Unsupervised text anomaly detection using a BERT-based language modeling architecture

Maria-Cristina Borza and Dragoș-Constantin Țânțaru
Supervisor: Andrei-Marian Manolache
Faculty of Mathematics and Computer Sciences, University of Bucharest, Romania

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

1. Introduction

Anomaly detection (AD) can be defined as the task of identifying examples that differ significantly from the norm. This problem has applications in a wide variety of domains such as **fraud detection**, **network intrusion**, **defect detection**, and **health monitoring**. Various approaches have been developed for anomaly detection, including **deviation analysis**, **unsupervised clustering methods** and **rule-based systems**.
Traditionally, kernel-based models have shown great results in this field, using methods such as the ones described by One Class SVMs or Support Vector Data Description, since it turns out that the kernel trick is very powerful for separating inliers from outliers. However, in the specific task of text anomaly detection, it seems natural to expect that leveraging the powerful properties of novel Transformer models could yield great results. Papers that explore this idea already exist, such as the DATE[1] paper, however our model uses *only* the language modeling pretext task for anomaly detection. We have also found a thresholding strategy that seems to improve the performance of our model, sometimes even dramatically so.

2. Datasets description

We have trained our model on two different datasets, **20newsgroup** and **AG News**. **AG News** contains 120k training samples, evenly distributed amongst the 4 classes, containing roughly 125k words per class. **20newsgroup** contains around 18k training samples which are **not** as evenly distributed, especially since we considered the grouping done in the CVDD[2] paper, in which the 20 different groups are themselves gathered in 6 thematic groups.
We have used each individual class' train set for the language modeling task, and considered their respective test sets for the anomaly detection task.

3. Model and Approach

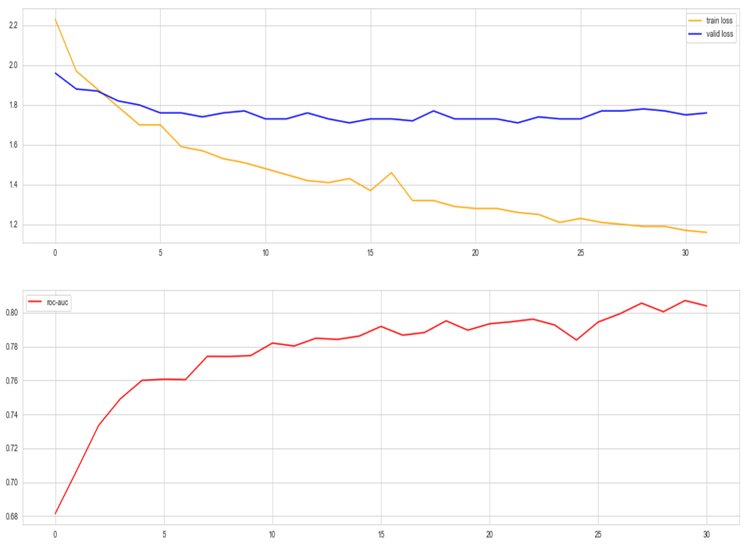
Our model consists of a BERT[3] language-modeling model. We used a pre-trained BERT base model that we fine-tuned on our task.
For **training**, we:

- masked 15% of the sequence's tokens
- passed the masked sequence through the model and calculated the loss on the predicted tokens

For **evaluation**, we:

- masked 15% of the sequence's tokens at a time
- calculated the probability of each ground truth token, by attempting to predict the masked tokens
- used the formula for the *anomaly_score*

Our model didn't showcase the usual behaviour of the AUROC% score curve of anomaly detection models, instead our model had a steady increase correlated with the decrease in our training loss, as shown in the following 2 graphs, on the *Sports* subset of the **AG News** dataset.

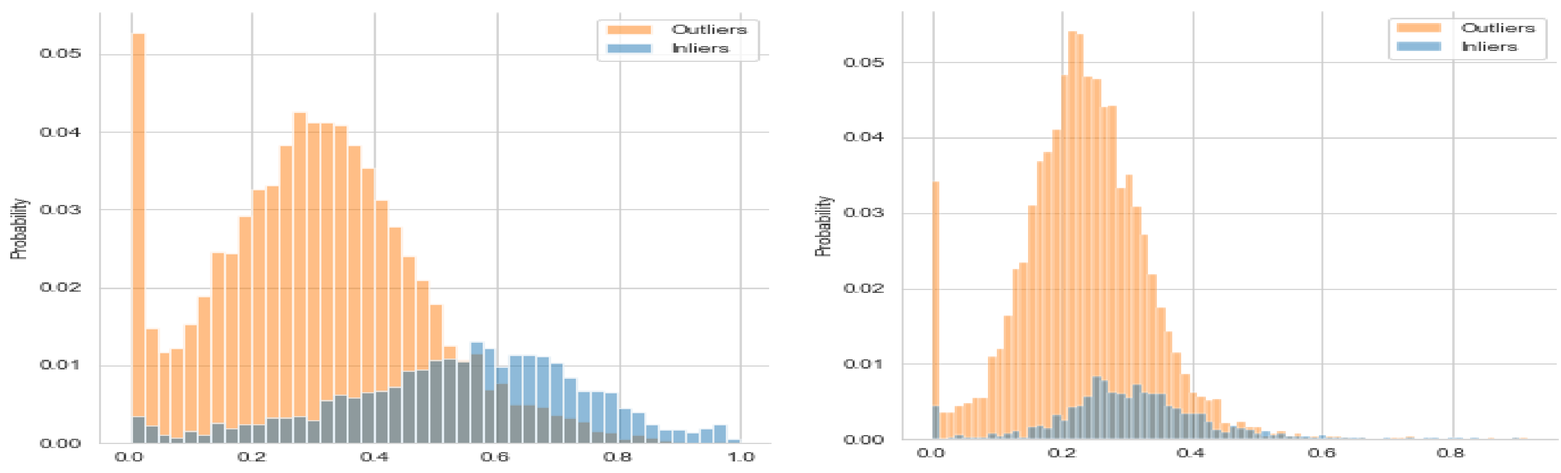


4.A Experiments and Results

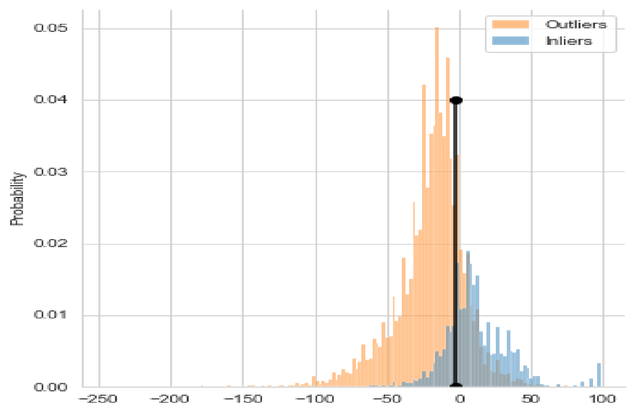
Our model has shown good performances on the **AG News** dataset and modest performances on the **20newsgroup** dataset. We suspect that the difference comes from the distribution of what we called *thematic words*, such as sports names for anomaly detection on a sports split etc. Since **20newsgroup** are forum posts, we suspect that they naturally contain less such words than headlines and article descriptions from **AG News**.
We have used the following formula for calculating a sequence's anomaly score:

$$anomaly\_score = \frac{1}{n}\sum_{i=1}^{n} P(x_i|context)$$

Even though we'd expect anomalies to have much lower scores than inliers, empirically it's not so simple, as shown in the graphs below, which show one of the most easily separable splits and one of the harder ones:



Due to training on a language modeling task, our model learns some of the distribution of the language itself on top of the distribution of *thematic words*. The consequence is that regularly used words will have a high probability *regardless* of the context they find themselves in (as long as it is grammatically and semantically correct). To circumvent this, we have introduce a **thresholding** technique, which gives to any word with a probability higher than a threshold a score of 1, and any word below a negative proportional score. This gives a boost of usually around $1 - 3\%$.



4.B Interesting Results table

AUROC Scores				
Split	Score	Score thr.	CVDD	DATE
Business	83.6%	85.7%	84.0%	90.0%
World	82.4%	82.9%	79.6%	90.1%
Sci	81.0%	81.3%	79.0%	84.0%
Sports	84.5%	89.2%	89.9%	95.9%
Comp	76.3%	79.6%	74.0%	92.1%
Misc	76.0%	78.6%	75.7%	86.0%
Rel	70.0%	70.7%	78.1%	86.1%

6. Conclusions and Future Work

We believe our work has shown promising results, even with the occasional hiccups on the **20newsgroup** dataset. We think that the thresholding technique might prove to be a useful technique for anomaly detection on text, if given perhaps a more mathematically-formal wording.

7. References

[1] Andrei Manolache, Florin Brad, and Elena Burceanu. DATE. 2021.

[2] L. Ruff et al. CVDD. 2019.

[3] Jacob Devlin et al. BERT. 2018.