Self-Supervised Losses for One-Class Textual Anomaly Detection

Anonymous ACL submission

Abstract

Current deep learning methods for anomaly detection in text rely on supervisory signals in inliers that may be unobtainable or bespoke architectures that are difficult to tune. We study a simpler alternative: fine-tuning Transformers on the inlier data with self-supervised objectives and using the losses as an anomaly score. Overall, the self-supervision approach outperforms other methods under various anomaly detection scenarios, improving the AUROC score on semantic anomalies by 11.6% and on syntactic anomalies by 22.8% on average. Additionally, the optimal objective and resultant learnt representation depend on the type of downstream anomaly. The separability of anomalies and inliers signals that a representation is more effective for detecting semantic anomalies, whilst the presence of narrow feature directions signals a representation that is effective for detecting syntactic anomalies.

1 Introduction

Anomaly detection is the task of identifying unusual samples relative to an exemplar inlier distribution. It has numerous applications in natural language processing (NLP), including fake news detection (Lee et al., 2021), spam detection (Crawford et al., 2015), and flagging atypical reviews (Ruff et al., 2019).

The difficulty of anomaly detection depends on the magnitude of difference between an anomalous representation and the distribution of inlier representations. Existing works in NLP focus on the far out-of-distribution (OOD) setting (Winkens et al., 2020) in which the anomalies are derived from a distinct dataset (Hendrycks et al., 2020; Arora et al., 2021; Li et al., 2021; Podolskiy et al., 2021; Zhou et al., 2021). For example, a model is trained on a sentiment classification dataset, and then that model is used to identify news articles as anomalies. These approaches also often assume the model is trained to classify the distinct inlier sub-classes. The anomaly scoring mechanisms typically leverage these supervisory signals by fitting a Mahalanobis distance (Lee et al., 2018) to each sub-class or by obtaining the highest probability in the softmax layer (Hendrycks and Gimpel, 2017). However, these supervisory signals may not always be available.

As an alternative configuration, we analyse the one-class anomaly detection setting on more challenging near-OOD anomalies. One-class anomaly detection assumes only inlier data are available at training time and only have one label. Instead of supervisory signals, we study the performance of fine-tuning a Transformer on the inlier data using various self-supervised objectives, and we use the loss as the anomaly score. We examine anomaly detection performance on two near-OOD anomaly types: semantic anomalies, which are created by partitioning a single dataset by class label, and syntactic anomalies, which are created by randomly shuffling inlier sentences. We find that fine-tuning on a pre-trained Transformer outperforms existing and more complex methods, boosting AUROC score on semantic anomalies by 11.6% and on syntactic anomalies by 22.8% on average.

Our findings also suggest that the separation of anomalies and inlier classes in the learnt representation space of the detectors is a strong signal for detecting semantic anomalies, whilst adversarially brittle features are a better indicator of performance in the syntactic anomaly detection setting. Overall, our results indicate the fine-tuning paradigm is a simple baseline that can achieve good results, and the self-supervised objectives used for fine-tuning exploit different cues to identify anomalies.

2 Approach

2.1 Models

Using the loss of a fine-tuned Transformer for anomaly detection is analogous to using an autoen-

coder's reconstruction error as an anomaly score in vision (Sakurada and Yairi, 2014). We anticipate that the fine-tuned models can learn the underlying characteristics of inlier data but not those of anomalies. Hence, the loss is used as the anomaly score as it should be higher for anomalous instances.

We analyse three self-supervised objectives in our experiments. To minimise the influence of architectural differences, we use the encoder from a pre-trained uncased $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) and append different heads depending on the objective. We fine-tune each model for a maximum of 30,000 steps on inlier data, employing early stopping based on the inlier validation set's loss.

Masked language modelling (MLM). We retain the default configuration for $\text{BERT}_{\text{BASE}}$ and randomly mask 15% of tokens. At inference time, we mask the same proportion of tokens in the test sentences and use the error between the predicted and true tokens as the anomaly score.

Causal language modelling (CLM). We fine-tune the model to predict the next token given previous tokens in the sequence and use perplexity as the anomaly score. Perplexity has been used to evaluate evidence-supported fact-checking (Lee et al., 2021) and far-OOD detection (Arora et al., 2021). Our work differs as it uses perplexity to evaluate more difficult anomalies and does not require auxiliary data.

Contrastive loss (SimCSE). Previous works in vision suggest a contrastive loss can help discriminate anomalies from inliers (Tack et al., 2020; Sehwag et al., 2021). However, these methods require data augmentations that are not directly transferrable to NLP.

SimCSE (Gao et al., 2021) resolves the data augmentation issue by applying different dropout masks to sentences and trains the model to select the same sentence from a minibatch of other sentence pairs. We fine-tune the model using the default dropout probability ($p = 0.1$) and temperature ($\tau = 0.05$) described in SimCSE and evaluate anomalies using the NT-Xent loss (Chen et al., 2020).

We compare the three fine-tuned models to four baselines:

Pre-trained BERT (Pre-trained). We evaluate MLM loss on $\text{BERT}_{\text{BASE}}$ without any fine-tuning. This configuration can be compared to MLM to examine the incremental benefit of fine-tuning. We disregard the auxiliary next-sentence prediction objective as we do not use sentence pairs for anomaly detection.

Other attention-based anomaly detectors. We compare our approach to two state-of-the-art methods which use attention. CVDD (Ruff et al., 2019) learns a set of compact context vectors to describe the inlier data using a multi-head self-attention mechanism. It evaluates a sentence through the average cosine distance of the sentence's contextual embedding to the context vectors.

DATE (Manolache et al., 2021) adapts ELECTRA (Clark et al., 2020) for the anomaly detection task. DATE includes an additional objective to predict which pre-defined pattern was used by the generator to mask the input tokens. At inference time, the input text is fed into the discriminator directly. The average probability of each token being uncorrupted serves as the anomaly score.

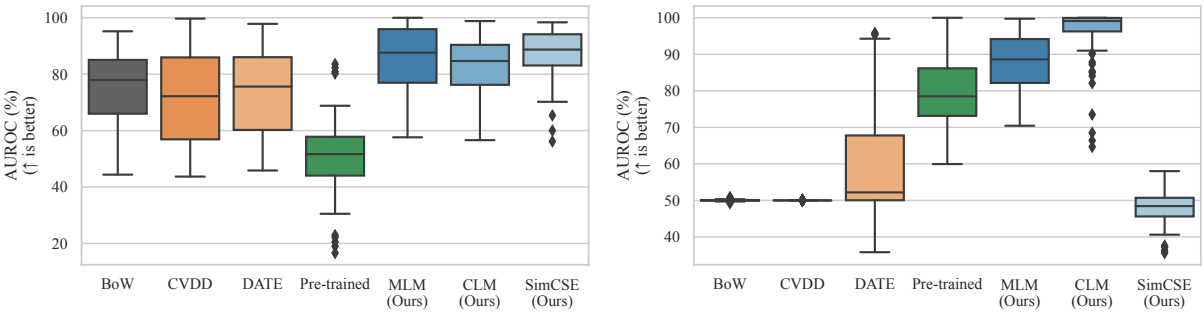
Bag-of-words models (BoW). We follow the approach in CVDD and compute the mean over word embeddings extracted from FastText (Bojanowski et al., 2017) to create a sentence embedding for each datum. We use these sentence embeddings to train linear OC-SVMs, which worked better than using k-NNs or Mahalanobis distances in our experiments.

2.2 Datasets and anomaly detection setup

To allow comparison with the baseline methods, we evaluate anomaly detection performance on 20 Newsgroups (Lang, 1995), Reuters-21578 (Lewis, 1997), AG News (Zhang et al., 2015) and IMDb Movie Reviews (Maas et al., 2011). We also perform experiments on Snopes (Vo and Lee, 2020) (a fact-checking dataset) and the Enron Spam Dataset (Metsis et al., 2006) to simulate more realistic anomaly detection applications. We pre-process each dataset by lowercasing text, stripping punctuation and removing stopwords as per Ruff et al. (2019).

We use the datasets' class labels to construct two setups for the inlier training data. This allows us to examine anomaly detection performance in the settings where the inliers are narrow and more diverse. For a dataset with m class labels:

- Unimodal normality: We construct the inliers using data from a single label.
- Multimodal normality: We construct the inliers using data from $m - 1$ labels.



(a) Semantic anomaly results.

(b) Syntactic anomaly results encompassing all n-grams.

Figure 1: Anomaly detection results aggregated by model.

Class	Sentence
Inlier	voip gaining ground despite cost concerns
Anomaly	concerns voip despite cost ground gaining

Table 1: Example of a syntactic anomaly derived from the AG News dataset. We look at n-grams ($n \in \{1, 2, 3, 4\}$) and shuffle them until each n-gram is no longer in its original position.

We use the test splits of each dataset to formulate two types of near-OOD anomalies:

- Semantic anomalies: Data belonging to the same original class label(s) as the training data are categorised as inliers whilst the remainder are categorised as anomalies.
- Syntactic anomalies: Inlier and anomaly data are derived from the same class of data used to construct the training set. Inlier data are unchanged; anomalies have shuffled word order. To create the anomalies, we implement the seeded random function algorithm in Sinha et al. (2021). This setup allows us to measure the anomaly detectors' sensitivity to the underlying syntactic information whilst fixing the word frequency statistics. We illustrate an example of a syntactic anomaly in Table 1.

3 Results

Figure 1 shows the overall anomaly detection results for both types of anomalies. The full results split by dataset and normality are in Appendix A.

Fine-tuning a pre-trained Transformer boosts anomaly detection performance. In the case of semantic anomalies, although the BoW performance suggests anomaly detection can be performed through analysing word frequency statistics, fine-tuning helps to give additional information about the nature of inliers. This observation aligns with observations in vision (Fort et al., 2021). Our approach also outperforms CVDD and DATE, particularly in the multimodal normality setting.

Fine-tuning also improves syntactic anomaly detection, where frequency statistics are insufficient for discrimination. SimCSE is an exception, and we attribute this to the NT-Xent loss considering the entire sentence representation at inference.

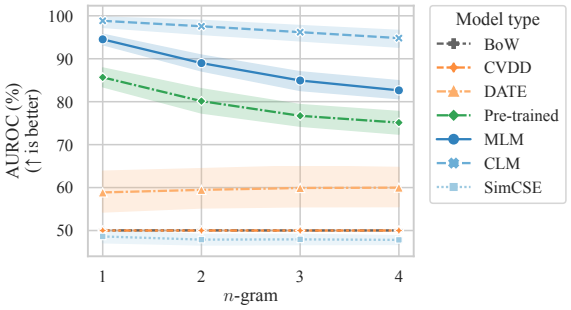


Figure 2: Mean AUROC across datasets on syntactic anomalies by n-gram level. Larger n-grams are more challenging to differentiate from inliers as fewer individual tokens are shuffled.

Density models are much better at detecting syntactic anomalies. We conducted an ablation study of performance under different permutation strengths. CLM is more stable under more challenging anomaly detection conditions (Figure 2), experiencing a decline of only 4% between 1-grams and 4-grams. Pre-trained and fine-tuned MLM experience similar drops (11%), which indicates the choice of objective for anomaly scoring is a core component for performance. As CLM calculates its score at the token level, it is more sensitive to syntactic changes compared to MLM, which considers spans of tokens through its masking mechanism.

In the following experiments, we extracted the embeddings at the last hidden BERT layer and mean-pooled over the positions to analyse the characteristics of the learnt embeddings.

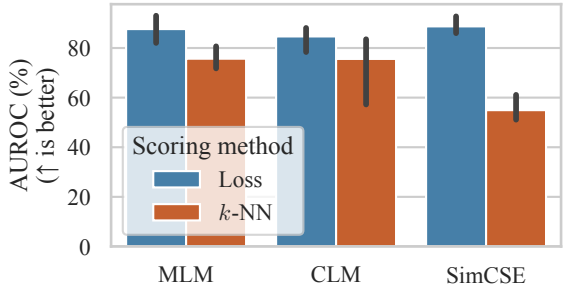


Figure 3: Comparison between using the loss as an anomaly score and k-NNs for semantic anomaly detection. We also experimented with OC-SVMs and Mahalanobis distances, but k-NNs performed best overall.

Using the loss combined with the embedding is better than using the embeddings as a feature extractor. Figure 3 shows the median semantic anomaly detection AUROC score when using the models end-to-end compared to extracting the embedding to train a k-NN. Although the raw embeddings are generally capable of performing anomaly detection, end-to-end use of the methods is more discriminative.

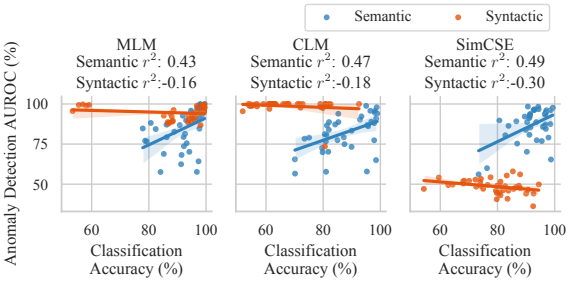


Figure 4: Scatter plot comparing classification accuracy of test inliers versus anomalies to anomaly detection performance across datasets.

Separability of inliers and anomalies is a stronger signal for better semantic anomaly detection. To examine the separability of embeddings for each learnt representation, we extracted both inlier and anomalous embeddings at the last hidden state and trained a logistic classifier. The correlation between classification accuracy and anomaly detection is more apparent for semantic anomalies (Figure 4), suggesting separability is a good indicator for better representations in this case, whereas there is no such relationship for syntactic anomalies. This pattern suggests there is another factor that influences syntactic anomaly detection.

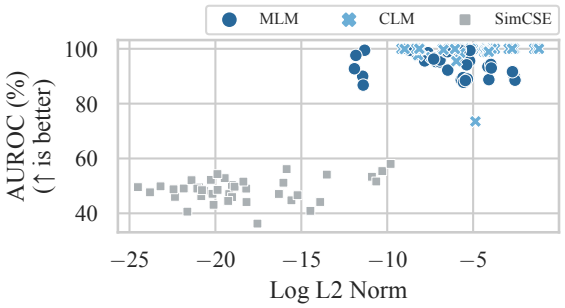


Figure 5: Scatter plot comparing average log L2 norms of the training inlier data to 1-gram syntactic anomaly detection performance. Higher norms are more brittle. The pattern is similar across different n-gram levels.

Syntactic anomaly detection performance is more correlated to brittle features. We hypothesise that a narrower[1] inlier representation is a better signal for syntactic anomaly performance as it provides more directions for anomalies to manifest.

We adopt the procedure in Mai et al. (2021) and calculate the average L2 gradient norms divided by the trace of the covariance matrix with respect to the training data. We observe similar behaviour across all datasets (summarised in Figure 5), whereby higher gradient norms clearly correspond to better anomaly detection performance.

Among the methods, CLM-based embeddings tend to be the most brittle and SimCSE the least. This corresponds with previous literature which states that autoregressive models like GPT (Radford et al., 2018) are highly anisotropic (Cai et al., 2021), and models such as SimCSE which are trained on contrastive objectives are more isotropic (Wang and Isola, 2020; Gao et al., 2021).

4 Conclusion

We studied the performance of fine-tuned Transformers using three self-supervised losses through a range of datasets and anomaly detection tasks. We show that this approach outperforms more complex methods, and employing the loss as an anomaly detector is better than using the learnt embeddings as a feature extractor. The best self-supervised loss depends on the nature of the anomalies, which suggests there is scope for analysing ensemble models or outlier exposure in future work.

[1]Narrow and brittle features refer to non-robust features as defined in adversarial machine learning literature (Ilyas et al., 2019).

Ethical considerations

Anomaly detectors are practical tools for indicating whether a system is working as intended and for flagging potential hazards (Hendrycks et al., 2021). An adversary may learn how to bypass systems by leveraging anomaly detection research. We restrict this by manually curating inliers and anomalies from publicly available datasets (as described in Section 2.2). By construction, our experiments are limited to the English language and may not represent features in other languages. We encourage extending our work to other domains and languages to investigate these differences.

References

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):1–24.

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

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. 2021. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 331–339. Morgan Kaufmann.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

David D. Lewis. 1997. Reuters-21578 text categorization test collection, distribution 1.0.

Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. 2021. kFolden: k-fold ensemble for out-of-distribution detection. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 3102–3115, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Kimberly T. Mai, Toby Davies, and Lewis D. Griffin. 2021. Brittle features may help anomaly detection. *Women in Computer Vision Workshop at the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Andrei Manolache, Florin Brad, and Elena Burceanu. 2021. DATE: Detecting anomalies in text via self-supervision of transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 267–277, Online. Association for Computational Linguistics.

Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes - which naive bayes? In *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.

Mayu Sakurada and Takehisa 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*, pages 4–11.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. 2021. SSD: A unified framework for self-supervised outlier detection. In *9th International Conference on Learning Representations*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33:11839–11852.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7717–7731, Online. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

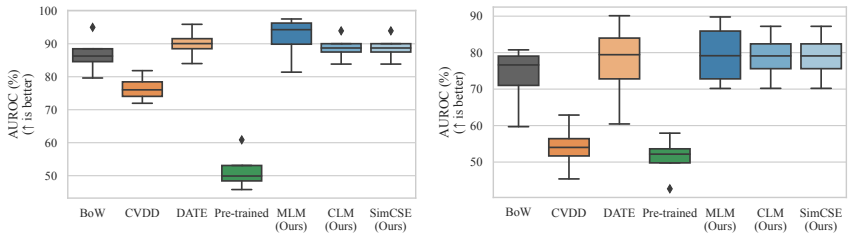
Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. 2020. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

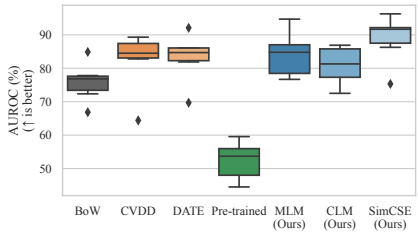
A Appendix

A.1 Semantic anomaly detection results

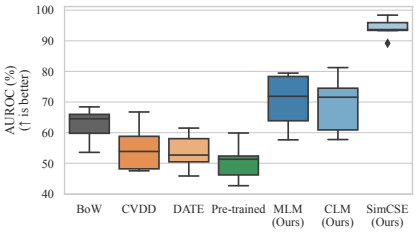


(a) AG News (Unimodal)

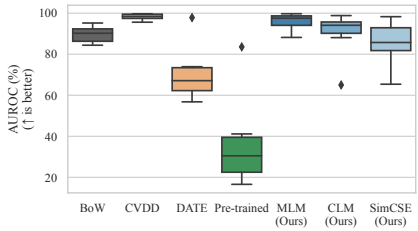
(b) AG News (Multimodal)



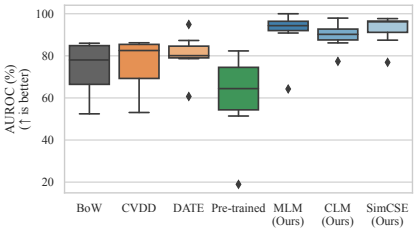
(c) 20 Newsgroups (Unimodal)



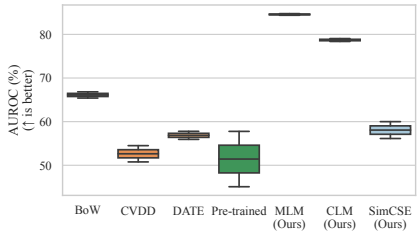
(d) 20 Newsgroups (Multimodal)



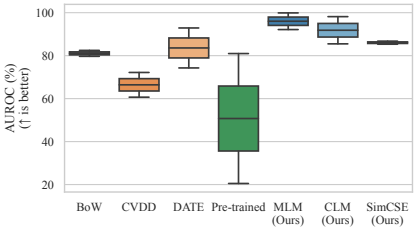
(e) Reuters-21578 (Unimodal)



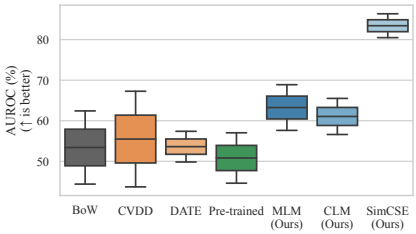
(f) Reuters-21578 (Multimodal)



(g) Snopes



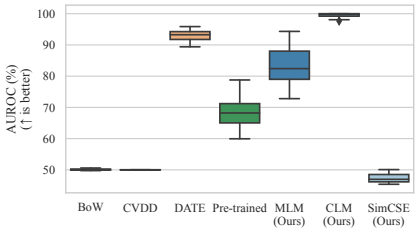
(h) Enron Spam



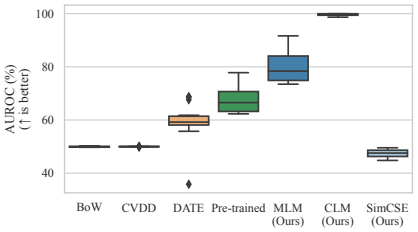
(i) IMDb

Figure 6: Semantic anomaly detection results split by dataset.

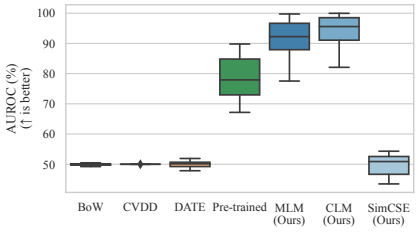
A.2 Syntactic anomaly detection results



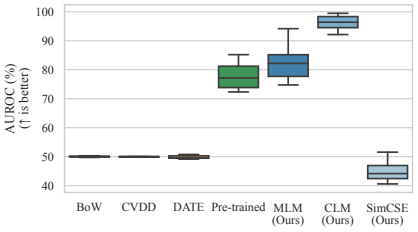
(a) AG News (Unimodal)



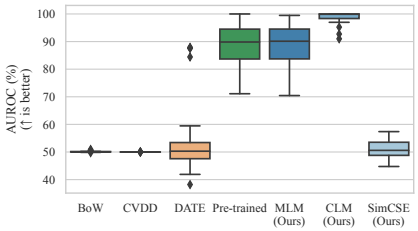
(b) AG News (Multimodal)



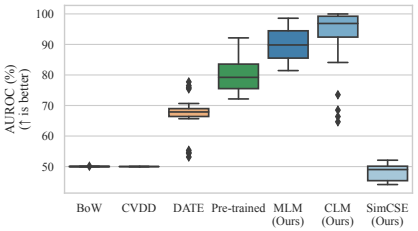
(c) 20 Newsgroups (Unimodal)



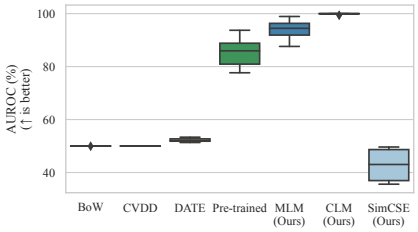
(d) 20 Newsgroups (Multimodal)



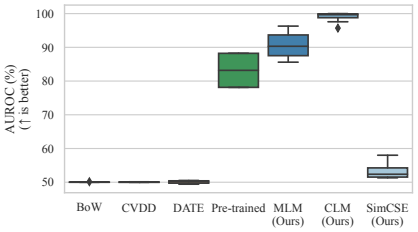
(e) Reuters-21578 (Unimodal)



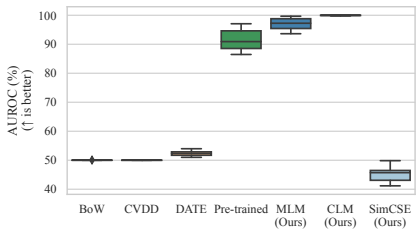
(f) Reuters-21578 (Multimodal)



(g) Snopes



(h) Enron Spam



(i) IMDb

Figure 7: Syntactic anomaly detection results split by dataset. The figures include all n-gram runs.

A.3 Contamination results

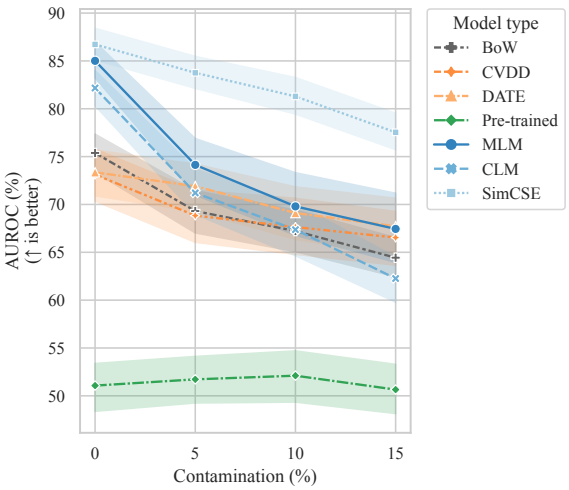


Figure 8: Mean AUROC scores across datasets by contamination percentage. Experiments are conducted using semantic anomalies.

We simulate a purely unsupervised anomaly detection setup by incorporating a set percentage of semantic anomalies $\{5\%, 10\%, 15\%\}$ into the training data. The self-supervised losses on average elicit higher AUROC scores compared to the other model types, and SimCSE appears to be the most robust approach.

A.4 Implementation details

We used an NVIDIA RTX Titan X and NVIDIA Tesla V100s to run our experiments depending on availability.

Model implementation. We used `Huggingface`'s[2] implementation of $\text{BERT}_{\text{BASE}}$ and `Sentence-Transformers`[3] for our Transformer experiments. In addition, we used `nltk`[4] for pre-processing, `spaCy`[5] for encoding the bag-of-words models, `Faiss`[6] to train the k-NNs, and `sci-kit learn`[7] for constructing OC-SVMs.

Dataset details. All of the datasets used in our paper are publicly available.

- 20 Newsgroups (Lang, 1995) is a collection of 20,000 newsgroup documents split across 20 different newsgroups. We use the six top-level subjects (*computer, recreation, science, miscellaneous, politics, religion*) to partition the classes. Partitioning by class label, there are 577-2859 training samples and 382-1909 test samples.

- Reuters-21578 (Lewis, 1997) is a collection of 10,788 news articles split across 90 topics. We only use a subset of data that have only one label (*earn, acq, crude, trade, money-fx, interest, ship*). Partitioning by class label, there are 108-2,840 training samples and 36-1,083 testing samples.

- AG News (Zhang et al., 2015) is a topic classification dataset gathered from more than 2,000 news sources over one year of activity. It contains four classes (*business, sci, sports, world*), each with 30,000 samples for training and 1,900 for testing.

- IMDb (Maas et al., 2011) is a sentiment classification dataset consisting of film reviews. It contains two classes (*pos, neg*), each with 25,000 samples for training and 25,000 for testing.

- Snopes (Vo and Lee, 2020) is a fact-checking dataset containing paired examples of tweets and a fact-checking article from *snopes.com*. There are four classes (*true, mostly true, mostly false, false*). We only use *true* (7,363) and *false* (21,256) tweets in our experiments and do not use the articles. We randomly partition 80% of this smaller dataset for training and use the remaining 20% for testing.

- The Enron Spam Dataset (Metsis et al., 2006) is derived from the Enron Email Dataset. There are two classes, *ham* (16,458) and *spam* (17,171) emails. We randomly partition 80% of the dataset for training and the remaining 20% for testing.

[2] `https://huggingface.co`
[3] `https://sbert.net`
[4] `https://nltk.org`
[5] `https://spacy.io`
[6] `https://faiss.ai`
[7] `https://scikit-learn.org`