract data, the other datasets were not provided with a Genre feature. In order to deal with the lack of labels, a labeling function was built using the spotify library, which uses the Spotify API in order to retrieve the genre of an Artist. The Spotify API returns a list of genres for one artist, so only the most common genre to be said artists dominant genre was considered. The results are shown in Tables 11 and 12.

¹⁴ Many thanks to Matei Bejan for collecting this dataset: *Kaggle multilingual lyrics for genre classification*

¹⁵ <https://www.sparktech.ro/textract-2018/>

Table 11: Semi-supervised AD Performance (AUC-ROC%) on the Music Genres datasets. We compare two classical models (Isolation Forest, One-Class SVM) with two neural models (CVDD, DATE). We are choosing the best score per split, as detailed in Sec. 4.2.3. DATE is using the same hyperparameters per dataset. Interestingly, every model is able to outperform DATE on the Hip-Hop subset, CVDD having the largest margin (+24.1% ROC-AUC).

	Inlier class	iForest	OC-SVM	CVDD	DATE
Music Genres	Rock	45.6	44.2	45.1	54.4
	Electronic	48.5	48.8	45.4	56.0
	Country	57.3	57.5	54.3	70.8
	Metal	55.9	46.6	48.5	56.6
	Indie	51.8	47.1	45.5	57.8
	Jazz	53.5	52.0	56.8	70.1
	Folk	49.5	49.5	47.0	53.4
	R&B	58.4	58.1	61.3	71.0
	Hip-Hop	65.0	72.9	84.3	60.2
	Pop	43.2	43.7	61.3	62.9
Mean		52.9	52.0	55.0	61.3

Context 1	Context 2	Context 3	Context 4	Context 5
'm	livin	to	the	que
i	turnin	can	of	mas
're	shakin	could	in	como
y'	waitin	make	's	porque
somebody	walkin	will	that	sabe
myself	keepin	would	is	desde
everybody	slippin	pray	this	boca

Table 12: CVDD context vectors. The second context is the most meaningful one when detecting anomalies. Notice that the second context contains the colloquial forms of verbs. The first context seems to cluster pronouns together. The third and forth contexts contain modal verbs and stopwords. The last context vector is clustering together words in Spanish.

Outlier	CVDD	DATE
Translations	43.2	55.4
Non-native	52.3	56.0

Table 13: Semi-supervised AD Performance (AUROC%) on the ENNTT dataset.
The performance is weak on both CVDD and DATE.

B.4 NATIVE LANGUAGE DETECTION

ENNTT ([Nisioi et al., 2016](#)) is an English corpus of original and human-translated texts extracted from the discussions that took place inside the European Parliament. The dataset consists of three types of texts, generated by either native English speakers, non-native English speakers or translated from another European language. It contains 116,341 native sentences, 29,734 non-native sentences and 738,597 translated sentences.

The results when using both CVDD and DATE for one-class classification on ENNTT are poor, as can be seen in Table [13](#).