Towards Identifying Local Content Deserts with Open-Source Large Language Models

Marianne Aubin Le Quéré
Cornell University
msa258@cornell.edu

Siyan Wang
Cornell University

Tazbia Fatima
Hearst Newspapers

Michael Krisch
Brown Institute for Media Innovation

ABSTRACT

News deserts have been defined as areas where residents do not have access to news and credible information. These are usually defined by whether an area has a physically proximate local news organization. In this project, we conceptualize *content deserts*: geographic areas that are systematically undercovered or not covered at all by the local press. We demonstrate an early approach to leveraging open-source large language models to identify article locations as well as key information about articles such as topic and community information need. We show that open-source language models can accurately identify the locations mentioned in a news article. When it comes to annotating local news articles, we show that the models perform well for tagging an article's topic, but that other local categorizations do not perform as well. We deploy the best-performing model and prompt on a set of 1,000 articles from two publication, and demonstrate how the annotations can help to identify *content deserts*. Looking forward, these methods will allow for the construction of auditing tools for journalists to view how their coverage differs by neighborhood along topical axes.

1 INTRODUCTION

As the local journalism crisis crescendos, existing disparities in access to crucial information sharpen and grow. Capping off long years of media layoffs, in January 2024 alone, over 500 journalists lost their jobs [8]. This worrisome trend stands to accentuate existing "news deserts," areas that have no or limited local outlets dedicated to their coverage. The proliferation of these news deserts has dire consequences: in areas without reliable news coverage, people vote in a more polarized fashion, communities feel more disconnected, and corporations are more prone to corruption. [6, 11, 17, 18].

Against this backdrop, the relationship between the news industry and the power players training the latest AI models is contested. On the one hand, the New York Times is at the helm of a movement seeking to repudiate OpenAI for surfacing paywalled and copyrighted content [9]. On the other hand, an increasing number of news aggregators and publishers are striking deals with AI companies to surface their content in searches and carve out their role in the new information-seeking landscape [15]. These dynamics lead to an industry that is mistrusting of private AI companies, yet is trying to reconfigure their business relationship with them. Thus, the need emerges to find solutions for journalistic systems that contend with this complex dynamic.

Simultaneously, large language models (LLMs) do have the potential to increase our capabilities for analyzing and understanding text content such as news articles at scale. In particular, scholars and journalists have used manual and automated tools to try and identify which locations have been mentioned in an article (e.g. [7, 20, 24]) Already, LLMs have been shown to be useful for categorizing text content for various tasks such as sentiment and misinformation detection [12, 25]. Looking forward, to further these lines of research, we must understand how well equipped LLMs are at performing local news analysis tasks specifically.

In this work, we seek to bring some of the potential benefits of LLMs for local news understanding to journalists and scholars through leveraging accessible open-source models. We describe how we created a set of human ratings of local news articles, and show the performance of various open-source AI models against them. Based on our preliminary findings, some key tasks such as identifying mentioned location and news topics are easy for open-source models to contend with, while others will require further optimization. We show the usefulness of this system through an analysis of on an exploratory dataset of 1,000 news articles. Ultimately, our system has the potential to contribute (a) an explicit conceptualization of *content deserts*, (b) an understanding of how well LLMs can identify locations and journalistic concepts in news articles, (c) an open-source auditing tool that local journalists can use to gain information about their own reporting and push for equity in coverage, and (d) a large-scale understanding of how socioeconomic trends contribute to nuanced journalistic coverage.

2 BACKGROUND

In this section, we review relevant literature to the areas of news deserts, content deserts, and review how LLMs have been applied to news analysis thus far.

The term *news deserts* has come to be the primary way that scholars and pundits define the ongoing local news crisis. The concept of *news deserts* emerged when it became apparent that many towns and communities were losing their local newspapers, and with them, their source of reliable community information [3]. Drawing on multiple definitions, Gulyas [10] defines a *news desert*

Computation + Journalism Symposium 2024, October 25-27, Boston, MA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

as "the lack of, or diminished availability, access or use of local news or media to a community in a geographical area." As of the end of 2023, over half of counties in the U.S. have "no or very limited access" to local news [23].

Improved automated location extraction and article availability have stirred scholars to conceptualize what we term *content deserts*. In contrast to news deserts, content deserts are defined using an audience-centered lens. In this work, we define *content deserts* as *geographic areas that are systematically undercovered or not covered at all by any local press.*

Thus far, most efforts to identify *content deserts* rely on manual or qualitative content analysis. One early such effort manually mapped the coverage of a Canadian newspaper to thirteen neighborhoods in Toronto, identifying that the coverage often reinforced stereotypes that could negatively impact the local community [16]. The most expansive effort to map *content deserts* is likely that performed by Napoli and Weber [20], who map the content production ecosystems of 100 communities. One of their central findings is that communities with a higher proportion of hispanic or latina residents are likely to have a journalistic output that is less robust. These findings demonstrate that a content-based analysis can reveal underlying inequalities in news coverage, however the manual nature of this content analysis has thus far limited generalizability.

Nonetheless, there have also been automated efforts to identify *content deserts*. For example, Vogler et al. [24] use a natural entity recognition pipeline to identify mentions of locations in a corpus of local news content based in Switzerland, and find that over time the number of unique place names mentioned appears to be in decline. Though not explicitly focused on identifying *content deserts*, Cai and Tian [5] also recognised the need to geo-locate local news articles and developed a contextually specific method for identifying locations in news articles. In light of these ongoing conversations, LLMs present an interesting new path forward for identifying locations mentioned in news articles.

There has also been interest by journalists to identify *content deserts* in their own coverage. These efforts have led to creations of tools to audit newspaper coverage. For example, the Philadelphia Inquirer launched a tool that allowed them to visualize their coverage by geography and analyze their coverage by demographic makeup of a neighborhood [2]. More recently, a New York City-based local newspaper experimented with using GPT-4 to map article coverage by neighborhood [7]. Despite interest from journalists, these tools currently require a lot of manual work to deploy at a particular location, and frequently require relying on private models.

There has been significant hype around deepening the understanding of news articles at scale with LLMs. In one of the earliest examples of using LLMs to tag qualitative data, Ziems et al. [25] demonstrate the ability to detect the presence of misinformation in a news corpus. Looking specifically at political news articles, others have found that GPT-4 can accurately tag components of text such as sentiment and ideology [12]. Overall, while active works in this area suggests that scholars are keen to use LLMs to tag news content, few have yet leveraged open-source models for these tasks.

To create AI tools for and with journalists, it is particularly critical to understand if open-source models can perform well at tagging news articles. In a comparison of various open-source models, Alizadeh et al. [4] found that they could perform better than crowd workers on average, and in some cases better than the closed OpenAI models. However, the authors relied on general tasks not specific to the local press, and did not look at location extraction. Here, we ask the following research questions:

- **RQ1:** How well do open-source language models perform at tagging news articles with relevant categories to the study of local news?
- **RQ2:** How well do open-source language models perform at identifying relevant locations mentioned in local news articles?
- **RQ3:** How does topical coverage in New York differ by borough and publication?

3 METHODS

In this section, we describe our methodology for comparing the performance of LLMs against that of humans.

3.1 Data

For this proof-of-concept, we focus our selection of articles on two New York City based outlets. The news outlets we chose were *The City*, which self-describes as "a nonprofit, nonpartisan, digital news platform dedicated to hard-hitting reporting that serves the people of New York" [1]. The second outlet we select is *The New York Post*, which touts itself as "America's oldest continuously-published newspaper," and is well-known for its tabloid format and sensationalist crime coverage. These were chosen to ensure we had diverse coverage of the same approximate geographic areas. The City's coverage was collected directly from their website, and comprises 4,810 articles between August 2018 and February 2024. To get the New York Post's coverage, we used an existing dataset called NELA-LOCAL, which includes approximately 94,776 articles published between January 2018 and February 2024 [13]. We additionally use a dataset of random, national news articles from 2019 as a control to avoid overoptimizing for the New York City use case [21].

3.2 Annotation Tasks

Our goal is to extract comprehensive information that is locally pertinent for each article. To be able to map articles, we want to identify *all locations* mentioned, as well as the *primary location*. To understand the degree to which outlets may be sensitive to different types of community needs, we also want to understand if an article is about a *specific community*, such as Asian-Americans or churchgoers. We also wanted to classify, broadly speaking, which *topic* an article is about. The list of topics was selected from a survey that compared the types of information the public prefers to receive from a local vs a national publication [22]. Finally, we also asked the crowd workers to label if an article met a community *information need*, since prior work has particularly honed in on the lack of information needs being met in news deserts [20].

3.3 Model selection and prompt development

Since we are seeking to deploy LLMs to tag articles with location and categorical information, we tested a variety of models, parameters, and prompts. The local news industry is often cash-constrained, so another goal of ours is to use cheap and accessible models that can be leveraged by journalists in the future. We select

three models for analysis: GPT-3.5-turbo, Llama3-8B, and Mistral3-7B. We use GPT-3.5-turbo as a baseline private model, while Llama3 and Mistral3 are currently two of the most frequently used and high-performing open-source models. However, we limit ourselves to the lower-parameter models; again, we are concerned with the potential accessibility of these tools and their ability to scale.

We implemented three types of prompts. The first type of prompt, which we refer to as the *static* prompt, uses the same instructions that were shared with human participants for article labeling. The second type of prompt asks the language model to imagine they are a data scientist and mapper (*role1*), while the third type of prompt asks the language model to imagine they are a local journalist (*role2*). For both the *role1* and *role2* prompts, we also provide a brief project purpose. We set the model temperatures to 0 for consistency.

3.4 Collecting ground truth

We use a crowdworker pipeline to generate a gold standard dataset against which to compare the LLM's performance. We compiled 150 total articles, 50 from the City, 50 from the New York Post, and 50 from the random news dataset.

In total, we solicited headline and article ratings from 127 Prolific participants. After training, each crowdworker was given the headline and first 250 words of five news articles. These articles are randomly selected from our the 150 sampled articles. We select only the first 250 words of an article to limit the loss of crowdworker attention. Additionally, we include javascript code to ensure that the crowdworkers cannot copy and paste the article text into ChatGPT. On average, participants took 30 minutes and 20 seconds to complete the task, and were compensated $4.90 for their time. These procedures were approved by Cornell University's IRB.

After collecting responses, we aggregated them to construct a ground truth for model comparison. Our goal was to have each article be rated by at least three participants to be considered usable. Each article was rated by between 3-9 raters, with the mean number of raters per article being 4.2. We consider a location to be "mentioned" if at least two crowdworkers mention the location. We consider a location to be "primary" if a majority of people agree that a location should be deemed primary. For all other questions, we default to the majority choice as the ground truth.

3.5 Calculating model accuracy

We first seek to compare the accuracy of the locations returned by the models to those returned by the human raters. Comparing locations is challenging since people may use different words to refer to the same entity, e.g. "NYC" versus "New York City." To tackle this problem, we geocode every location using the Google Maps API and save its Google Maps ID. We consider a location to be equivalent if the Google Maps ID are the same. Occasionally, locations may not return a Google Maps ID, such as when a location is too general (e.g. "subway"). If a location does not return a Google Maps ID, we consider exact lowercase string matches to be equivalent.

We use different techniques for creating the "gold standard" data to compare against depending on the annotation task. The most complex is the *all locations* task, where we asked the participants and the model to provide a comma-separated list of locations mentioned in an article. After matching location strings to their Google Maps ID, we subset the list of all locations provided by any participants to a list of *valid locations*: these are locations that have been listed by at least two participants. We select locations that have been listed by at least two participants to ward against instances where a participant may have mistakenly included a location in a list. We then rate each model's performance by calculating how many of the locations returned by the model have a Google Maps ID that matches one in the *valid locations* list. To compare the categorical data, we use accuracy between the category returned by the models and the modal category identified by the human raters. If more than one category is the modal category, any of the modes may be considered correct. Comparing the *primary location* is treated similarly to the categorical data.

3.6 Exploratory Analysis

To demonstrate how these techniques can be used to identify *content deserts*, we ran a larger sample of articles through this pipeline to explore initial findings. From our super-set of articles, we randomly sampled 500 articles from the City and 500 articles from the NY Post for further analysis. For simplicity, in this short analysis, we focus on only *locations* mentioned and *topics* identified.

From the accuracy numbers, we identified that the best-performing model for our use case is the Llama3 model using the local journalist prompt. We deployed this model to annotate these 1,000 sampled news articles. We then extracted, from the list of returned *locations* mentioned by an article, if an article explicitly mentions one of the five boroughs, or the location of "New York." In the results section, we present a high-level analysis of how divergent topics are covered by these two publications across boroughs.

4 RESULTS

In this section, we evaluate how well the models perform against our annotation tasks, and showcase an exploratory analysis.

4.1 Model Performance

RQ1 pertains to how well open-source models can annotate news articles with relevant categorical concepts. We display the overall accuracy of each model against the human "gold standard" dataset in Table 1. At a high level, the best-performing model remains GPT-3.5-turbo, which is the best performing model for a majority of metrics. In particular, this model performs very well at categorizing the topic of an article, where the static prompts achieves a 78% accuracy. Llama3 also classifies the news topic well, achieving a 70% accuracy at this task with the Role2 prompt. The information needs are identified quite well by GPT-3.5.-turbo, and while Llama3 and Mistral3 lag behind, Llama3 with role2 still manages a 0.64 raw accuracy score. Notably, none of the models perform well at identifying if the model was focused on a specific community. Though the accuracy may look comparable to the others, this is a binary classification task rather than a multiclass task, and as such an accuracy of 50% is expected.

Our second research question concerns how well the open-source models are able to retrieve locations mentioned in a news article. In Table 1, we identify two metrics for how well a model identifies mentioned locations. The models perform quite high in terms of the % of locations matched, meaning 74% of the locations Llama3

Annotation Task	Metric	GPT-3.5-turbo			Llama3			Mistral3		
		Static	Role1	Role2	Static	Role1	Role2	Static	Role1	Role2
All locations	% matched	0.68	0.71	0.71	0.73	0.72	0.74*†	0.65	0.66	0.67
All locations	% missing	0.16*	0.18	0.20	0.20	0.18†	0.19	0.28	0.27	0.27
Primary location	accuracy	0.67*	0.65	0.63	0.58	0.60	0.58	0.63†	0.61	0.59
Specific community	accuracy	0.65*	0.65	0.65	0.59	0.62	0.60	0.64†	0.63	0.64
Topic	accuracy	0.78*	0.73	0.75	0.70	0.66	0.70†	0.60	0.68	0.67
Information Needs	accuracy	0.69*	0.69	0.67	0.55	0.57	0.64†	0.54	0.53	0.54

Note: * best-performing; † best-performing open-source

Table 1: Table shows the accuracy metrics between three LLMs and the "gold standard" human-aggregated dataset. GPT-3.5-turbo seems to perform the best overall, achieving the highest accuracy for a majority of measures. Of the open-source model, Llama3 outperforms Mistral3 for a majority of tasks, with the Role2 (Local Journalist) prompt performing particularly well.

identified as being mentioned in an article are also identified by at least two humans. Conversely, the percentage of locations identified by at least two humans and not mentioned by the models is 18% for the best-performing open-source model. While these metrics still leave room for improvement, they are promising in terms of being able to correctly identify mentioned locations in the future.

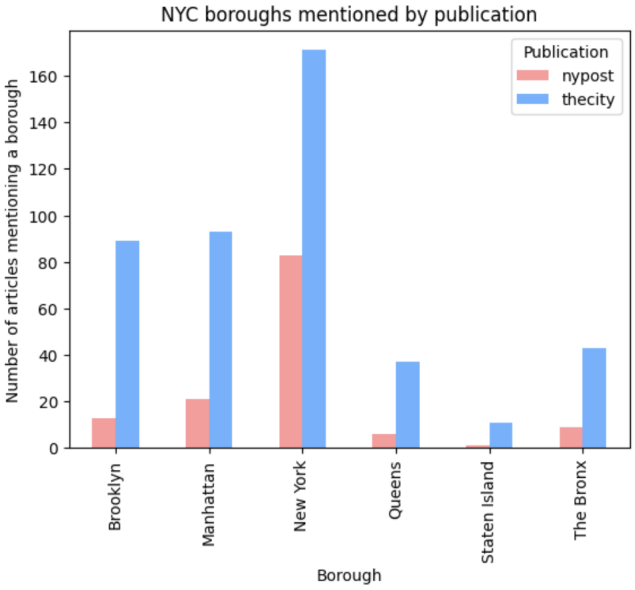

Figure 1: Mentions of NYC boroughs for 1,000 sampled articles from the City and the New York Post. While the location 'New York' is mentioned most by both outlets, a higher proportion of the City's coverage mention specific boroughs.

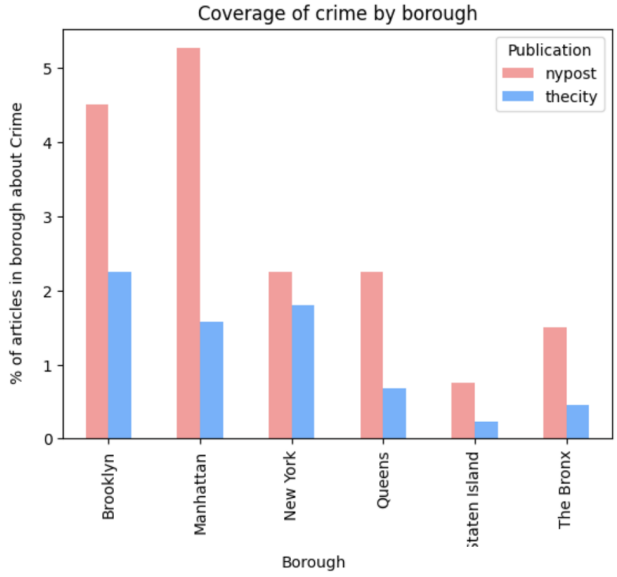

Figure 2: Bar graph shows, for each borough and publication, what percentage of the coverage pertains to the topic of crime. While coverage of crime is overall not that high, the NY Post covers crime proportionally more than the City for the individual boroughs, but not for New York in general.

4.2 Identifying Content Deserts

To answer RQ3, we demonstrate one application of our location matching and tagging for identifying equitable coverage and *content deserts*. First, we show how much each borough is mentioned by both publications in Figure 1. While both publications mention "New York" the most often, the NY Post's coverage mentions the other boroughs proportionally less when compared with The City. Additionally, The City appears to mention Manhattan and Brooklyn proportionally more than Queens, the Bronx, and Staten Island.

We further explore how much certain topics are covered within each borough. For this analysis, we opted to compare the *crime* topic, since coverage of crime is a crucial component of local news coverage, but can also have implications for the equitable representation of specific communities. Figure 2 visualizes what percentage of the stories in each borough by each publication is related to *crime*. Looking at Figure 2, we can see that while the *crime* topic in "New York" is covered approximately equally by both publications, the individual boroughs are more likely to be framed in terms of crime by the NY Post than by The City.

5 DISCUSSION

In this work, we demonstrate the potential for open-source LLMs to identify *content deserts* in local coverage. Although there is room for improvement, our findings suggest that these models can accurately identify locations mentioned in an article, when compared to a set of human ratings. Our results contribute a new systematic approach to identifying locations in news articles, which others have tried to implement using other NLP approaches [5, 14, 19]. Although other tools have been developed that demonstrate the potential for LLMs to tag article locations [7], our approach is the first to present findings on the robustness of this endeavor and using open-source models. In future work, we will continue to optimize approaches to disambiguate between local place names.

We have also explored the use of open-source LLMs for tagging local news content. Mirroring prior work, we find that news topics can be identified quite reliably with nimble open-source zero-shot models as annotators [4]. The open-source models perform moderately well at identifying local information needs. These needs have been fundamental to previous approaches that map *content deserts* [20], and being able to identify these at scale has not been possible until now. Despite these wins, the models perform quite poorly at identifying if an article refers to a specific community or group, implying that some questions of importance for the study of local news cannot be used as easily out of the box. In the future, we will continue to refine our approach for identifying these categories, for example by experimenting with finetuning these models.

Finally, we demonstrated a first approach to identifying *content deserts* with our tagging pipeline. At a high level, we show that both the NY Post and The City tend to cover Brooklyn and Manhattan more than Queens, the Bronx, and Staten Island. Although we only have two publications comprised in this initial analysis, and we do not yet control for population, such findings might suggest that Queens, the Bronx, and Staten Island are systematically undercovered. These types of findings could be directly used by newspapers to balance which locations they report on and spend resources on.

6 CONCLUSION

Journalists and scholars need reliable ways to audit newspaper coverage by location. New technologies of LLMs offer promising avenues for extracting precise locations from text for such an analysis. However, their capabilities must be balanced with the risk of sharing proprietary data with contested AI companies, and the relative overhead of costly models that require excessive GPUs. In this work, we test the potential for accessible, cheap, and open-source language models to be used to annotate local news coverage and locations. We find that largely, these open-source models can reliably identify article locations and news topics. However, the performance on more local news-specific tasks such as identifying which articles cover specific communities, remains less stable. Our proposed pipeline lays the groundwork for robust, cheap, and generalizable local news audits.

7 ACKNOWLEDGEMENTS

We thank Ben D. Horne for kindly supplying us with the data for this work, and Mark Hansen and Mor Naaman for their guidance.

REFERENCES

[1] [n. d.]. *About Us.* http://www.thecity.nyc/about-us/
[2] [n. d.]. *Considering geography, equity, and representation in news - Lenfest Institute.* http://www.lenfestinstitute.org/solutions-resources/the-opportunities-and-limitations-of-considering-geography-as-an-aspect-of-equity-and-representation-in-local-news/
[3] Penelope Muse Abernathy. 2016. *The rise of a new media baron and the emerging threat of news deserts.* University of North Carolina.
[4] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179* (2023).
[5] Guoray Cai and Ye Tian. 2016. Towards geo-referencing infrastructure for local news. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*. 1–10.
[6] Joshua P Darr, Matthew P Hitt, and Johanna L Dunaway. 2018. Newspaper closures polarize voting behavior. *Journal of Communication* 68, 6 (2018), 1007–1028.
[7] Tazbia Fatima. 2024. *We Asked an AI to Map Our Stories Across NYC.* http://www.thecity.nyc/2024/02/29/chatgpt-map-stories-nyc/
[8] Kerra Frazier. 2024. *Over 500 journalists were laid off in January 2024 alone - POLITICO.* https://www.politico.com/news/2024/02/01/journalism-layoffs-00138517
[9] Michael M. Grynbaum and Ryan Mac. 2023. *New York Times Sues OpenAI and Microsoft Over Use of Copyrighted Work - The New York Times.* https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
[10] Agnes Gulyas. 2021. Local news deserts. In *Reappraising local and community news in the UK*. Routledge, 16–28.
[11] Danny Hayes and Jennifer L Lawless. 2018. The decline of local news and its effects: New evidence from longitudinal data. *The Journal of Politics* 80, 1 (2018), 332–336.
[12] Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics* 11, 1 (2024), 20531680241236239.
[13] Benjamin D Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P Wihbey, and Sibel Adalı. 2022. NELA-Local: A dataset of US local news articles for the study of county-level news ecosystems. In *Proceedings of the international AAAI conference on web and social media*, Vol. 16. 1275–1284.
[14] Morteza Karimzadeh and Alan M MacEachren. 2019. GeoAnnotator: a collaborative semi-automatic platform for constructing geo-annotated text corpora. *ISPRS international journal of geo-information* 8, 4 (2019), 161.
[15] Melissa Koenig. 2024. *OpenAI Makes Rapid-Fire Deals With Media Companies: Publication List | Observer.* https://observer.com/2024/05/openai-news-media-deal-list/
[16] April Lindgren. 2009. News, geography and disadvantage: Mapping newspaper coverage of high-needs neighbourhoods in Toronto, Canada. *Canadian Journal of Urban Research* 18, 1 (2009), 74–97.
[17] Ted Matherly and Brad N Greenwood. 2021. No news is bad news: Political corruption, news deserts, and the decline of the fourth estate. *Academy of Manage-ment Proceedings* 1 (2021).
[18] Nick Mathews. 2022. Life in a news desert: The perceived impact of a newspaper closure on community members. *Journalism* 23, 6 (2022), 1250–1265.
[19] Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–27.
[20] Philip M Napoli and Matthew S Weber. 2020. Local Journalism and At-Risk Communities in the United States. In *The Routledge Companion to Local Media and Journalism*. Routledge, 368–378.
[21] Alina Petukhova and Nuno Fachada. 2023. MN-DS: A multilabeled news dataset for news articles hierarchical classification. *Data* 8, 5 (2023), 74.
[22] Stacy Rosenberg. [n. d.]. *5. The importance of local news topics often does not align with how easily the public can find information about them.* https://www.pewresearch.org/journalism/2019/03/26/the-importance-of-local-news-topics-often-does-not-align-with-how-easily-the-public-can-find-information-about-them/
[23] Medill-Northwestern University. 2023. *More than half of U.S. counties have no access or very limited access to local news - Medill - Northwestern University.* https://www.medill.northwestern.edu/news/2023/more-than-half-of-us-counties-have-no-access-or-very-limited-access-to-local-news.html
[24] Daniel Vogler, Morley Weston, and Linards Udris. 2023. Investigating News Deserts on the Content Level: Geographical Diversity in Swiss News Media. *Media and Communication* 11, 3 (2023), 343–354.
[25] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.