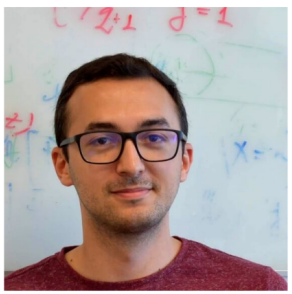
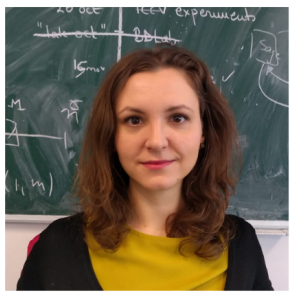
DATE: Detecting Anomalies in Text via Self-Supervision of Transformers

Andrei Manolache, Florin Brad, Elena Burceanu

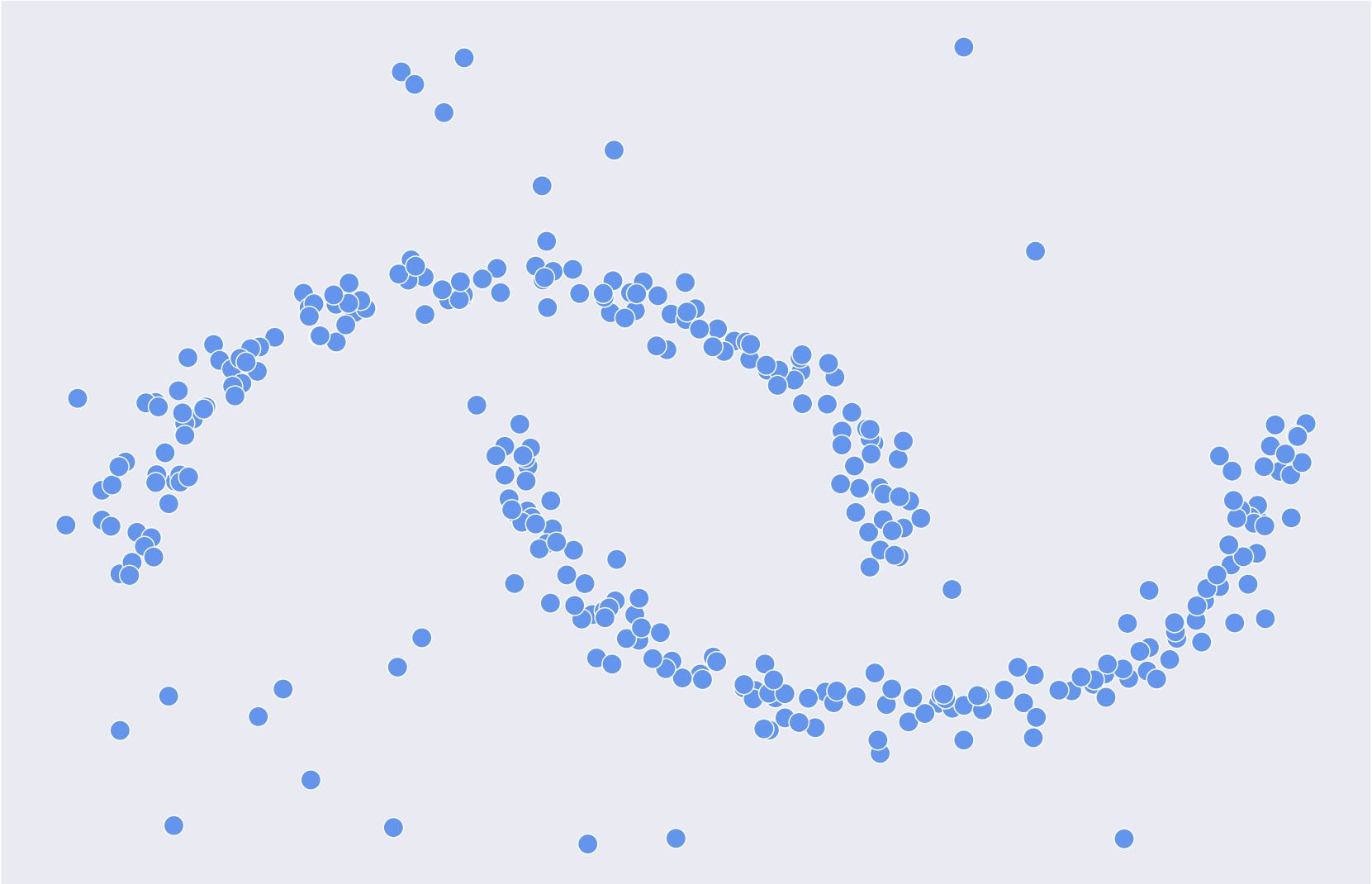
{amanolache, fbrad, eburceanu}@bitdefender.com

Anomaly Detection

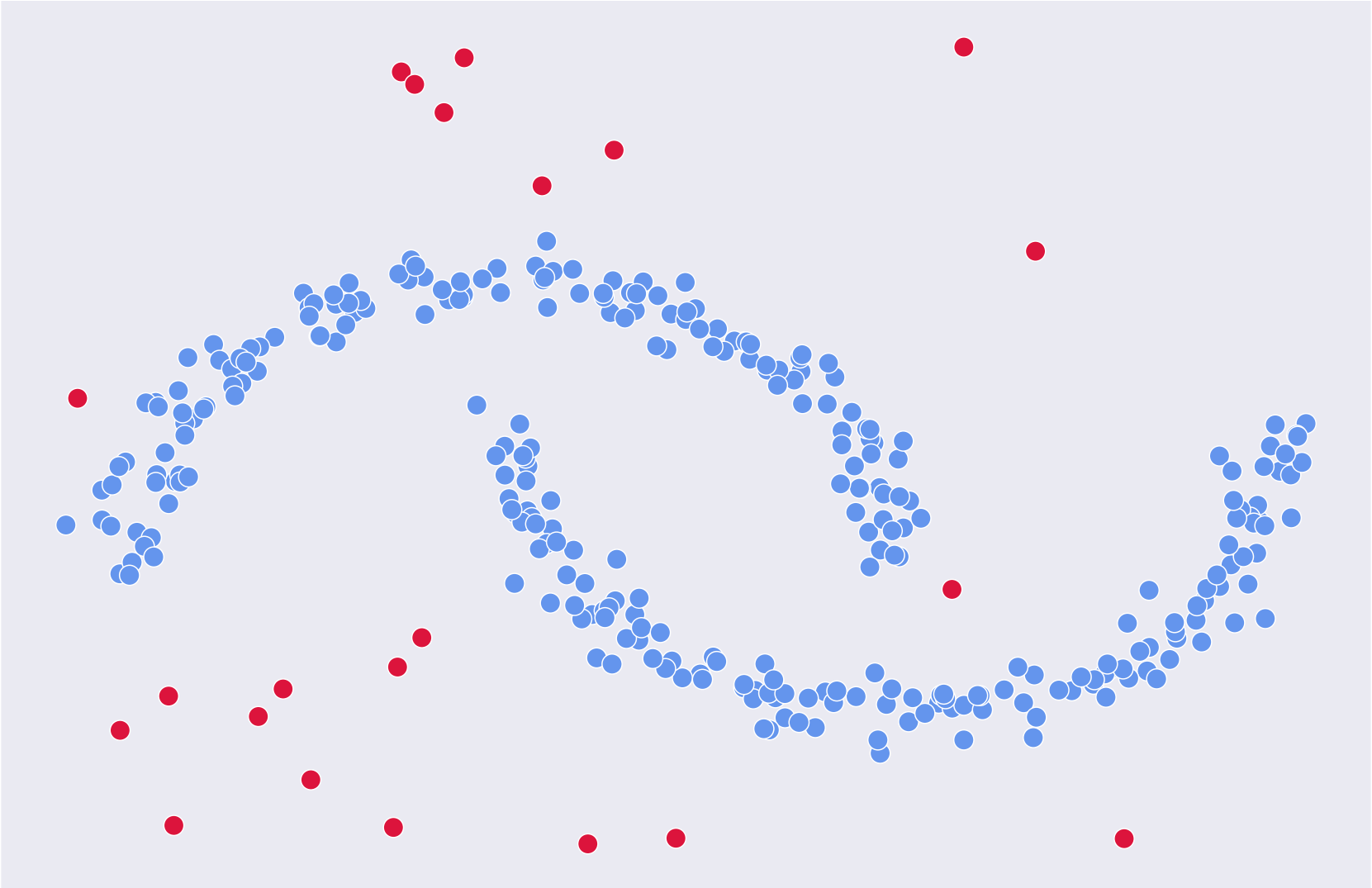
DATE

Contributions

What is an anomaly?

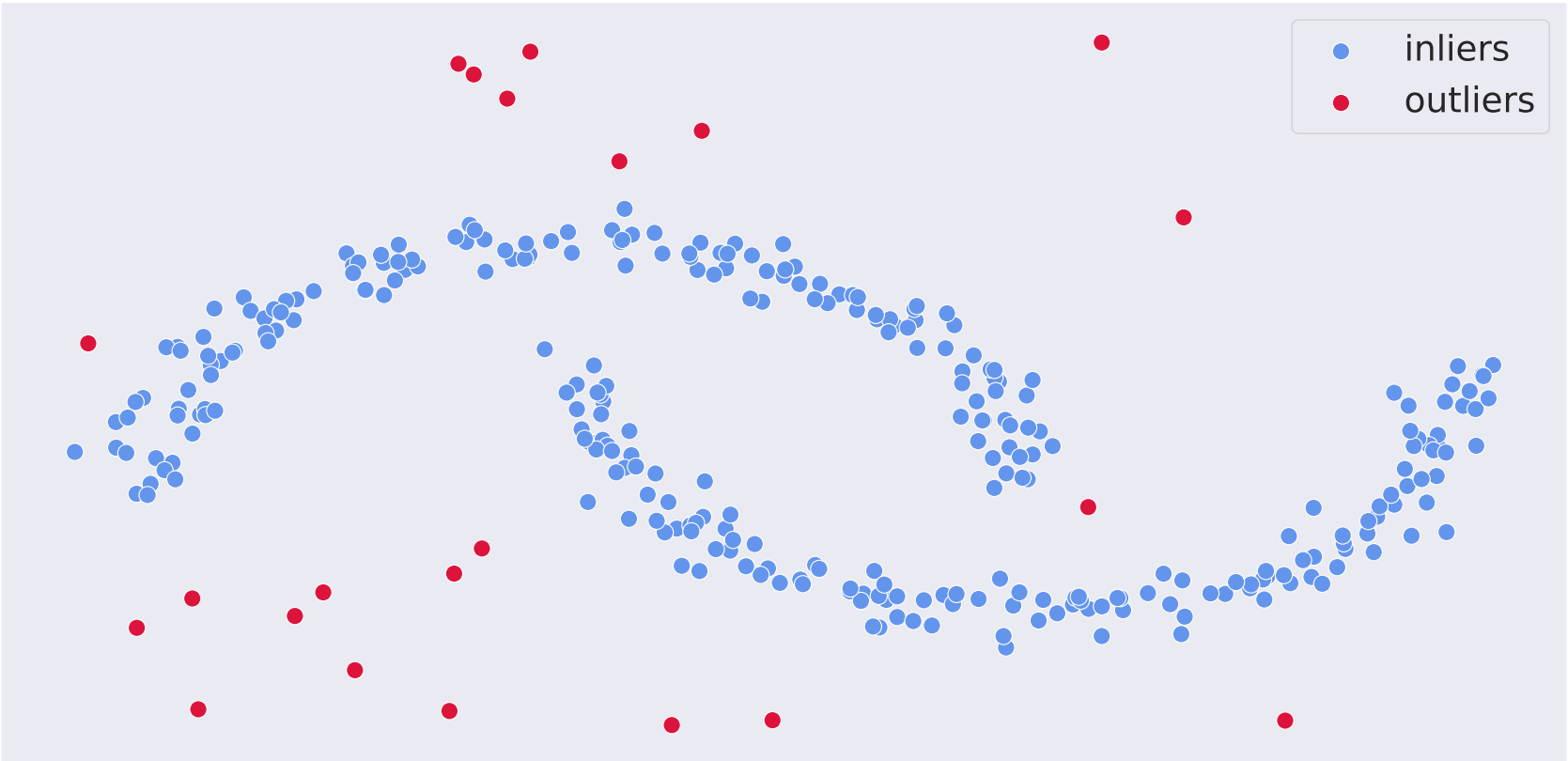


What is an anomaly?



What is an anomaly?

An outlier (*anomaly, deviant*) is an observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism [6].



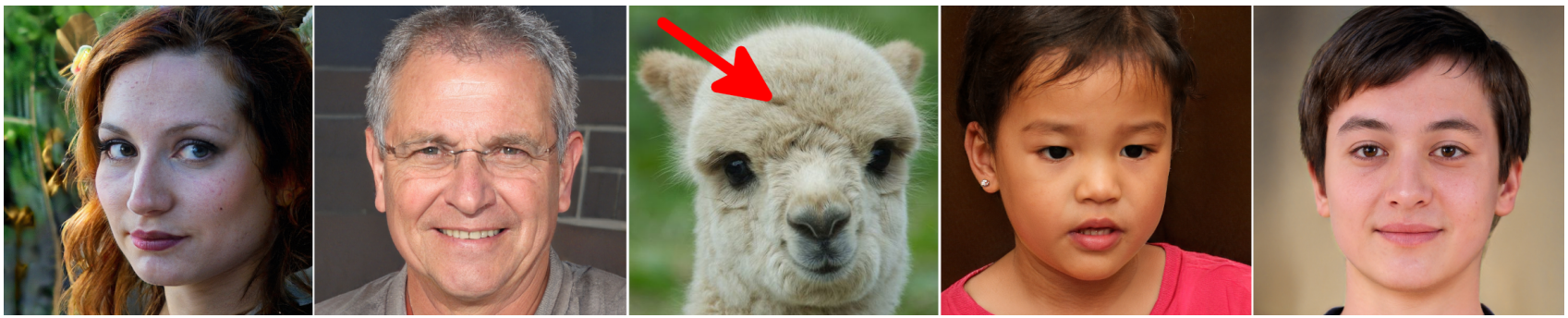
Anomalies in Computer Vision

Quite obvious:

More subtle:

Anomalies in Computer Vision

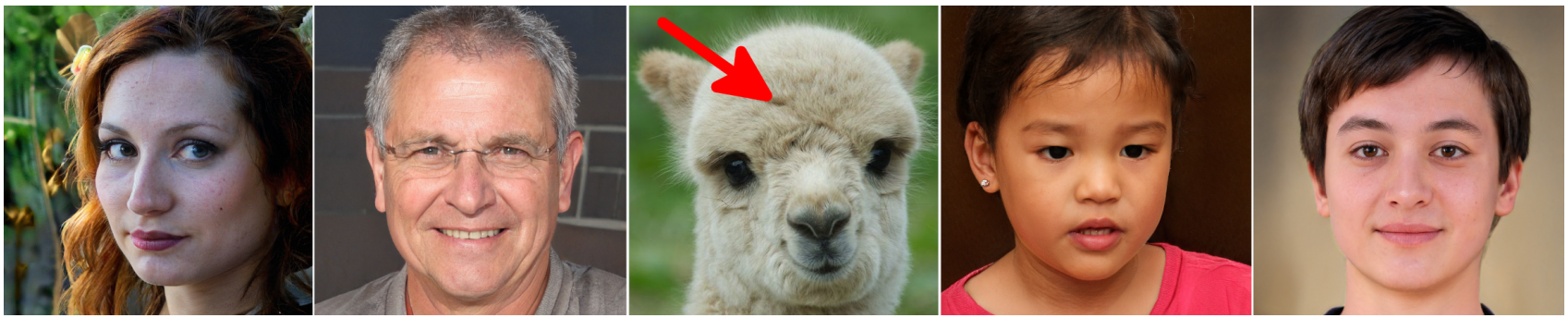
Quite obvious:



More subtle:

Anomalies in Computer Vision

Quite obvious:



More subtle:

Anomalies in Text - Phishing and Spam

amazon

Account Service

Dear andrei_m[redacted]

The suspicious login process on your account, as well as the data we get;

Date: 1/9/2021 12:25:03 PM
Device: Amazon Shopping App for Android
Near: Texas, United States

Your account has been locked because we discovered suspicious activity that was fatal. Please log into your account immediately and fill in all the data we provide to restore your account

Login to Account

https://kunyukbeber.com/L[redacted]o

This message is specially crafted by Amazon to :
andrei_m[redacted]

EH

emilia hunt <emiliahunt16@gmail.com>
Wed 1/13/2021 3:21 PM

Salut ce mai faci.
Sunt Emilia, îmi place să fiu prietenul tău, te rog să-mi răspunzi mulțumesc

Reply | Reply all | Forward

Table of contents

DATE

Main idea:

DATE

Main idea:

1. Learn how a normal text looks like ...

DATE

Main idea:

1. Learn how a normal text looks like ...
2. ... in an unsupervised or semi-supervised fashion ...

DATE

Main idea:

1. Learn how a normal text looks like ...
2. ... in an unsupervised or semi-supervised fashion ...
3. ... by carefully designing a self-supervised task for Transformers.

The DATE Architecture

Our network has two components:

The DATE Architecture

Our network has two components:

1. A generator that samples tokens from a word distribution and corrupts the input.

The DATE Architecture

Our network has two components:

1. A generator that samples tokens from a word distribution and corrupts the input.
2. A discriminator with two heads:
   - Replaced Token Detector [2]: detects **which token** is corrupt and which one is original.
   - Replaced Mask Detector: detects **which masking pattern** was applied over the input text.

DATE: training

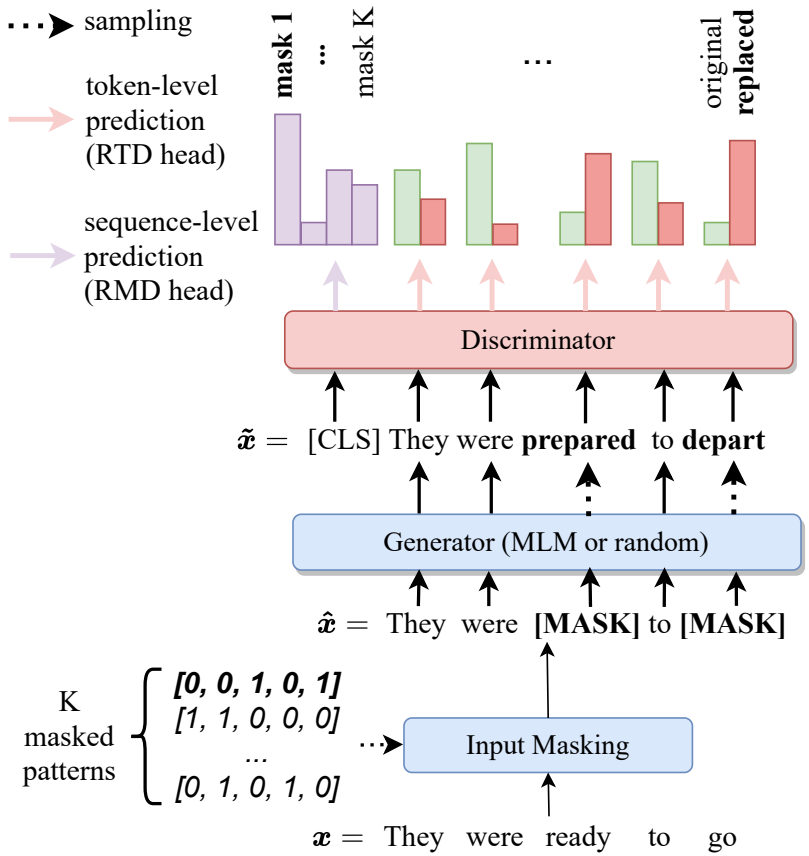


Figure: DATE: training.

DATE: training

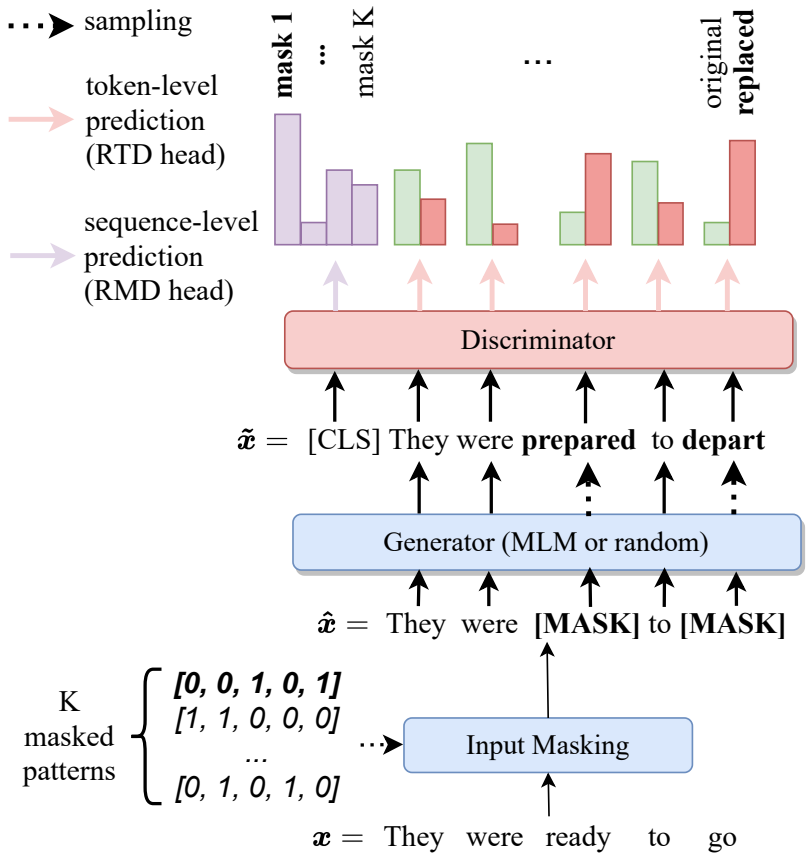


Figure: DATE: training.

1. Sample a masking pattern from a collection of K pre-generated masks.

DATE: training

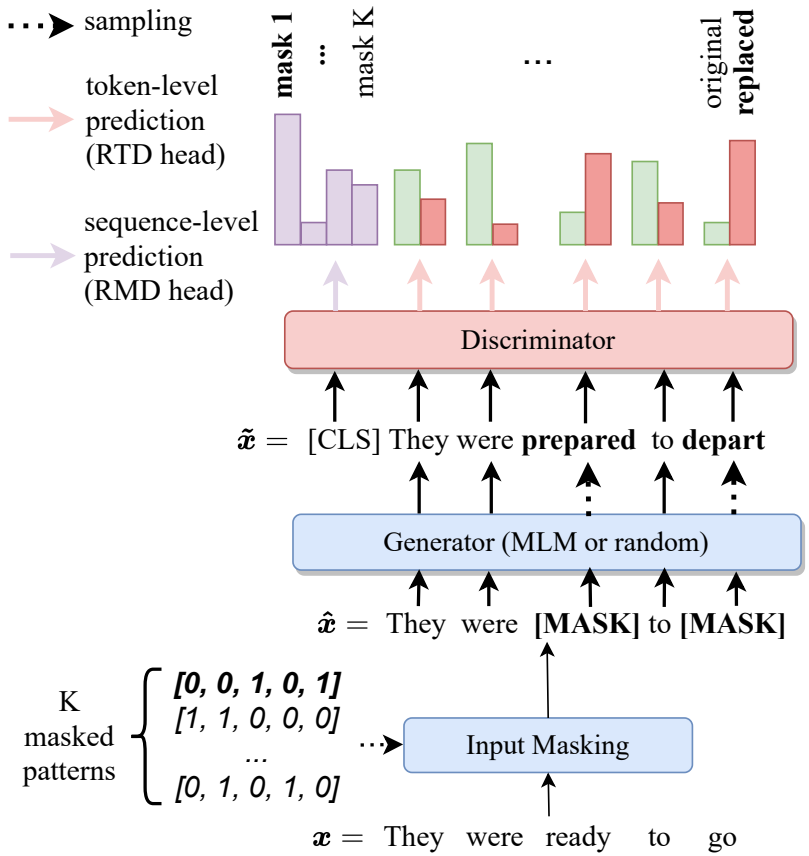


Figure: DATE: training.

1. Sample a masking pattern from a collection of K pre-generated masks.
2. Apply the masking pattern on the original input, and corrupt the masked tokens.

DATE: training

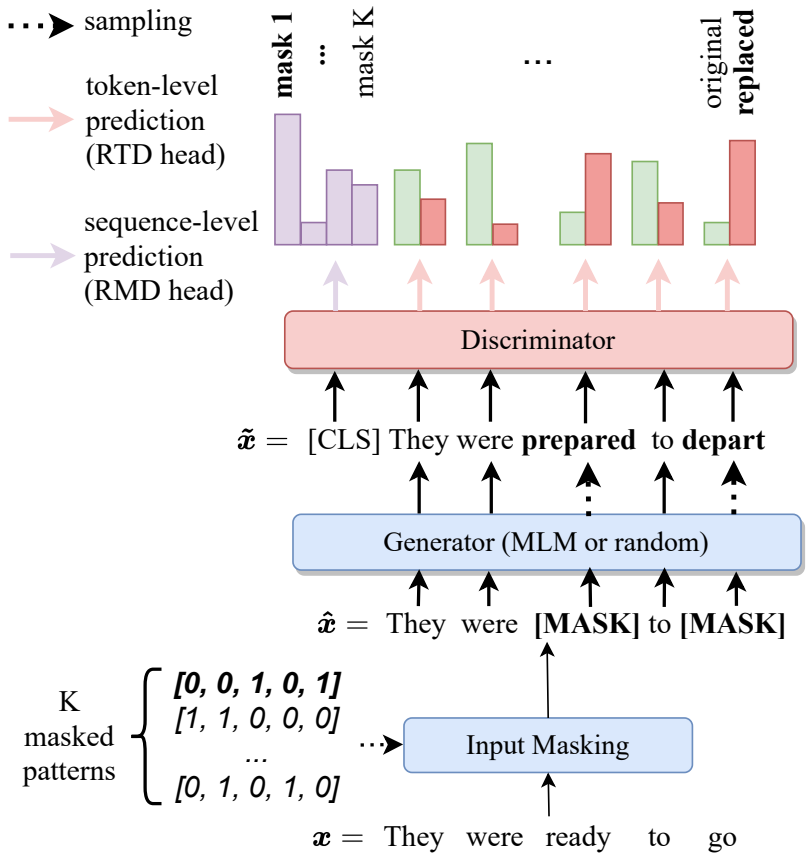


Figure: DATE: training.

1. Sample a masking pattern from a collection of K pre-generated masks.
2. Apply the masking pattern on the original input, and corrupt the masked tokens.
3. Guess which tokens were replaced (RTD head).

DATE: training

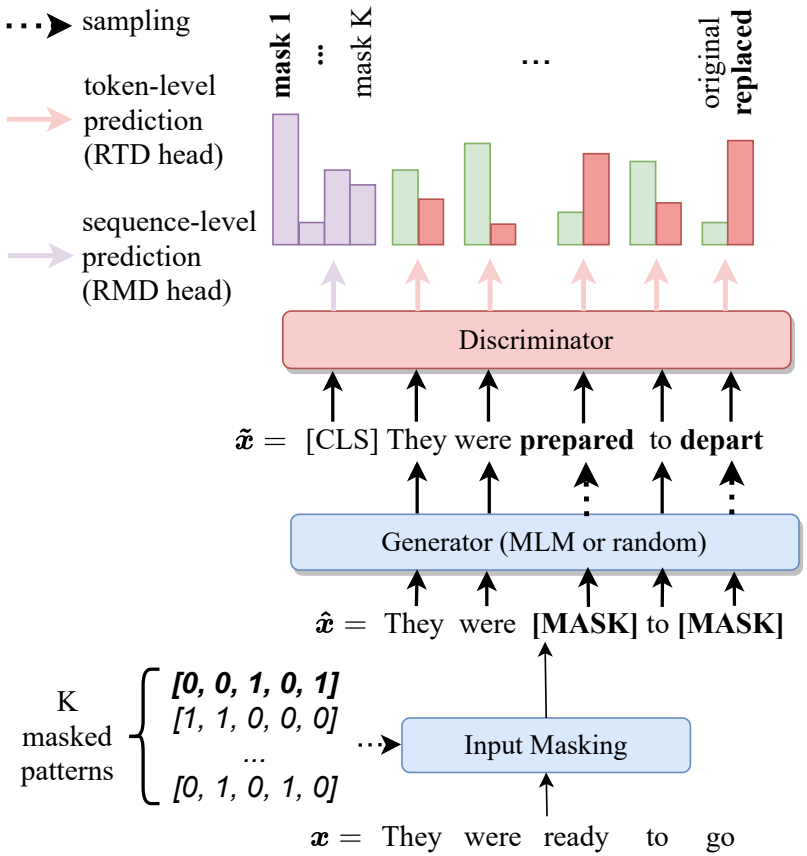


Figure: DATE: training.

1. Sample a masking pattern from a collection of K pre-generated masks.
2. Apply the masking pattern on the original input, and corrupt the masked tokens.
3. Guess which tokens were replaced (RTD head).
4. Guess what masking pattern was applied (RMD head).

DATE: training

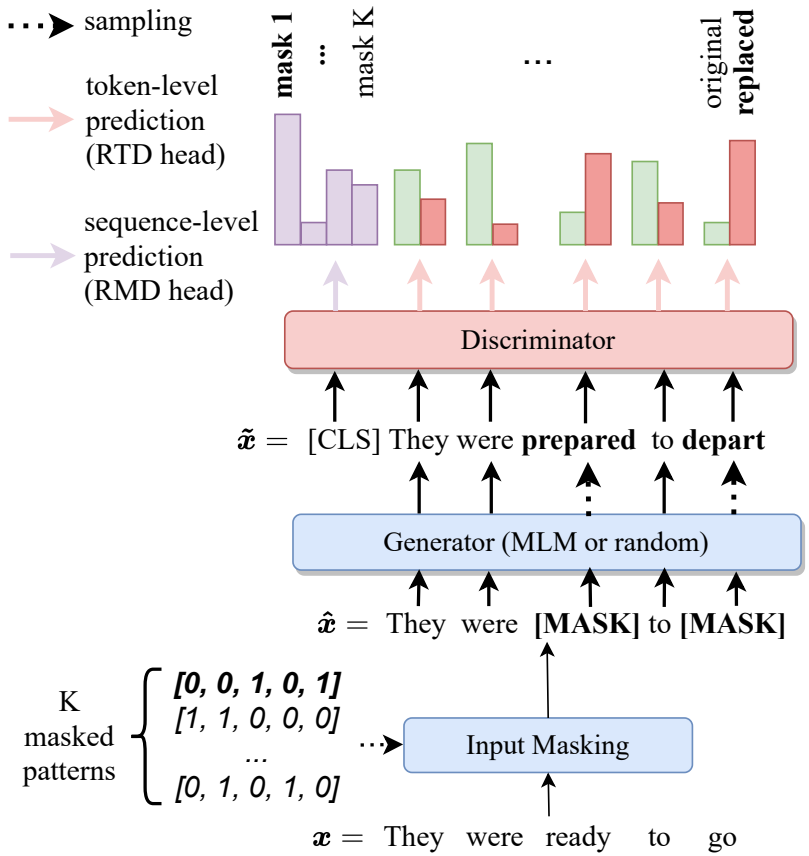


Figure: DATE: training.

1. Sample a masking pattern from a collection of K pre-generated masks.
2. Apply the masking pattern on the original input, and corrupt the masked tokens.
3. Guess which tokens were replaced (RTD head).
4. Guess what masking pattern was applied (RMD head).
5. Optimize the network based on 3. and 4..

DATE: inference

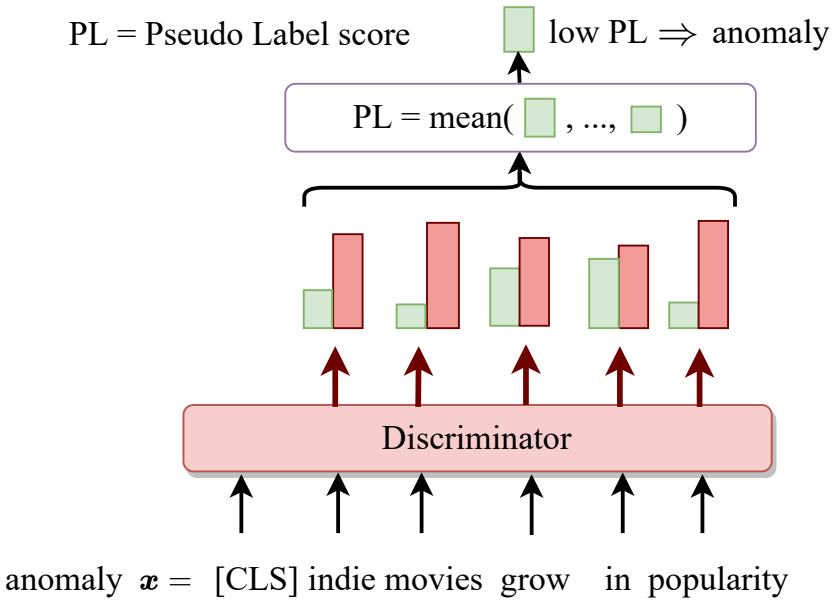


Figure: DATE: inference.

DATE: inference

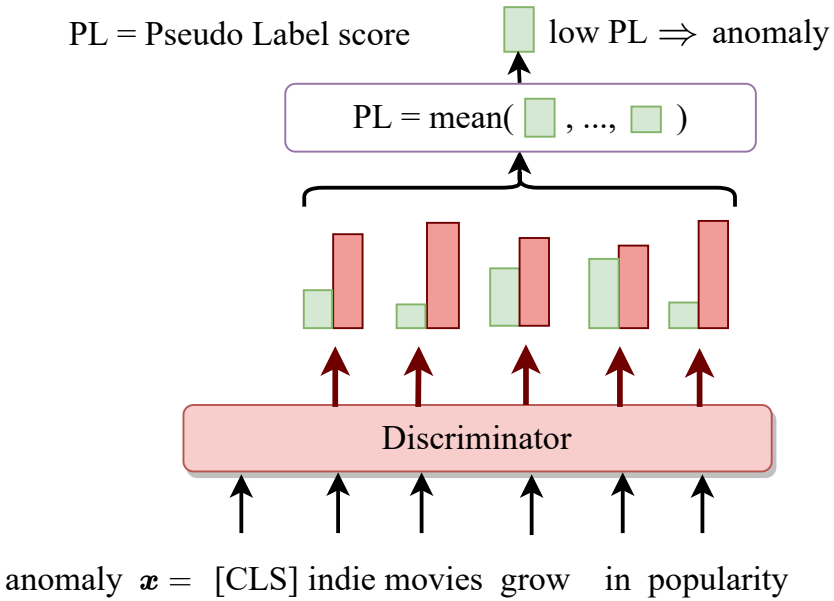


Figure: DATE: inference.

1. We throw away the *generator* and feed the uncorrupted sequence directly into the *discriminator*.

DATE: inference

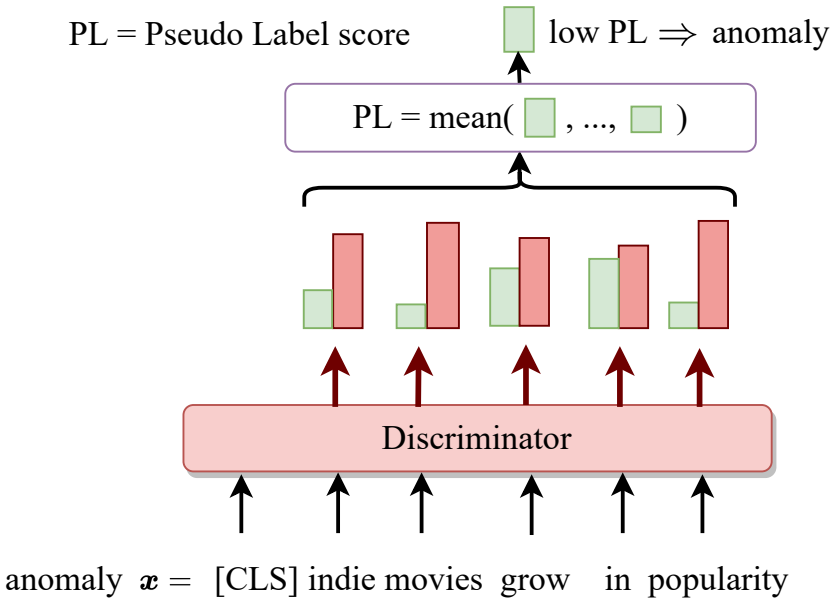


Figure: DATE: inference.

1. We throw away the *generator* and feed the uncorrupted sequence directly into the *discriminator*.
2. For every token *i* in the sequence, we compute the PL score - this tells us the likelihood that a token is **not corrupted**.

DATE: inference

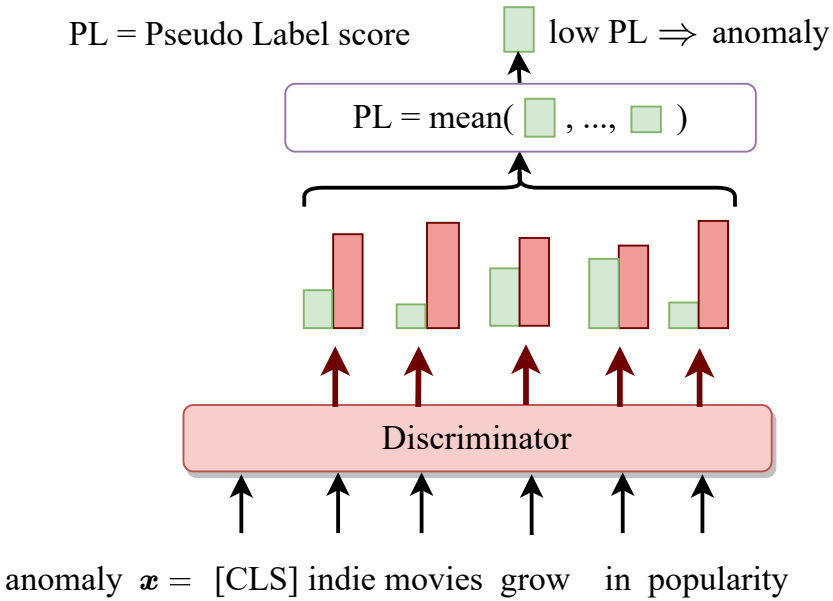


Figure: DATE: inference.

1. We throw away the *generator* and feed the uncorrupted sequence directly into the *discriminator*.
2. For every token i in the sequence, we compute the PL score - this tells us the likelihood that a token is **not corrupted**.
3. We average the PL score over the entire sequence.

DATE: train

Formally, we optimize the combined loss for the RTD and RMD heads:

$$\underbrace{\mathbb{E}\left[\sum_{i=1}^{T} -\log P_D(m_i|\widetilde{x}(m);\theta_D)\right]}_{\mathcal{L}_{RTD}} + \underbrace{\mathbb{E}\left[-\log P_M(m|\widetilde{x}(m);\theta_D)\right]}_{\mathcal{L}_{RMD}}$$

DATE: inference

Assumption: the network is more likely to predict **corrupt** for tokens of an *outlier* sample (even if we don't corrupt any!) - this gives a **lower PL score for outliers** and a **higher PL score for inliers**:

$$PL_{RTD}(x) = \frac{1}{T}\sum_{i=1}^{T} P_D(m_i = 0|\widetilde{\boldsymbol{x}}(\boldsymbol{m}^{(0)}); \theta_D)$$

DATE: Experimental Setup (semi-supervised)

We use two text classification datasets:

1. 20 Newsgroups: news articles - around 577-2856 training samples per split.
2. AG News: news articles - 30000 training samples per split.

We're training the network on the *clean samples* (one split) and consider every other split as being *anomalous*.

DATE: Quantitative results

	Inlier class	IsoForest best	OCSVM best	CVDD best	**DATE (Ours)**
20 News	comp	66.1	78.0	74.0	**92.1**
	rec	59.4	70.0	60.6	**83.4**
	sci	57.8	64.2	58.2	**69.7**
	misc	62.4	62.1	75.7	**86.0**
	pol	65.3	76.1	71.5	**81.9**
	rel	71.4	78.9	78.1	**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: Semi-supervised AD Performance (AUROC%)

Ablation Study

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
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 func	RTD only	89.4 ± 4.4
		RMD only	85.9 ± 4.1
D.	Masking	5 masks	87.5 ± 4.5
	patterns	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: AUROC mean and std are computed over the AG News splits.
Ours: **A**. PL_{RTD}; **B**. Rand **C**. RTD + RMD; **D**. 50 masks; **E**. 50%.

DATE: Quantitative results

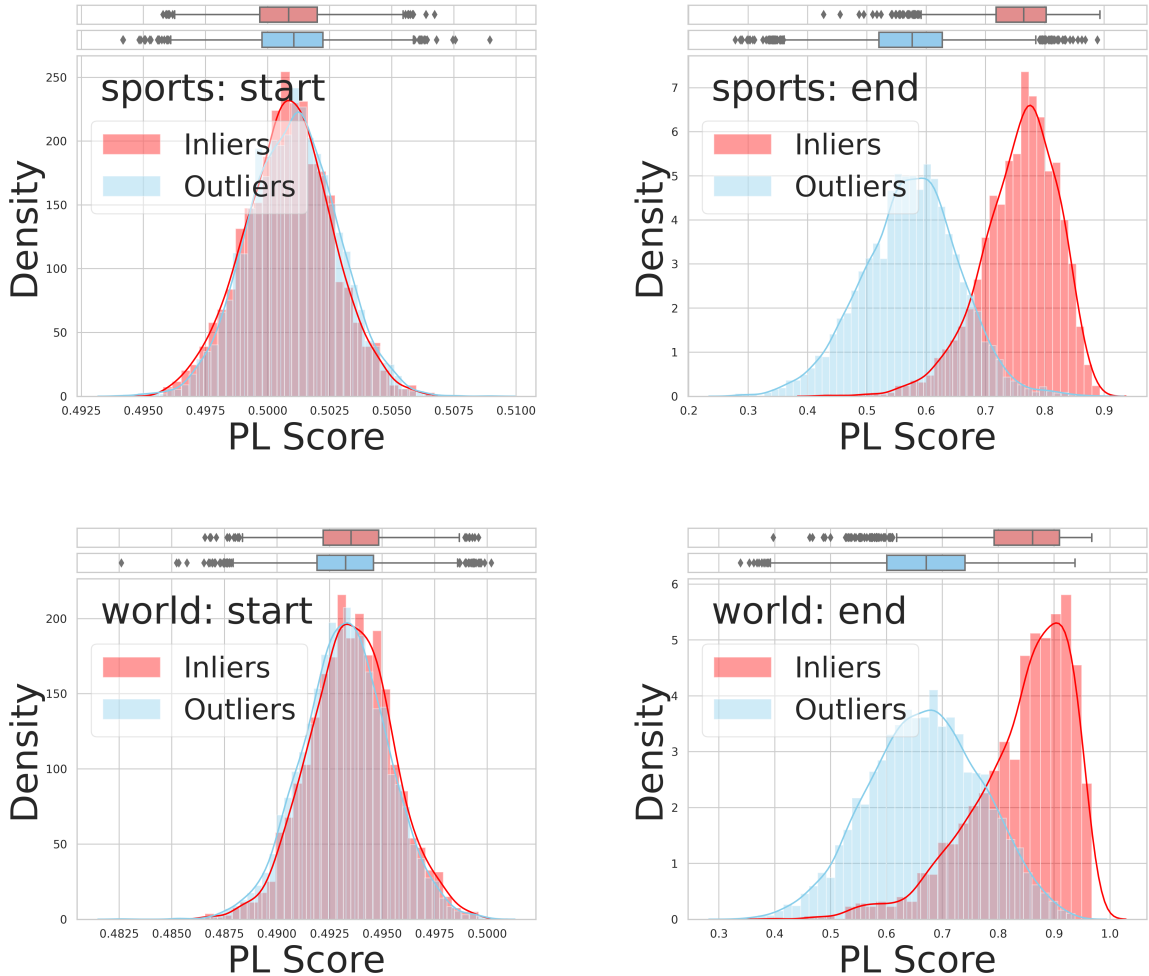


Figure: Normalized Histograms for Anomaly Score (start of training vs. end of training).

DATE: Unsupervised Learning

We contaminate our training dataset with 0% – 15% outliers to test our method in the *unsupervised* scenario.

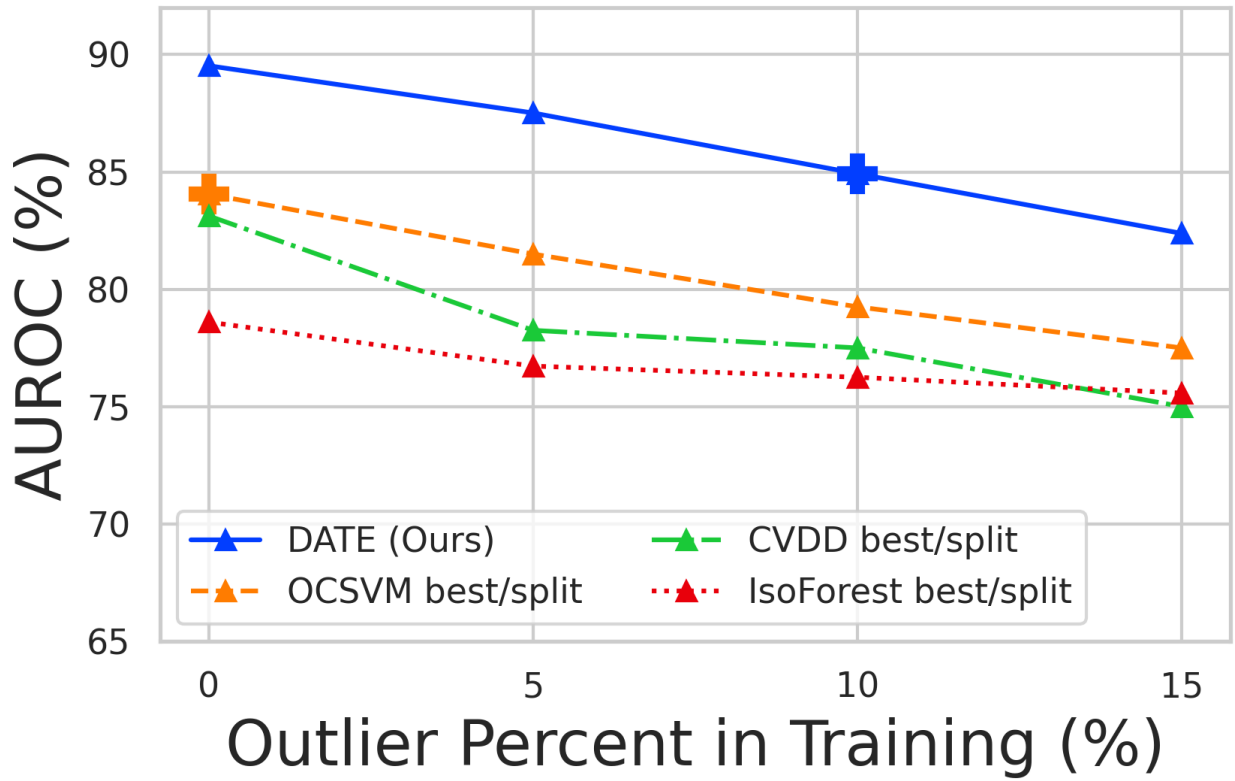


Figure: We outperform the competition trained on 0% contaminated data even at 10% contamination.

DATE: Qualitative results

Inlier	Label	Pred	Sample (BERT tokens)
Sports	Outlier (World)	Outlier	jail ##ing democrat china politically motivated af ##p af ##p hong kong democrats accused china jail ##ing one members trump ##ed prostitution charges bid disgrace political movement beijing feud ##ing seven years
Sci	Outlier (World)	Outlier	panama flooding kills nine people least nine people seven children died flooding capital panama authorities say least people still missing heavy rainfall caused rivers break banks
Business	Inlier	Inlier	motorola cut jobs new york reuters telecommunications equipment maker motorola inc hr ##ef http www investor reuters com full ##qu ##ote as ##p ##x tick ##er mo ##t target stocks quick ##in ##fo full ##qu ##ote mo ##t said tuesday would cut jobs take related charges million focus wireless business

Figure: Qualitative examples.

Contributions

- We introduce a novel self-supervised Transformer-based model for Anomaly Detection in text.

Contributions

- We introduce a novel self-supervised Transformer-based model for Anomaly Detection in text.
- We augment the model with a sequence-level self-supervised task called Replaced Mask Detection.

Contributions

- We introduce a novel self-supervised Transformer-based model for Anomaly Detection in text.
- We augment the model with a sequence-level self-supervised task called Replaced Mask Detection.
- We compute an efficient Pseudo-Label score for anomalies that makes our model efficient and interpretable.

Contributions

- We introduce a novel self-supervised Transformer-based model for Anomaly Detection in text.
- We augment the model with a sequence-level self-supervised task called Replaced Mask Detection.
- We compute an efficient Pseudo-Label score for anomalies that makes our model efficient and interpretable.
- We outperform state-of-the-art models by a very large margin on two datasets.

References I


Brown, T. et al.
Language Models are Few-Shot Learners.
NeurIPS '20.


Clark, K. et al.
ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.
ICLR '20.


Vaswani, A. et al.
Attention Is All You Need.
NeurIPS '17.


Devlin, J. et al.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
NAACL '19.

References II

Chen, J. et al.
Big Self-Supervised Models are Strong Semi-Supervised Learners.
NeurIPS '20.

Hawkins.
Identification of outliers.
volume 11, Springer.

Zisserman, A.
Self-Supervised Learning, 2018.
https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf