Infrastructure Ombudsman: Mining Future Failure Concerns from Structural Disaster Response

Md Towhidul Absar Chowdhury[1] , **Soumyajit Datta**[1] , **Naveen Sharma**[1] and **Ashiqur R. KhudaBukhsh**[1]

[1]Rochester Institute of Technology

{mac9908, sd3528, naveen.sharma, axkvse}@rit.edu

Abstract

Current research concentrates on studying discussions on social media related to structural failures to improve disaster response strategies. However, detecting social web posts discussing concerns about anticipatory failures is under-explored. If such concerns are channeled to the appropriate authorities, it can aid in the prevention and mitigation of potential infrastructural failures. In this paper, we develop an *infrastructure ombudsman* – that automatically detects specific infrastructure concerns. Our work considers several recent structural failures in the US. We present a first-of-its-kind dataset of 2,662 social web instances for this novel task mined from Reddit and YouTube.

1 Introduction

On January 28, 2022, at 6.39 a.m. EST, the Fern Hollow Bridge in Pittsburgh, Pennsylvania collapsed. Due to the timing of the failure, thankfully, fewer vehicles were on the bridge and only ten people were injured with no fatalities. Pittsburgh, also known as the City of Bridges, was getting ready for a visit from President Biden that day. Biden visited the collapse site and assured federal assistance to rebuild the bridge on the spot. This infrastructural failure, coinciding with a high-profile political visit and a push towards passing the Build Back Better infrastructure bill, attracted considerable media attention to the flailing infrastructural health in the US.

As we were sifting through the social web discussions surrounding this issue, broad themes such as words of compassion for the victims and typical responses in social web political discourse such as political name-calling, conspiracy theories, and partisan mud-slinging emerged. However, apart from these expected social web reactions, we noticed a small minority of interactions that talked about anticipatory failures of other bridges in the US. Table 1 lists a few illustrative examples.

The comments in Table 1, if mined efficiently and surfaced to the appropriate authorities, can present an effective path to intercept structural failure concerns. Failures that are yet to happen - but a responsible citizen is worried that they might.

Social Media Post
There is a bridge in Lowell Massachusetts, it goes over the Merrimack river and it is rusted strait through. It won't be long before we suffer major injuries because that bridge is always bumper to bumper traffic!
I'm surprised the New Kensington bridge wasn't the first to go. Haha. Terrible condition.
The bridge on 1-81 that spans the Potomac between WV and MD that could be next. It has a lot of 18 wheelers beating it up. I kept a hammer in my car to get out if it collapsed when I was on it.

Table 1: Illustrative examples indicating concerns over other vulnerable bridges following the collapse of the Fern Hollow Bridge in Pittsburgh.

Vulnerabilities that they perhaps noticed before, but the sudden, exogenous shock in the form of a structural failure gives them an outlet to voice their concerns. Extant research on disaster response discourse focuses on a diverse set of tasks that include: efficient distribution of relief [Varga *et al.*, 2013], crisis management, handling emergencies etc [Horita *et al.*, 2017]. Much of this research has focused on natural disasters such as typhoons [Zou *et al.*, 2023], earthquakes [Sakaki *et al.*, 2012], floods [Feng *et al.*, 2020], and accidents with severe fatalities [Liu *et al.*, 2020]. Analyzing social media responses to structural failures such as bridges or building collapses is rather new and looking beyond the immediate, and focusing on future potential crises has no prior literature to the best of our knowledge.

This paper[1] presents *infrastructure ombudsman*, an automated social media listener tool that surfaces infrastructure concerns. Via a curated corpus of 2,662 instances (271 positives and 2391 negatives)[2], we demonstrate that state-of-the-art NLP and AI methods can be harnessed to build such tools that can aid humans in identifying infrastructure concerns. While the Pittsburgh Bridge collapse was an exogenous shock that got people discussing other potential vulnerabilities, it is possible that our findings point to a broader human pattern where structural collapses trigger similar thoughts about an-

[1]Paper has been published in The ACM Web Conference (WWW) 2024

[2]https://github.com/towhidabsar/InfrastructureOmbudsman

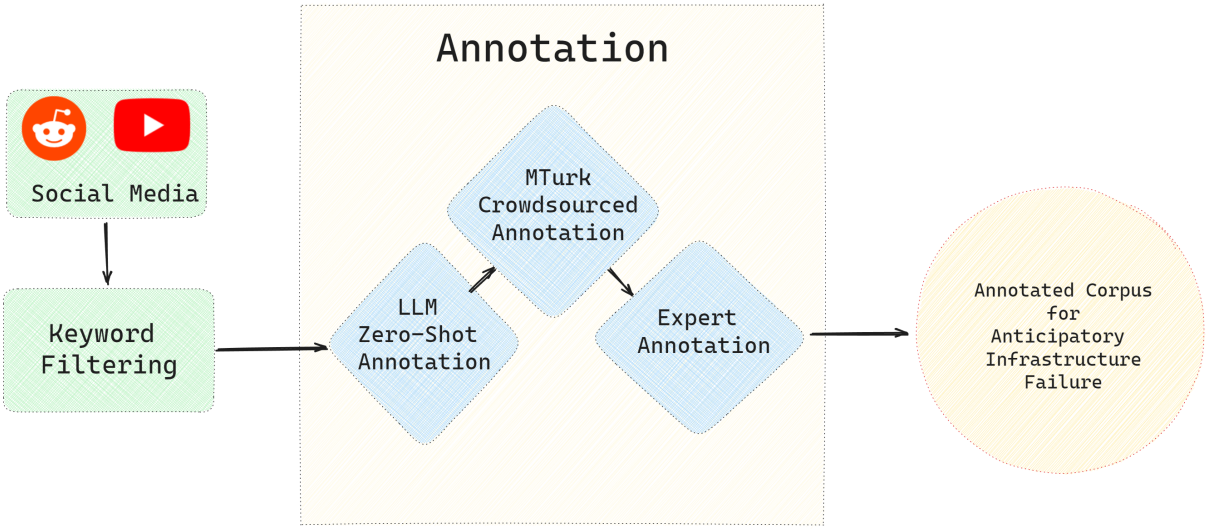

Figure 1: Dataset Creation Pipeline

ticipatory failures. Our dataset thus considers several recent infrastructure failures in the US and combines effective rare-class mining methods to present a more holistic resource to this novel task.

Our contributions are the following:

Novel task: detecting infrastructure concerns: We define a new task of detecting infrastructure concerns from social web posts. Our task presents a marked departure from existing disaster response literature where our goal is to identify citizen concerns about possible, future structural failures.

New resource: We release a dataset of 2,662 instances (271 positives and 2,391 negatives) of infrastructure concerns mined from millions of Reddit posts and YouTube comments. We present a suite of strong baselines trained on this dataset and demonstrate the feasibility of automatically detecting infrastructure concerns with reasonably high precision and recall. On unseen data, we demonstrate that our content classifier can effectively aid humans in detecting potential candidates that express infrastructure concerns.

2 Dataset

We curate the first dataset for this novel task by collecting social media instances from Reddit and YouTube by employing a multi-step pipeline to filter the data and identify the most likely candidates expressing infrastructure concerns. The process begins with keyword filtering using a set of search terms related to various structural failures. This high-recall approach [Halterman *et al.*, 2021; Dutta *et al.*, 2022] yields an initial corpus of over 140,326 Reddit posts and 416,009 YouTube comments. Next, we prune the data using natural language inference (NLI) where we only retain instances where an off-the-shelf NLI system predicts entailment between the post and the hypothesis *"There is a growing infrastructure concern somewhere."* The filtered data is then annotated using a large language model (PaLM 2) [Chowdhery *et al.*, 2022] in a zero-shot setting. We designed a prompt to guide the annotation process [Ziems *et al.*, 2023], aligning the model's responses with the desired labels. This machine annotation step further narrows the dataset to 243 Reddit examples and 2,419 YouTube examples. Each step of our data creation process, visualized in Figure 1, focused on filtering down the sample to just the most likely positive instances of anticipatory infrastructure concerns. After collecting an initial dataset of over 2 million social media discussions from both Reddit and YouTube, we filtered it down to over 10,251 positives with textual entailment. Another round of LLM-based zero-shot annotation lowered our sample size to ∼3,000 possible positives. The first round of partisan MTurk annotation yielded 1,000 possible comments labeled as future infrastructure concerns by at least two crowd-sourced annotators. The two expert annotators then reduced this to just 271 highly confident positive examples, compared to 2,391 confirmed challenging negative examples (as it passed through our three filters including crowdsourced workers), for a total annotated corpus of 2,662 data points.

3 Experiments

A key contribution of our work is developing natural language processing methods to automatically identify discussions expressing anticipatory infrastructure concerns. We frame this as a binary text classification problem, where the goal is to determine whether a given text contains evidence of concerns about potential infrastructure failures. In what follows, we design and evaluate a suite of classifiers trained on this binary classification task.

3.1 Zero-Shot Classification

We select three well-known LLMs (`PaLM 2` [Chowdhery *et al.*, 2022], `GPT-3.5-Turbo` [noa,], and `Mistral AI` [Jiang *et al.*, 2023]) to perform zero-shot classification as our baselines. As already mentioned, this prompt has been designed following best practices suggested by [Ziems *et al.*, 2023]. We present the social web post following the prompt.

Model	Precision	Recall	F1	Accuracy
BERT$_{nomask}$	0.79±8e-5	0.82±1e-4	0.81±5e-4	0.92±1e-5
BERT$_{mask}$	0.80±2e-4	0.83±2e-4	0.81±1e-4	0.93±2e-5
RoBERTa$_{nomask}$	0.78±8e-5	0.83±1e-4	0.80±3e-5	0.92±2e-5
RoBERTa$_{mask}$	**0.82±1e-5**	**0.83±3e-4**	**0.82±9e-5**	**0.93±3e-5**
LLAMA2$_{nomask}$	0.83±2e-5	0.78±5e-5	0.80±2e-5	0.93±4e-5
LLAMA2$_{mask}$	0.81±1e-4	0.79±1e-4	0.80±1e-4	0.93±4e-5
Mistral$_{nomask}$	0.77±9e-6	0.75±5e-5	0.76±8e-5	0.91±6e-5
Mistral$_{mask}$	0.83±2e-5	0.79±1e-4	0.80±5e-5	0.93±3e-5
Mistral$_{zero}$	0.54	0.57	0.31	0.32
GPT3.5-Turbo$_{zero}$	0.53	0.57	0.35	0.38
PaLM2$_{zero}$	0.58	0.63	0.59	0.80

Table 2: The results of the classifier performances on our dataset. We split our corpus into training and validation sets (70:30). All values are the macro average results evaluated across a 5-fold run over the validation set except for the zero-shot classification. We report the mean along with the variance.

3.2 Supervised Classification

We consider multiple classifier models including smaller language models with around 700M parameters (BERT [Devlin *et al.*, 2018] and RoBERTa [Liu *et al.*, 2019]) as well as large language models with 7B parameters (Mistral and Llama 2 [Touvron *et al.*, 2023]). We fine-tuned each model on our dataset using a 70/30 train/test split for 5 epochs using Adam optimizer [Kingma and Ba, 2017]. We consider standard machine learning performance metrics precision, recall, F1, and accuracy on the held-out validation set.

Most of the instances present in our dataset mention physical locations. All these locations can be clubbed into a placeholder <LOCATION> which might benefit a text classifier not to attend to irrelevant information. To isolate the impact of locale-specific references, we ran experiments under two settings: (1) masked locations, and (2) no masked locations. Under the masked location setting, we utilized named entity recognition (NER) to identify locations and geopolitical entities like cities and states within each comment using spaCy [Montani *et al.*, 2023], replacing them with the generic token <LOCATION>. With no masked locations, we left the comments unmodified. This enabled us to evaluate whether classifiers rely heavily on local references to identify infrastructure concerns, or can infer these concerns solely from high-level semantic content. Evaluating performance with masked and unmasked locations can reveal opportunities to improve model robustness. Our experiments aim to determine the degree to which anticipatory infrastructure concerns can be detected from language patterns alone, without relying on localization signals.

4 Results

Table 2 summarizes our supervised solutions' performance. We observe all supervised solutions attaining reasonable precision, recall, and F1 score. We do not observe any across-the-board discernible benefit in masking or not masking the location information indicating that our models are most likely learning from the semantic content of the infrastructure concern rather than being fixated on location mentions.

In contrast with the fine-tuned models, we observe that the zero-shot models perform considerably poorly, with PaLM 2 emerging as the winner among the zero-shot LLMs. This is possibly due to the nuanced nature of our task.

4.1 Performance In The Wild

To be practically useful, the infrastructure ombudsman, i.e., our content classifier, will have to effectively mine infrastructure concerns in the wild. To evaluate in-the-wild performance, we employ the top-performing classifier from Table 2 (RoBERTa$_{mask}$) to previously unseen $\mathcal{D}_{in\text{-}the\text{-}wild}$.

Running on this data, our classifier flagged 2,116 comments as positive anticipatory infrastructure concerns, and 7,884 as negative. To estimate precision and recall on this unlabeled set, we manually annotated a random sample of 100 predicted positives and 100 predicted negatives.

On these manually annotated samples, our classifier achieved a macro precision of 0.82, recall of 0.85, F1 score of 0.85, and an accuracy of 0.82, in identifying true infrastructure concerns outside of the training data. This demonstrates promising generalization to unseen data, with a high precision indicating most flagged instances were true positives.

5 Conclusion and Discussions

This paper presents a novel direction in mining anticipatory concerns from social media following structural failures (e.g., bridge or building collapses). To this end, we present the first dataset resource to perform this novel task. We present a suite of strong baselines relying on the recent advancements in large language models. Our automated anticipatory infrastructure concern mining tool, dubbed infrastructure ombudsman, performs effectively in the wild.

Beyond infrastructure concerns: Our study reveals an interesting pattern in human behavior where we observe that people often discuss potential vulnerabilities following a structural failure. There could be broader generalizability to this phenomenon. Following a brawl in a bar that results in a shooting incident can trigger discussions on other bars where violence might happen. Similar methods can be employed to intercept such concerns.

References

[Chowdhery *et al.*, 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and others. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dutta *et al.*, 2022] Sujan Dutta, Beibei Li, Daniel S. Nagin, and Ashiqur R. KhudaBukhsh. A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance. In *Proceedings of the Thirty-First IJCAI*, pages 5059–5065, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization.

[Feng *et al.*, 2020] Yu Feng, Claus Brenner, and Monika Sester. Flood severity mapping from Volunteered Geographic Information by interpreting water level from images containing people: A case study of Hurricane Harvey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:301–319, 2020. Publisher: Elsevier.

[Halterman *et al.*, 2021] Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. Corpus-Level Evaluation for Event QA: The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253, Online, August 2021. Association for Computational Linguistics.

[Horita *et al.*, 2017] Flávio EA Horita, João Porto de Albuquerque, Victor Marchezini, and Eduardo M Mendiondo. Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems*, 97:12–22, 2017. Publisher: Elsevier.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. arXiv:2310.06825 [cs].

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. Publisher: arXiv Version Number: 1.

[Liu *et al.*, 2020] Tiezhong Liu, Huyuan Zhang, and Hubo Zhang. The impact of social media on risk communication of disasters—a comparative study based on sina weibo blogs related to tianjin explosion and typhoon pigeon. *International journal of environmental research and public health*, 17(3):883, 2020. Publisher: MDPI.

[Montani *et al.*, 2023] Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. explosion/spaCy: v3.7.2: Fixes for APIs and requirements, October 2023.

[noa,] GPT-3.5 Turbo fine-tuning and API updates.

[Sakaki *et al.*, 2012] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE transactions on knowledge and Data Engineering*, 25(4):919–931, 2012. Publisher: IEEE.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv:2307.09288 [cs].

[Varga *et al.*, 2013] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1619–1629, 2013.

[Ziems *et al.*, 2023] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science?, April 2023. arXiv:2305.03514 [cs].

[Zou *et al.*, 2023] Lei Zou, Danqing Liao, Nina SN Lam, Michelle A Meyer, Nasir G Gharaibeh, Heng Cai, Bing Zhou, and Dongying Li. Social media for emergency rescue: An analysis of rescue requests on Twitter during Hurricane Harvey. *International Journal of Disaster Risk Reduction*, 85:103513, 2023. Publisher: Elsevier.