Comparing conversational dialog data to local news

Keywords: local news, NLP, conversation data, participatory journalism

Extended 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, and as an input into public dialogues with the mayoral candidates. 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 in the US. 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.

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. 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 [2], and hand-crafted keyword lists are subject to availability and other biases of their creators. Keyword lists created for one text domain may also miss terms that are more common in another text domain, including colloquial terms like "cop" are common in spoken language, but less common than "police" in formal writing. With this in mind, we use word embedding models [3] trained on news [1]to expand seed sets of keywords and terms in order to improve computational topic capture in a lightweight, interpretable fashion that efficiently elicits topic insight across different types of text.

Using this method, we identify differences in the topic distributions of the two corpora: conversation transcripts and news. These differences reflect a mismatch between how much attention the city's largest news source gives to historically underheard residents and their expressed needs and concerns. Figure 1 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 to rent here?" Questions similar to this one came up many times in the conversations, as captured here in our topic analysis.

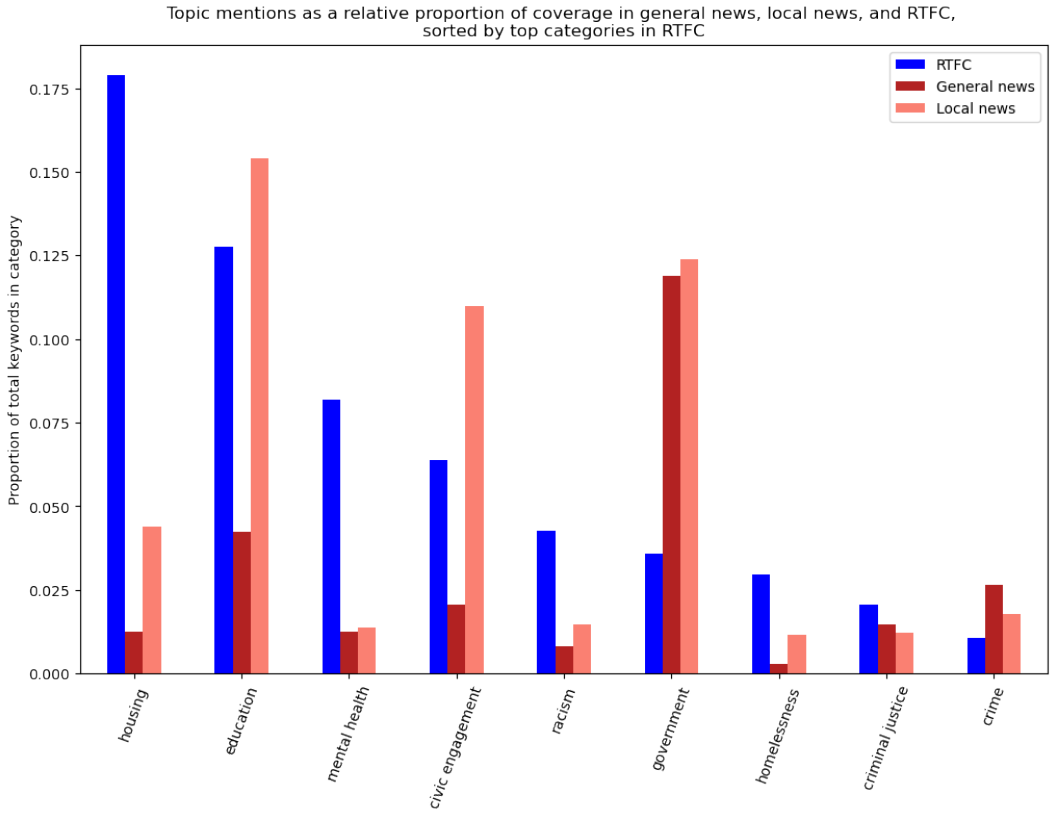
Of the most frequently discussed topics in RTFC, which topics co-occur most regularly as community members discuss these issues? We compare topic co-occurrence at the speaker turn level in RTFC to an analysis on news at the article level, with each topic binarized for presence within the document. In 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 sensemakers back to the RTFC community discussions for additional insight–indicating a potential feedback loop between quantiative insights from the text data and quotes from the stories that make journalism salient. Topic analysis surfaces exemplary quotes from the RTFC transcript to further illuminate how community members think about these issues, like 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 the described 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.

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, in addition to surfacing differences in what citizens care about and what the news covers. hearken Follow on work is analyzing how large language models can help with this kind of comparison across different kinds of text, and we are analyzing those capabilities and comparing them to the method presented here as ongoing work. We are also scaling analysis to a larger corpus of conversations, which will be analyzed using several sensemaking methods.

Figure 1: Topic distribution



References

[1] Doug Beeferman, William Brannon, and Deb Roy. "Radiotalk: A large-scale corpus of talk radio transcripts". In: *arXiv preprint arXiv:1907.07073* (2019).

[2] Wallace Chafe and Deborah Tannen. "The Relation between Written and Spoken Language". In: *Annual Review of Anthropology* 16 (1987), pp. 383–407. ISSN: 00846570. URL: http://www.jstor.org/stable/2155877 (visited on 05/09/2022).

[3] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/ARXIV.1301.3781. URL: https://arxiv.org/abs/1301.3781.

Figure 2: Visualization of a RTFC conversation with labeled topics, speakers, and speaker turn distributions.