Comparing open-ended community dialogue with local news

Hope Schroeder[1], Doug Beeferman[1], Deb Roy[1]
[1]MIT Center for Constructive Communication
{hopes, dougb5, dkroy}@mit.edu

Abstract

In the lead-up to Boston's local elections in November 2021, the Real Talk For Change project hosted small group conversations in which local residents shared their experiences of living in Boston in response to the prompt "What is your question about the future of Boston and your place in it? What experience led you to that question?". Over 370 people from 21 of the 23 neighborhoods of Boston participated, often sharing deeply personal stories. The conversations were recorded, analyzed, and used as a basis for sharing themes and stories with the Boston community via a public portal, and as an input into public dialogues with the mayoral candidates. In this study, we apply topic modeling to a subset of the RTFC corpus of transcribed conversations to surface the inferred agenda emerging from conversations, here expressed as a distribution over topics. We compare this to the distribution of topics covered in a time-matched sample of news stories published in *The Boston Globe*. We apply a semi-supervised keyword extraction method to enable quantitative analysis across the conversation and news corpora. Significant differences in the topic distributions of the two corpora reflect a mismatch between how much attention the city's largest news source gives to historically underheard residents and their expressed needs and concerns. The methodology points towards a systematic way for local news organizations to consider community experiences as an input for which topics they cover and how to cover them.

Introduction

In an era of growing distrust in media (Brenan, 2021) and social fragmentation, local news organizations have the potential to play a key role in helping strengthen democracy through new forms of community engagement. New journalism curricula like CUNY's engagement journalism program, tools like Hearken, and organizations like The 19th are emerging to meet the ideal of more community-engaged journalism. A key question in these efforts, which motivates the current study, is whether local news organizations are able to reflect the diverse voices, perspectives and concerns of all members of the community.

We present a preliminary comparison of the emergent topic focus of local residents as expressed in open-ended community dialogues to the agenda of the main local news organization. The study is based on the Real Talk For Change project (RTFC), which in the lead-up to Boston's local elections in November 2021 hosted small group conversations in which local residents shared their experiences of living in Boston in response to the prompt "What is your question about the future of Boston and your place in it? And what experience led you to that question?". RTFC prioritized participation from traditionally underheard voices – BIPOC low-income residents of Boston – by collaborating with several community-based organizations with trusted community networks. As an initial point of comparison, we analyze a corpus of news articles from *The Boston Globe*, the most widely-read local news source in Boston, published over the same time period as the RTFC conversations.

Although we do not expect the agenda of the *Globe* to match the attentional focus of the inherently biased sample of RTFC participants, we believe the automated topic analysis methodology we have developed combined with the community dialogue methods of RTFC can provide a useful new input for newsrooms that seek to better reflect the perspectives of the communities they cover and serve.

Methodological motivation

With a large, complex media ecosystem that includes linguistic corpus types as diverse as formal written news articles, casual spoken language on radio, and high-variance short-form text on social media, automated topic analysis methods that scale and generalize well across language domains are critical.

Spoken and written language are notoriously different (Chafe and Tannen, 1987). Discrepancies in measures like lexical diversity as well as skew towards different word distributions based on text domain presents a challenge for traditional topic modeling methods such as LDA. Structural topic models can allow for the inclusion of covariates in a model, but they do not solve the issue of cross-domain style differences that can negatively affect results. Keyword lists made using noun extraction alone often omit named entities that are essential for news. Popular keyword lists like LIWC cover functional and emotion word lists and have some general noun categories, but they lack the ability to automatically extend into current events with named entities.

Subject-specific hand-crafted keyword lists intended for use at scale are easy to create and update, but they are subject to availability and other biases of their creators, and can also be time-consuming to thoroughly create. Keyword lists created for one text domain may also miss terms that are more common in another text domain. For example, colloquial terms like “cop” are common in spoken language, but less common than “police” in formal writing. With this in mind, we propose and test using word embedding models trained on news to expand seed sets of keywords and terms in order to improve computational topic capture in a lightweight, interpretable fashion that can efficiently elicit topic insight across different types of text.

Data

Real Talk for Change

The Real Talk For Change (RTFC) project created a new opportunity for residents of Boston typically underheard in civic processes to participate in small group recorded conversations. Sixty-nine Real Talk for Change conversations, hosted on the Local Voices Network platform, occurred between August and the end of November of 2021 and involved 322 community members as participants from 21 different neighborhoods. Strong community partnerships with leaders and organizations known for uplifting voices of underheard neighbors enabled a participant recruitment strategy that prioritized some of the most marginalized communities in Boston. As a result of these efforts, 47% of participants identified as Black or African American, 19% identified as Hispanic or Latino, and over the half of participants identified as women. Many lived in neighborhoods with high proportions of Black residents, including Dorchester (38% of all participants), Roxbury (8%), and Mattapan (6%).

Conversations were facilitated by trained community members as well as leaders and organizers from partner organizations. Many of the facilitators were leaders in civic groups or community-based and grassroots organizations. Conversation participants were recruited through partner organization networks, direct recruitment from facilitators, and open registration online. The conversations were held in person or over zoom, and were recorded and transcribed with consent from participants.

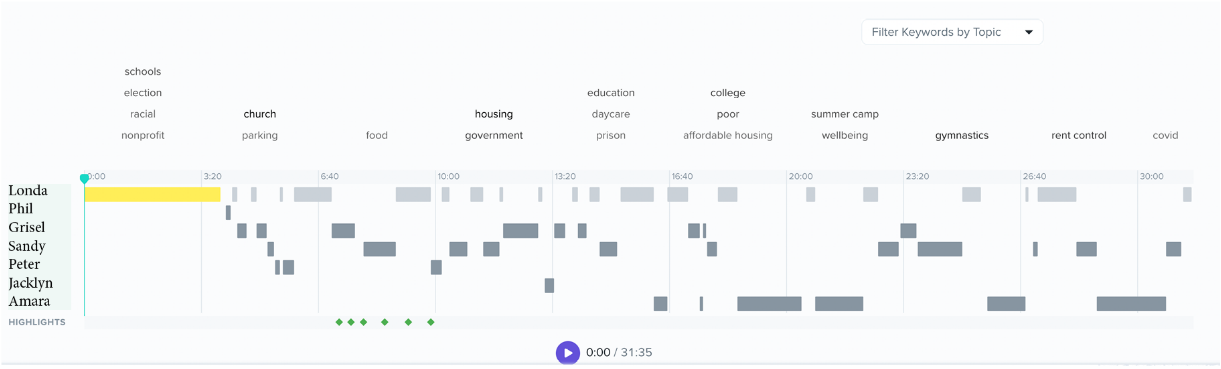
Conversations were facilitated by trained conversation hosts, beginning with a discussion of purpose, consent, and norm-setting. Norms encouraged participants to engage in active listening and speak from personal experience. The first section of the conversation asked participants to share in response to the following prompt: “What is your question about the future of Boston and your place in that future? What experience led you to that question?” An example response someone shared is: “My question about the future of Boston is about education, since I’ve got two four-year-old daughters. What is the educational system going to look like five years from now? 10 years from now? 20 years from now? Because I want them to have the best education possible, so that’s my question about the future of Boston. Where I see myself in it? I do work within the school system a little bit, so that’s where I’m at. I want to have them thrive and do the best that they can do.”

The average length of a conversation was 51 minutes but ranged between 29 and 79 minutes. Conversation audio for each conversation was manually transcribed and unique speaker turns within the conversation were recorded. Facilitators encouraged turn-taking to share time within the conversation. The average number of participants in a conversation was between 6 and 7 but ranged between 4 and 12. Specially constructed data trust agreements grant community members shared ownership over the conversation data set and control of data access. Honoring these agreements, 41 RTFC conversations were included in this analysis. Excerpts from these conversations are made available on the project’s public portal. Figure 1 provides a visualization of participants’ audio within a conversation and shows how main topics were distributed throughout the conversation.

News data *The Boston Globe* was selected for comparison to RTFC as the most widely-read local news publication in the Boston area. According to the *Globe*’s website, the average household income of the *Daily Globe* readers is $118,000 (over double the median household income of $55,777 for all of Boston (BPDA, 2019)), and 40% of *Daily Globe* readers have a net worth of over $1 million. In comparison, a 2015 study (Muñoz et al., 2015) reported that the median net worth of Black families, who make up around 20% of Boston’s population (Fatima, 2021), is just $8. These disparities make comparison between the BIPOC-focused conversations in RTFC and the *Globe*’s topic agenda potentially insightful. “Boston Metro” is *The Boston Globe*’s newspaper section with a local focus, described on its website as “the region’s latest local news and commentary from the State House to the cities, towns, and neighborhoods.” We consider the Boston Metro section as a subset of *Globe* coverage to see if it reflects local priorities to a greater or different degree than the general *Globe* coverage does.

In order to gather a data set of news articles, we first accessed all tweeted news stories from *The Boston Globe* and Boston Metro Twitter accounts between the beginning of August and the end of November, 2021. There were 9,822 unique stories posted by *The Boston Globe* account and 312 unique stories tweeted by the Boston Metro account during our period of interest, of which 297 were cross-posted on the main *Globe* account. We treat these 312 stories as a subset of *The Boston Globe*’s total coverage for this analysis. We retrieved the full text of the news stories and found that the average story length was 4,800 characters for *The Boston Globe* and 5,600 for stories in the Boston Metro section. Using Twitter as a method of story capture missed articles that were never tweeted, but it does capture all the news stories that the *Globe* decided to promote through its sizable Twitter audience (over 800,000 followers).

Figure 1: Visualization of a RTFC conversation with labeled topics, speakers, and speaker turn distributions.



Methods

Semi-supervised topic list creation

After experimenting with a variety of methods, we deployed a method of keyword set expansion using a hand-selected seed set of terms and a Word2Vec model (Mikolov et al., 2013). Here, we use a Word2Vec model that is trained on a third data source: a large, diverse corpus of talk radio data called RadioTalk (Beeferman, Brannon, and Roy, 2019). It is appropriate for this purpose because of its wide coverage of news topics, but it has shared qualities with spoken conversation due to the radio format. We query this word embedding model with a seed set of terms created around a hand-crafted topic, and it finds additional terms that are jointly similar to the seed set terms according to a cosine similarity measure applied to their vector representations.

Topic name	**Seed terms**
Education	special education, education, class size, IEPs, schools, public schools, school district

Table 1: Example topic and seed terms

Running query expansion over this seed set of education-related terms, the model helps us generate an additional 48 terms, including "teacher", "truancy", and "remote learning." We can then manually read and approve the suggested expansions. This was repeated for another several dozen seed topic lists. Some example word lists expanded using this method are seen in Table 2.

We finalize a total of 33 word lists. Code for the topic expansion method is available here.

For each article, we calculated how many times a term occurred within a document. We then summed counts within a topic across all documents to get a count at the publication level of topic coverage over the whole time period. For RTFC conversation transcripts, we treated speaker turns as "documents" and summed across all topic lists. For both news and conversation transcripts, we then normalized each absolute topic count by the sum of all keywords in the corpus to get a relative distribution of topics across all conversations or across all articles. We then compared coverage of each topic by relative percentage across the three data sets of interest.

Energy (n = 63)	**Business** (n = 53)	**Poverty** (n = 36)
'alternative energy', 'biofuel', 'biofuels', 'carbon neutral', . . . 'wind power', 'windmills'	'big business', 'black business', 'brokerage firm', . . . 'venture capitalists', 'wall st', 'wall street'	'abject poverty', 'child poverty', . . . 'underprivileged', 'uninsured', 'wealth inequality', 'welfare benefits'

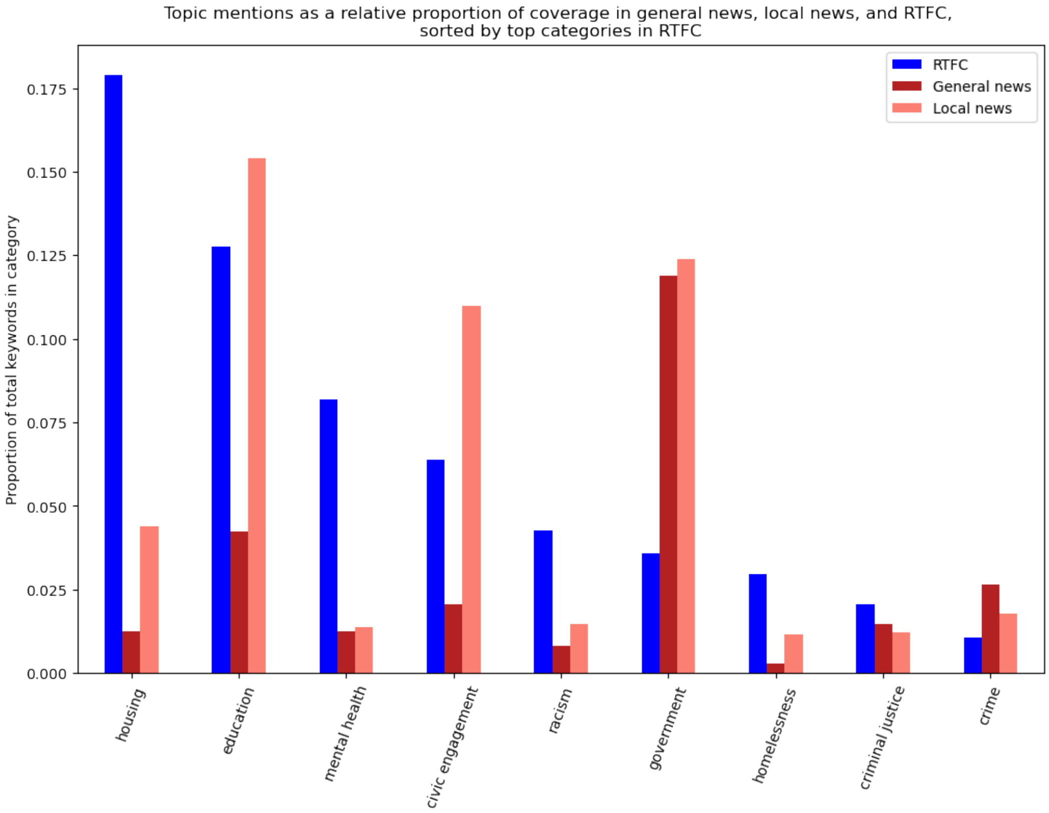
Table 2: Examples of expanded topic lists

Results

Figure 2 visualizes the distribution of top topics in RTFC across the corpora. Housing, homelessness, criminal justice, mental health, and racism are discussed more frequently in RTFC conversations compared to news. On the topics of education and civic engagement, Boston Metro coverage matches and even exceeds levels of community discussion in RTFC compared to coverage levels in *The Boston Globe*, indicating strong levels of local coverage by the Boston Metro section on these issues. Crime was covered more in the news compared to RTFC, but the issue of criminal justice in particular was covered more in RTFC compared to the news.

The emphasis on housing and homelessness in RTFC conversations replicates our team's qualitative observations of these community conversations to this point. Over 17% of keyword occurrences within RTFC across speaker turns come from the housing topic. Salient highlights of conversation pulled by our team on the public portal further illustrate the attentional focus on housing issues in Boston. Doug, a RTFC participant from Hyde Park, presented the following as his opening question: "Is there a place in Boston for me in the near future and people like me, black and brown folks? . . . I'm not even talking about owning. Can I even afford

Figure 2: Topic distribution



to rent here?" Questions similar to this one came up many times in the conversations, as captured here in our topic analysis.

Of the most frequently discussed topics in RTFC, which topics co-occur most regularly as community members discuss these issues? We compare topic co-occurrence at the speaker turn level in RTFC to an analysis on news at the article level, with each topic binarized for presence within the document. In RTFC, housing was the most frequent topic, and when it co-occurred with any topic, it did so with the civic engagement topic 15% of the time. An example "question" that covered both these topics in RTFC was "What is the next mayor going to do about the housing situation?" In the news, discussion of housing co-occurred with government 10% of the time and civic engagement 6% of the time.

Homelessness was a topic related to housing that came up more regularly in RTFC than news. In RTFC, discussions of homelessness co-occurred with housing discussions 31% of the time and with mental health 25% of the time. In the news, homelessness was most often concurrently covered with government topics (10% of the time), and was covered with mental health issues just 4% of the time. In RTFC, when mental health was mentioned with any other topic, it was with housing (17% of the time) and homelessness (15% of the time). In the news, mental health occurred with discussions of government 10% of the time and housing just 3% of the time. Taken together, these findings yield insight into the deep ways in which community members view housing issues as entwined with mental health issues, perhaps to a different degree than is portrayed in the news.

Insights from automated topic co-occurrence analysis can then guide us back to the RTFC community discussions for additional insight. We can surface exemplary quotes from the RTFC transcript to further illuminate how community members think about these issues. A data point from RTFC containing both the housing and mental health topics came from a community member who said the following in their own words: "My question about the future of Boston is: what will happen to our mentally ill residents who are homeless? Will there be more housing for this population? And the reason I got to ask this question is because that's me. I have mental health issues and I was homeless before, and I like to stop stigma... I'm very passionate about that community." Findings from topic analysis help us surface personal stories like this one from the RTFC corpus, creating an opportunity for public understanding of underrepresented issues as well as opportunities for journalists to follow stories that might not ordinarily be surfaced by traditional reporting practices.

For the most frequent topics in RTFC, we also analyzed which terms in the topic list drive the difference in frequency between RTFC and the news. Overall, we see differences in within-topic word frequency as largely indicative of differences of spoken conversation and written news, which we take as a signal that grouping these topic terms has happened at the right level of resolution. For example, the term "mental health" is by far the most common term in the RTFC corpus within the "mental health" topic list, and is 60% relatively more common as a term within the RTFC topic list than it is in the news. However, the terms "depression" and "anxiety" are relatively more common within this topic list in the news compared to RTFC. Depression and anxiety are specific terms which, taken together, are descriptors of "mental health" issues. As such, we interpret differences in relative distribution of terms within topics to indicate that the expanded topic list is doing its job by grouping such terms to construct a measure of the topic incidence across both RTFC and news.

Discussion

News serves a variety of purposes, including to report and inform on issues its audience cares about. Is *The Boston Globe* coverage reflecting the the issues that low-income BIPOC Bostonians represented in the RTFC corpus care most about?

Several findings are of note. First, we see a relative dearth of coverage on housing, homelessness, and mental health in the *Globe* compared to the major attentional focus on this issue in RTFC conversations. This may reflect a difference between the interests and concerns of RTFC participants, of which many were from underprivileged neighborhoods, and stories *The Boston Globe* covers as it bears its own audience in mind. We began this analysis expecting differences between RTFC and the *Globe*, at least in part due to baseline differences in the purpose of news coverage compared to open-ended community conversation within marginalized communities. However, documenting differences between the emergent RTFC agenda and the *Globe*'s agenda can bring out opportunities for increased journalistic coverage on issues that matter most to marginalized Bostonians. Future work could use similar topic analysis to see if local minority press like the BayState Banner more closely mirrors the local concerns of these Bostonians.

The topic analysis of RTFC conversations also appears to echo some findings of a quantitative study (Cox and Poepsel, 2020) of Hearken, a tool designed to support participant-driven journalism by eliciting feedback and topics of interest from members of a community in order to shape news coverage. The study conducted a manually coded topic analysis of participant-driven news stories that used Hearken and compared them to stories that did not use Hearken as part of the news cycle. Participant-driven stories covered "lifestyle issues and evergreen news" more often than traditional media did, an indication that people are hungry for news to reflect public experience back to them at a higher rate.

Open-ended community conversations like those in RTFC provide another such avenue to bring community priorities to light for the journalistic and policy communities. Better methods of automated topic analysis on the differences between local priorities and media coverage can help media organizations better notice gaps in coverage and tailor their coverage to issues of community interest.

Conclusion

Journalism has an obligation to portray reality by holding up a mirror to society (McQuail, 2013). One way of doing so is through the distillation of a community agenda in order to understand community priorities, needs, and interests. Using computational techniques to perform topic analysis of the RTFC corpus, we have revealed how these priorities do and do not align with the reporting agenda of *The Boston Globe*. We believe this initial exploration of the use of automatic topic analysis and community conversation points towards a powerful new method for newsrooms to listen to diverse communities and move towards more responsive journalism.

Acknowledgments

We thank the partner and community organizations involved in Cortico, LVN, and Real Talk for Change, and we are grateful to Ceasar McDowell, Dimitra Dimitrakopoulou, and Maggie Hughes for their valuable feedback.

References

Beeferman, D.; Brannon, W.; and Roy, D. 2019. Radiotalk: A large-scale corpus of talk radio transcripts. *arXiv preprint arXiv:1907.07073*.

BPDA. 2019. Neighborhood profiles. Technical report, Boston Planning Development Agency Research Division.

Brenan, M. 2021. Americans' trust in media dips to second lowest on record.

Chafe, W., and Tannen, D. 1987. The relation between written and spoken language. *Annual Review of Anthropology* 16:383–407.

Cox, J. B., and Poepsel, M. A. 2020. Deep participation in underserved communities: A quantitative analysis of hearken's model for engagement journalism. *Journalism Practice* 14(5):537–555.

Fatima, S. 2021. Boston is losing its black population, new census data show, even as it could elect its first black mayor.

McQuail, D. 2013. Journalism and society. *Journalism and Society* 1–256.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space.

Muñoz, A. P.; Kim, M.; Chang, M.; Jackson, R. O.; Hamilton, D.; and Jr., W. A. D. 2015. The color of wealth in boston. Technical report, Federal Reserve Bank of Boston.