



# A Comparison of Topic Modeling Approaches Using Networked Discussion Forum Posts From the City-data.com Corpus

RYAN M. OMIZO 

RESEARCH PAPER

 ubiquity press

## ABSTRACT

The [City-Data.com](#) Corpus provides over 15,000 discussion forum posts scraped from [city-data.com](#)--a website that hosts information about cities across the United States. Like the 20 Newsgroups dataset, the [City-Data.com](#) Corpus is weakly labeled by forum topics and thread titles and can be used to trial natural language processing techniques or be used to stage lessons in digital textual analysis in digital humanities pedagogy.

## CORRESPONDING AUTHOR:

**Ryan M. Omizo**

Department of English, Temple University, Philadelphia, Pennsylvania, United States

[tuk35906@temple.edu](mailto:tuk35906@temple.edu)

---

## KEYWORDS:

forum data; digital humanities; topic modeling; natural language processing

## TO CITE THIS ARTICLE:

Omizo, R. M. (2024). A Comparison of Topic Modeling Approaches Using Networked Discussion Forum Posts From the [City-data.com](#) Corpus. *Journal of Open Humanities Data*, 10: 16, pp. 1–12. DOI: <https://doi.org/10.5334/johd.182>

## (1) CONTEXT AND MOTIVATION

With the expansion of digital humanities and cultural analytics methodologies into textual analysis fields such as English literature and rhetoric and composition (Ridolfo & Hart-Davidson 2015), the need for data that supports computational text analysis pedagogy and projects has become an ongoing exigency for instructors. The City-Data Corpus addresses these needs by providing instructors and students with textual data that is amenable to a range of different text analysis methods and is well-suited for methodological experimentation. To dramatize the affordances of the City-Data.com Corpus for benchmarking novel computational text analysis techniques, I stage a comparison between a novel topic modeling method employing sentence embeddings and latent Dirichlet allocation (LDA) approaches.

For qualitative researchers, the City-Data.com Corpus provides opportunities for studying asynchronous online interactions about civic issues concerning the city of Philadelphia, PA. The Coronavirus discussion thread, for example, provides a snapshot of early reactions to the COVID-19 pandemic. The *Plan*, *Crime*, and *Retail* discussion threads, on the other hand, evidence sustained conversations over several years.

The City-Data.com Corpus can serve as the basis for digital humanities pedagogical lessons and experiments, including network modeling, timeseries analysis, processing HTML for text analysis, and the full gamut of computational text analysis techniques such as term weighting, word and sentence embedding, machine classification, named entity extraction, and topic modeling. The scope of the City-Data.com Forum data is small enough to be read fully, enabling analysts to reconcile human assessment with statistical results.

Lastly, the titles of discussion threads make the City-Datam.com Corpus amenable to classification tasks like Ken Lang's 20 Newsgroups Dataset. I leverage the weakly annotated forum posts to benchmark a novel topic modeling method that employs kmeans clustering and sentence embeddings against the widely used latent Dirichlet Allocation (LDA).

### (1.1) LIMITATIONS AND CAVEATS

City-Data.com forum rules prohibit flaming, hate speech, and trolling; however, posts do feature frank depictions of violence and potentially controversial statements that some may find offensive.

While the discussion board format of City-Data.com provides users with a direct means to quote previous posts when replying, this is not a strict requirement. Some replies attend to the content of previous posts with explicit attribution. Thus, some replies are not accessible by the presence of a "Quoted" field. Moreover, the quote\_body and quote content may not always present the full content of the original post. Users can be selective about their citations. Consequently, quote\_body and quote content may not always recover original posts. Lastly, because forum users can post across forum, quoted post content may have not accessed the originating page, thus, leaving gaps in quote\_body and quote content.

## (2) DATASET DESCRIPTION

The City-Data.com Corpus consists of five discussion forum threads from the site, city-data.com (Advameg, Inc., n.d.a) scraped using the BeautifulSoup Python library (Richardson 2007). City-Data.com aggregates geographic, demographic, and historical information about major cities across the United States and Canada from public and governmental sources. The site provides data visualizations that enable users to compare cities. As such, City-Data.com targets audiences who are traveling or relocating, real estate professionals, and advertisers. City-Data.com also hosts discussion forums by region and topicality (e.g., Classified Ads or Food and Drink). City-Data.com forums are moderated and prohibit trolling, hate speech, spam, doxing, and cross-posting (Advameg, Inc. n.d.b). Moreover, posters are advised to stay on topic and avoid personal attacks.

At the time of this writing, the City-Data.com forums house 2,940,053 threads, 62,355,814 posts authored by 2,476,620 members (Advameg, Inc., n.d.c). In comparison with other social media sites enlisted for data modeling, machine learning, and network analysis, City-Data Forums has a small footprint. Reddit, by contrast, boasts of over 13 billion posts and comments (Reddit Inc, 2023).

The [City-Data.com](#) Corpus contains 15,008 across five forum topics from the US > Pennsylvania > Philadelphia forums. [Table 1](#) presents forum title, post count, date ranges, and summary descriptions of the [City-Data.com](#) Corpus.

THREAD TOPIC TITLE	LABEL	SAMPLES	DATE RANGE	SUMMARY
How's everyone doing amongst the Coronavirus shut down? (home, movies) ( <a href="#">Advameg, Inc 2020</a> )	coronavirus	481	2020/03/16 – 2020/07/27	Posts discuss experiences with public shutdowns in the opening months of the COVID-19 pandemic in Philadelphia. Content dwells on governmental interventions and the status of public services than the etiology or effects of COVID-19.
“Official Greater Philadelphia Area Crime Thread” ( <a href="#">Advameg, Inc, 2013</a> )	crime	2,402	2013/04/11 – 2020/01/15	Posts discuss and share information about crime in Philadelphia.
Official Philadelphia Metro Crime Thread (York, Chester: apartment complexes, houses, unemployment) ( <a href="#">Advameg, Inc, 2012a</a> )	crime	1,284	2012/01/12 – 2013/03/11	Posts discuss and share information about crime in the Philadelphia metro area.
Philadelphia 2035 (Houston: foreclosure, neighborhoods, wage) ( <a href="#">Advameg, Inc. 2011</a> )	plan	6,796	2011/06/14 – 2020/01/14	Posts discussing Philadelphia's 2035 civic renovation plan authored by the Philadelphia City Planning Commission ( <a href="#">2023</a> ).
Retail coming to Philadelphia (Penn, Burlington: real estate, house, buying) ( <a href="#">Advameg, Inc. 2012b</a> )	retail	4,045	2012/11/27 – 2020/01/20	Posts discuss new retail business developments in Philadelphia.

**Table 1** City-Data Corpus post count, word count, and date range of postings per forum.

Post-level data captured includes the following fields (see [Table 2](#)):

POST_ID	POST_BODY	POST	DATETIME	QUOTE_ID	QUOTE_BODY	QUOTE	FORUM
Numerical post identifier	Post HTML	Post text	YYYY-MM-DD HH:MM:SS	Numerical post identifier	Quoted reply HTML	Quoted reply text.	forum title label

**Table 2** Tabular data model for [City-Data.com](#) forum posts.

## OBJECT NAME

[City-Data.com](#) Corpus

## FORMAT NAMES AND VERSIONS

CSV

## CREATION DATES

07/14/2022 – 08/14/2022

## DATASET CREATORS

Ryan M. Omizo (Temple University)

## LANGUAGE

English

## LICENSE

Creative Commons Attribution 4.0 International

## REPOSITORY NAME

Zenodo [10.5281/zenodo.10086354](https://zenodo.org/doi/10.5281/zenodo.10086354)

### (3) METHOD

Topic modeling is a method through which the latent structure of documents is inferred from lists of terms (topics) that collocate with high statistical significance. For example, latent Dirichlet allocation (LDA) (Blei et al. 2003; Steyvers & Griffiths 2007; Hoffman et al. 2010) creates a topic model by generating probability distributions over words. These probability distributions designate the word features that would most likely generate the documents in under the model. While some topic modeling procedures like latent semantic indexing (Deerwester et al. 1990) do not produce human-interpretable topic models, most current topic modeling procedures like LDA produce term lists that order the most influential features of a topic. There are general limitations to LDA, however. As Vayansky and Kumar (2020) note in their review of topic modeling methods, LDA performance suffers when applied to short texts. Moreover, bag-of-words document representations used in many traditional LDA approaches are less informative than more advanced embedding-based representations, which can capture the dense relationships between words in context. For this reason, embedding-based approaches have been explored (Bhatia et al. 2016; Grootendorst 2022; Angelov 2020; Aharoni & Goldberg 2020; Bianchi et al. 2021a; Bianchi et al. 2021b; Zhang et al. 2022; Limwattana & Prom-on 2021). Unlike bag-of-words approaches, which will capture the presence or absence of term in documents, sentence embeddings encode word contexts derived from large language models pretrained on vast corpora. The richness of sentence embedding representations recommend their incorporation in topic modeling approaches. The density of information should lead to more nuance when grouping documents into topics, which should lead to more relevant topical term lists. My method for performing sentence embedding-based topic models (henceforth, SE-Topics) follows Grootendorst's (2022) pipeline that involves transforming texts into sentence embeddings, clustering sentence embeddings, and extracting representative word lists or topics. To group the thread embeddings, I use scikit-learn's (Pedregosa et al. 2011) kmeans clustering implementation.

To extract representative words, I adapt Grootendorst's (2022) "class based TFIDF" approach. Posts assigned to the same cluster are merged into a single document. However, prior work with these approaches indicates that frequency counts lead to better topical quality than TFIDF for SE-Topic and LDA topic modeling. This topic modeling approach emphasizes what Grootendorst (2022) describes as pipeline "modularity." Although the step to derive significant textual features is separate from the embedding and clustering steps, this modularity enables people to replace or extend a phase in the topic modeling process. For example, dimensionality reduction can be applied to sentence embeddings to accelerate clustering; clustering algorithms can be exchanged (e.g., spectral or density-based clustering can be used); frequency distributions can be replaced with TFIDF vectorization when deriving term topics.<sup>1</sup>

#### (3.1) EXPERIMENTAL TRIALS

I conduct 16 topic modeling trials designed to leverage the networked structure of the City-Data.com Corpus. I evaluate topic quality of models trained on post-level segments, thread-level segments, and topics guided by prior information (see Li et al. 2018; El-Assady et al. 2019; Popa and Rebedea 2021; Gourru et al. 2018). Guided topic modeling uses prior information about the data to center model priorities. Post and thread-level topic modeling test how the unitization of City-Data.com's networked content influences topics modeling and is primarily a data preparation step. Guided topic modeling intervenes in the topic modeling process. To guide the kmeans clustering process (the basis of the sentence embedding-based topic model),

---

<sup>1</sup> The embedding-based topic modeling approach illustrated here departs from Grootendorst's (2022) BERTopic significantly. First, kmeans clustering used to generate SE-Topics produce flat clusters; BERTopic's default implementation uses hierarchical density-based clustering (HDBSCAN, McInnes et al., 2017). HDBSCAN forms clusters around points of density in semantic space. Data not proximal enough to these points of density are labeled as outliers. The lack of outliers among SE-Topics produces a topic model more comparable to LDA. Moreover, at the time of this writing, BERTopic's "guided topic modeling" (equivalent to the use of topical priors) was inoperable due to dependency issues. Thus, BERTopic could not offer results comparable to LDA or the Sentence Embedding-Based routine tested in this study. Despite this offset, I include BERTopic's modeling of City-Data.com Corpus Threads and Posts with discussion in Appendix B.

I manually designate initial cluster centers. This initial positioning will guide subsequent re-centerings as the model converges to reduce the distance between intra-cluster datapoints (Arthur & Vassilvitskii 2007). To guide the LDA topic modeling, I adapt Li et al.'s (2018) method of injecting seed words extracted from the data into the topic modeling process.<sup>1</sup> In one experiment conducted on the 20 Newsgroups dataset, Li et al. (2018) used forum label text (e.g., *talk.politics.guns*) as seeds with the assumption that forum labels provide distinguishing categorical information about potential topics.

I test three types of topic seeds:

1. Topic titles – Following Li et al. (2018), I use thread topic titles as seeds (see Table 1).
2. Initial forum posts – Initiating posts declare the horizons of participating in forum discussions. Empirical work by Sobkowicz and Sobkowicz (2010), See Jagarlamundi et al. 2012 for example, demonstrates that social medial discussions evidence strong first mover advantage similar to scientific papers (Newman 2009). Papers that appear early in the rise of a discipline will outpace the citation rate of newer papers. Here, the hypothesis is that posts that appear first gain more engagement and this increased engagement will condition the content of a sizable portion of the thread. (see Table 7 in Appendix A).<sup>2</sup>
3. Posts with the highest degree – Leveraging the network properties of the City-Data.com corpus, I create a graph of each forum and extract posts with the most incoming and outgoing links or node degree (Gerlach et al. 2018; Duan et al. 2021; Yang et al. 2016). Posts that receive numerous quoted replies and/or are replying to other posts serve as proxies for engagement. Like the intuition behind the use of initial forum posts, posts that are bound up in more extensive conversations, condition more content because respondents must stay on topic to sustain discourse (see Table 8 in Appendix A).

I calculate topical coherence and diversity scores to measure model quality. Topical coherence includes several measures that indicate how well topical term lists reflect the underlying data. In this study, I use Mimno et al.'s (2011; see also Hinneburg et al. 2014) UMASS method. UMASS coherence measures the probability that co-occurring words in the topical term list occur in documents divided by the total number of documents. Unlike other coherence measures such as UCI-coherence, which compare topical terms to a large reference corpus like Wikipedia dumps, UMASS coherence is derived from the original dataset. I balance coherence scores against topical diversity scores (Mimno et al. 2011). I employ Gensim's coherence measures (Řehůřek, R., & Sojka 2011) to calculate UMASS coherence. Topical diversity refers to a family of metrics that indicate the variability of topical terms, thus, the range of data explainable by topical term lists. Quality topics are both coherently related to the underlying data, but also distinctive enough to offer thorough faceting (Dieng et al. 2020). For this study, I employ four diversity measures (Terragni 2023):

1. Proportion of unique words (PUW) – PUW determines the ratio of unique topical terms for all topics (Dieng et al. 2020). Scores closer to 1 indicate diverse topics; scores closer to 0 indicate repetitive topical terms.
2. Jaccard Distance between topical term lists (JD) – Proposed by Tran et al. (2013), this diversity measure evaluates the Jaccard distance between topical term lists. Greater distances between topical terms indicate more topical diversity (Terragini 2023).
3. Word embedding centroid distance (WE-CD) (Bianchi et al. 2020b) – WE-CD calculates the distance between collocations in the topical term list to a reference corpus of word embedding. This metric determines how diverse topical term lists are in comparison to generalized usage in embedding models of large volumes of texts like Wikipedia or the Common Crawl Corpus of internet sites. For this paper, I use the FastText Common Crawl word embedding model with 300 dimensions and 2 million subword vectors (Bojanowski et al. 2016; Joulin et al. 2016a; Joulin et al. 2016b).<sup>3</sup>

In all, topical diversity measures used in this study determine intra-topical term list diversity, inter-topical term list diversity, and generalized topical term list diversity when compared to a reference corpus. To calculate each measure, I use the top-25 terms per topic.

---

<sup>2</sup> Due to the length of some posts, I provide the post\_id in lieu of the full text.

<sup>3</sup> I employ Terragni's (2023) suite of diversity scripts to measure PUW, JD, and WE-CD. See also Röder (2015a; 2015b), Stevens et al. (2012), Terragni et al. (2021).

Because the Metro forum also discusses criminal activity, I set the topic number at 4 to correspond to the following thematic categories for topic modeling: *coronavirus*, *crime*, *plan*, and *retail*.<sup>4</sup>

Tables 3 and 4 illustrate the SE-Topics and LDA modeling coherence and diversity scores across different data segmentations.

TOPIC MODEL TYPE	UMASS	PUW	JD	WE-CD
SE-Topics post	-4.10	0.74	0.85	0.06
SE-Topics guided topic titles	-2.76	0.54	0.63	0.06
SE-Topics guided initial posts	-2.56	0.52	0.62	0.07
SE-Topics guided high degree	-2.53	0.54	0.63	0.06
SE-Topics threads	-5.46	0.74	0.86	0.12
SE-Topics guided threads topic titles	-4.84	0.69	0.79	0.06
SE-Topics guided threads initial posts	-4.57	0.67	0.78	0.21
SE-Topics guided threads high degree	-4.26	0.67	0.78	0.21
<b>MEAN</b>	-3.88	0.63	0.74	0.11
<b>MAX</b>	-2.53	0.74	0.85	0.21
<b>Q3</b>	-2.66	0.71	0.82	0.17
<b>MEDIAN</b>	-4.18	0.67	0.78	0.07
<b>Q1</b>	-4.70	0.54	0.62	0.06
<b>MIN</b>	-5.46	0.52	0.62	0.06

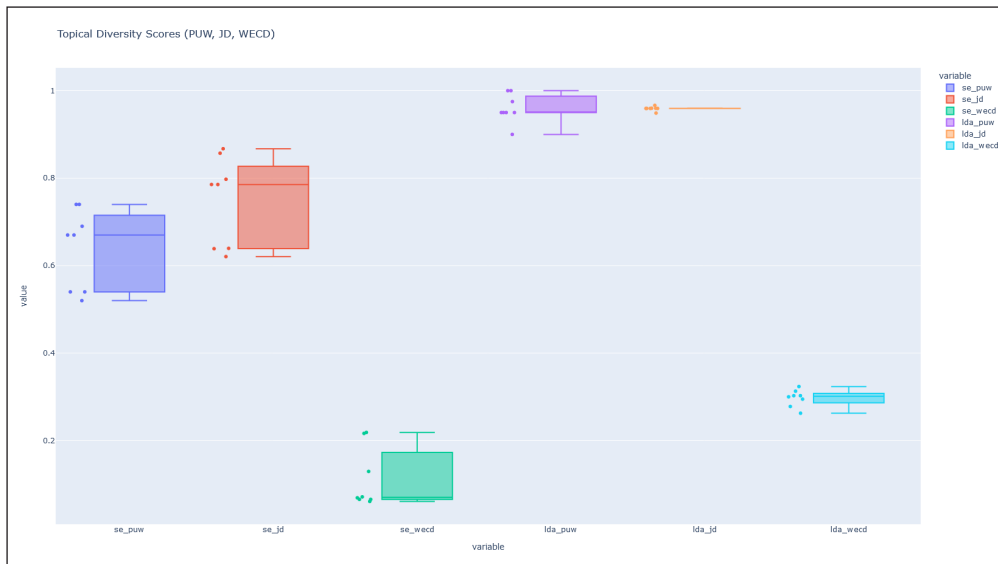
**Table 3** SE-Topics Coherence and Diversity Scores.

TOPIC MODEL TYPE	UMASS	PUW	JD	WE-CD
LDA posts	-2.59	1.000	0.95	0.30
guided LDA (topic titles)	-2.63	1.000	0.95	0.27
guided LDA (high degree)	-2.10	0.950	0.95	0.29
guided LDA (initial posts)	-2.36	0.975	0.96	0.26
LDA threads	-2.10	0.950	0.95	0.30
LDA threads (high degree)	-2.15	0.950	0.95	0.31
LDA threads (topic titles)	-2.32	0.950	0.95	0.30
LDA threads (initial post)	-2.22	0.900	0.94	0.32
<b>MEAN</b>	-2.31	0.95	0.95	0.29
<b>MAX</b>	-2.10	1.00	0.96	0.32
<b>Q3</b>	-2.12	0.98	0.95	0.30
<b>MEDIAN</b>	-2.27	0.95	0.95	0.30
<b>Q1</b>	-2.47	NaN	0.95	0.28
<b>MIN</b>	-2.63	0.9	0.94	0.26

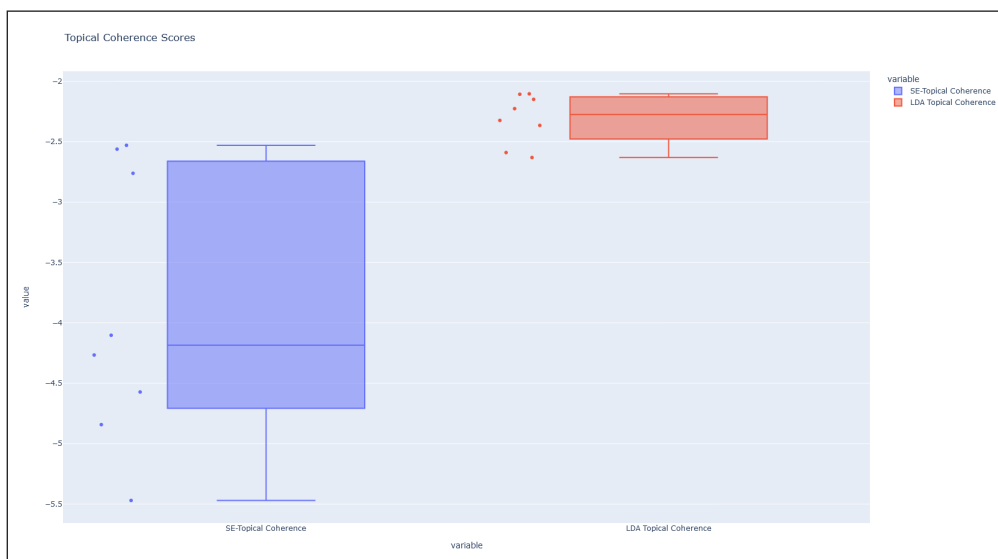
**Table 4** LDA Topic Modeling Coherence and Diversity Scores.

In general, LDA topic modeling produced better topical coherence and diversity scores compared to sentence embedding approaches with an average UMass coherence of -2.31 and PUW and JD diversity scores near 1.0. Mean WE-CD for LDA topic models is also 0.12 points greater, signifying that LDA topic models are more semantically distant from the reference corpus than SE-Topics with high coherence. Box and whisker plots of topical coherence and diversity scores (see Figures 1 and 2) also indicate that LDA topic modeling results are more consistent across segment types. SE-Topical coherence and diversity scores indicate a wider

<sup>4</sup> Text processing, topic modeling, and evaluations scripts available at <https://github.com/rotemple/city-data-com-corpus-scripts/blob/main/sentence-embedding-tm.ipynb>.



**Figure 1** Boxplots of SE-Topics and LDA Coherence Scores.



**Figure 2** Boxplots of SE-Topics and LDA Diversity Scores.

interquartile range between the best performing segments (high degree posts) and the lowest performing segments with a difference of 2.93. On the other hand, there is only a 0.53 difference in coherence scores between the best performing LDA topics (high degree posts and threads) and the worst performing topic model (topic titles).

Put another way, the best scoring SE-Topics—guided initial posts and guided high degree —perform as well as the two worst scoring LDA topic models.

The discrepancies in coherence and diversity scores, however, are complicated by qualitative assessments.

Comparing SE-Topics and LDA topics guided by high degree posts, we can discern close similarities among term collocations within individual topics. Both SE-Topics and LDA Topics represent the Retail thread and the development rhetoric of the Plan thread; however, SE-Topics have agglomerated crime and coronavirus discourse (see Topic 0 in Table 5). The LDA topic model seeded by high degree posts (see Table 6), however, has captured discussions

**Table 5** SE-Topics Guided Post (high degree nodes).

TOPIC	1	2	3	4	5	6	7	8	9	10
0	people	white	crime	black	time	year	murder	think	city	police
1	city	philly	philadelphia	people	street	area	year	think	center	neighborhood
2	building	city	street	project	tower	think	center	market	development	broad
3	store	retail	city	mall	retailer	market	walnut	center	think	location



TOPIC	1	2	3	4	5	6	7	8	9	10
0	money	state	local	people	issue	care	neighborhood	income	help	white
1	store	retail	mall	shopping	retailer	location	center	shop	gallery	also
2	city	think	philadelphia	philly	people	year	time	even	area	much
3	street	market	building	walnut	space	chestnut	center	east	block	south

about government restriction of public services that ensued in the first months of the 2020 COVID-19 pandemic in Philadelphia with terms such as *money*, *state*, *local*, *care*, and *help*.

**Table 6** LDA (High Degree) Topic Model.

Thread segmentation modestly improves LDA topical coherence (mean  $-2.2$  compared to mean  $-2.4$ ). Coherence scores for SE-Topics, on the other hand, worsen with thread segmentation (mean  $-4.41$  compared to mean  $-3.49$ ).

SE-Topics models benefited the most from the injection of prior information. The results of the SE-Topics informed by the embeddings bundled with the most quoted replies are particularly interesting when compared with the generally poor performance of thread-based segments. Both permutations depend on the networked structure of the [City-Data.com](#) Corpus to structure topics. Threads utilize the chained structure of posts and quoted replies to crystalize texts that emerge through interaction; high degree node priors utilize the linkages of posts with other posts. Although more information is encoded when threads are joined into single texts, the centrality of posts within conversations lends better guidance to the formation of more cohesive topics. This makes intuitive sense in that posts with numerous linkages will anchor common content. Threads, on the other hand, are more intrinsically diverse segments, striated by conversational turns and/or dissensus among interlocutors.

Guided topic modeling returned better topical coherence and diversity scores than unguided models (with the sentence embedding model guided by initial posts the exception). Posts with the most incoming and outgoing linkages (degree) produced the best scoring topic models. This finding suggests that network structure can influence the development of topical content in extended asynchronous conversations. Messages enmeshed in replies are more likely conserved as interlocutors attend to given information as they add commentary.

#### (4) REUSE POTENTIAL

In this paper, I have explored the potential of sentence embedding-based topic modeling against LDA benchmarks. LDA approaches yielded better topical coherence and diversity scores in comparison to sentence embeddings. Inspection of topical term lists suggests that qualitative distances between methods are less pronounced, though (See Appendix B for a truncated comparison to BERTopic ([Grootendorst 2022](#))).

That said, final topical coherence and diversity scores are less important than the different topical permutations that the [City-Data.com](#) Corpus allowed us to test. The networked structure of the [City-Data.com](#) Corpus enables the unitization of data into posts or threads as well as principled means to designate topical priors based on posts with the highest engagement. Consequently, the [City-Data.com](#) Corpus is conducive to evaluating the strength text analysis algorithms as well as aspects of research design such as text segmentation and the incorporation of topical guidance. Illustrating the effects of different modelling parameters can be a boon to data-driven pedagogies because students can witness how different data selection choices impact their topic modeling results.

Along these pedagogical lines, the [City-Data.com](#) Corpus provides opportunities for students to practice other data processing techniques such as scraping, cleaning, and parsing HTML data as well as other methods that rely upon labeled data such as machine classification.



## APPENDIX A. INITIAL FORUM POSTS AND HIGH DEGREE POST IDS

FORUM	POST ID
coronavirus	758124
crime	908813
metro	251839
plan	958364
retail	710692

**Table 7** Initial post ids per City-Data.com Corpus forum.

FORUM	POST ID	DEGREE
coronavirus	57681616	5
crime	50175634	5
metro	22518391	5
plan	38332691	5
retail	38055575	5

**Table 8** High degree City-Data.com Corpus forum posts used for guided topic modeling.

## APPENDIX B. BERTOPIC RESULTS ON CITY-DATA.COM CORPUS POSTS AND THREADS

For completeness, I trialed BERTopic's topic modeler on City-Data.com Corpus posts and threads. Because BERTopic assigns isolated datapoints to an outlier cluster, I generate 5 topics for each trial to yield at least 4 cohesive topics. Topics were required to store a minimum of 50 textual units (posts or threads).

Table 9 illustrates the coherence and diversity scores of BERTopic models of posts and thread units. Both BERTopic models yield UMass coherence and diversity scores comparable to SE-Topics method. Qualitatively, the BERTopic Post Model produces the most legible topics. The Topic -1 (the outlier topics) is characterized by general references to *philadelphia* and *people* (see Table 10). Topic 0 captures references to city infrastructure found in the plan forum. Topic 1 includes references to race and crime (*white*, *crime*, *murders*) indicative of the *crime* and *metro* forums.

TOPIC MODEL TYPE	UMASS	PUW	JD	WE-CD
BERTopic posts	-5.12	0.52	0.78	0.12
BERTopic threads	-5.59	0.49	0.79	0.28

**Table 9** BERTopic Coherence and Diversity Scores for City-Data.com Corpus.

TOPIC	0	1	2	3	4	5	6	7	8	9
-1	city	like	people	just	philly	philadelphia	think	dont	street	center
0	city	like	think	new	just	people	building	philadelphia	dont	philly
1	white	people	city	dont	like	crime	im	year	just	murders
2	like	inga	news	just	dont	people	article	think	does	thread
3	bau	hello	nasty	bart	fancy	update	haha	awful	ok	finally

**Table 10** BERTopic Post Topic Model. Note that Topic -1 indicates outliers.

The BERTopic Thread Model is similar to the Post Model, although it more clearly features references to terms in the *retail* forum such as *retail*, *store*, and *stores*. Both BERTopic Post and Thread Topic Models feature low quality topics (Topics 3 and 2, respectively). Neither conveys interpretable information about the content of the forums (see Tables 10 and 11), indexing two short posts in the corpus.

However, I would not argue that the SE-Topic Modeling method discussed above is a clear improvement over BERTopic. These results do not represent the full range of BERTopic parameters but are useful to highlight how little tuning the SE-Topic method requires to produce legible topics that yield similar coherence and diversity scores.

0	1	2	3	4	5	6	7	8	9	
-1	city	like	people	philly	just	think	dont	philadelphia	new	street
0	city	like	new	store	think	just	philadelphia	retail	stores	people
1	people	crime	city	dont	just	like	white	im	know	year
2	hmmm	hello	fancy	update	haha					
3	inga	like	just	news	article	dont	read	writing	people	does

**Table 11** BERTopic Thread Level Topic Model. Topic -1 indicates outliers.

## FUNDING INFORMATION

Ryan Omizo’s work on this project received no special funding.


## COMPETING INTERESTS

The author has no competing interests to declare.

## AUTHOR CONTRIBUTIONS

- Conceptualization
- Data curation
- Formal Analysis
- Investigation
- Methodology
- Project administration
- Resources
- Software
- Visualization
- Writing – original draft
- Writing – review & editing

## AUTHOR AFFILIATIONS

**Ryan M. Omizo**  [orcid.org/0000-0002-2796-6281](https://orcid.org/0000-0002-2796-6281)  
 Department of English, Temple University, Philadelphia, Pennsylvania, United States

## REFERENCES

- Arthur, D., & Vassilvitskii, S.** (2007). K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035).
- Advameg, Inc.** (2011). Philadelphia 2035 (Houston: Foreclosure, neighborhoods, wage)—Pennsylvania (PA)—City-Data Forum. *City-Data.Com*. <https://www.city-data.com/forum/philadelphia/1304227-philadelphia-2035-a.html>
- Advameg, Inc.** (2012a). Official Philadelphia Metro Crime Thread (York, Chester: Apartment complexes, houses, unemployment)—Pennsylvania (PA)—Page 10—City-Data Forum [Forum]. *City-Data.Com*. <http://www.city-data.com/forum/philadelphia/1470248-official-philadelphia-metro-crime-thread-10.html>
- Advameg, Inc.** (2012b). Retail coming to Philadelphia (Economy, Penn: 2013, tenant, shop)—Pennsylvania (PA)—Page 3—City-Data Forum [Forum]. *City-Data.Com*. <https://www.city-data.com/forum/philadelphia/1740992-retail-coming-philadelphia-3.html>
- Advameg, Inc.** (2013). Official Greater Philadelphia Area Crime Thread (York, Mars: Leasing, condominium, place to live)—Pennsylvania (PA)—Page 267—City-Data Forum [Forum]. *City-Data.Com*. <https://www.city-data.com/forum/philadelphia/1839911-official-greater-philadelphia-area-crime-thread-267.html>

- Advameg, Inc.** (2020). How's everyone doing amongst the Coronavirus shut down? (Philadelphia, York: Restaurants, bus)—Pennsylvania (PA)—Page 37—City-Data Forum [Forum]. *City-Data.Com*. <https://www.city-data.com/forum/philadelphia/3137059-hows-everyone-doing-amongst-coronavirus-shut-37.html>
- Advameg, Inc.** (n.d.a). *City-Data.com*—Stats about all US cities—Real estate, relocation info, crime, house prices, cost of living, races, home value estimator, recent sales, income, photos, schools, maps, weather, neighborhoods, and more. Retrieved January 26, 2024, from <https://www.city-data.com/>
- Advameg, Inc.** (n.d.b). *City-Data.com* Forum: Relocation, Moving, General and Local City Discussions. Retrieved January 26, 2024, from <https://www.city-data.com/forum/>
- Advameg, Inc.** (n.d.c). Terms of Service—City-Data Forum. Retrieved October 22, 2023, from <https://www.city-data.com/forumtos.html>
- Aharoni, R., & Goldberg, Y.** (2020). Unsupervised Domain Clusters in Pretrained Language Models (arXiv:2004.02105), Cornell University, arXiv. <http://arxiv.org/abs/2004.02105>. DOI: <https://doi.org/10.18653/v1/2020.acl-main.692>
- Angelov, D.** (2020). Top2Vec: Distributed Representations of Topics.
- Bhatia, S., Lau, J. H., & Baldwin, T.** (2016). Automatic Labeling of Topics with Neural Embeddings. DOI: <https://doi.org/10.48550/arXiv.1612.05340>
- Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E.** (2020a). Cross-lingual Contextualized Topic Models with Zero-shot Learning (arXiv:2004.07737). arXiv. <http://arxiv.org/abs/2004.07737>. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.143>
- Bianchi, F., Terragni, S., & Hovy, D.** (2020b). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. arXiv preprint arXiv:2004.03974. DOI: <https://doi.org/10.18653/v1/2021.acl-short.96>
- Blei, D. M., Ng, A., & Jordan, M.** (2003). Latent dirichlet allocation. (Jan), 993–1022.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T.** (2016). Enriching Word Vectors with Subword Information. arXiv Preprint arXiv:1607.04606. DOI: [https://doi.org/10.1162/tacl\\_a\\_00051](https://doi.org/10.1162/tacl_a_00051)
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R.** (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391–407. <https://www.cs.csubatn.edu/~mmartin/LDS/Deerwester-et-al.pdf>. DOI: [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9)
- Dieng, A. B., Ruiz, F. J. R., & Blei, D. M.** (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453. DOI: [https://doi.org/10.1162/tacl\\_a\\_00325](https://doi.org/10.1162/tacl_a_00325)
- Duan, Z., Xu, Y., Chen, B., Wang, D., Wang, C., & Zhou, M.** (2021). TopicNet: Semantic Graph-Guided Topic Discovery (arXiv:2110.14286). arXiv. DOI: <https://doi.org/10.48550/arXiv.2110.14286>
- El-Assady, M., Kehlbeck, R., Collins, C., Keim, D., & Deussen, O.** (2019). Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections (arXiv:1908.00475). arXiv. <http://arxiv.org/abs/1908.00475>. DOI: <https://doi.org/10.1109/TVCG.2019.2934654>
- Gerlach, M., Peixoto, T. P., & Altmann, E. G.** (2018). A network approach to topic models. *Science Advances*, 4(7), eaq1360. DOI: <https://doi.org/10.1126/sciadv.aq1360>
- Gourru, A., Velcin, J., Roche, M., Gravier, C., & Poncelet, P.** (2018). United We Stand: Using Multiple Strategies for Topic Labeling. In M. Silberztein, F. Atigui, E. Kornysheva, E. Métails, & F. Meziane (Eds.), *Natural Language Processing and Information Systems* (Vol. 10859, pp. 352–363). Springer International Publishing. DOI: [https://doi.org/10.1007/978-3-319-91947-8\\_37](https://doi.org/10.1007/978-3-319-91947-8_37)
- Grootendorst, M.** (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint arXiv:2203.05794. DOI: <https://doi.org/10.48550/arXiv.2203.05794>
- Hinneburg, A., Rosner, F., Pessler, S., & Oberländer, C.** (2014, November). Exploring document collections with topic frames. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 2084–2086). DOI: <https://doi.org/10.1145/2661829.2661857>
- Hoffman, M., Bach, F., & Blei, D.** (2010). Online learning for latent dirichlet allocation. *Advances. Neural information processing systems*, 23. URL: [https://papers.nips.cc/paper\\_files/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf](https://papers.nips.cc/paper_files/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf)
- Jagarlamudi, J., Iii, H. D., & Udupa, R.** (2012). Incorporating Lexical Priors into Topic Models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp.204–213). Association for Computational Linguistics URL. <https://aclanthology.org/E12-1021>.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T.** (2016a). FastText.zip: Compressing text classification models. arXiv Preprint arXiv:1612.03651. DOI: <https://doi.org/10.48550/arXiv.1612.03651>
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T.** (2016b). Bag of Tricks for Efficient Text Classification. arXiv Preprint arXiv:1607.01759. DOI: <https://doi.org/10.48550/arXiv.1607.01759>; <https://doi.org/10.18653/v1/E17-2068>
- Li, C., Chen, S., Xing, J., Sun, A., & Ma, Z.** (2018). Seed-Guided Topic Model for Document Filtering and Classification. *ACM Transactions on Information Systems*, 37(1), 9:1–9:37. DOI: <https://doi.org/10.1145/3238250>

- Limwattana, S., & Prom-on, S.** (2021). Topic Modeling Enhancement using Word Embeddings. *18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–5. DOI: <https://doi.org/10.1109/JCSSE53117.2021.9493816>
- McInnes, L., Healy, J., & Astels, S.** (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11), 205. DOI: <https://doi.org/10.21105/joss.00205>
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A.** (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp.262–272) Association for Computational Linguistics URL: <https://aclanthology.org/D11-1024>
- Newman, M. E.** (2009). The first-mover advantage in scientific publication. *Europhysics Letters*, 86(6), 68001. DOI: <https://doi.org/10.1209/0295-5075/86/68001>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. DOI: <https://doi.org/10.48550/arXiv.1201.0490>
- Philadelphia City Planning Commission.** (2023). About | Philadelphia2035. <https://www.phila2035.org/>.
- Popa, C., & Rebedea, T.** (2021). BART-TL: Weakly-Supervised Topic Label Generation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1418–1425. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.121>
- Reddit Inc.** (2023). Homepage—Reddit. <https://www.redditinc.com/>.
- Řehůřek, R., & Sojka, P.** (2011). Gensim—statistical semantics in python. Retrieved from [genism.org](http://genism.org). URL: <https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf>
- Richardson, L.** (2007). Beautiful soup documentation.
- Ridolfo, J., & In Hart-Davidson, W.** (2015). *Rhetoric and the digital humanities*. University of Chicago Press. DOI: <https://doi.org/10.7208/chicago/9780226176727.001.0001>
- Röder, M., Both, A., & Hinneburg, A.** (2015a). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. DOI: <https://doi.org/10.1145/2684822.2685324>
- Röder, M., Both, A., & Hinneburg, A.** (2015b). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. DOI: <https://doi.org/10.1145/2684822.2685324>
- Sobkowicz, P., & Sobkowicz, A.** (2010). Dynamics of hate based networks. *The European Physical Journal B*, 73(4), 633–643. DOI: <https://doi.org/10.1140/epjb/e2010-00039-0>
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D.** (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. URL: <https://aclanthology.org/D12-1087>
- Steyvers, M., & Griffiths, T.** (2007). Probabilistic topic models. In Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, & Walter Kintsch (Eds.), *Handbook of latent semantic analysis*, 427(7), (pp. 424–440)
- Terragni, S.** (2023). A collection of Topic Diversity measures for topic modeling. [Python]. <https://github.com/silviatti/topic-model-diversity> (Original work published 2020).
- Terragni, S., Fersini, E., & Messina, E.** (2021, June). Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems* (pp. 33–45). Cham: Springer International Publishing. DOI: [https://doi.org/10.1007/978-3-030-80599-9\\_4](https://doi.org/10.1007/978-3-030-80599-9_4)
- Tran, N. K., Zerr, S., Bischoff, K., Niederée, C., & Krestel, R.** (2013). Topic Cropping: Leveraging Latent Topics for the Analysis of Small Corpora. In T. Aalberg, C. Papatheodorou, M. Dobrev, G. Tsakonias, & C. J. Farrugia (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 297–308). Springer. DOI: [https://doi.org/10.1007/978-3-642-40501-3\\_30](https://doi.org/10.1007/978-3-642-40501-3_30)
- Vayansky, I., & Kumar, S. A.** (2020). A review of topic modeling methods. *Information Systems*, 94. DOI: <https://doi.org/10.1016/j.is.2020.101582>
- Yang, W., Boyd-Graber, J., & Resnik, P.** (2016). A Discriminative Topic Model using Document Network Structure. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 686–696. DOI: <https://doi.org/10.18653/v1/P16-1065>
- Zhang, Z., Fang, M., Chen, L., & Namazi-Rad, M.-R.** (2022). Is Neural Topic modeling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics (arXiv:2204.09874). arXiv. DOI: <https://doi.org/10.48550/arXiv.2204.09874>; <https://doi.org/10.18653/v1/2022.naacl-main.285>

**TO CITE THIS ARTICLE:**

Omizo, R. M. (2024). A Comparison of Topic Modeling Approaches Using Networked Discussion Forum Posts From the *City-data.com* Corpus. *Journal of Open Humanities Data*, 10: 16, pp. 1–12. DOI: <https://doi.org/10.5334/johd.182>

**Submitted:** 10 November 2023

**Accepted:** 10 January 2024

**Published:** 07 February 2024

**COPYRIGHT:**

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.