Polly McKim & Matt Fein

Extra! Extra! Read All Articles About It! Topic Modeling & Clustering for News Aggregation

Abstract:

Now more than ever, people rely on written news to remain informed about important world issues. However, multiple sources routinely report on the same subject or event; and, due to word choices and semantic differences, coverage varies between media sources. The ability to aggregate news coverage about the same event across different sources would be beneficial to multiple sectors, and has important use cases in politics, business, and media consumption. By testing a modified form of BertTopic and LDA modeling combined with the pyLDAvis visualization library, this project explored which method works best toward achieving the overall goal of automated news aggregation. Both methods retrieved interpretable results, but each method also had separate drawbacks. Overall, BERTtopic modeling worked better for the stated objectives, while the pyLDAvis library enabled more efficient analysis of results. Additional research and refinement is needed to create a model that can detect and create hyper localized clustering of the same specific news event in a reliable manner.

Introduction:

In the modern information age written news content comes to readers at rates never before seen in history. However, media outlets often cover the same event in different ways. Automatic detection of articles that pertain to the same event across new sources would be especially useful in today's society where misinformation and news bias has helped foster a deep sense of cultural division. By using natural language processing techniques, such as topic modeling and clustering, to identify articles that cover the same event, this project explores the feasibility of this idea and works towards a fully automated process.

Business Case:

Automated detection and collection of event coverage across different media sources provides benefits in many contexts, because it nullifies the need to assemble and read massive amounts of information to get a full understanding of a news topic. This type of aggregation enables politicians to track public opinion, target messaging, and modify voting behavior accordingly; or allow businesses to quickly synthesize important information about their own company, competitors, or business impacting events. Additionally, news aggregation could help combat media bias and misinformation by giving readers a more holistic picture of an event, and exposing them to media sources that project different perspectives from their own news silos.

Data Sources:

The dataset was obtained by querying the APIs for the online component of two news sources: The Guardian, a British news outlet; and the New York Times, an American news outlet, then selecting

only articles published from 10/1/2020 through 12/5/2020 that were designated as either political or foreign news. The resulting dataset was then cleaned for extensive formatting issues, and short articles (ie. articles that only said “Election Updates Live”) were removed. The final dataset contained 3073 entries.

Processes and Tools:

For this project, two different topic modeling and clustering techniques were compared to examine which worked better. First, the BERTTopic model was tried, which uses the BERT language model to create embeddings, umap for dimension reduction, and HDBSCAN for clustering. BERTtopic was selected because BERT is strong at determining language within context and HDBSCAN does not require all elements to be assigned to a cluster or necessitate designating a predetermined number of clusters. The combination of these approaches aligned with the purposes of this project because the articles included in the dataset did not necessarily have a corresponding article from the other source or often only had one article written about the event covered.

The first attempt at using the BertTopic model as an off-the-shelf package did not produce usable results, as the initial trial resulted in too few clusters and overly broad topics. However, tweaking the parameters of the model created a modified version of the BERTtopic model which successfully retrieved results better suited to the goals of this project. Work started by converting the articles into embeddings using the “Sentence-Transformers" package and the “distilbert-base-nli-mean- tokens” sentence transformer. UMAP dimension reduction was then performed to simplify the content embeddings and to facilitate suitable clustering and graphing. Next, HDBSCAN was used to identify clusters within the documents and then graphed the results. The graph (*figure two*) shows an example of the clusters determined by the algorithm using the parameters shown in fig one. In the visualization, different topic clusters generated by the model are represented by different colors. The gray dots are articles that were not classified by the model and are considered outliers. After the clusters were determined, the algorithm used a class based form of TF-IDF at the article level to get the importance of the words based on the previous embedding and clustering. Finally, the overall topics were reduced by comparing the TF-IDF vectors among topics, merging the most similar ones, and then re-calculating the TF-IDF vectors to update the distribution and finalize the topics returned.

Next, gensim’s LDA LdaMulticore model was tested. This method worked in conjunction with pyLDAvis, a library which enables the creation of dynamic, highly readable visualizations of topics found by an LDA model. Ideally, the traditional LDA model, paired with pyLDAvis, would obtain easily interpretable results for the dataset. Since LDA is a more commonly used model, details of the preprocessing steps, how the LDA model works, or the various parameters used are omitted; for additional details on all of these aspects of our project, refer to the two LDA Jupyter notebooks.

The two different processes were run on both the dataset in its entirety, then on smaller subsets of the data based on date, to determine top news stories for specific periods in time. Due to space constraints, only the results for the two weeks, denoted in the dataset as weeks 4 and 5, around election day in the US will be discussed. This subset of the data consisted of 697 articles.

BERTtopic Results (Whole Dataset):

Several combinations of parameters for BERTtopic model were tried, but ultimately the ones shown in *figure one* were chosen, as these parameters provided the optimal clusters. Using these parameters originally resulted in 283 topics before final reduction (modeled in *Figure two*) and 273 finalized topics after final reduction of mostly intelligible clusters. *Figure three* in the appendix shows the number of articles in each of the clusters after final topic reduction. The table shows a truncated selection of clusters to save space. Articles in the "-1" cluster are outliers and were not assigned to a cluster.

Since the methods used for this model relied on unsupervised techniques, finding ways to test the accuracy of this method in the aggregate was difficult. Therefore, the results required manual review. While some of the generated topics had no clear semantic meaning, the majority of the clusters did depict discernable events that have recent media coverage. *Figure four* in the appendix shows the result returned for the largest cluster out of the entire dataset, which contained articles related to economic policy news in the UK (see additional details in appendix). Other topics covered in the clusters include the assassination of the Iranian nuclear scientist, Morales's return to Bolivia, recent hurricanes in Lousiana, election coverage, pandemic related news, unrest in Ethiopia, earth quakes in Turkey, among other topics. It is important to also note that some of the clusters returned multiple similar, but distinct events, for example, recent helicopter crashes in Egypt and Azerbaijan were clustered together (*Figure five*).

BERTtopic Results (Election Weekly):

Testing the BERTtopic model on smaller date ranges also provided interesting results. Retrieving readable results required tweaking the parameters from those used for the whole dataset, as our initial clusters appeared to be overly broad. For example, the election in the US was at first conflated with articles detailing the elections happening in Myanmar around the same time frame. Once the parameters were tweaked to facilitate smaller clusters (*Figure six*), the model returned more distinct clusters, although some conflation issues were unavoidable, similar to the full dataset results. Running the specified data range through the model with the new parameters resulted in 79 topics initially (graphed in *Figure seven*), which was then reduced to 59 (*Figure eight*) final topics. Similar to the full dataset iteration results, running the model on this subset of dates yielded mostly intelligible clusters that were traceable to a particular event. The largest cluster included articles regarding the suspension of Jeremy Corbyn from the UK's Labour Party over perceived antisemitism (*Figure nine*). Interestingly, the model also returned UK politician Mike Gove in a cluster along with the word "childcare." Some online

investigation revealed that Mike Gove misspoke in March regarding pandemic child care restrictions. Articles written in early November related to updated pandemic restrictions frequently made mention of this previous gaffe, which in turn led our model to place it into its own topic cluster (*Figure ten*).

LDA Results (Entire Dataset and Election Weeks Subset):

Running the LDA model required less trial and error than the BERTtopic model because this model can be at least partially validated by using coherence scores, which indicated that 152 was the ideal number of topics for the whole dataset (far fewer than the number of topics returned by BERTtopic). The LDA results were easier to interpret due to the readability of the visualizations created by this method (visualization example in *Figure eleven*). One can easily click through and see the top 30 words returned by the various clusters and even each topic relates to another based on the distance between and overlapping of the clusters. However, in reviewing the accuracy of the model, it was determined that most of the topics retrieved were either nonsensical or overly broad and could only at best give a general idea of events happening. Unlike the results returned by the BERTtopic model, there were not any identifiable news events that were previously unknown by reviewers that could be deduced purely by searching the top topic words returned for clusters determined by the LDA model. Testing this method on the election timeframe data subset returned similar vague or nonsensical results (*Figure twelve*). Overall, this method did not perform as well as the BERTtopic model with regards to the overall goal of this project of achieving discernable news clusters. However, this method's visualization component was far superior to the effectively nonexistent visualization capabilities compatible with the BERTtopic model method.

Conclusion and Possible Next Steps:

Overall, this project was successful as the models developed clusters based on topics, and the output often corresponded with recent events. However, there is still a lot of improvement needed to achieve the overarching aim of creating automated hyper local news article clustering. Each model had benefits and drawbacks but overall, the BERTopic model is more suited to our needs as it allows for more flexibility and returns more readable clusters for both the entire dataset and date related subsets. The accuracy and usability of the model would likely benefit from exploring parameter optimization using the current clustering and topic modeling methods or by exploring additional methods that are better suited to this task. This project would also benefit from building a visualization package for the final model that works similarly to the pyLDAvis package for LDA, as this would enable easier analysis of detected clusters. Additionally, in the future, combining the resulting model with article summarization, sentiment analysis, and similarity measurement capabilities could be used to detect differences in how the same story is portrayed across different media outlets and how events and their portrayal in media evolves over time.

Appendix

Sources and Works Consulted:

- NYTimes API
- Guardian API
- https://github.com/MaartenGr/BERTopic
- https://github.com/bmabey/pyLDAvis
- https://radimrehurek.com/gensim/

Results from BERTtopic News Aggregator Model (whole dataset)

Fig One: The parameters chosen to run the model. Min_cluster_size was highly important as it enabled us to set lower cluster sizes and create smaller more focused clusters.

UMAP Parameters	
N_neighbors	4
n_components	6
HDBSCAN Parameters	
Min_cluster_size	3

Fig Two: Clusters detected by BERT Topic Modeling and HDBSCAN after Dimension Reduction prior to final TF-IDF reduction

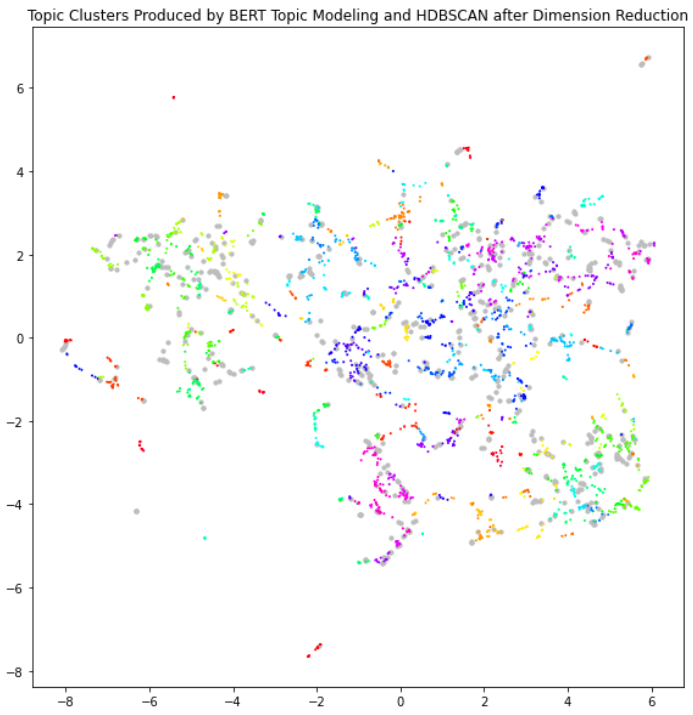

Fig Three: Size of top ten topics generated after final TF-IDF reduction. The topic numbers will change if the model is rerun.

	Topic	Size
0	-1	663
168	167	39
37	36	37
272	271	30
2	1	25
...	...	...
36	35	4
118	117	4
22	21	4
107	106	4
164	163	4

Fig Four: This code snippet has key words associated with the largest cluster in the entire dataset. This topic has to do with tax, budget, and spending policies in the UK. There have been a lot of articles about this topic recently. They often discussed "Rishi Sunak", who holds a chancellor position equivalent to a finance minister, and his response to the economic problems caused by the pandemic, including housing and goods pricing issues. His wife "Akshata Murty", the daughter of an Indian billionaire, is mentioned in many articles as well.

```
[
 "spending",
 0.01344825494254602
],
[
 "sunak",
 0.012705153030166939
],
[
 "tax",
 0.01013738022900249
],
[
 "chancellor",
 0.00834252558929291
],
[
 "murty",
 0.007505228461413092
],
[
 "prices",
 0.007397734339899093
],
[
 "investment",
 0.00699183135961955
],
[
 "fund",
 0.006946648687227484
],
[
 "housing",
 0.0069230306621033515
],
[
 "budget",
 0.006861224557995196
]
```

Fig Five: This code snippet shows an example of a cluster that conflated two similar (at least in terms of keywords likely detected) but distinct events. Here, the American helicopter crash in Sinai, Egypt is conflated with an incident around the same time when a Russian helicopter was shot down by Azerbaijan near the border of Armenia. The helicopter was hit by a missile launched in Shusha.


```
 [
  "helicopter",
  0.06589792056710667
 ],
 [
  "mfo",
  0.053321795665832734
 ],
 [
  "azerbaijan",
  0.04340175128806479
 ],
 [
  "egypt",
  0.04067544288439231
 ],
 [
  "crash",
  0.04015559873466052
 ],
 [
  "sinai",
  0.03554786377722182
 ],
 [
  "crashed",
  0.03551201655560585
 ],
 [
  "armenia",
  0.03302475047841557
 ],
 [
  "shusha",
  0.03159830719283467
 ],
 [
  "russian",
  0.03035191130762229
 ]
]
```


Results from BERTtopic News Aggregator Model Run on a weekly basis (Election Week)

Fig Six: Parameters chosen to run model

UMAP Parameters	
N_neighbors	3
n_components	5
HDBSCAN Parameters	
Min_cluster_size	3

Fig Seven: Clusters detected by BERT Topic Modeling and HDBSCAN after Dimension Reduction prior to final TF-IDF reduction

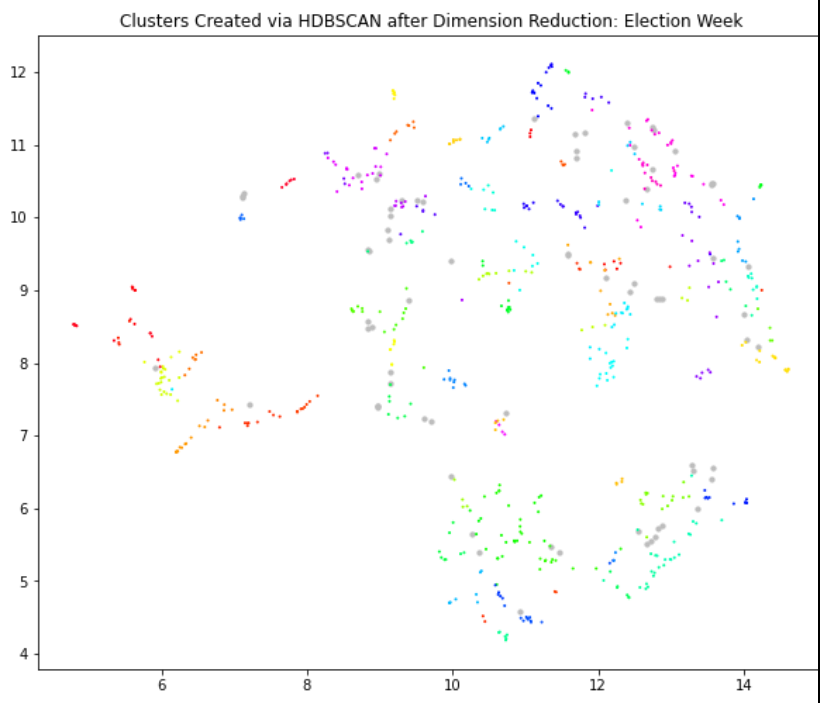

Fig Eight: Size of top ten topics generated after final TF-IDF reduction

	Topic	Size
0	-1	126
14	13	23
33	32	18
5	4	16
10	9	16
18	17	16
2	1	15
46	45	15
42	41	15
26	25	15
38	37	14
27	26	14

Fig Nine: Top output during the election weeks time period. This cluster refers to Jeremy Corbyn's suspension from the Labour party due to perceived antisemitism. "Keir Starmer" is the current head of the leader party.	` "corbyn", 0.02240343816823932], ["antisemitism", 0.01799626503846685], ["jeremy", 0.013922247233702745], ["ehrc", 0.012622377361450849], ["starmer", 0.012453622394948977], ["suspension", 0.012181017349184016], ["report", 0.010658747756127542], ["labour", 0.010492347447324761], ["complaints", 0.008589582364436032], ["jewish", 0.008127330331812442

Fig Ten: Output of a cluster that includes articles which mention Mike Gove's childcare policy gaffe from March. As restrictions were extended in the UK in early November, several recent articles mentioned his previous misspeak. In the same articles regarding the new restrictions "Mark Drakeford", who currently serves as First Minister of Wales, was often mentioned along with "firebreak," which is a word used by the British press to describe the level of shutdown occurring in Wales at the time.


```
[
 [
  "childcare",
  0.011770434442802877
 ],
 [
  "gove",
  0.009573454472723812
 ],
 [
  "household",
  0.009058339006803764
 ],
 [
  "continuity",
  0.008766804306312
 ],
 [
  "agreements",
  0.008691950046428747
 ],
 [
  "turn",
  0.0077949156874479895
 ],
 [
  "firebreak",
  0.007726177819047775
 ],
 [
  "scheme",
  0.007520072792636587
 ],
 [
  "drakeford",
  0.007225065126259075
 ],
 [
  "wales",
  0.007151587859309454
 ]
]
```


LDA Results Entire Dataset

Fig Eleven: Example of visualization created from LDA model by pyLDAvis library. Please refer to the accompanying LDA full dataset .Jupyterb notebook to explore full functionality as visualization is dynamic; please note, notebook must be run as github version does not show visualization.

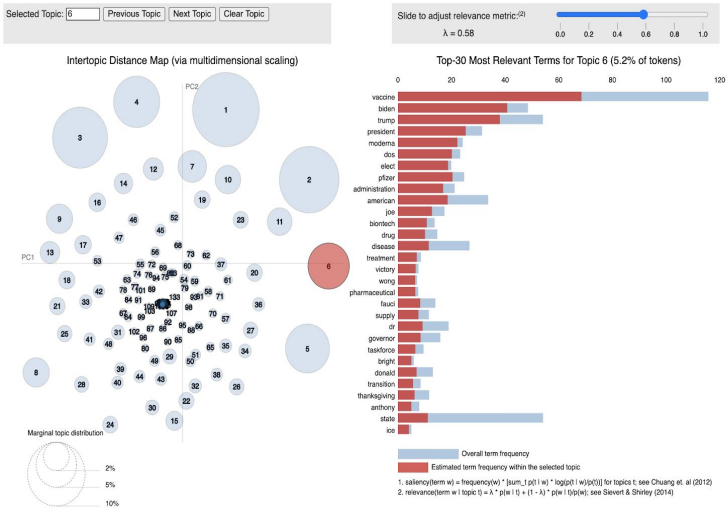

Fig Twelve: Example of visualization created from LDA model by pyLDAvis library. Please refer to the accompanying LDA election dataset Jupyter notebook to explore full functionality as visualization is dynamic; please note, notebook must be run as github version does not show visualization.