Unsupervised Anomaly Detection in Parole Hearings using Language Models

Graham Todd
Stanford University
gdrtodd@stanford.edu

Catalin Voss
Stanford University
catalin@stanford.edu

Jenny Hong
Stanford University
jyunhong@stanford.edu

Abstract

Each year, thousands of roughly 150-page parole hearing transcripts in California go unread because legal experts lack the time to review them. Yet, reviewing transcripts is the only means of public oversight in the parole process. To assist reviewers, we present a simple unsupervised technique for using language models (LMs) to identify procedural anomalies in long-form legal text. Our technique highlights unusual passages that suggest further review could be necessary. We utilize a *contrastive perplexity score* to identify passages, defined as the scaled difference between its perplexities from two LMs, one fine-tuned on the target (parole) domain, and another pre-trained on out-of-domain text to normalize for grammatical or syntactic anomalies. We present quantitative analysis of the results and note that our method has identified some important cases for review. We are also excited about potential applications in unsupervised anomaly detection, and present a brief analysis of results for detecting fake TripAdvisor reviews.

1 Introduction

California houses America's largest "lifer" population, with 25% of its 115,000 prisoners serving life sentences. Each year, the Board of Parole Hearings (BPH) conducts thousands of parole hearings to decide whether to grant prisoners early release. As California has enacted legislation to reduce its prison population, the number of hearings is scheduled to double this year and continue to rise for the foreseeable future. While each hearing is transcribed into about 150 pages of dialogue and sent to the BPH and governor's office for review, capacity constraints mean that, in practice, only grants of parole are reviewed. Legal scholars who painstakingly analyzed small subsets of transcripts have found that parole decisions are sometimes made in an arbitrary and capricious manner (Bell, 2019), but they lack the resources for ongoing review.

PRESIDING COMMISSIONER: Let me ask you a question, Mr. [REDACT]. Are you angry?
INMATE [REDACT]: No.
PRESIDING COMMISSIONER: You seem kind of like you're a smart ass. I don't mean to say that rudely, but are you a smart ass?

Figure 1: Example of a semantic anomaly

To help alleviate these capacity constraints and allow for greater review of parole denials, we propose an automatic anomaly detection system that allows reviewers to focus their attention on the most anomalous portions of text in each hearing.[1] The lack of gold anomaly labels precludes the use of many supervised anomaly detection techniques, so instead we propose using language models trained on the parole transcripts to perform unsupervised anomaly detection.

Defining an "anomaly" in this context is challenging. There are many ways in which a piece of text might be unusual without constituting grounds for additional review. We distinguish primarily between *non-semantic*, *semantic*, and *procedural* anomalies. We define a *non-semantic* anomaly as an irregularity in the linguistic structure of a piece of text (for instance, a sentence fragment). A *semantic* anomaly, by contrast, is one caused by the meaning of the text. In the context of a parole hearing, a conversation that deviates substantially from the typical topics of discussion would constitute a semantic anomaly. Finally, a *procedural anomaly* is an irregularity that indicates the hearing differed substantively from the prescribed guidelines. Often, a procedural anomaly will also be a semantic

[1] Our project raises ethical questions about the use of technology in criminal justice review procedures. We provide a statement about the ethical implications of our work in Appendix A.

https://doi.org/10.18653/v1/P17

anomaly. Figure 1 represents such a case, as it both includes language atypical for a parole hearing and, more generally, indicates a breakdown in communication between the commissioner and the parole candidate. We note that there are also, of course, legal anomalies that do not manifest as atypical language.

A language model (LM) provides an organic way to identify unusual text through its perplexity score. We hypothesize that many procedural anomalies can be identified by examining statistical anomalies in the texts of transcripts, which would seemingly allow for their detection by an LM. However, most instances of unusual text found by a naive LM are *non-semantic*, consisting of typos, ungrammatical sentences, etc. To solve this problem, we instead use a pair of language models. We define our anomaly metric, the *contrastive perplexity score*, as the scaled difference between the perplexity of one LM, which has been fine-tuned on the target domain, and the perplexity of another LM, which has only been pre-trained on out-of-domain text. Non-semantic anomalies will have high perplexity under both LMs (and thus low *contrastive perplexity*), so the second LM acts as a "normalizer" for non-semantic content. We present our results on a human-annotated subset of the parole data. Our method recalls 71% of human-labeled *procedural* anomalies while only asking experts to review 50% of the text of each transcript. We also show that our method can be extended to other domains where a large labeled corpus of anomalous text is unavailable, namely the task of opinion spam detection in TripAdvisor reviews.

2 Related Work

Anomaly detection (AD) techniques cover a range of problem settings. Schölkopf et al. (1999); Hodge and Austin (2004); Chandola et al. (2009); Sakurada and Yairi (2014); Ruff et al. (2018); Schlegl et al. (2017) present general techniques for out-of-sample anomaly detection, with an increasing interest in deep unsupervised AD.

Text is a challenging regime for AD because of the importance of domain-dependence: what is shocking in one case might be mundane in another. Few, if any, universal features for AD exist. General approaches for text AD include non-negative matrix factorization (Kannan et al., 2017) and the use of "selectional preferences" (Dasigi and Hovy, 2014). One notable approach, studied in the discourse coherence literature, is to focus on local abnormalities in topics. Li and Jurafsky (2017) and Lin et al. (2011) present deep models for identifying incoherent passages of text, but discourse coherence studies much shorter text than parole hearings. To address longer text, our approach, like that of Guthrie et al. (2008), splits each document into segments ranked by anomaly score.

Our strategy of using LMs for AD has precedents, but primarily much simpler LMs, and for AD contexts that require more supervision than is available in the parole hearing setting. Rieck and Laskov (2006, 2007) and Aktolga et al. (2011) use n-gram LMs to identify anomalous sections and documents in a corpus of American bills presented before Congress. Axelrod et al. (2011) and Xu et al. (2019) also explore using a "baseline" LM for translation and discourse coherence, respectively.

3 Approach

Our model uses GPT-2, a transformer-based LM pre-trained on WebText, a corpus scraped from the internet (Radford et al., 2019; Vaswani et al., 2017). The following three observations motivate our approach to identifying anomalous text: (1) The perplexity of a *fine-tuned* LM on a target domain yields a score that measures both genre-specific semantic anomalies and general language anomalies (e.g. ungrammatical inputs, misspellings, incoherence). (2) The perplexity of an LM only *pre-trained* on many domains represents solely general language anomalies. (3) Putting the two together, the difference in perplexity between a fine-tuned language model and a pre-trained language model gives a "semantic anomaly score" of a piece of text.

We define the *contrastive perplexity* LM anomaly score to be the scaled difference in perplexities observed from two models. One model, the *fine-tuned LM*, is fit to a target corpus of text, without any supervision on which passages are anomalies. The other model, the *normalizer LM*, is the out-of-the-box GPT-2 model (Radford et al., 2019; Vaswani et al., 2017).

$$\textbf{LM anom.} = \text{pplx}_{\text{fine-tuned}} - \beta \cdot \text{pplx}_{\text{normalizer}}.$$

For a mundane piece of text, both $\text{pplx}_{\text{fine-tuned}}$ and $\text{pplx}_{\text{norm.}}$ are low. For a non-semantic anomaly, both are high. In both cases, contrastive perplexity is low. However, for a semantic anomaly, we expect $\text{pplx}_{\text{fine-tuned}}$ to be high, because of its sensitivity to the text's context domain, and

$\text{pplx}_{\text{norm.}}$ to be low, because the text may not otherwise be unusual in general English, leading to high contrastive perplexity.

Because the fine-tuned LM achieves a lower perplexity, we use β to re-scale the perplexity output of the normalizer and ensure the models operate at the same scale. While β can be tuned as a hyperparameter, a reasonable and balanced choice is the ratio between the mean perplexities achieved by the fine-tuned model and the normalizer model on a validation dataset.

$$\beta = \frac{\sum_{x \in \text{val}} \text{pplx}_{\text{fine-tuned}}(x)}{\sum_{x \in \text{val}} \text{pplx}_{\text{normalizer}}(x)}$$

3.1 Anomaly Aggregation

We can use our LM anomaly score to identify the top k chunks of anomalous text for a given set of documents directly. In a completely unsupervised setting, with no labels as to which documents (or chunks) are anomalies, there is no way to associate the absolute contrastive perplexity scores with the predictive target. However, if given a *clean dataset* (i.e. a validation set that is labeled and known not to contain anomalies) we can instead anchor the scores to the clean dataset and detect anomalies by performing an out-of-distribution test.

4 Experimental Setup

4.1 Baselines

We compare our model to a number of unsupervised baseline models.

Within AD, most existing algorithms are unsuitable for our task (e.g. due to the need for supervision, incompatibility with long-form text). The most straightforward baseline is simply the fine-tuned GPT-2 model alone. We also compare our work to an unsupervised topic-modeling baseline that should also be agnostic to non-semantic anomalies, like Misra et al. (2008). We fit a latent Dirichlet allocation (LDA) model (Blei et al., 2003) to our train-corpus, then compute the mean representation and covariance matrix over topics, over a held-out portion of data. At prediction time, we compute the LDA representation for some text $f(x)$ and use its Mahalanobis distance from the mean representation as our anomaly score: $\sqrt{(f(x) - \mu_T)^T \Sigma_T^{-1} (f(x) - \mu_T)}$, where μ_T and Σ_T are the sample mean and covariance over the topic mixture embeddings, respectively.

4.2 Parole Hearings

Our analysis is performed over the complete[2] set of parole hearing transcripts in California between January 2007 and July 2018, which totals 30,734 transcripts. Each document is a transcript of an hours-long conversation between the parole board and a candidate (other parties are occasionally also present), which ends in a decision from the parole board. Transcribed, each hearing is roughly 27,000 tokens long.

We train our model on a train corpus of 27,577 transcripts, each split into non-overlapping chunks of 1024 tokens. We fit β on a validation corpus of 1,963 transcripts, with chunksize 256. The training chunksize was selected to maximize efficiency of the underlying GPT-2 model, while the smaller validation chunksize better matches the scale at which we expect to observe linguistic anomalies. We collected a held-out test corpus of anomalies over 315 transcripts by asking undergraduate and law students to label instances of anomalous language. Out of 82,959 chunks, students found 179 anomalies. An experienced parole attorney checked the anomalies and confirmed 68. Student reviewers were asked to identify *semantic* anomalies and the expert was asked to determine which of those were also *procedural* anomalies. While we believe that this offers a viable estimate of the true set of procedural anomalies, this leaves out anomalies that are not manifested by irregular language. To evaluate our model's recall, we investigate the tradeoff between the share of the expert's "true anomalies" we recover, and the number of chunks human reviewers must read. We asked the parole attorney to review our model's predictions at a fixed threshold. We compute the mean reciprocal rank (MRR) (**?**), rather than precision, because a single anomaly suffices to flag a whole transcript for review: only the rank of the highest scoring anomaly affects reviewer time. Details are given in Appendix B.

4.3 Hotel Reviews

Our second experiment is performed over the Deceptive Opinion Spam dataset (Ott et al., 2011, 2013). The dataset consists of 1,600 short human-generated reviews of 20 hotels in the Chicago area. 800 of these reviews were scraped from TripAdvisor and are marked "authentic"; the remaining 800 reviews are marked "anomalous" and were gen-

[2] The Dept. of Corrections withheld a a few hundred transcripts from that period, citing "confidential information."

Model	k=20	k=50
LDA Baseline	0.103	0.426
Fine-tuned LM	0.235	0.573
Contrastive Perplex.	0.279	0.676

Table 1: True anomaly recall achieved by reviewing the top-k chunks for each document. The average document has 105 chunks in this sample.

erated by Mechanical Turk workers. In order to fine-tune GPT-2, we use a collection of TripAdvisor reviews collected by (Wang et al., 2010).[3] We only include the 171,016 reviews that were shorter than 1024 tokens and longer than 30 tokens. Additionally, we hold out 10,000 reviews to fit μ and Σ for the LDA baseline.

4.4 Model & Training

We use the GPT-2 base model for all of our experiments, trained for 48 hours using the Adam optimizer with an initial learning rate of 10^{-5} and linear decay.

5 Results & Analysis

5.1 Parole Hearings

Our fine-tuned and normalizer model achieve mean perplexities of 9.22 and 22.99 ($\beta = 0.40$), respectively, on the validation set with fixed chunksize 256. Figure 2 describes the tradeoff between recall and the percentage the transcript human reviewers must read for our model and baselines as we vary the model. Contrastive perplexity outperforms all baselines, but overall recall is low. We also observe that the LM anomaly score produced by our model is not well-conserved across documents. Rather than using a global threshold for our model, we can instead ask reviewers to always use top k predictions for each document. Table 1 shows recall for different values of k.

We evaluate our model's precision at the threshold that yields an average of 55 chunks per document (corresponding to about 52% of average transcript length) and recall of 0.68, marked on the plot. At this threshold, our model achieves an MRR of 0.227. Student annotators achieve 0.264 precision (note that, because the ratings from the students were not ranked, it is not possible to compute their MRR). The low human precision underscores the intrinsic difficulty of the task and the level of disagreement between human annotators over what constitutes an anomaly.

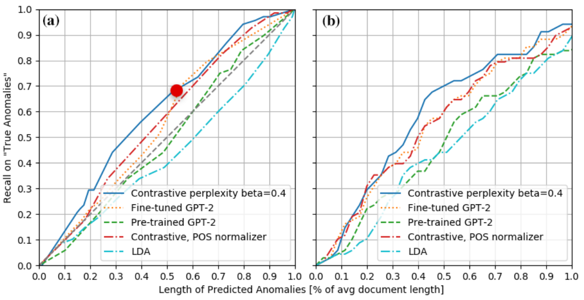

Figure 2: Recall on true anomalies vs. the amount of reading required of the reviewer; (a) by varying the threshold, (b) by fixing k chunks per document.

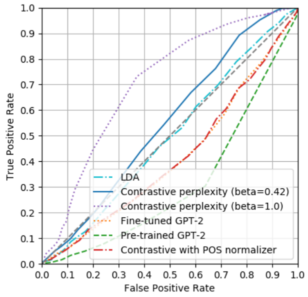

Figure 3: ROC curve for unsupervised fake review prediction on TripAdvisor dataset. The un-tuned $\beta = 0.42$ is outperformed by $\beta = 1$.

5.2 Hotel Reviews

Our fine-tuned and normalizer model achieve mean perplexities of 22.62 and 53.60 ($\beta = 0.42$) on the validation set of "real" TripAdvisor reviews. Figure 3 shows the ROC curve of our model compared to baselines, using our unsupervised LM anomaly measure as a "fake review classifier" on the Deceptive Opinion Spam dataset. Our model achieves an F1 of 0.537 at the optimal threshold. With manually tuned $\beta = 1.0$, we achieve 0.679. While below the 0.898 F1 achieved by the best fully supervised models (Ott et al., 2011), this indicates that our model is a promising unsupervised predictor.

6 Discussion & Conclusion

We present a novel contrastive perplexity-based approach for unsupervised anomaly detection. We define semantic and non-semantic anomalies, and present evidence that our model can distinguish between them better than other unsupervised baselines. Detecting procedural anomalies in legal

[3] We ensured that there is no overlap in between the reviews used for fine-tuning and the Deceptive Opinion Spam dataset.

cases is easier with structured data, but that data is often not readily available. Our approach seeks to support legal decision makers in identifying anomalous cases for review when structured records are unavailable.

Our experiments on an unexplored dataset of 30,734 parole hearing transcripts have identified troubling cases for review. However, our quantitative evaluations also show the difficulty of defining a semantic anomaly consistently. Our results on detecting fake hotel reviews indicate that our approach becomes more powerful when anomaly-free documents are available to perform an out-of-distribution test.

In future work, we seek to use conditional LMs to bridge the gap between our unsupervised method and settings in which some structured data is available.

References

Elif Aktolga, Irene Ros, and Yannick Assogba. 2011. Detecting outlier sections in us congressional legislation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 235–244, New York, NY, USA. ACM.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristen Bell. 2019. A stone of hope: Legal and empirical analysis of california juvenile lifer parole decisions. *Harv. CR-CLL Rev.*, 54:455.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58.

Pradeep Dasigi and Eduard Hovy. 2014. Modeling newswire events using neural networks for anomaly detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1414–1422, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

David Guthrie, Louise Guthrie, and Yorick Wilks. 2008. An unsupervised probabilistic approach for the detection of outliers in corpora. In *LREC 2008*.

Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

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

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 997–1006, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hemant Misra, Olivier Cappé, and François Yvon. 2008. Using LDA to detect semantically incoherent documents. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 41–48, Manchester, England. Coling 2008 Organizing Committee.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Konrad Rieck and Pavel Laskov. 2006. Detecting unknown network attacks using language models. In *Detection of Intrusions and Malware & Vulnerability Assessment*, pages 74–90, Berlin, Heidelberg. Springer Berlin Heidelberg.

Konrad Rieck and Pavel Laskov. 2007. Language models for detection of unknown attacks in network traffic. *Journal in Computer Virology*, 2(4):243–256.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the*

35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402, Stockholmsmässan, Stockholm Sweden. PMLR.

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*, MLSDA'14, pages 4:4–4:11, New York, NY, USA. ACM.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery.

Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 582–588, Cambridge, MA, USA. MIT Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 783–792, New York, NY, USA. ACM.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model.