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

Andrei Manolache[1,2], Florin Brad[1], Elena Burceanu[1,2,3]

[1]Bitdefender, [2]University of Bucharest,

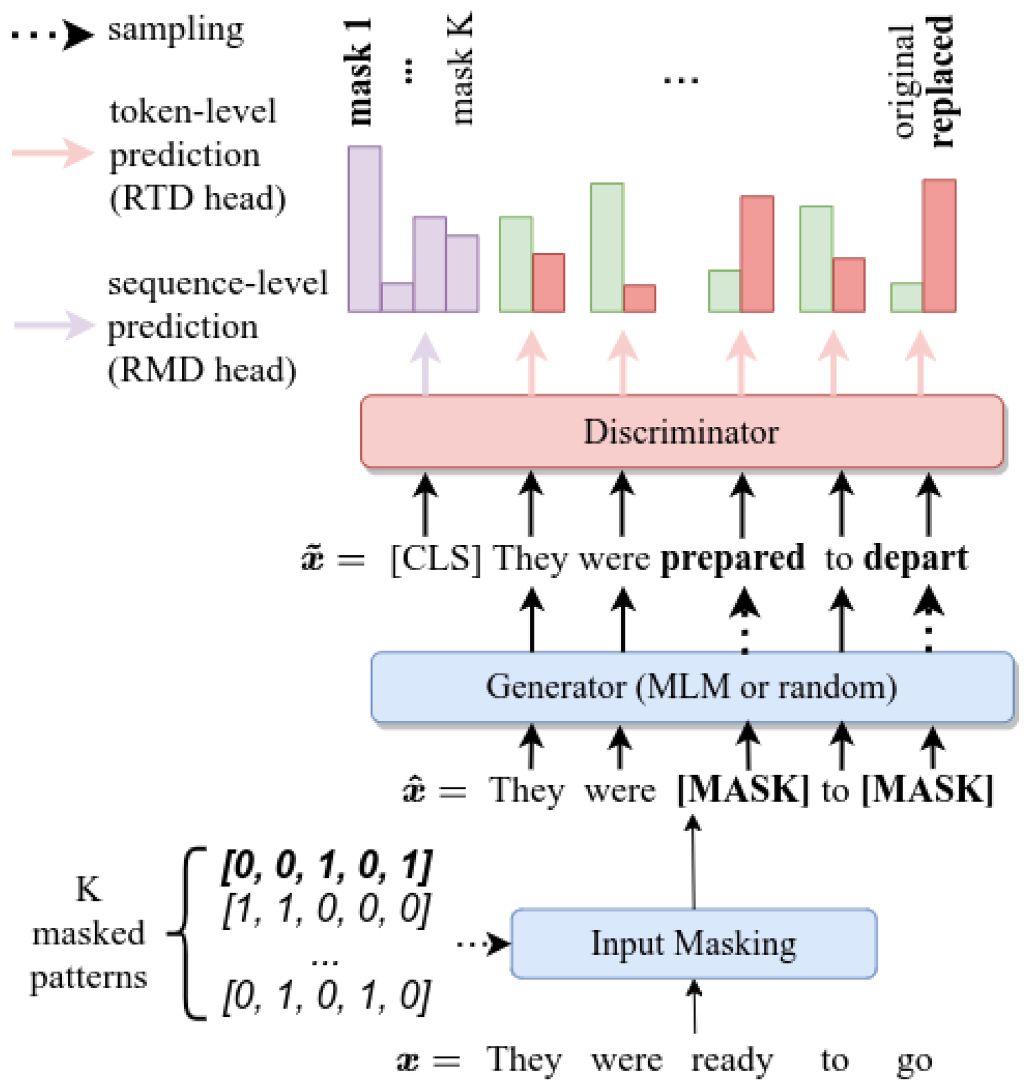
[3]Institute of Mathematics of the Romanian Academy

Contribution

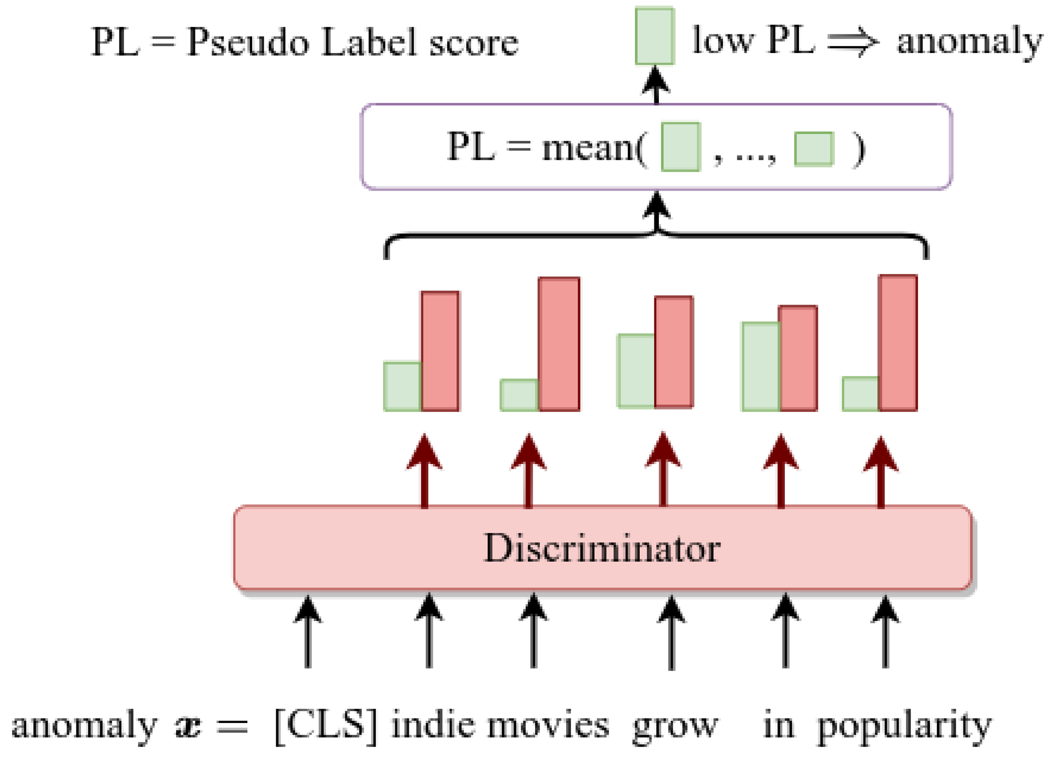
1. **Self-supervised task formulation** tailored for Anomaly Detection in text using Transformer models by combining the **Replaced Token Detection (RTD)** [1] and the novel **Replaced Mask Detection (RMD)** losses.
2. **Efficient anomaly scoring** by adapting the **Pseudo-Label (PL)** [2] score to the text domain, allowing it to work directly on individual token probabilities. This makes out model faster and its results more interpretable.
3. **Outperforming the state-of-the-art** on the **20Newsgroups** and **AG News** datasets **by a large margin** in both **semi-supervised** and **unsupervised** settings.

1. Training and Inference

- **Training**: The masked tokens are replaced with tokens sampled from a generator. The discriminator solves the RMD and RTD tasks.



- **Inference**: The input sequence is fed directly to the discriminator. The resulting token-level probabilities for the **normal class** are aggregated into an anomaly score.
- **Assumption**: when given an *outlier*, the network will detect normal tokens as being corrupted.



2. Task Formulation and Anomaly Score

- **Sample a mask** from a pre-defined collection of K masks $m \in \{m^{(1)}, ..., m^{(K)}\}$.
- **Replace the tokens** in the masked positions with tokens sampled from a generator and obtain the sequence $\tilde{x}(m)$.
- **Compute** $\mathcal{L}_{RMD}$, $\mathcal{L}_{RTD}$, and **optimize** the network w.r.t. $\mathcal{L}_{DATE} = \mu\mathcal{L}_{RMD} + \lambda\mathcal{L}_{RTD}$, where μ, λ are the weights of the loss components.

Formally, the loss components have the following expressions:

$$\mathcal{L}_{RMD} = \mathbb{E}\left[-\log P_M(\boldsymbol{m}|\tilde{\boldsymbol{x}}(\boldsymbol{m}); \theta_D)\right] \qquad \mathcal{L}_{RTD} = \mathbb{E}\left[\sum_{\substack{i=1;\\ x_i \neq \texttt{[CLS]}}}^{T} -\log P_D(m_i|\tilde{\boldsymbol{x}}(\boldsymbol{m}); \theta_D)\right]$$

For a sample x, the **Pseudo-Label (PL)** anomaly score is defined as:

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

where $\boldsymbol{m}^{(0)} = [0, 0, ..., 0]$ effectively leaves the input unchanged.

3A. Qualitative Results

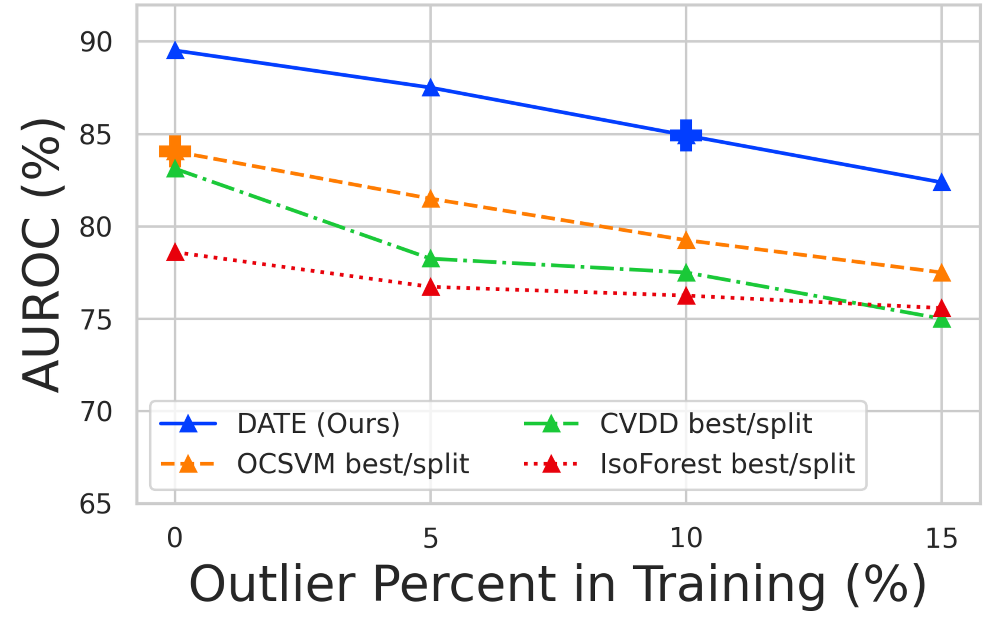
Inlier	Label	Pred	Sample (BERT tokens)
Sports	Outlier (World)	Outlier	jail ##ing democrat china politically motivated af ##p af ##p hong kong democrats accused china jail ##ing one members trump ##ed prostitution charges bid disgrace political movement beijing feud ##ing seven years
Sci	Outlier (World)	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

- 1st example: words from politics are flagged as anomalous for sports.
- 2^{nd} example: words describing natural events are outliers for technology.
- 3^{rd} example: few words have higher anomaly scores, but the model correctly classifies the sample as not being anomalous.

3B. 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: AUROC (%) scores for the AG News and 20Newsgroups datasets.



References

[1] Clark et al., ICLR 2020
[2] Wang et al., NeurIPS 2019