Can LLMs Find a Needle in a Haystack? A Look at Anomaly Detection Language Modeling

Leslie Barrett, Vikram Sunil Bajaj and Robert Kingan
Bloomberg
{lbarrett4, vbajaj11, rkingan}@bloomberg.net

Abstract

Anomaly detection (AD), also known as Outlier Detection, is a longstanding problem in machine learning, which has recently been applied to text data. In these datasets, a textual anomaly is a part of the text that does not fit the overall topic of the text. Some recent approaches to textual AD have used transformer models, achieving positive results but with trade-offs in pre-training time and inflexibility with respect to new domains. Others have used linear models which are fast and more flexible but not always competitive on certain datasets. We introduce a new approach based on Large Pretrained Language Models in three modalities. Our findings indicate that LLMs beat baselines when AD is presented as an imbalanced classification problem regardless of the concentration of anomalous samples. However, their performance is markedly worse on unlabeled AD, suggesting that the concept of "anomaly" may somehow elude the LLM reasoning process.

1 Introduction

Traditional machine learning approaches to Anomaly Detection (AD) include proximity-based models where outliers are points that are separated from the rest of the data by a certain distance. High dimensional data was a challenge for traditional models and AD applied to textual data has gotten attention only fairly recently, with Non-negative Matrix Factorization (NMF) approaches performing best (see Barrett et al. (2022), Barrett et al. (2019), Kannan et al. (2017)).

Most recently, Transformer models (Manolache et al., 2021) and word embeddings with multi-head self-attention (Ruff et al., 2019) have been applied in textual AD models, surpassing previously top-performing reconstruction-based approaches using Non-negative Matrix Factorization, although certain updated NMF-based approaches remain competitive (Barrett et al., 2022).

LLMs have recently been used in AD applied to anomalous tabular values (Li et al., 2024), and anomalies in log file data (Ott et al., 2021) but so far not to detect topically anomalous text segments consistent with the present research on AD in text.

We propose a new Large Language Model (LLM)-based approach to detecting these types of anomalies using three testing modalities with three different prompting techniques without fine-tuning. We show results for three LLM models on three datasets. Our best LLM model surpasses current best-performing non-LLM models on two of the three modalities.

We address the following research questions:

R1. How do LLMs compare with the current leading (non-LLM) models on textual anomaly detection?

R2. Do LLMs perform better when the task is framed as highly imbalanced classification, or unlabeled general textual anomaly detection?

This paper is organized as follows: Previous approaches are discussed in Section 2, Data and Methods are discussed in Section 3, our results are in Section 4 and the Conclusion and plans for future work in section 5.

2 Past Approaches

The majority of recent textual AD studies (Manevitz and Yousef (2002), Kannan et al. (2017), Barrett et al. (2019), Ruff et al. (2019), Manolache et al. (2021)), treat textual anomalies as topical intrusions, where the texts from one topic constitute the "inliers" and a smaller set of intrusion texts constitute the "outliers". We use this data definition for our anomaly detection task.

Among such models, the currently best-performing is the transformer approach in Manolache et al. (2021), a discriminator-generator model that outperformed the previously top-performing OCSVM approach in Ruff et al. (2019). NMF-based models used in (Barrett et al., 2022) were also found to be competitive with the transformer approach on certain datasets. Both approaches have consistently outperformed traditional AD models like Isolation Forests (Désir et al., 2013) on text and at this point represent the state of the art. A similar approach to the distribution-based approaches using Gaussian Mixture models (Ait-Saada and Nadif, 2023) outperformed the transformer approach on French short-text corpora.

LLMs have been used for text classification with mixed results. Recent studies have improved results with CoT prompting (Wu et al., 2024), progressive-reasoning-based prompting (Sun et al., 2023) and LLM ensembles (Zhang et al., 2024). In Liu et al. (2023) performance on Out-of-Distribution detection for near-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6428–6435
November 4-9, 2025 ©2025 Association for Computational Linguistics

OOD samples improves greatly with model size but falls short of far-OOD samples, where results are state of the art. However, strong baselines for traditional models, especially on multiclass text classification, remain hard to beat. On AG News for example (Yang et al., 2019) XLNet results reported there remain competitive. Also, in Zhang et al. (2024) the authors note that imbalanced datasets tend to be associated with worse performance, even with their boosting approach which performs well overall. This tends to suggest that LLMs may have difficulty with AD, especially in some of the lower outlier concentrations.

A recent study on textual AD, (Yang et al., 2024) provides results in a zero-shot setting with task descriptions including anomalous and baseline examples. This model showed promising results but did not compare directly with non-LLM models or test fully unlabeled AD. Further details on previous approaches can be found in the Appendix.

Our study focuses on how well LLMs are able to find text that "does not belong", contrasting scenarios where information on the inlier and outlier classes is provided with scenarios where it is not. We compare with traditional baselines.

3 Proposed Methods

LLMs should in theory be well suited to the task of detecting topically anomalous sentences in a text. Autoregressive models such as GPT (Radford, 2018), since they are trained to recognize likely text sequences, should assign low probabilities to unlikely ones. Despite an intuitive fit, LLMs have been used on AD only in limited contexts and, as noted in previous research, may have difficulty with imbalanced data.

For an LLM approach, the problem can be treated as a classification problem with highly imbalanced data. It is unclear that the likelihoods of the outlier classes need to be taken into account as in traditional approaches to such problems. On the other hand, traditional approaches don't have the risk of hallucinated labels. We therefore use three experimental modalities in our tests in order to learn whether the concept of "textual anomaly" is something that a large language model can "understand" or whether AD in text is best handled with LLMs as a simple classification problem. We further try an unlabeled test to see if the LLM can identify text samples that seem "unusual" in a large context of samples.

3.1 Prompting for Anomalies

Our experimental design included three modalities. In each, we prompted the model separately for each dataset and concentration level of anomalous samples. The three modalities were defined as follows:

1. In the *binary outlier detection*, or “binary” modality, we prompted the model to output an "inlier" or "outlier" label for all of the text in the three datasets, given a list of the inlier and outlier topics.

2. In the *multi-class topic detection*, or “multi” modality, we prompted the model to select from a set of labels (topics) for each without specifying which topics were inliers or outliers.

3. In the *unlabeled outlier discovery*, or “unlabeled” modality, we used all the texts together (stratified samples keeping topic concentrations the same as before) and prompted the model to return the samples that did not belong to the corpus in general.

We run GPT-4o and Claude 3.5 Sonnet in each of the three modalities against two non-LLM baselines. OpenAI o1-preview was additionally used only for the unlabeled modality. The choice of LLMs was driven by overall performance and throughput. For the baselines, the first is the DATE model (Manolache et al., 2021), representing the current best performance on AD text data, and the other is the R-NMF model (Barrett et al., 2022), the best among the NMF-based models. We use three textual datasets in three outlier concentrations each.

3.2 Prompting Details

Prompting is done in a zero-shot setting for the binary outlier detection modality, assessing the ability to reason with predefined inlier and outlier topic structure. For the multi-class topic detection modality, we use a few-shot setting with two sample texts for each topic label, testing generalization from minimal examples without anomaly framing. For the unlabeled outlier discovery modality, we use a zero-shot setting, evaluating purely contextual anomaly reasoning without topic framing.

All of our data is used for for the unlabeled outlier discovery modality, and sampled using stratified sampling to address prompt size limitations (GPT-4o has a context window of 128,000 tokens, while Claude 3.5 Sonnet supports 200,000 tokens). We prompt the model to return samples that do not fit the theme of the text. We sample 500 texts from each dataset while keeping the topic concentrations the same as the original dataset.

Temperature is set to 0.1 for GPT-4o and Claude 3.5 Sonnet and top_p (nucleus sampling) is set to the model defaults. Temperature is set to 1 for OpenAI o1-preview. Prompt examples are shown in the Appendix.

4 Experimental Results

Below we describe the datasets and preparation. All models were run on three public datasets representing distinct genres (listserv, news, and wiki). We used three outlier-inlier concentrations for each.

4.1 Data and Experimental Design

Each dataset we use is identical to the extent possible in terms of content and concentration to those used in (Barrett et al., 2022). We note that these are also used in previous work on textual AD. For example, 20Newsgroups is used in Manolache et al. (2021) and both Reuters and 20Newsgroups are used in Kannan et al. (2017) and Ruff et al. (2019).

20Newsgroups is a publicly available listserv collection of approximately 20,000 newsgroup documents organized into 20 topical subgroups[1]. Some subgroups are similar (e.g., IBM/Mac hardware), while others are highly dissimilar (e.g., For Sale/Christian religion). The PC/mac hardware sections are used as inliers and selections from either comp.os or comp.windows are outliers.

Reuters-21578 is a publicly available dataset of news stories appearing on Reuters in 1987[2]. Outliers are taken from Interest and Trade and inliers from Earnings and Acquisitions (combined together as a single inlier corpus).

WikiPeople (Guan et al., 2019) is the subset of the English language Wikipedia dump consisting of the 945,662 articles in the category "living people". Samples from the "Life" section of selected biographies are the inliers and the "Career" sections are used as outlier texts.

For each dataset, the inlier classes listed in Table 2 are blended with a sample from the outlier class to achieve three concentrations: .01, .025 and .05. The size of these concentrations is based on rare event analysis where such events have a chance of occurrence of < 0.05. The .01 concentrations were included except in cases where there are less than 100 samples. Again, the data selections are taken from (Barrett et al., 2022) wherein the selection strategy avoids highly diverse samples. The authors claim this strategy favors samples that are topically similar so as to create a robust test of the compared approaches. To date however, we have not seen any work addressing the issue of how topic "distance" may affect AD results or giving a rigorous definition to topical similarity.

For the R-NMF model (Barrett et al., 2022), we parse the input text into word count vectors using sklearn's CountVectorizer with all default parameters. Following Barrett et al. (2022) we call the factorization routine on the sparse word-document matrix to obtain the low-rank matrices and outlier matrix. We use the ℓ_2 norm of each column in the outlier matrix as the outlier score for every document. We use 3 CPU cores with 8GB RAM.

We train the DATE model (Manolache et al., 2021) on our data as a benchmark, as it represents the current SOTA on textual AD. We use the code provided by the authors[3] to run experiments. We use a learning rate of $1e^{-5}$ and sequences of maximum length 128. Training is stopped at convergence, which occurs after 5000 steps on average. We use the same evaluation framework as proposed by the authors to report results. For the DATE experiments, we use 2 Tesla V100 GPU nodes each with 32 GB RAM and 6 CPU cores.

[1]https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[2]https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[3]https://github.com/bit-ml/date

For our LLM experiments we use GPT-4o and Claude-3.5-Sonnet in the three modalities on all pairings in the three datasets, and OpenAI o1-preview on the unlabeled outlier discovery modality only. Table 2 shows the dataset details.

4.2 Model Results

We list the AUROC results for each dataset for each sample and concentration. Results are in Table 1. For background on reporting AD model quality as AUROC see Aggarwal (2016).

The results for the R-NMF model are the best from a sweep of eight values of the hyper-parameter k within the range [1,128] and 5 values of alpha within the range [1,16] following (Barrett et al., 2022). The DATE model was run with the parameters discussed in the previous section. All models were run on all concentrations for each dataset.

4.3 Results Analysis

The results overall indicate that the outlier concentration has little effect on any model. The effects that do appear seem to depend on both the dataset and the model. It is possible that due to the sample sizes, concentration differences of this magnitude simply do not have much effect. It is also possible that the within-sample variance has some effect, since the outliers chosen for each pairing are re-sampled, not added to the previous sample. Thus, although one blend has a higher concentration, that particular blend may be more difficult inherently. This in fact would be consistent with the findings in (Liu et al., 2023), where performance on near-OOD samples differed from far-OOD samples. Also a recent study (Tajwar et al., 2021) showed that OOD detection methods are inconsistent across datasets. It is certainly possible that some topic-pairings' word distributions are more divergent than others.

For the binary outlier detection, both the LLM models surpass previous results except for the Wikipeople dataset 4, where Claude scores slightly below R-NMF. Among the LLMs themselves however, the binary outlier detection modality is more challenging than the multi-class topic detection modality. Within this group, LLMs perform best on the Reuters datasets, and worst on the Wikipeople datasets. Wikipeople in fact is challenging for all models in all modalities. For the multi-class topic detection modality, both models perform very well on all concentrations and we see none of the effects found in Zhang et al. (2024) for imbalanced multiclass classification.

For all LLM models the unlabeled outlier discovery modality is associated with the worst performance, and baselines do considerably better. The results of 0.49 AUC across many datasets for the LLM unlabeled outlier discovery runs results from the model failing to capture any outliers at all (0.0 Precision) for any of those datasets and concentrations, but getting most inliers (TNs) correctly with a few FPs. This common

Table 1: AUROC Results for LLMs (GPT: GPT-4o, Claude: Claude 3.5 Sonnet, OpenAI o1-preview: AI_o1) across all Modalities (B: Binary, M: Multi, U: Unlabeled)

Dataset	R-NMF	DATE	GPT_B	Claude_B	GPT_M	Claude_M	GPT_U	Claude_U	AI_o1_U
rtr01	0.769	0.691	0.976	0.986	0.992	0.986	0.713	0.857	0.500
rtr02	0.766	0.712	0.977	0.986	0.996	0.966	0.538	0.580	0.538
rtr03	0.777	0.725	0.977	0.986	0.997	0.973	0.563	0.632	0.496
rtr10	0.889	0.886	0.979	0.993	0.998	0.974	0.700	0.698	0.797
rtr11	0.859	0.905	0.979	0.992	0.998	0.982	0.684	0.763	0.763
rtr12	0.877	0.894	0.978	0.992	0.998	0.966	0.832	0.830	0.998
ng01	0.592	0.650	0.922	0.904	0.925	0.908	0.495	0.494	0.497
ng02	0.559	0.767	0.922	0.926	0.931	0.897	0.516	0.496	0.498
ng03	0.595	0.691	0.954	0.973	0.998	0.957	0.494	0.495	0.498
ng04	0.555	0.712	0.954	0.973	0.979	0.948	0.511	0.496	0.497
wp03	0.694	0.548	0.797	0.973	0.992	0.894	0.529	0.529	0.496
wp04	0.707	0.617	0.802	0.688	0.983	0.899	0.491	0.714	0.712

Table 2: Data Details for Pairings

Name	Source	Inlier Topics	Outlier topic	Outlier fraction	Samples
rtr01	Reuters	earn+acq	interest	0.01	5895
rtr02	Reuters	earn+acq	interest	0.025	5951
rtr03	Reuters	earn+acq	interest	0.05	5994
rtr10	Reuters	earn+acq	trade	0.025	5947
rtr11	Reuters	earn+acq	trade	0.05	6095
rtr12	Reuters	earn+acq	trade	0.01	5855
ng01	20Newsgps	pc/mac hardware	comp.os	0.025	1993
ng02	20Newsgps	pc/mac hardware	comp.os	0.05	2043
ng03	20Newsgps	pc/mac hardware	comp.winds.x	0.025	1993
ng04	20Newsgps	pc/mac hardware	comp.winds.x	0.05	2043
wp03	WikiPeople	life	career	0.05	5250
wp04	WikiPeople	life	career	0.025	5125

behavior may be caused by missing the idea of a topical outlier entirely.

The performance of LLMs in the unlabeled outlier discovery modality is unexpected given recent experiments on the similar task of authorship attribution. In (Huang et al., 2024), LLMs performed well with certain prompting techniques and without fine-tuning. In our study, however, while there were exceptions for some datasets, all models generally fall short of the baselines. We provide some sample false negatives and false positives in the Appendix. In each case, the model focuses on aspects other than topicality in its reasoning, suggesting that a topical "anomaly" may not be something it inherently understands, despite recognizing other anomalous aspects of the texts. In one case for example, it focuses on the mention of a particular currency. While that mention might be frequent, it is not associated with the topic of the article. It seems likely that the model's idea of "topic" is overly biased by lexical frequency, and fails to pick up on recurring themes. Further analysis of model reasoning reveals that models are attending to stylistic rather than topic cues, again possibly as a result of over-sensitivity to lexical overlap.

It seems possible that giving the model a better idea of what a topic is may lead to greater improvement than providing hints as to what a textual anomaly is. Some approaches in the area of topic-change detection, which have not so far been widely explored in the context of LLMs, might be useful here.

5 Conclusion and Future Work

We have run AD experiments in three modalities, on three datasets in multiple anomaly concentrations. We found that both experimental models, GPT-4o and Claude-3.5-Sonnet beat baselines when in binary outlier detection and multi-class topic detection modes, but all three LLMs performed quite poorly in unlabeled outlier discovery mode. This unexpected result suggests that LLMs might be a poor choice for AD in contexts where anomalous text genres may vary and where prompting may not be able to include typical inlier and outlier samples.

We did not find the sensitivity to the concentrations of anomalous texts that previous studies suggested. Instead we found that both experimental models are highly robust text classifiers even without fine tuning and despite the extreme imbalance characteristic of AD datasets.

Overall, we would continue future research on larger datasets, perhaps including real-world AD scenarios. We would also conduct further experiments with additional prompting techniques to gain a better understanding of why these models seem unable to identify text that "does not belong".

6 Ethical Considerations

Anomaly detection is a type of classification model, which may have imperfect Precision and Recall. As such it should be subject to human review in contexts of high risk. Deployment in the context of a real listserv or subscription media products could lead to users being banned due to false positive outputs as well as unwanted or offensive posts being allowed due to false negatives.

7 Limitations

Our experimentation and analysis focused on comparing three LLMs on specific datasets. These datasets were fairly small in size, and future work might benefit from experimentation on larger and more varied datasets. Using datasets with different topical mixes and outlier concentrations might yield different results. Additionally, adding more recently released LLMs to the mix would allow for more detailed results analysis.

References

Charu C. Aggarwal. 2016. *Outlier Analysis*, 2nd edition. Springer Publishing Company, Incorporated.

Mira Ait-Saada and Mohamed Nadif. 2023. Unsupervised anomaly detection in multi-topic short-text corpora. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1392–1403, Dubrovnik, Croatia. Association for Computational Linguistics.

Leslie Barrett, Sidney Fletcher, Robert Kingan, Mrinal Kumar, Anu Pradhan, and Ryon Smey. 2019. Textual outlier detection and anomalies in financial reporting. KDD '19, New York, NY, USA. Association for Computing Machinery.

Leslie Barrett, Robert Kingan, Alexandra Ortan, and Madhavan Seshadri. 2022. A lightweight yet robust approach to textual anomaly detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 62–67, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Chesner Désir, Simon Bernard, Caroline Petitjean, and Laurent Heutte. 2013. One class random forests. *Pattern Recogn.*, 46(12):3490–3506.

Saiping Guan, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2019. Link prediction on n-ary relational data. In *Proceedings of the 28th International Conference on World Wide Web (WWW'19)*, pages 583–593.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *ArXiv*, abs/2403.08213.

Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. 2017. Outlier detection for text data: An extended version.

Aodong Li, Yunhan Zhao, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. 2024. Anomaly detection of tabular data using LLMs.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yu Feng, Lei Xue, and Xiao-Ming Wu. 2023. How good are LLMs at out-of-distribution detection? In *International Conference on Language Resources and Evaluation*.

Larry M. Manevitz and Malik Yousef. 2002. One-class SVMs for document classification. *J. Mach. Learn. Res.*, 2:139–154.

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.

Harold Ott, Jasmin Bogatinovski, Alexander Acker, Sasho Nedelkoski, and Odej Kao. 2021. Robust and transferable anomaly detection in log data using pre-trained language models. *2021 IEEE/ACM International Workshop on Cloud Intelligence (CloudIntelligence)*, pages 19–24.

Alec Radford. 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.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. 2021. No true state-of-the-art? OOD detection methods are inconsistent across datasets. *CoRR*, abs/2109.05554.

Hui Wu, Yuanben Zhang, Zhonghe Han, Yingyan Hou, Lei Wang, Siye Liu, Qihang Gong, and Yunping Ge. 2024. Quartet logic: A four-step reasoning (qlfr) framework for advancing short text classification. *ArXiv*, abs/2401.03158.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Lin, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, Xia Hu, and Yue Zhao. 2024. AD-LLM: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of LLM capacity for text classification.

A Appendix: LLM Prompts

This section shows the prompts for the three aforementioned modalities that were used for both GPT-4o and Claude 3.5 Sonnet and the unlabeled outlier discovery modality for OpenAI_o1-preview. In the context of the binary outlier detection and multi-class topic detection modality prompts below, "text" is a single sample from the dataset, while for the unlabeled outlier discovery modality prompt, samples from the dataset were provided as a list of "sentences".

Binary Outlier Detection Modality Prompt

```
You are provided a text and your goal is
to classify it into one of two categories:
"inlier" or "outlier".
You are also given the list of topics that
must be used to determine the category.

Inlier topics:
-  topic_1
-  topic_2
- ...
-  topic_n
Outlier topic:
-  outlier_topic

Return your response for the text below in
JSON format with keys:  "classification":
<"inlier" or "outlier">, "reasoning":
<reasoning for selecting classification>,
"topic": <topic that the text belongs to from
the aforementioned topics>

text:  input_text
```

Multi-class Topic Detection Modality Prompt

```
You are provided a text and your goal is to
classify it into one of the listed topics
below.

Topics:
-  topic_1
-  topic_2
- ...
-  topic_n

Some examples to assist with the
classification:

- Topic:  topic_1
- Examples:
-  example_1a
-  example_1b
- ...

- Topic:  topic_2
- Examples:
-  example_2a
-  example_2b
- ...

...

Return your response for the text below in JSON
format with keys: "classification": <topic>,
"reasoning": <reasoning for selecting topic>

text:  input_text
```

Unlabeled Outlier Discovery Modality Prompt

```
You are provided a numbered list of sentences
and your goal is to detect sentences from
this list that look like they do not belong,
i.e., they are outliers when compared to the
other sentences.

Do not consider grammatical features or
sentence length when deciding if a sentence
is an outlier.  Focus more on the content
of the sentences—their meaning, topics, and
themes.  Think carefully while making your
selections.  Do not take the ordering of
sentences into account when determining
outliers.

Return your response as a JSON object with
the key "outlier_sentences" and the value
is a list of the sentence indices from the
provided list that you think are outliers.
For these outliers, also return a reasoning
for why you think the sentence is an outlier.
You can use the format: {"outlier_sentences":
[{"sentence_index": <outlier sentence index>,
"reasoning": <reasoning>}]}

Do not make up sentence indices.  Return
sentence indices as they appear in the list.

Sentences:
0.  sentence_0
1.  sentence_1
...
```

B Appendix: Error Examples

B.1 Binary Outlier Detection Modality

Model: Claude, **Dataset:** wf04
Topic: plot, **Predicted topic:** production
Error type: False Negative (outlier predicted as inlier)
Text: Hajj, a rascally beggar on the periphery of the court of Baghdad, schemes to marry his daughter to royalty and to win the heart of the queen of the castle himself. Production Warner Bros. spared no expense in making this picture. They spent $600,000 in producing it, and the extravagance of the film was noted by every reviewer. The film played in ten cities across the United States in the wide-screen Vitascope (65mm) version, while the rest of the country (which did not yet have theaters capable of playing widescreen films) were provided with standard 35mm prints. Otis Skinner at 73 was up in age when he made the film. He was younger at 53 when he premiered the Broadway play in 1911, and it was the hit of his long career.
Reasoning: The text primarily focuses on the production aspects of a film, including its budget ($600,000), the production company (Warner Bros.), technical details about film formats (Vitascope 65mm vs 35mm prints), and distribution information. While there is a brief mention of the plot at the beginning, the majority of the content discusses production-related elements.

Model: GPT, **Dataset:** rtr01
Topic: earn, **Predicted topic:** interest
Error type: False Positive (inlier predicted as outlier)
Text: DATA GENERAL DGN TO TAKE 2ND QTR CHARGE WESTBORO, Mass., April 7 - Data General Corp said it expects to record an extraordinary pretax charge of about 18.2 mln dlrs in the second quarter ended March 28. The company said the charge will result from a redemption of debt. On May 18, Data General said, it will redeem 150 mln dlrs of outstanding 12-3/8 pct sinking fund debentures due 2015. The redemption price will be 111.138 pct of the total principal amount of notes then outstanding plus accrued interest to the redemption date, it said. No interest will accrue or will be payable on the debentures starting May 18, the computer and communications company said. Data General reported a loss of 1.8 mln dlrs or six cts a share for the second quarter ended March 29, 1986. The loss included a charge of 3.8 mln dlrs from redemption of debentures. In the year-ago quarter, revenues fell less than one pct to 318.8 mln dlrs. Reuter
Reasoning: The text primarily discusses financial details related to debt redemption and interest on debentures, which aligns with the 'interest' topic. It does not focus on acquisitions or earnings.

B.2 Multi-class Topic Detection Modality

Model: GPT, **Dataset:** wf01
Topic: plot, **Predicted topic:** production
Error type: False Negative (outlier predicted as inlier)
Text: The film plays out with three first-season episodes edited together into a single story: "The Benefit", "Breaking the Lease", and "The Ballet", with new footage included between episodes to help transition the episodes into one coherent storyline. As the series routinely took the format of filming scenes in chronological order, this adds to the "show within a show within a show" format of the film, as viewers watch the cast perform the episodes live. The film itself ends with a "curtain call", as the cast comes out and Arnaz thanks the audience for their support.
Reasoning: The text describes how the film was created by editing together episodes from a TV series and adding new footage to create a coherent storyline. It discusses the filming format and the structure of the film, which are aspects related to the production process rather than the narrative or storyline of the film.

Model: Claude, **Dataset:** rtr01
Topic: earn, **Predicted topic:** interest
Error type: False Positive (inlier predicted as outlier)
Text: HOMESTEAD FINANCIAL HFL UPS CASH DIVIDEND BURLINGHAME, Calif., March 5 - Homestead Financial Corp said it has increased the dividend on its Class A common shares to 6-1/4 cts a share, from five cts prior, while declaring an initial dividend on its Class B common stock of 3-3/4 cts a share. Homestead said the two classes of stock emerged as part of its recapitalization plan which also doubled the number of authorized common, comprising both classes, to 11 mln shares. Homestead said the dividends are payable on April 6, to shareholders of record March 20. Homestead also said that class b shareholders who want to increase their dividends may exchange their shares for an equal number of class A shares on or before March 20. Reuter
Reasoning: The text discusses dividend payments and changes in dividend structure, which is related to financial interest payments to shareholders. The article focuses on Homestead Financial Corp's announcement of increased cash dividends for Class A shares and new dividend payments for Class B shares, along with details about payment dates and share exchange options. This type of financial news about dividend payments and rates falls under the 'interest' category as it deals with returns on investment to shareholders.

B.3 Unlabeled Outlier Discovery Modality

Model: Claude, **Dataset:** rtr02
Error type: False Positive (inlier predicted as outlier)
Text: TRACTORS MALAYSIA HOLDINGS BHD KUALA LUMPUR, March 7 - Six months to December 31 SHR 8.8 cts vs 0.5 ct Interim dividend 12.5 cts vs nil Group net 9.5 mln ringgit vs 0.6 mln Pre-tax 11 mln vs 1.1 mln Turnover 88.9 mln vs 70.8 mln Note - dividend pay May 15, register April 17. REUTER
Reasoning: This sentence discusses Malaysian company earnings in ringgit currency, which is notably different from the predominantly U.S.-focused corporate news in the dataset.

Model: GPT, **Dataset:** rtr09
Error type: False Positive (inlier predicted as outlier)
Text: UNITED ARTISTS UACI.O SUSPENDS MERGER TALKS DENVER, Oct 20 - United Artists Communications Inc and United Cable Television Corp said they have indefinitely suspended negotiations on a proposed merger of their companies previously announced. The companies cited a combination of extraordinary market conditions and unresolved terms of the merger as contributing to the action. Reuter
Reasoning: This sentence discusses the suspension of merger talks between United Artists Communications Inc and United Cable Television Corp due to market conditions and unresolved terms, which is unrelated to the financial performance or transactions of companies that the other sentences focus on.

B.4 Details of Previous Approaches

Matrix factorization (MF) models find outliers through a reconstruction process that isolates outlier documents as residual noise. In this approach, a term-document matrix is created where terms correspond to rows and documents correspond to columns. In (Barrett et al., 2022) there is a term-topic matrix and a topic-document matrix. A residual matrix is intended to capture outliers depending on the configuration of its norm. The idea is that if a document is not representable as a linear combination of topics, the corresponding column in the matrix will have more entries. The quality of the result depends on manipulating norms on both the residual matrix and the low-rank approximation of the input matrix. These methods use sparse representations as input to the model. See (Kannan et al., 2017; Barrett et al., 2022) for details on MF-based AD.

Word-embedding-based representations used with one-class modeling approaches and leveraging self-attention (Ruff et al., 2019) are competitive with the latest MF models, relying on the more detailed dense representations coupled with attention to identify the outlier classes.

Transformer-based models like DATE (Manolache et al., 2021) take yet a different approach. Such methods train a model to discriminate between different transformations applied to the data, using the output to compute an anomaly score. The DATE application in particular uses a novel self-supervised task for text, called Replaced Mask Detection (RMD). This comprises a discriminative task to create training data by transforming an existing text using one out of K given operations on the sample. The transformation then does two things; it masks some input words through a pre-defined pattern and it replaces the masked words with other words. In this way, the model computes the probability of a token being an original vs a "corrupted" token.