Classification and Geotemporal Analysis of Quality-of-Life Issues in Tenant Reviews

Adam Haber
JLL Technologies
adam.haber@eu.jll.com

Zeev Waks
JLL Technologies
zeev.waks@eu.jll.com

Abstract

Online tenant reviews of multifamily residential properties present a unique source of information for commercial real estate investing and research. Real estate professionals frequently read tenant reviews to uncover property-related issues that are otherwise difficult to detect, a process that is both biased and time-consuming. Using this as motivation, we asked whether a text classification-based approach can automate the detection of four carefully defined, major quality-of-life issues: severe crime, noise nuisance, pest burden, and parking difficulties. We aggregate 5.5 million tenant reviews from five sources and use two-stage crowdsourced labeling on 0.1% of the data to produce high-quality labels for subsequent text classification. Following fine-tuning of pretrained language models on millions of reviews, we train a multi-label reviews classifier that achieves a mean AUROC of 0.965 on these labels. We next use the model to reveal temporal and spatial patterns among tens of thousands of multifamily properties. Collectively, these results highlight the feasibility of automated analysis of housing trends and investment opportunities using tenant-perspective data.

1 Introduction

The use of artificial intelligence in commercial real estate investing has grown given the availability of new data modalities. Motivated by the potential for new insights and improving investment decisions in the large real estate market, recent efforts have used cellular network data (Pinter et al., 2020), satellite images (Law et al., 2019), building permits (Lai and Kontokosta, 2019), interior and exterior photos for luxury estimation and automated appraisal (Poursaeed et al., 2018), and construction of new retail stores for predicting future rent growth (Humphries and Rascoff, 2015), among others. However, one mostly untapped, yet highly informative, data source, is online tenant reviews.

Online tenant reviews of the properties in which tenants reside present a unique source of information in the multifamily domain due to their distinctive, tenant-perspective view (Fradkin et al., 2015). In recent years, the popularity of such reviews has grown such that there are now millions of newly generated reviews annually, with some properties garnering hundreds and even thousands of reviews over time. Nonetheless, as they are rarely constrained to a specific format and can drastically vary in length and linguistic style, classifying reviews for detection of quality-of-life issues is a challenging task.

Text classification refers to the process of categorizing textual data into a set of defined classes. Classical approaches to text classification rely on feature extraction techniques such as n-grams, Bag-of-Words, and TF-IDF, a potential dimensionality reduction step, followed by learning a classification model such as Logistic Regression, Naive Bayes, Support Vector Machines, Latent Dirichlet Allocation, and Nearest-Neighbours algorithms (Kowsari et al., 2019; Kiatkawsin et al., 2020). More recently, deep-learning-based language models that are trained using contextualized word representations have been used to achieve state-of-the-art results on a wide range of natural language benchmarks and datasets, including text classification (Devlin et al., 2019; Lewis et al., 2020; Liu et al., 2019; Minaee et al., 2021; Sanh et al., 2020).

Deep-learning language models generally require large training data, use up to billions of parameters, and are costly to train. Fortunately, language models pretrained on large corpora such as Wikipedia or Common Crawl can be adapted to perform tasks in diverse domains, very effectively and with little labeled data (Sun et al., 2020).

The above process is referred to as fine-tuning or transfer learning and entails modifying the parameters of the pretrained model to adapt to the statistical properties of the new corpus. Fine-tuning

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2541–2553

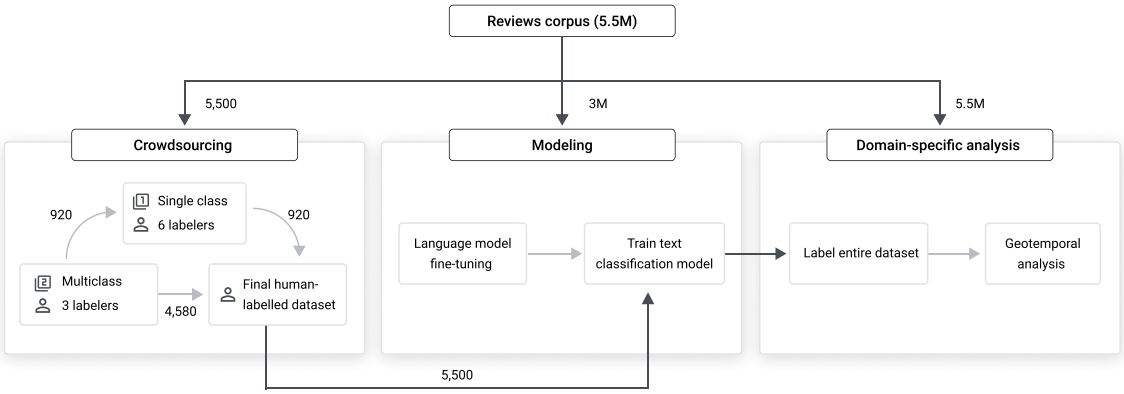

Figure 1: **Study workflow.** 5.5M reviews were collected, of which a small subset were manually labeled via crowdsourcing using multiple labelers per review. A larger set of reviews was used for language model fine-tuning, and the full set was used for uncovering domain-specific insights.

has been shown to improve learned representations and consequently downstream predictions on numerous domain-specific corpora without requiring large-scale labeling (Elwany et al., 2019; Lee et al., 2020), thus opening the possibility of employing these techniques in different applications with relative ease, including tenant reviews classification.

Prior NLP-based efforts on online reviews have used both classical (Hu and Liu, 2004; de Kok et al., 2018) and deep learning-learned representations (Xu et al., 2019) to extract sentiment polarity and/or classify reviews (Pontiki et al., 2014a,b). One popular group of methods, known as aspect-based sentiment analysis (ABSA), attempts to combine these two tasks by evaluating sentiment polarity with respect to specific aspects (Poria et al., 2020). One notable example in the real-estate domain performed a local analysis of 7,673 neighborhood-level reviews in New York City using ABSA and topic modeling (Hu et al., 2019).

A commonality across many review classification efforts is that the review classes are generally broadly defined. However, carefully-tuned class definitions are often of high value to practitioners. For such cases, an approach that goes beyond coarse-grained classification may be beneficial.

In this paper, we analyze a dataset of nearly 5.5 million tenant reviews from multiple online sources, covering tens of thousands of multifamily properties in the US. After analyzing the textual characteristics of this unique corpus, we describe an iterative crowdsourcing-based approach to ensure accurate labeling of a random sample of reviews for multiple, non-mutually exclusive classes. We then show how, using state-of-the-art NLP techniques, we label millions of reviews using a model that was trained on a few thousand annotated samples, and that the labeled corpus provides important insights on spatiotemporal trends affecting the real estate market (Fig 1).

2 Corpus

The data used in this study consisted of 5,468,037 online tenant reviews gathered from five different sources, covering approximately 96,134 different US multifamily properties[1] and spanning 21 years from 2000 - 2020 (Table 1). The total number of words in the corpus was 536,702,874, amounting to 14% of the size of Wikipedia as determined on April 1st, 2021. The contribution of the five sources to the total number of reviews varied from 2.3% to 52% of the corpus, with the largest two sources accounting for 91% of the reviews. 99.2% of the reviews in the corpus are written in English as estimated using the langdetect Python library[2].

The data for each review consisted of the review body text and metadata containing the date and the specific property associated with the review. The distribution of reviews per property was highly skewed as was the distribution of words per review (Fig 2a and 2b). The majority of the reviews (66%) were from recent years (2015-2020), consis-

[1]Since the data is aggregated from multiple sources, the exact number of properties might be slightly different due to entity resolution inaccuracies.

[2]https://github.com/Mimino666/langdetect

Source	Properties	Reviews
A	10,862	126,609
B	22,293	169,539
C	38,524	345,164
D	41,660	1,839,530
E	68,819	2,987,195
Total	96,134	5,468,037

Table 1: **Number of properties and reviews for each source**. A single property may have associated tenant reviews in multiple sources.

tent with the increasing popularity of online media and the digitization of commercial real estate (Fig 2c). Geographically, reviews showed nation-wide coverage, with Texas having the largest number of reviews, both in absolute and relative (per-capita) terms (Fig 2d).

The reviews varied significantly in their sentiment and linguistic style. While the majority of the reviews were positive - "*The [property name] staff are great and the residents are nice. It is a quiet and safe place to live*", some expressed anger and frustration with the property, its surroundings, or its management - "*This place Is horrible I would not alow my dogs to live their, drugs being sold and apartments getting robbed stay away from these people*".

We randomly sampled 500 reviews and 500 Wikipedia articles of similar lengths to measure the statistical discrepancy between the reviews corpus and a more general corpus such as Wikipedia. Correspondingly, we obtained 1000 document embeddings using fastText (Joulin et al., 2017), for which we computed the pairwise Euclidean distance matrix between embeddings (Fig 2e). The block-diagonal structure of the resulting dissimilarity matrix implied that the model representations of reviews were clustered compared to random articles, reflecting their statistical and linguistic idiosyncrasies. This suggested the importance of fine-tuning a pretrained language model to the reviews corpus - see Section 4.

3 Data Labeling

We labeled 0.1% (5,500) of the reviews in order to train models that can detect four detrimental quality-of-life issues. If accurate, these models may enable domain-specific analysis of the entire corpus, especially when paired with property-level geographical and temporal metadata.

3.1 Label Selection

We decided to focus on four issues which are of high interest to real estate professionals after consultation with multiple domain experts. The selected issues are often hard to identify using traditional data sources and are typically difficult and expensive to remedy. The four chosen labels were:

- *Crime and violence*: Have violent or severe crimes occurred at the property or very close by?
- *Noise issues / thin walls*: Are there constant noise issues at the property, either due to environmental or structural reasons?
- *Pests / vermin*: Are pests, roaches and vermin a significant and constant concern for residents?
- *Parking*: Are there not enough parking spaces for residents in the property and its immediate surrounding?

As a single review can contain more than one label, or none at all, this postulates a multilabel classification problem.

3.2 Crowdsourcing

As accurate manual labeling all of the reviews was impractical due to the size of the corpus, we randomly sampled a subset of 5,500 reviews (0.1% of the corpus) with the intention of generating a small amount of high-quality labels. We considered labels to be high-quality when they were precisely aligned with both the detailed definitions given above as well the specific positive and negative examples provided to the labelers. These labels would later be used for downstream model training and evaluation.

We first conducted a series of single-label crowdsourcing experiments, each with 1000 reviews, to refine the exact instructions provided for each label and to choose a labeling vendor. The experiments comprised multiple labeling vendors, had between three to nine labelers per review, and were conducted using the AWS GroundTruth platform. Disagreements between different label providers were assessed to detect systematic differences (Fig S1). As an example, in one pilot experiment, labelers were instructed to label reviews that mention break-ins; while labelers from one vendor interpreted this as solely apartment break-ins, other vendors

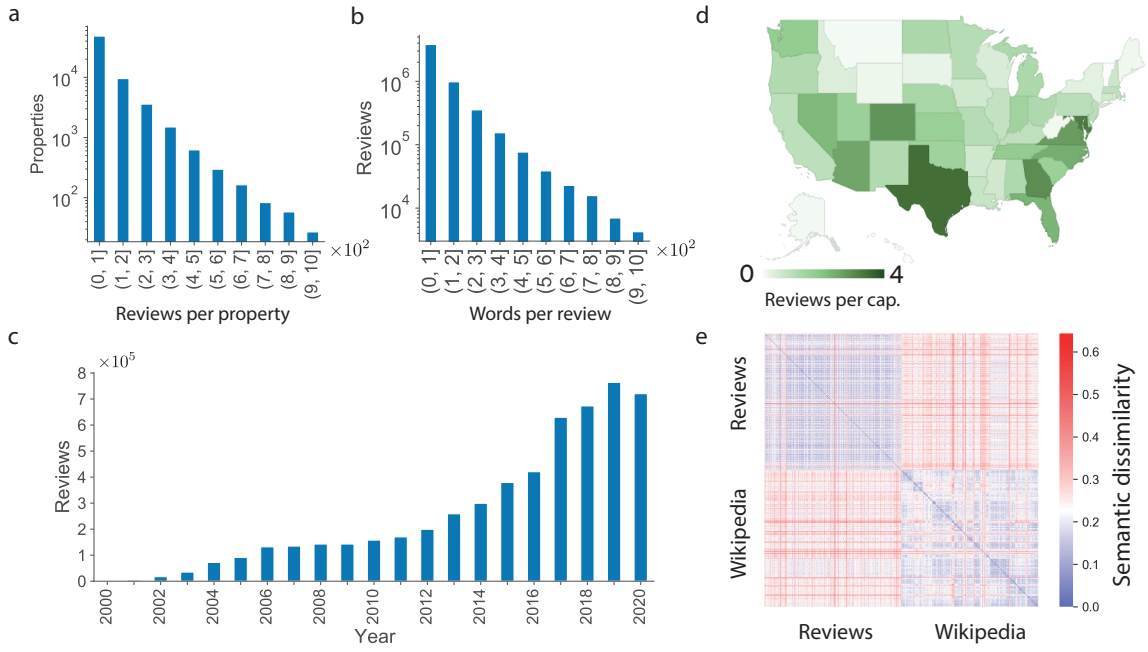

Figure 2: **Statistical properties of the tenant reviews corpus. (a)** Number of reviews per property. **(b)** Number of words per review. **(c)** Number of reviews per year. **(d)** Reviews per capita. The number of reviews per 100 people in each state from 2000 to 2020 is shown. **(e)** Sentence dissimilarity, as measured by Euclidean distance between document embeddings, between 500 randomly sampled reviews and 500 randomly sampled Wikipedia articles. Reviews are generally more similar to other reviews, and statistically different than random articles.

also included reviews that refer to vehicle break-ins. These discrepancy comparisons enabled to detect ambiguities in our instructions and helped refine subsequent experiments. Afterwards, we conducted multilabel pilots with three top performing vendors, as assessed by consensus labeling and manual review of discordantly labeled reviews in the single label pilots, to choose the vendor with which we will proceed.

We next designed a two-stage crowdsourcing pipeline to ensure label quality (Fig 1). In the first stage, all 5,500 reviews were seen by three different labelers that provided an annotation for each of the four classes. 4,580 (83%) of the reviews had consensus among the three labelers in all four classes, for example all three labelers agreed that there was no crime, no noise, there were pest issues, and there were no parking issues. To gain more confidence in the remaining 920 reviews that were not unanimously labeled, we passed them through to a second crowdsourcing stage with six additional labelers, focusing on the specific label(s) in which there was disagreement. The final label in the 2-stage scenario was given by a majority vote among the nine labelers. This iterative approach was cost-effective as reviews for which there was a consensus were pruned, thus more labeling resources were placed on ambiguous reviews. Table 2 shows the distribution of the crowdsourced labels, of which 88.8% were None.

	Crime	Noise	Pests	Parking	None
Labels	215	139	246	91	4888
Fraction	3.9%	2.5%	4.4%	1.6%	88.8%

Table 2: **Abundance of each positive label within the set of labeled reviews.** Total unique reviews - 5,500. Some reviews can have more than one label and thus the percentages sum to slightly more than 1.

4 Results

4.1 Modeling Details

We trained the review classifier in two steps using the 5,500 labeled reviews. First, we fine-tuned a pretrained model for 10 epochs (Adam optimizer, batch size 8, learning rate 10^{-5}). The pretrained model was either RoBERTa (Liu et al., 2019) or DistilBERT (Sanh et al., 2020). Each model was trained (unsupervised) on a random sample of 3M reviews that did not overlap with the 5,500 labeled reviews using a single GPU on

an AWS ml.p3.2xlarge instance. Pretrained models were based on HuggingFace implementations (Wolf et al., 2020), and the training was done using PyTorch (Paszke et al., 2019). Second, we trained a multilabel classifier downstream to the fine-tuned model on the set of 5,500 labeled reviews without freezing the encoder layers. The classifier consisted of a dense layer with a hyperbolic tangent activation function and 768 hidden units, a dropout layer (p=0.1), and another dense output layer with one output neuron for each label. We used binary cross-entropy (logit scale) as our loss function, averaged over the different labels. Model results were evaluated via 5-fold cross-validation.

4.2 Modeling Results

We computed the cross-validated area under the receiver operating characteristic curve (AUROC) for each of the labels to estimate model predictive accuracy. The AUROC scores stabilized for 3 out of 4 labels at around 3000 samples, as shown via learning curves (Fig S2). Due to the sparsity of the labels, there was variability between folds, with fine-tuning improving both the average and the variance across folds. The plateauing AUROC suggested diminishing returns for obtaining additional labeled reviews. Finally, the neural models had a strong tendency to overfit the train set as observed by fitting the models to permuted labels, stressing the importance of cross-validation in performance estimation (Fig S3).

Interestingly, despite the fact that the model was trained on binary labels (chosen via majority voting between labelers), model prediction were highly correlated with labeler uncertainty (Fig 3). This suggests that the model predicted probabilities may be used to learn the inherent ambiguity in label definitions.

In Table 3, we provide the AUROC, as well as average precision and F1 score for different models trained on our labeled dataset. Numbers represent the average cross-validated scores using the probabilistic, not thresholded, predictions, except for F1 in which we chose the optimal threshold (separately for each model and label). Fine-tuned models outperformed the base model for both DistilBERT and RoBERTa, and were also better calibrated, as evident by Brier score (see Table S1). As baselines, we also provide comparisons to fastText, an efficient C++ implementation of a Bag-of-Words-based classification algorithm (Joulin et al., 2016),

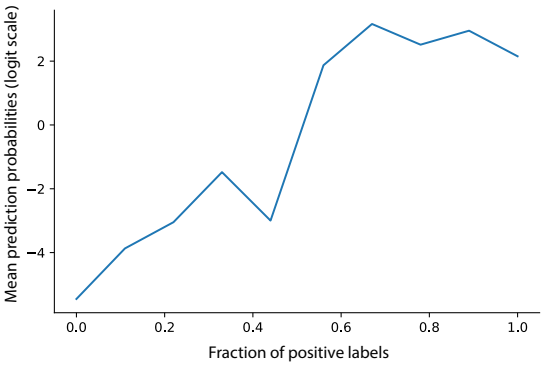

Figure 3: **Model predicted probabilities match labelers uncertainty.** The mean predicted probability (logit scale) is plotted against the ratio of labeler disagreement, as defined by the fraction of positive labels (ranging from 0/9 to 9/9), averaged over all 5,500 reviews and 4 labels. 0 or 1 on the x-axis indicates full agreement among labelers.

and to a BERT-based, ABSA classification model[3]. The latter model is composed of a HuggingFace implementation of a BERT model (Wolf et al., 2020), pretrained on SemEval 2014, Task 4 (Pontiki et al., 2014a), a subsequent dropout layer, and a dense classification layer, and was not post-trained on the crowdsourced labels. Negative sentiment was evaluated on four aspects corresponding to the labels "crime", "noise", "pests", and "parking", and serves as a benchmark for the performance of an unsupervised approach.

Fine-tuning the pretrained base models improved results across all four labels, both for DistilBERT and RoBERTa. This suggests the presence of differences in statistical properties between our corpus and the concatenation of Wikipedia and the Toronto Book Corpus, on which both DistilBERT and RoBERTa were trained. In contrast, there was no substantial difference in results between fine-tuned RoBERTa and fine-tuned DistilBERT when considering all labels.

We conducted error analysis by manual examination of the subset of the 5,500 labeled reviews with the highest disagreement between model output scores and labeler annotations. For each label, we investigated the 10 highest model output scores in which the annotation was negative and the 10 lowest model scores with positive annotations. We found no systematic bias among these reviews, and generally agreed with the labels given by human

[3] https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis

	AUROC				Average Precision				F1			
	Crime	**Noise**	**Pests**	**Parking**	**Crime**	**Noise**	**Pests**	**Parking**	**Crime**	**Noise**	**Pests**	**Parking**
ABSA (unsup.)	0.75	0.71	0.73	0.73	0.08	0.06	0.07	0.07	0.14	0.13	0.13	0.13
fastText	0.87	0.83	0.83	0.81	0.19	0.10	0.16	0.06	0.29	0.19	0.25	0.13
DistilBERT	0.98	0.94	0.98	0.92	0.82	0.45	0.80	0.32	0.72	0.46	0.74	0.21
DistilBERT (f.t.)	**0.99**	**0.96**	**0.99**	**0.95**	**0.83**	**0.56**	**0.83**	0.45	**0.79**	**0.53**	0.78	**0.47**
RoBERTa	0.98	0.94	0.98	0.93	**0.83**	0.49	0.82	0.43	0.71	0.45	0.74	0.45
RoBERTa (f.t.)	**0.99**	**0.96**	**0.99**	0.92	**0.83**	0.53	**0.83**	**0.50**	0.77	0.44	**0.79**	0.46

Table 3: **Classification results across different labels and model combinations.** Bold numbers represent the best score per column. The dotted line separates unsupervised vs supervised models. f.t. - fine-tuned model, unsup. - unsupervised, see main text.

annotators, especially for reviews with positive annotations.

After verifying the accuracy of the model, we proceeded to use the RoBERTa fine-tuned model to predict the labels of all 5.5M reviews. This created what is, to the best of our knowledge, the largest labeled reviews dataset in the field of commercial real estate.

4.3 Association of Model Predictions with Property and Demographic Data

Model predictions on the review corpus, together with review metadata, enabled us to analyze nationwide multifamily housing trends from a tenant-perspective. Below are select examples that demonstrate associations between automatically identified issues in reviews and property-level or geographic level data.

One natural question to ask was to what extent model scores correlated with established property quality metrics. One commonly used metric is asset grade, which ranges from A (best) to D (worst), and reflects where the property falls across the quality spectrum relative to its U.S. Census-defined geographic area (source: Axiometrics). We computed the mean scores per asset grade for all properties in which an asset grade was obtainable (23,912 properties). Higher grade properties were found to have less crime and pest issues in their reviews, as expected (Fig 4a). In contrast, no strong association existed between noise or parking scores and asset grade. A similar behavior was observed when comparing model scores to property expense ratios, which refers to the ratio of operating expenses to gross revenue (sources: Fannie Mae and Freddie Mac) (Fig 4b).

We additionally investigated whether the tenant reviews reveal geotemporal trends in the data. We compared predicted review scores against the year built of each property in our dataset as newer properties are typically of higher quality. The analysis was conducted for 64,810 properties that were built after 1970 (sources: multiple). Indeed, we found that newer properties had fewer issues across all labels, however the improvement only commenced in the past decade for noise and parking issues, in contrast to crime and pest problems (Fig 4c). Spatially, we compared per-city average crime scores from the reviews (mean predicted crime score across all the reviews from 2015-2017 for properties in a given city) against nationwide public FBI crime reports from 2017[4], which are at the city level. The FBI report covered 4 different types of violent crimes and 4 different types of property-specific crimes, and there was a strong positive correlation between levels of various crime categories across cities (mean Pearson correlation between different crime types is 0.6). Fig 4d shows an example for a single crime category, motor vehicle theft.

5 Discussion

In this study, we applied NLP-techniques to investigate a unique dataset of millions of online tenant reviews. We demonstrated that tenant reviews have idiosyncratic textual and statistical properties, differentiating them from other commonly used textual datasets. We further presented a resource-effective multi-labeling approach, and showed that using a limited set of high quality labels can achieve excellent results in a previously little studied domain. Finally, we illustrated that NLP-based scores are informative, as verified by domain-specific validations, and can be used to study financial, demographic, geographical and temporal trends in a quantitative way.

Our work is in line with prior observations that with a relatively small number of labels, fine-tuned language models can be trained to accurately predict human annotations in novel corpora (Yu et al., 2018). Although we focused on four key labels of interest, we expect this approach will generalize

[4] https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/tables/table-8/table-8.xls/view

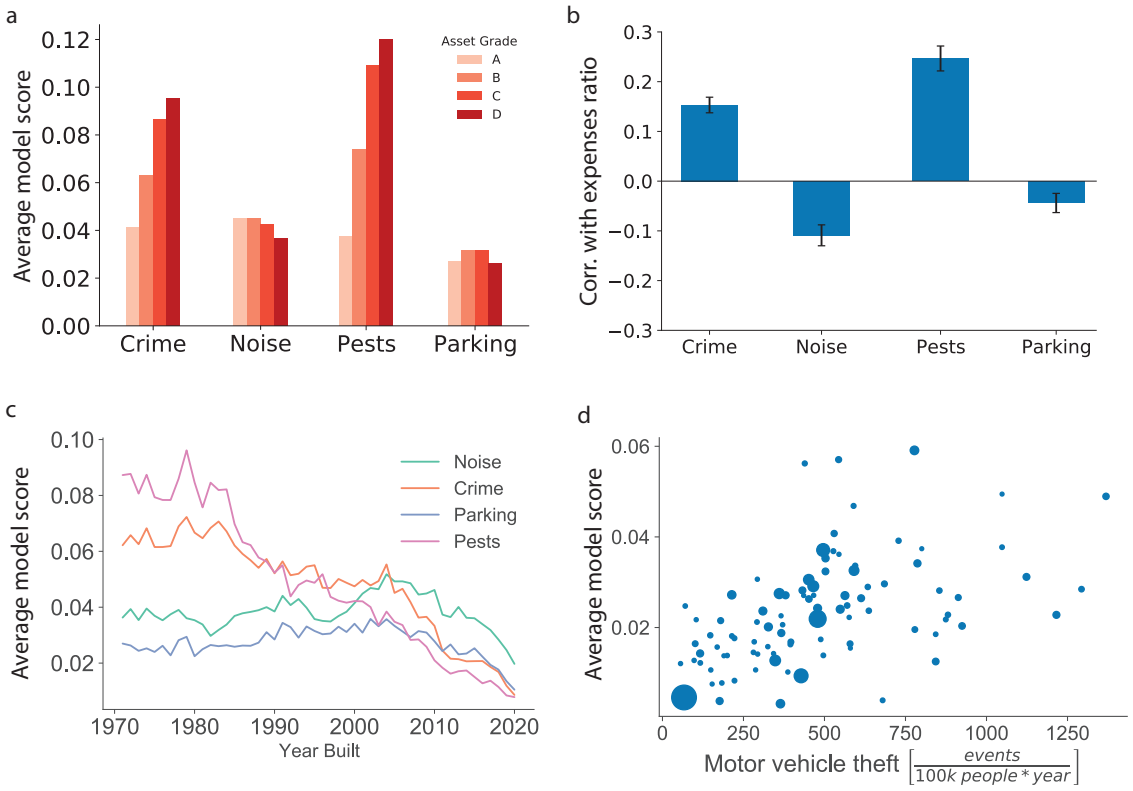

Figure 4: **Demographic analysis using review scores from model predictions.** **(a)** Average score per year built, averaged over all properties from a specific year. **(b)** Spearman correlation between the expense ratio of each property and model predictions. Error bars represent standard deviations using data from 2014 to 2020. **(c)** Average predicted review score per label for each asset grade (A = best, D = worst). **(d)** Average crime score, using reviews from 2015-2017 only, versus annual motor vehicle theft rate per 100,000 people (Spearman correlation = 0.58, $p<0.01$). Each dot represents a single city, with dot size corresponding to city population (source: Federal Bureau of Investigation, Uniform Crime Reports, 2017).

to other informative labels such as maintenance issues, management-related concerns, and renovation needs. Additionally, while our analysis bears similarity to aspect-based sentiment analysis (Xu et al., 2019), the class definitions used are more precise. For example, a review that mentions a single event of a pest sighting in a property might demonstrate a negative sentiment towards pests, but is not necessarily indicative of a recurrent problem in the property as we defined in labeling instructions.

Domain-specific validation serves as an orthogonal means for validating model usefulness. Encouragingly, model predictions often correlated with prior domain knowledge: crime and pest issues were higher in lower grade properties, all four labels improved in newer properties, and cities with higher crime rates had a higher amount of crime-related reviews. These serve as secondary validations that strengthens our conviction in the value of model predictions.

Our results reveal differences between crime and pests issues versus parking and noise issues in relation to external, non-review data. Whether this is an artifact, for instance due to the latter two being sparser labels, or whether it is a true real estate phenomenon warrants further investigation. One potential explanation may be variation in tenant base. For example, tenants in grade A properties may be more sensitive to noise and parking issues, and thus lower noise levels may receive increased mention. Construction-wise, the evolution of building standards may be associated with the differences in pest, noise, and parking issue mentions in newer buildings. Finally, demographic changes may also be linked to the strong reduction in crime mentions with newer year builds.

One concern when analyzing online reviews is the potential presence of fake or solicited reviews.

Non-authentic reviews can bias the average score of a given property, in turn compromising the accuracy of downstream inferences. While online review sites have made large efforts to ensure review authenticity, there is nonetheless a risk. Initial results indicate that NLP-based analysis might help in identifying these reviews (Abri et al., 2020); applying this to our dataset and investigating the sensitivity of the results to such preprocessing is a potentially exciting future direction.

6 Conclusion

The use of AI and non-traditional data in commercial real estate is expected to have far-reaching implications. Our work contributes to this broader scope by highlighting how online tenant reviews, which have become ubiquitous, can uncover valuable insights that support both real estate investment decisions and research into local and nationwide housing trends.

Acknowledgements

We would like to thank our colleagues at JLL Technologies and those previously at Skyline AI for their contribution to this work. We would like to thank Gavan Gooch, Vincent Menechino, Benjamin Sternberg, Gil Siedner, and Christopher Sameth for providing real estate domain expertise. We would like to thank Or Hiltch, Shahar Shechter, and Omri Bromberg for assisting with data collection. Finally, we would like to thank Dana Rapoport, Hai Rozencwajg, Dotan Davidovich, Yana Volovik, Eyal Ella, and Iddo Israely for helpful feedback and research suggestions throughout the project.

References

Faranak Abri, Luis Felipe Gutierrez, Akbar Siami Namin, Keith S. Jones, and David R. W. Sears. 2020. Fake Reviews Detection through Analysis of Linguistic Features. *arXiv:2010.04260 [cs]*. ArXiv: 2010.04260.

Sophie de Kok, Linda Punt, Rosita van den Puttelaar, Karoliina Ranta, Kim Schouten, and Flavius Frasincar. 2018. Review-level aspect-based sentiment analysis using an ontology. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, page 315–322, New York, NY, USA. Association for Computing Machinery.

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.

Emad Elwany, Dave Moore, and Gaurav Oberoi. 2019. BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding.

Andrey Fradkin, Elena Grewal, Dave Holtz, and Matthew Pearson. 2015. Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, page 641, New York, NY, USA. Association for Computing Machinery.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Yingjie Hu, Chengbin Deng, and Zhou Zhou. 2019. A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers*, 109(4):1052–1073.

Stan Humphries and Spencer Rascoff. 2015. *Zillow Talk: The New Rules of Real Estate*. Grand Central Publishing.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Kiattipoom Kiatkawsin, Ian Sutherland, and Jin-Young Kim. 2020. A comparative automated text analysis of airbnb reviews in hong kong and singapore using latent dirichlet allocation. *Sustainability*, 12(16).

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information*, 10(4):150. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Yuan Lai and Constantine E Kontokosta. 2019. Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems*, 78:101383.

Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a look around: Using street view and satellite images to estimate house prices. *ACM Trans. Intell. Syst. Technol.*, 10(5).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Gergo Pinter, Amir Mosavi, and Imre Felde. 2020. Artificial intelligence for modeling real estate price using call detail records and hybrid machine learning approach. *Entropy*, 22(12).

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014a. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014b. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Omid Poursaeed, Tomáš Matera, and Serge Belongie. 2018. Vision-based real estate price estimation. *Machine Vision and Applications*, 29(4):667–676.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification? *arXiv:1905.05583 [cs]*. ArXiv: 1905.05583.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.

A Assessing vendor disagreement

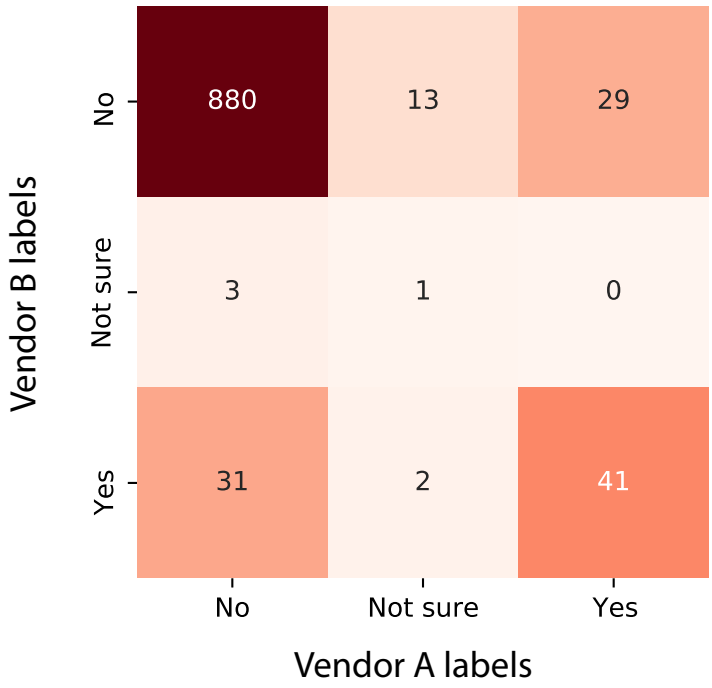

Figure S1: **Assessing vendor disagreement.** An example of vendor comparison for a single pilot crowdsourcing experiment with 1000 reviews. The final label for each review was chosen using a majority vote between the labelers. In the case of a tie among 3 labelers the final label was set as "Not sure" (the case of 1 "Yes", 1 "No" and 1 "Not sure"). Manual analysis of vendor differences focused on reviews that were majority labeled as "Yes" by one vendor and "No" by the other vendor, which in this experiment was 29 and 31 reviews (top right and bottom left in the figure).

B Diminishing effect of increasing train set size (learning curves)

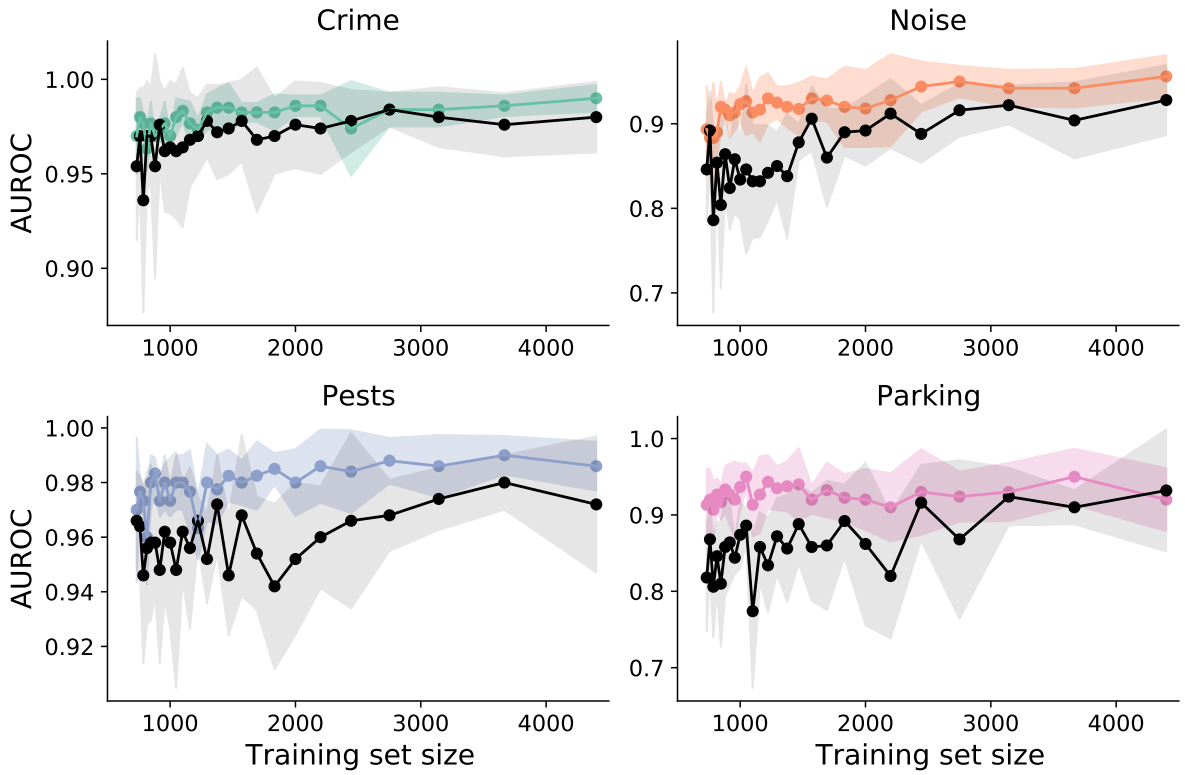

Figure S2: **Diminishing effect of increasing train set size (learning curves).** We trained the fine-tuned RoBERTa text classification model using increasing amounts of training examples (from 150 to 4,400), while keeping the test set size fixed at 1,100 and using the same test reviews in each case. 5-fold CV was used for evaluation (220 test samples per fold). The filled area represents standard deviation over 5 folds. The black dots and gray area represent means and standard deviations in the non fine-tuned model. While the variability between folds is large likely due to test set size, the benefit of increasing the train set size beyond 3000 samples appears small for 3 out of 4 labels (results for "Parking" were too noisy to infer this).

C ROC curves on permuted labels

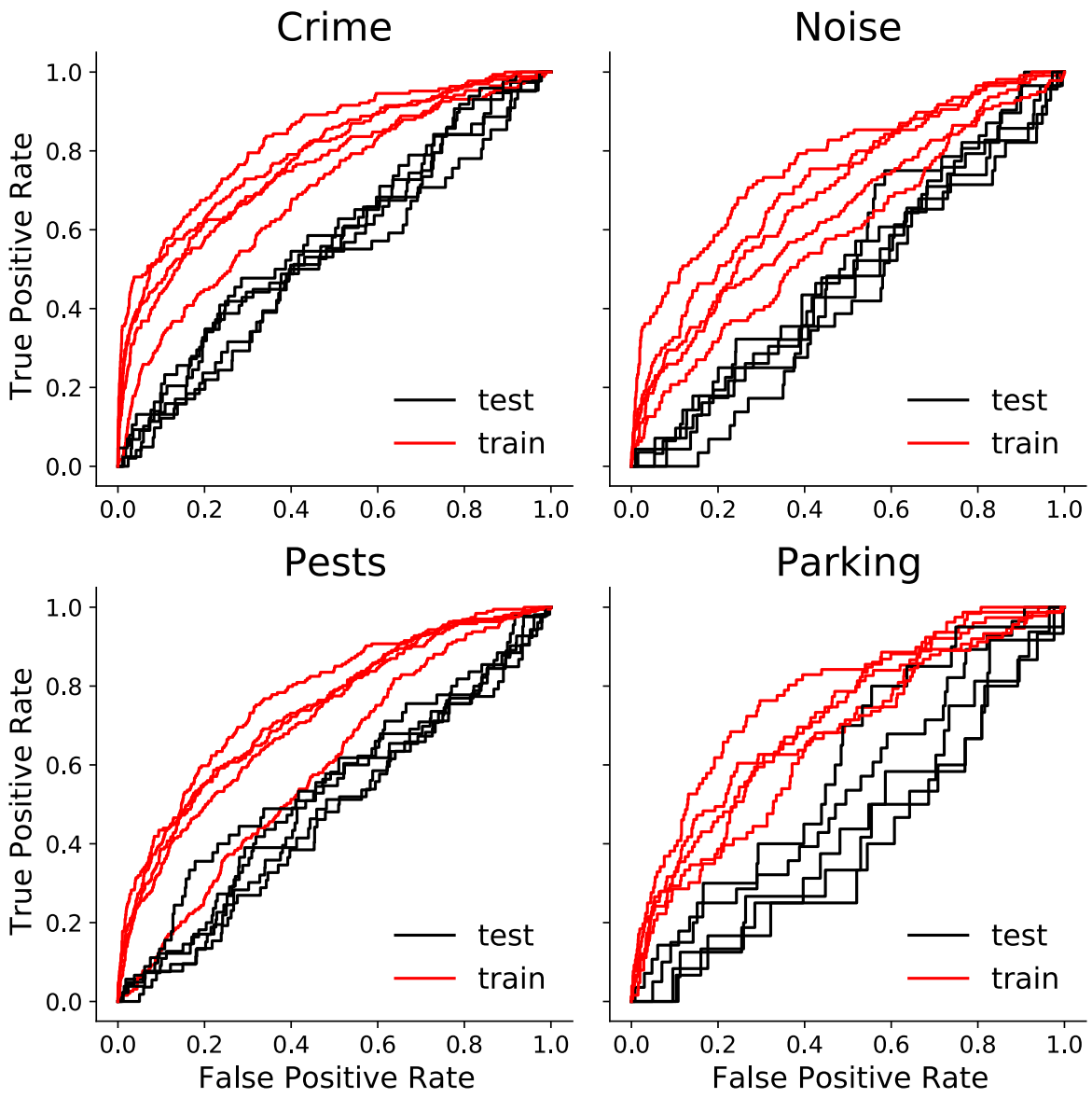

Figure S3: **ROC curves on permuted labels.** We trained the fine-tuned RoBERTa text classification model for 5 epochs (all other parameters are as described in the main text) on permuted labels (each label was permuted differently). Red lines correspond to ROC curves on the training set (for 5 different folds), black lines - test set. The model shows significant overfitting to the train set already after 5 epochs.

D Brier loss per model

	Brier loss			
	Crime	**Noise**	**Pests**	**Parking**
fastText	0.028	0.021	0.03	0.016
DistilBERT	0.018	**0.018**	0.018	0.014
DistilBERT (f.t.)	0.016	**0.018**	0.018	0.014
RoBERTa	0.016	0.022	0.018	**0.012**
RoBERTa (f.t.)	**0.014**	**0.018**	**0.016**	**0.012**

Table S1: **Brier loss per model.** Loss is averaged across 5 folds (see main text).