A Case Study in an A.I.-Assisted Content Audit

Rahul Bhargava
r.bhargava@northesatern.edu
Northeastern University
Boston, Massachusetts, USA

Elisabeth Hadjis
hadjis.e@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Meg Heckman
m.heckman@northesatern.edu
Northeastern University
Boston, Massachusetts, USA

ABSTRACT

This paper presents an experimental case study utilizing machine learning and generative AI to audit content diversity in a hyperlocal news outlet, The Scope, based at a university and focused on underrepresented communities in Boston. Through computational text analysis, including entity extraction, topic labeling, and quote extraction and attribution, we evaluate the extent to which The Scope's coverage aligns with its mission to amplify diverse voices. The results reveal coverage patterns, topical focus, and source demographics, highlighting areas for improvement in editorial practices. This research underscores the potential for AI-driven tools to support similar small newsrooms in enhancing content diversity and alignment with their community-focused missions. Future work envisions developing a cost-effective auditing toolkit to aid hyperlocal publishers in assessing and improving their coverage.

1 INTRODUCTION

There is growing interest in auditing news content to assess diversity of ownership, brand production, content, and consumption [16]. This can be connected to recent struggles to respond to persistent accusations of coverage bias [21], renewed questioning of what objectivity means in community-focused reporting [29], and the continual iterative growth in the capabilities of computational text analysis. For newsrooms that cover small to mid-sized communities, content diversity audits can demonstrate how they are, or are not, adequately representing the areas and populations they serve vis-a-vis aspects of topic, identity, place, and more. This can, in turn, lead to changes in editorial guidelines and beats covered by reporters, feeding into a virtuous cycle between news organizations, the communities they serve, and automated analysis of their content.

This paper presents an experimental case study using machine learning and generative AI to assess the diversity in coverage of themes, places, and people quoted in The Scope, a university-based hyperlocal news outlet [12]. Founded in 2017 to amplify underrepresented voices in Boston, The Scope aims to connect communities by fostering civic life. It is one of a number of outlets that demonstrate how student journalists are filling local news gaps with university-based hyperlocal publications, addressing news deserts while building skills [15]. These initiatives extend coverage to underserved communities and offer students practical experiences and digital storytelling experimentation [8]. They can also serve as testing grounds for emerging technologies like generative AI in local journalism [15]. This study innovates cost-effective methods to leverage off-the-shelf tools in support of partially automating analysis of content from The Scope, shares results and editorial reflections and plans, and lays the groundwork for a proposed audit toolkit to support similar hyperlocal newsrooms.

2 RELATED WORK

This case study brings together several research topics, including our evolving understating of identity-based news deserts, existing approaches to assessing news diversity, modern techniques of natural language parsing and topic detection, and experimental applications of chat-based interfaces to large language models.

The rise of news deserts, areas lacking credible local news, has become a pressing concern in journalism and the media industry [23, 10]. Rebuilding the local information ecosystem requires attention not just to geographic focus areas but also to the representation of diverse communities [34]. Despite challenges, there is potential for revitalization through innovation, community engagement, and targeted interventions, with universities and student-staffed news organizations playing key roles in providing local coverage [9].

Thus, content diversity is a growing area of concern for newsrooms covering communities made up of residents who share multiple identities. Journalists often prioritize the voices of government officials, business leaders and other institutional authorities over ordinary citizens [19, 33]. The Gender Gap Tracker found that men are quoted about three times as frequently as women across various news outlets and time intervals. This imbalance exists despite efforts by journalists to provide a more diverse set of voices, suggesting that structural changes are necessary to achieve gender parity in news quoting [3]. Student journalists also follow this pattern, though somewhat less so [32].

A small but growing number of news organizations are adding source audits to their workflows. For instance, National Public Radio launched software in 2021 to track details ranging from the race and ethnicity of the source to their age and geographic location [13]. Local outlets are building their own simple workflows to track sources via manual data entry [24]. Other outlets employ commercial services such as SourceMatters or AllSides.

In parallel, the capabilities of computational analysis to support semi-automated auditing of news content has grown significantly. Named entity recognition (NER) is a foundational task in natural language processing (NLP) that involves identifying and categorizing specific types of content within a body of text: peoples, organizations, dates, places, etc. Off-the-shelf software libraries such as SpaCy make NER quite simple to integrate when you have large sets of text documents needing analysis [17]. Similarly, automatically labeling a news story against a set of pre-determined topic labels is a well-established approach, with well-performing examples that leverage a variety of machine learning techniques [11].

Quotes are also a standard form of sharing information from sources in news articles, driven in part by various practices related to power and influence in journalism [4]. The computational tasks of extracting quotes, attributing quotes (including anaphora resolution), and disambiguating speakers (including coreference resolution) in news content is generally approached via heuristic

[27], machine-learning [14], or combined approaches [22]. While Quootstrap [26] and the Stanford CoreNLP dcoref QuoteAnnotator [20] are freely available, neither is both ready-to-use and performant in news audit settings.

Outside of established machine learning approaches, the rise of generative AI tools such as ChatGPT present new opportunities for even the smallest newsrooms to conduct computer-assisted source audits (driven by the novel chat-based interface and the promise of solving a broader range of content analysis needs). University-based publications like the one in this study have the potential to be venues for experimentation with emerging digital tools such as generative AI — a technology that both fascinates and concerns news executives and journalism scholars [7].

3 METHODOLOGY

Our venue for this study is The Scope, a university-based hyperlocal news organization covering core neighborhoods in Boston. The Scope aims to serve geographic areas and identity-based communities traditionally overlooked by legacy media. Its mission statement notes its commitment to inclusive coverage; this is also repeatedly emphasized by faculty advisors and supervising editors. The authors of this paper are affiliated with The Scope, two as faculty members and one as a former student who is now working as the full-time editor-in-chief. This positionality gave us access to the publication's content management system and knowledge of the publication's editorial routines.

The Scope is published on a custom hosted installation of the WordPress content management system. Since launching in 2017 it has published over 700 stories. Via the built-in WordPress XML export feature, we downloaded article content and prepared it for analysis programmatically in a Jupyter Python, converting it from HTML to text using the BeautifulSoup library. After removing non-English stories (n=24), stories that were too long to process via the libraries and services selected (more than 20,000 characters, n=17), stories with esoteric parse errors (n=5), and stories that were short placeholders or announcements (less than 1,000 characters, n=13), we were left with 654 stories.

3.1 Entity Extraction

The SpaCy library allows for relatively easy programmatic extraction of entities. Using a simple server-based API wrapper that was created and deployed for other article analysis work, we processed stories to extract the named people, organizations, and places from each article [6].

This processing yielded 6,873 mentions of 857 unique potential geographic entities. Based on editorial guidance we aggregated results to count the number of stories each place was mentioned in, limited by frequency to places mentioned in at least 5 stories, and categorized the resulting locations by whether there in the local metropolitan region or not. This yielded 34 total "relevant" places for analysis (at different scales of geography).

3.2 Topic Labeling

Historically printed newspapers published content in sections such as "sports", "health", and so on. Using an existing server-based API to a custom multi-label machine learning classifier, we labelled each story with the high-level "topic" it was determined to be about [30, 28]. The training data for that model is the top 600 most-used labels in the NYT Annotated corpus [31], combined with manually created mappings from the NYT labels to traditional sections. This yielded 1,059 topic labels on stories. To further support high-level analysis, we assigned each story a "top" topic based on the topic with the highest confidence score (0 to 1).

3.3 Quote Extraction and Attribution

Existing ready-to-use quote-related solutions did not perform in early testing, so we looked to new chat-based LLMs to extract and attribute quotes. Based light experimental evaluation of prototypes we selected ChatGPT's "gpt-3.5-turbo" model. The first prompt focused on quote extraction and attribution by name and pronoun; the second prompt focused on speaker disambiguation. We relied on reporting norms associated with pronouns use to avoid ethical concerns related to automating analysis of gender in text corpora [18, 25]. For validation we used both random sampling and qualitative review of results. The full iterative development and final data processing via the OpenAI API cost USD $6. After some data cleaning this left us with 1,388 speakers quoted across this corpus of stories. See [5] for more details on these methods.

4 RESULTS & IMPACT

The corpus includes 654 stories published over almost seven years. As shown in Figure 1, the rate of publication has tracked with the ebb and flow of student writing and staff editing capacity. The largest spike of stories occurs in late 2018, attributable to a special project that yielded dozens of interview-based stories about residents of a single neighborhood near campus. These were published in a short window, creating the spike.

Figure 1: The quantity of stories published each week.

4.1 Relevant Places

After geographic entity extraction, aggregation, and filtering for relevance, we were left with 411 (63%) stories that mentioned a relevant location (703 mentions across those stories). Figure 2 shows the frequency of relevant places mentioned across all stories.

The heavy concentration of stories mentioning the Mission Hill neighborhood is unsurprising, giving a special project that was launched in that geographic area. The overall breakdown of source geography does show some under coverage of areas that are defined as poor neighborhoods vis-à-vis the mission statement of The Scope, namely Mattapan, Nubian Square, Roxbury and Fenway.

4.2 Topical Coverage

Figure 3 shows the frequency of top topics by story volume. The top three topics are culture, politics, and economics. This reflects the

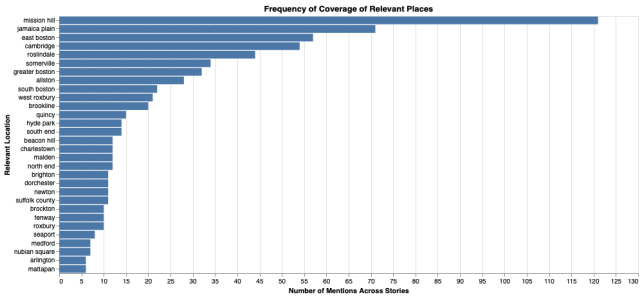

Figure 2: The most-mentioned relevant places.

Boston area's vibrant and multifaceted cultural scene as well as The Scope's commitment to providing regular coverage of local politics. Reviewing sample headlines from stories labelled as "economics" reveals key themes included housing costs, unemployment and food insecurity. The three least covered topics — transportation, sports and technology — point to possible areas for more intense coverage in the future especially as these topics intersect with the outlets focus on race and class.

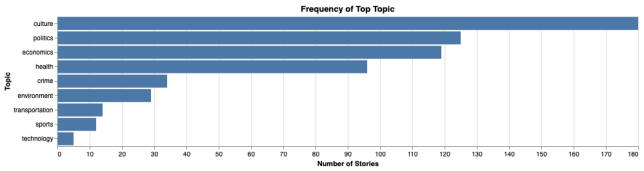

Figure 3: The number of stories about each topic.

Discussions with editorial staff led us to create an analysis of topic covered by location that provides a more granular look at The Scope's coverage patterns (Figure 4). This multi-dimensional picture of what was covered and where allows for more detailed understanding of reporting patterns. Reviewing sample headlines and contextual knowledge, we can see that some of these results are a function of local geography. For instance, the Massachusetts statehouse is in the Beacon Hill neighborhood, hence the concentration of politics stories in that area. The high volume of stories about economics in Nubian Square reflects a focus on the small business community in the heart of the historically Black Roxbury neighborhood. Culture is more evenly represented across neighborhoods likely because of the high volume of stories about that topic. This may also illustrate The Scope's commitment to covering neighborhoods' vibrancy as well as their challenges. The absence of stories about transportation in Mattapan, Quincy, Roxbury and other socioeconomically disadvantaged areas is a coverage gap The Scope might consider addressing in the future.

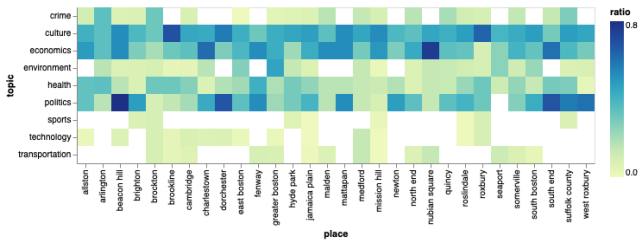

Figure 4: Topic coverage by relevant place.

4.3 Use of Quoted Sources

The corpus of stories includes 5,084 quotes from 1,388 speakers. The frequency of quoted sources matches The Scope newsroom editorial guidance and norms from syllabi at the hosting university (2-3 sources per article). This presents us with an initial piece of evidence indicating that editorial guidelines on representation are being followed. There was a short-term increase in quotes during March of 2022 (Figure 5), for which we can find no obvious explanation beyond normal variance based on authors at the time.

Figure 5: The number of speakers quoted each week.

Looking at the specific people quoted, we see that most quoted sources named in the publication appear by far in only one story. Just 7% (n=107) of the quoted sources are quoted in more than one story. The most quoted person is the current mayor (and former city councilor), who appears in 31 stories.

Returning to the stated mission of The Scope, we see the finding that so few sources are quoted in more than one story as one way to evaluate the goal to "amplify voices that are often overlooked or mischaracterized." While we didn't specifically characterize "overlooked" or "mischaracterized" populations, the fact that the vast majority of quoted speakers are only quoted once suggests that The Scope is engaging with and representing a variety of residents in their coverage. Although the most quoted person is a high-profile politician, she appears in just under 5% of the stories analyzed. Against this metric about the sheer number of quoted speakers, The Scope appears to be performing well against its underlying goal (from our position as engaged parties in the publication).

4.4 Gender of Quoted Speakers

Figure 6 presents the breakdown of pronouns most used with each disambiguated speaker. Based on the extracted and attributed pronouns we find that women (57%) were quoted more than men (38%). People using the pronoun "they" represented 5% of sources quoted. (Note: The Scope launched in late 2017, shortly after the Associated Press Stylebook added guidance for limited use of "they" as a singular pronoun.)

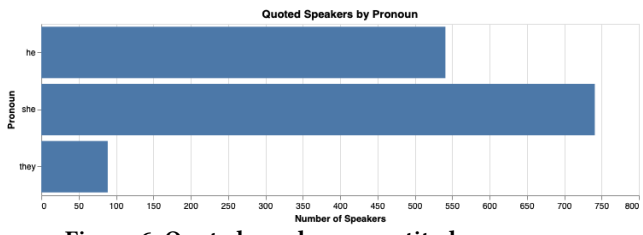

Figure 6: Quoted speakers quantity by pronoun.

Several factors may be at play when considering the relatively high number of women quoted in The Scope stories. During the time analyzed, women were well represented in on the city council

and in a mayoral race that received intense coverage from The Scope. In addition, it is relevant to note that most of the reporters writing for The Scope identify as women; related work has shown that women are more likely to quote other women in digital [2], traditional newspaper [1], and broadcast [35] settings.

We see the over-representation of women as an example of The Scope fulfilling its goal of justice-informed reporting. One can consider the dominance of female quoted sources here as a corrective effort to remedy long-standing bias towards male perspective and voices in news [3]. It also echoes previous research that found student journalists, especially women, were somewhat more likely than their professional counterparts to quote female sources [32].

4.5 Editorial Planning

The Scope's mission centers around reporting about the city's core neighborhoods and their diverse populations. Understanding coverage patterns can help inform The Scope's editorial planning and newsroom operations by indicating which regions are underserved by their reporting.

Geographically, the results of this study indicated that coverage is lower for some neighborhoods, including Mattapan, Nubian Square, Roxbury and Fenway. The Scope plans to use this data as guidance for where to focus staff attention during the next academic year. This data also suggests a new shift in The Scope's organizational structure and assignment process. In the current model, staff are assigned stories based on their beat (city politics, climate and environment, etc.) or medium interests (podcast, text, video etc.) rather than by neighborhood or area. In response to the undercoverage of certain parts of the city, The Scope's editorial team is considering restructuring its staff around geographies. This could create more consistent coverage patterns across neighborhood coverage while allowing reporters to develop deeper community-level source networks.

Regarding topical coverage, the study points to a need for increased reporting on topics of public importance such as transportation, among others. This study also prompts the idea of launching new audience engagement practices, such as running a readership poll to better understand what topics the audiences hope to see covered. This would help shape and refine The Scope's editorial mission

Regarding quoting practices, we learned that very few sources were quoted more than once across different stories. These insights reflect The Scope's goal of amplifying marginalized and underserved voices. In contrast to the prior findings, which provoke changes and new ideas, this finding reinforces that the minimum source number recommended – two to three per story – is effective and should remain part of the publication's editorial guidelines vis-a-vis goals.

On the gender front, the data showed that women are quoted more frequently than men and much more frequently than those using they/them pronouns. While the over-representation of women's voices serves The Scope's diverse source mission and reinforce editorial guidelines, these results point to a need for improvement in the amplification of nonbinary voices. This baseline data serves as a reference point for assessment in future iterations of an audit after policy changes have been rolled out.

5 CONCERNS & LIMITATIONS

There are several limitations to this experimental case study worth considering. First, geographic mentions, topical focus, and source representation are just parts of the news product (albeit ones that can reflect broader newsroom norms). We engaged these metrics for our audit based on an understanding of The Scope's ethics and mission statement, combined with readily available technology capabilities. The results of any source audit should of course be considered as a launchpad for larger conversations, ideally in collaboration with the community, about representation in coverage as opposed to a standalone solution. Such audits should also be understood in the context of other forms of news diversity [16].

Similarly, gender is just one aspect of larger identities influenced and mediated by social systems of power. Other factors such as race, ability, religion, social status, and class should also be considered in conversations about news representation. As noted previously, some newsrooms are collecting this type of self-identified demographic data from sources as part of their standard interview process [13]. The Scope has not taken this step but is considering adding that as a new reporting practice to both produce the data for analysis and build awareness and practices with the student reporters.

Technically, each method we employed raises concerns about both ethics and accuracy. Entity extraction is well-established and validated for English-language corpora, but nonetheless needed hand-tweaking that could alter results is left un-supervised. In our case, a notable example was how the "Jamaica Plain" location was sometimes fully extracted, but other times simply identified as "Jamaica" despite consistent use in the story text. The topic detection engine we used was built on the NYTimes corpora and Google's word2vec, and therefore is both dated and incorporates traditional news biases. Sample-based validation of these off-the-shelf tools could potentially reveal shortcomings in more depth, but would be challenging to perform at scale to catch potential edge cases. Using OpenAI for the quote-related data creates another level of concern due to the oft-cited risks generative AI news stories pose to journalism writ-large. Our research team's affiliation with The Scope also influenced this work, driven by our desire to engage in a participatory design process. This positionality allowed us to make informed and grounded decisions about data collection and analysis, but it may have also limited our perspectives in ways we are not best suited to determine.

6 CONCLUSION & FUTURE WORK

Content audits play a critical role in helping media publishers assess how their coverage is meeting their editorial goals and mission statement. Small and mid-sized student-staffed newsrooms play a critical role in filling geographic and identity-based news deserts. At the intersection of these two phenomena lies the opportunity for novel experimental AI assisted content audits like the case study we've shared in this paper. In the case of The Scope, our findings both validate that the publication is meeting some of its goals and illuminate opportunities to improve alignment between the coverage and the mission statement.

In future work we envision a combination of established machine learning techniques and experimental chat-based AI approaches

could be used to create an "auditing toolkit" available to publishers of similar scale. A hosted web-based service that could feature WordPress content import (as demonstrated here), sitemap-based ingestion, or other approaches, and return raw results for review by editorial staff. While similar commercial services do exist, our initial explorations show an opportunity to innovate on both overall cost and, more importantly, data and visualizations created to support editorial decision-making based on multiple measures. The analysis of quoted speakers, and the connection between topics and geography, included here are examples. Leveraging innovative computational analysis can support newsrooms internally and externally demonstrate some of the ways that their content represents the concerns and identities of the communities they serve, and opportunities to realign coverage with the same.

7 ACKNOWLEDGEMENTS

The authors would like to thank all the writers, editors, and faculty who have contributed to The Scope over the years, especially former editors in chief Catherine McGloin, Ha Ta and Lex Weaver. Since its launch in 2017, The Scope has received several internal grants from Northeastern's College of Arts, Media and Design as well as external funding via Stand Together Fellowships

REFERENCES

[1] Cory L. Armstrong. 2004. The influence of reporter gender on source selection in newspaper stories. *Journalism & Mass Communication Quarterly*, 81, 1, (Mar. 1, 2004), 139–154. Publisher: SAGE Publications Inc. DOI: 10.1177/107769900408100110.

[2] Claudette G Artwick. 2014. News sourcing and gender on twitter. *Journalism*, 15, 8, (Nov. 1, 2014), 1111–1127. Publisher: SAGE Publications. DOI: 10.1177/1464884913505030.

[3] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: using natural language processing to measure gender bias in media. *PLOS ONE*, 16, 1, (Jan. 29, 2021), e0245533. Publisher: Public Library of Science. DOI: 10.1371/journal.pone.0245533.

[4] Dan Berkowitz. 2019. Reporters and their sources. In *The Handbook of Journalism Studies*. (2nd ed.). Routledge, 165–179. ISBN: 978-1-315-16749-7.

[5] Rahul Bhargava, Elisabeth Hadjis, and Meg Heckman. 2024. Testing generative AI for source audits in student-produced local news. In Association for Education in Journalism and Mass Communication (AEJMC). Philadelphia, PA, (Aug. 2024). DOI: 10.31235/osf.io/7hc2d.

[6] [SW] Rahul Bhargava and Harini Suresh, News Entity Server version v2.4.3, July 2024. URL: https://github.com/dataculturegroup/news-entity-server.

[7] K. Boyle. 2023. Artificial intelligence and the future of journalism: where do we go from here? *Newspaper Research Journal*, 44, 1, (Mar. 1, 2023), 3–5. DOI: 10.1177/07395329231158682.

[8] Center for Community News. 2023. More than 14 Million Views: The Impact of Student Reporting. University of Vermont, (Sept. 8, 2023). Retrieved May 6, 2024 from https://www.uvm.edu/sites/default/files/Center-for-Community-News/pdfs/Impact_summary_-_Updated_9.8.23.pdf.

[9] Ellen Clegg and Dan Kennedy. 2024. *Beacon Press: What Works in Community News*. Beacon Press, (Jan. 9, 2024). ISBN: 978-0-8070-0994-9.

[10] McKay Coppins. 2021. A secretive hedge fund is gutting newsrooms. *The Atlantic*, (Oct. 14, 2021). Section: Business Volume Title: November 2021. Retrieved May 6, 2024 from https://www.theatlantic.com/magazine/archive/2021/11/alden-global-capital-killing-americas-newspapers/620171/.

[11] Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba, Robertas Damaševičius, and Abdul Sattar. 2023. Topic classification of online news articles using optimized machine learning models. *Computers*, 12, 1, (Jan. 2023), 16. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/computers12010016.

[12] Teri Finneman, Meg Heckman, and Pamela E. Walck. 2022. Reimagining journalistic roles: how student journalists are taking on the u.s. news desert crisis. *Journalism Studies*, 23, 3, (Feb. 17, 2022), 338–355. Publisher: Routledge _eprint: https://doi.org/10.1080/1461670X.2021.2023323. DOI: 10.1080/1461670X.2021.2023323.

[13] Angela Fu. 2021. New tool allows NPR to track source diversity in real time. Poynter. (Aug. 12, 2021). Retrieved May 6, 2024 from https://www.poynter.org/reporting-editing/2021/new-tool-allows-npr-to-track-source-diversity-in-real-time/.

[14] João Daniel Fernandes Godinho. 2018. *Extraction, Attribution, and Classification of Quotations in Newspaper Articles*. M.S. Técnico Lisboa, Lisobon, Portugal, (Oct. 2018).

[15] Kristen Hare. 2021. Can j-schools fix news deserts? Poynter. (July 21, 2021). Retrieved May 6, 2024 from https://www.poynter.org/reporting-editing/2021/can-j-schools-fix-news-deserts/.

[16] Jonathan Hendrickx, Pieter Ballon, and Heritiana Ranaivoson. 2022. Dissecting news diversity: an integrated conceptual framework. *Journalism*, 23, 8, (Aug. 1, 2022), 1751–1769. Publisher: SAGE Publications. DOI: 10.1177/1464884920966881.

[17] [SW] M. Honnibal and I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing 2017.

[18] Os Keyes. 2018. The misgendering machines: trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2, (Nov. 1, 2018), 88:1–88:22, CSCW, (Nov. 1, 2018). DOI: 10.1145/3274357.

[19] Stephen Lacy and David Coulson. 2000. Comparative case study: newspaper source use on the environmental beat. *Newspaper Research Journal*, 21, 1, (Jan. 2000). DOI: https://doi.org/10.1177/07395329000210010.

[20] H. Lee, Y. PeirsMan, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system. In *Proceedings of the CoNLL-2011 Shared Task*. Fifteenth Conference on Computational Natural Language Learning. Portland, Oregan, (June 2011).

[21] Ric Neo. 2022. Fighting fire with fire? relegitimizing strategies for media institutions faced with unwarranted "fake news" accusations. *Social Media + Society*, 8, 1, (Jan. 1, 2022). Publisher: SAGE Publications Ltd. DOI: 10.1177/20563051221077014.

[22] C. Newell, T. Cowlishaw, and D. Man. 2018. Quote extraction and analysis for news. In *Proceedings of the Workshop on Data Science, Journalism and Media*. KDD.

[23] Rasmus Kleis Nielsen. 2015. The uncertain future of local journalism. In *Local journalism: The decline of newspapers and the rise of digital media*. Rasmus Kleis Nielsen, (Ed.) I.B. Tauris & Co., Rochester, NY, 1–30. Retrieved May 6, 2024 from https://papers.ssrn.com/abstract=2614334.

[24] North Country Public Radio. 2022. FY2022 Diversity Statement. Canton, New York. Retrieved May 6, 2024 from https://www.northcountrypublicradio.org/pdfs/2022DiversityStatement.pdf.

[25] Vedika Pareek. 2019. Non-binary gender and data. In *Handook of Gender and Open Data*. (Dec. 28, 2019). Retrieved May 7, 2024 from https://cis.pubpub.org/pub/non-binary-gender-data/release/7.

[26] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. In *Proceedings of the 12th International Conference on Web and Social Media*. ICWSM. Retrieved May 6, 2024 from arXiv: 1804.02525[cs].

[27] Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2017. Automatic detection of quotations in multilingual news. In Recent Advances in Natural Language Processing. Incoma Ltd., Shoumen (Bulgaria), (Nov. 2017), 487–492. https://publications.jrc.ec.europa.eu/repository/handle/JRC37835.

[28] Hal Roberts et al. 2021. Media cloud: massive open source collection of global news on the open web. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, (May 22, 2021), 1034–1045. DOI: 10.1609/icwsm.v15i1.18127.

[29] Sue Robinson and Kathleen Bartzen Culver. 2019. When white reporters cover race: news media, objectivity and community (dis)trust. *Journalism*, 20, 3, (Mar. 1, 2019), 375–391. Publisher: SAGE Publications. DOI: 10.1177/1464884916663599.

[30] Yasmine Rubinovitz. 2017. *News Matter : embedding human intuition in machine intelligence through interactive data visualizations*. Thesis. Massachusetts Institute of Technology.

[31] Evan Sandhaus. 2008. The new york times annotated corpus LDC2008t19. (2008). Retrieved June 19, 2014 from https://catalog.ldc.upenn.edu/LDC2008T19.

[32] K. Shine. 2021. Gender and sourcing in student journalism from australia and new zealand. *Australian Journalism Review*, 43, 2.

[33] Rodney Tiffen et al. 2014. Sources in the news : a comparative study. *Journalism Studies*, 374–391. DOI: 10.1080/1461670X.2013.831239.

[34] Nikki Usher. 2023. The real problems with the problem of news deserts: toward rooting place, precision, and positionality in scholarship on local news and democracy. *Political Communication*, 40, 2, (Mar. 4, 2023), 238–253. Publisher: Routledge. DOI: 10.1080/10584609.2023.2175399.

[35] Geri Zeldes and Frederick Fico. 2010. Broadcast and cable news network differences in the way reporters used women and minority group sources to cover the 2004 presidential race. *Mass Communication and Society*, 13, (Nov. 1, 2010), 512–527. DOI: 10.1080/15205430903348811.