Using Embedding-Based Topic Modeling Techniques to Investigate Themes in Open-Ended Response Questions in a Federal Survey

Arezou Koohi, Haley Hunter-Zinck
Center for Optimization and Data Science, U.S. Census Bureau

May 16th, 2024
AAPOR

Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau or the National Center for Education Statistics (NCES). The NCES Disclosure Review Board has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release

United States® Census Bureau

Background: NTPS and TFS

The Department of Education’s “**Teacher Follow-Up Survey** (TFS) is a follow-up survey of public and private elementary and secondary school teachers who participated in the National Teacher and Principal Survey (NTPS) during the previous school year. The purpose of the survey is to determine how many teachers remained at the same school, moved to another school, or left the profession.

The major objectives of the TFS are to:

- measure the attrition rate for teachers;
- examine the characteristics of teachers who stayed in the teaching profession and those who changed professions or retired;
- obtain activity or occupational data for those who left the position of a K-12 teacher;
- obtain reasons for moving to a new school or leaving the K-12 teaching profession; and
- collect data on job satisfaction.”

National Teacher and Principal Survey - Overview (ed.gov)

Problem: open response question

What are some ways the coronavirus pandemic affected your teaching experience?

This can include any challenges you faced or enhancements you made in areas such as new teaching methods, classroom management strategies, communications, and technology.

- Objective: What are some of the themes in responses that can be derived automatically?
- We apply topic modeling to the responses to discover interpretable topics
- We apply sentiment analysis to further assess the tone of responses within each topic

TFS teachers and responses

The 2021–22 TFS sampling frame included about 43,900 teachers with the overall weighted response rate of 43.9% for the public schools and 33.8% for the private schools. *

Teacher status for the year 2021-2022 *

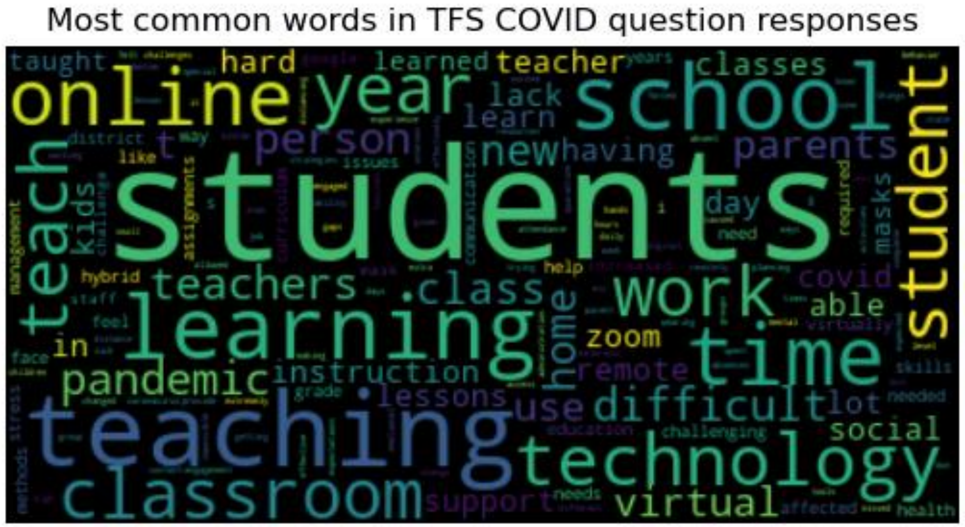
School	Current(%)	Former(%)
Public	92%	8%
Private	88%	12%

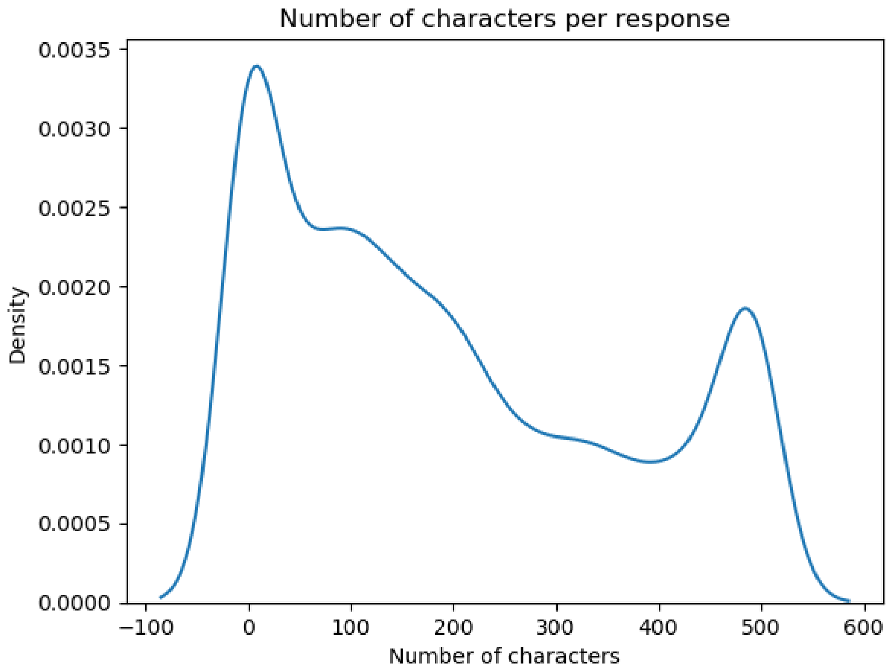
Number of responses to the TFS COVID question

Number of responses	Number of filled responses
7500	6100

* Teacher Attrition and Mobility Results From the 2021–22 Teacher Follow-up Survey to the National Teacher and Principal Survey
[Teacher Attrition and Mobility: Results From the 2021–22 Teacher Follow-up Survey to the National Teacher and Principal Survey (ed.gov)]

Response exploratory analysis





For example, this sentence has 95 characters → Plot is truncated due to empty responses or the responses that had filled the space completely.

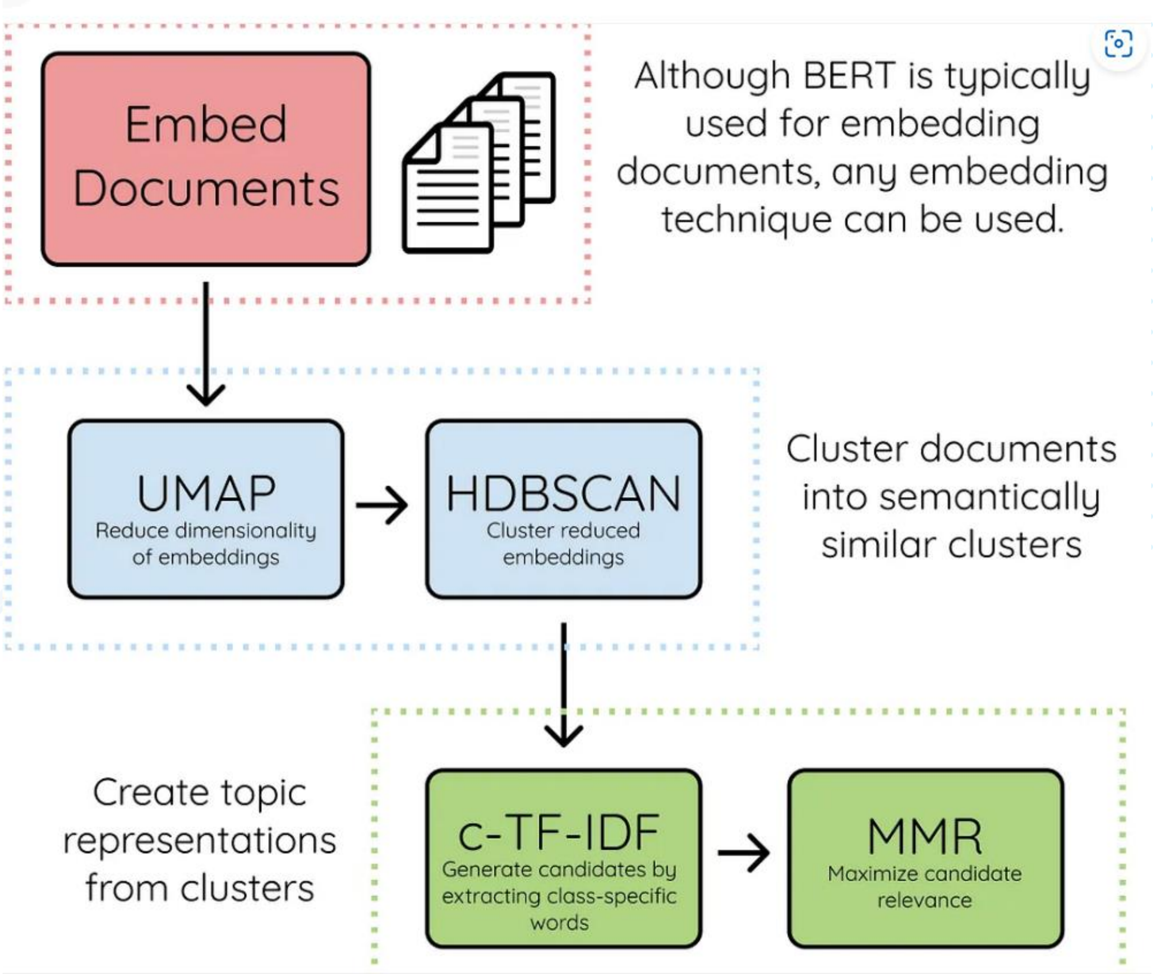
Methods: embedding-based topic modeling with BERTopic

1. Embeddings
2. Dimensionality reduction
3. Clustering
4. Association between words and topics

Embedding: [#, #, #, ..., # ,#]$_{384}$

Dimensionality reduction

[#, #]


Embed Documents
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

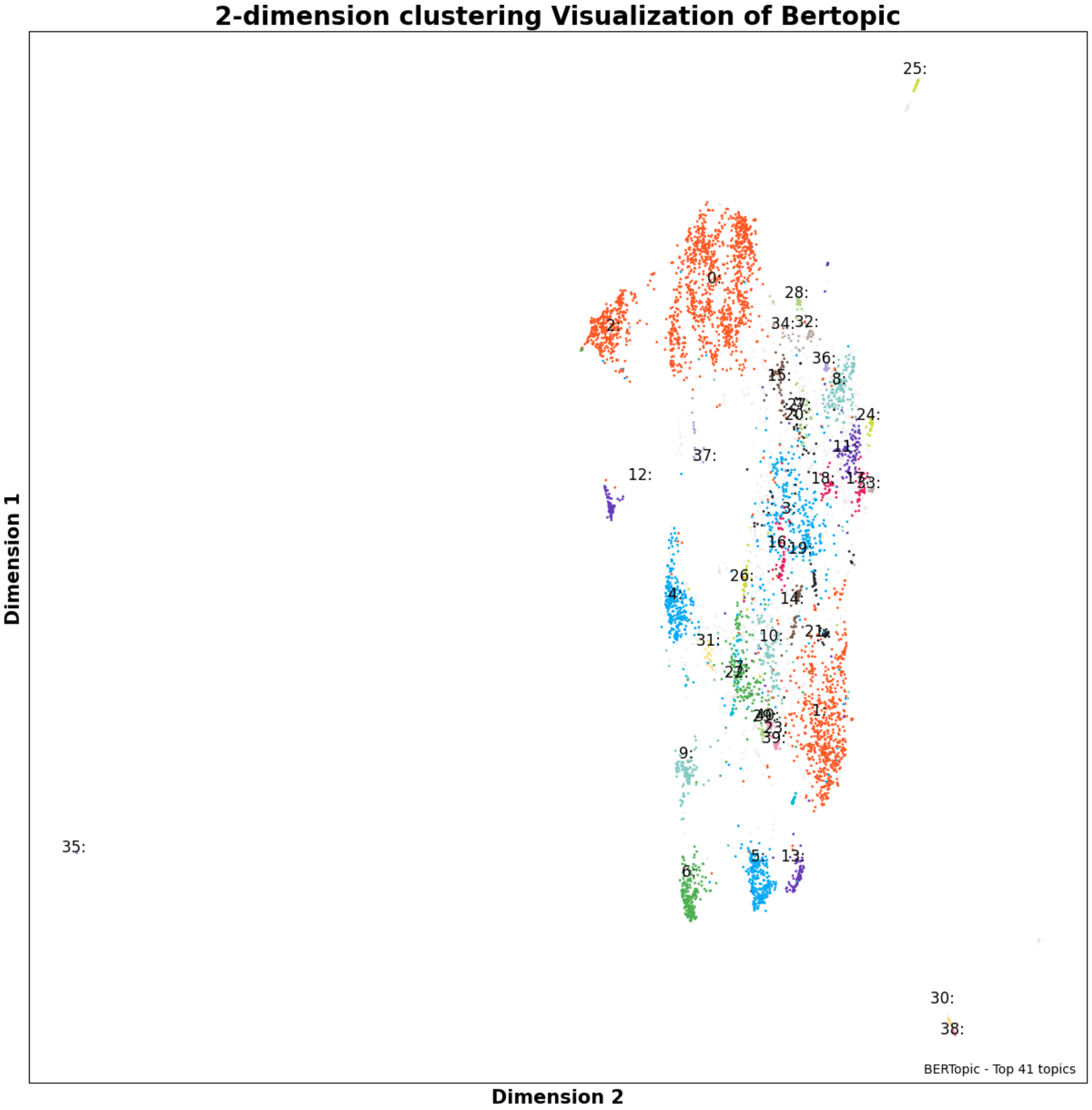
Although BERT is typically used for embedding documents, any embedding technique can be used.

Cluster documents into semantically similar clusters

Create topic representations from clusters

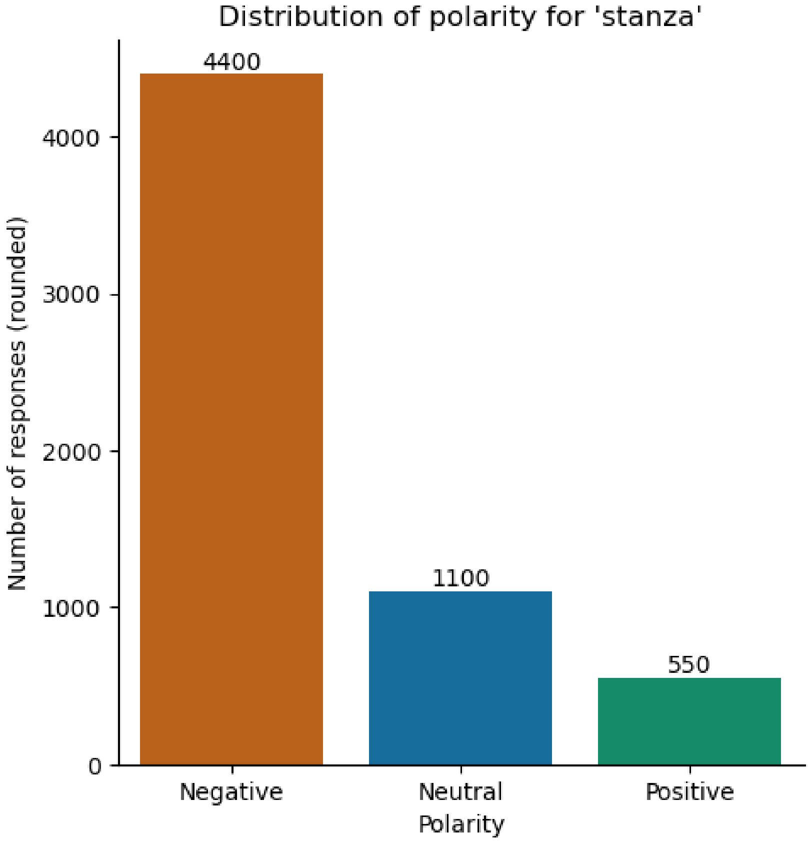
BERTopic: topic modeling as you have never seen it before | by Elisa Distante | Data Reply IT | DataTech | Medium

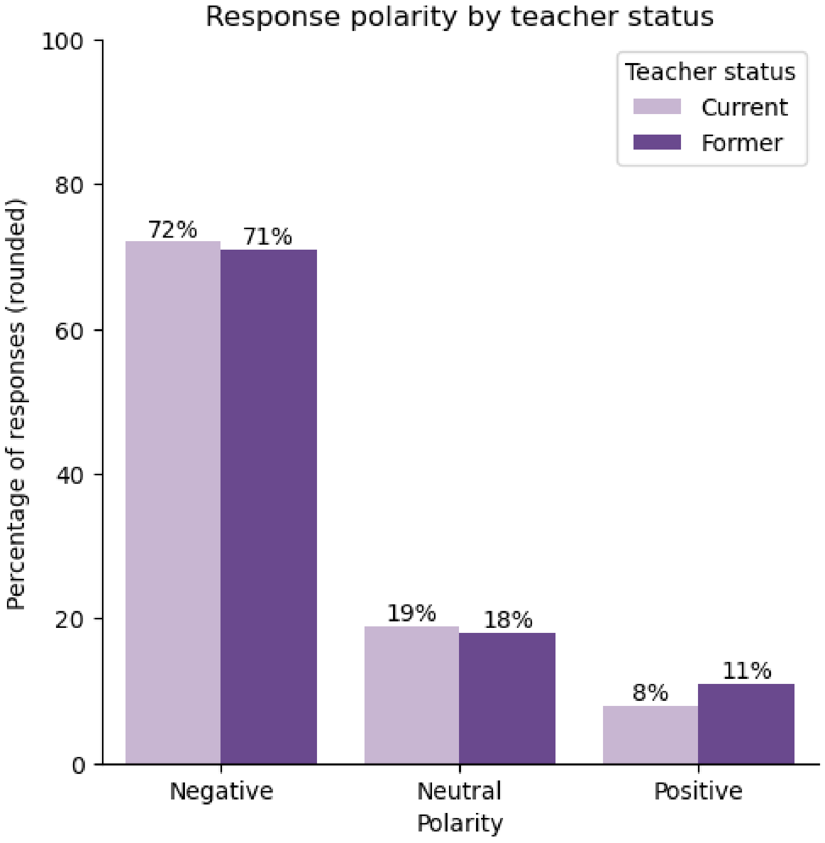
BERTopic finds 41 topics



Methods: sentiment analysis

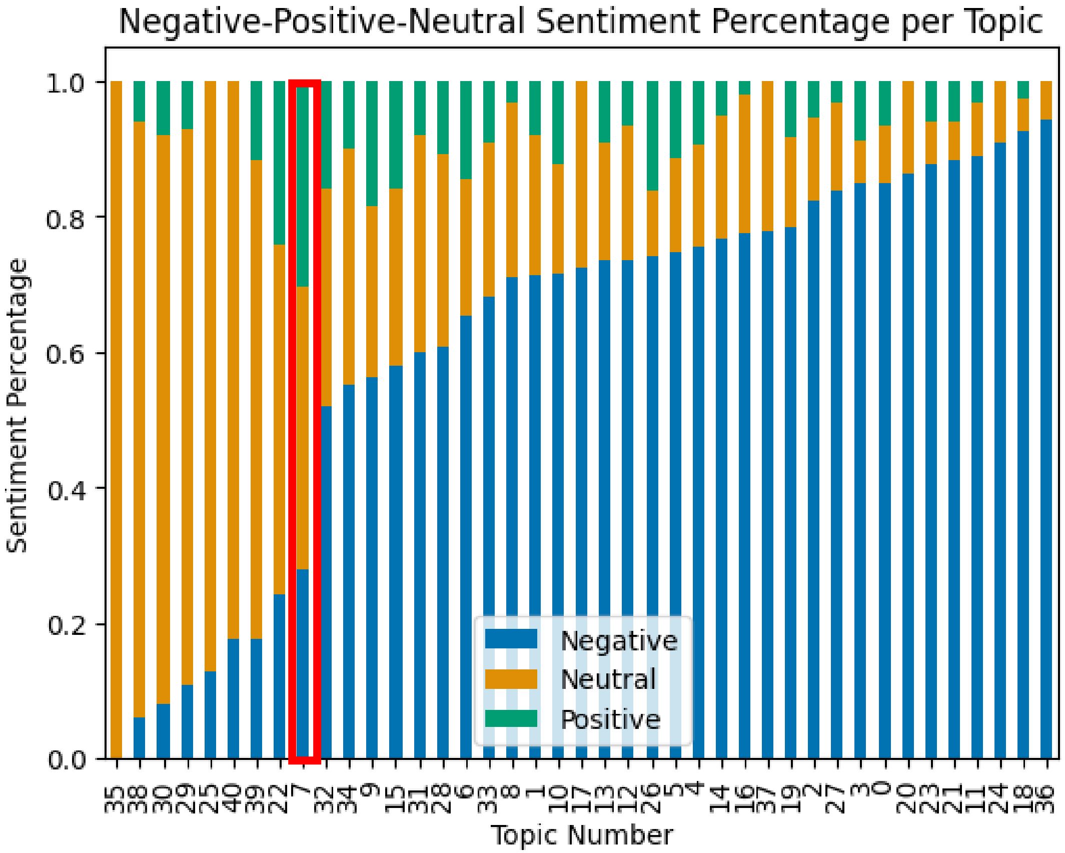
Sentiment analysis is the process of analyzing digital text to determine if the emotional polarity of the message is positive, negative, or neutral.





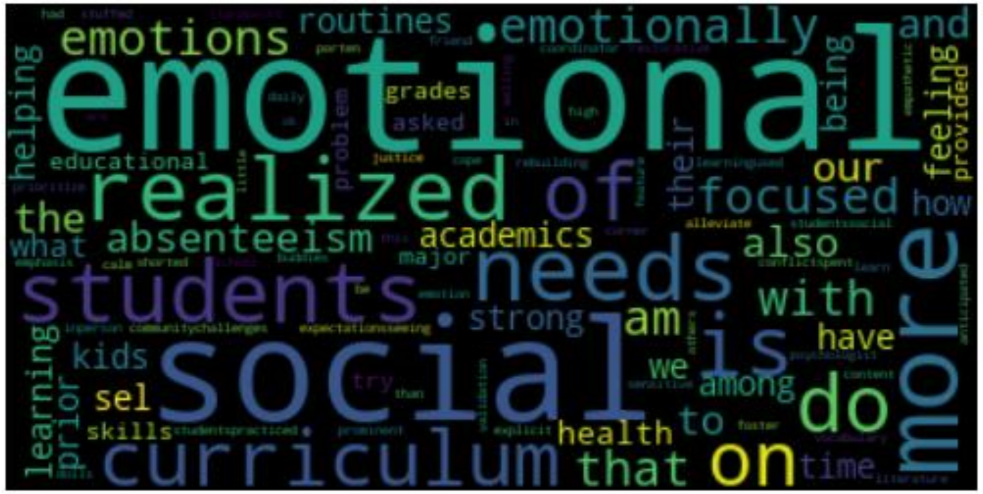
https://stanfordnlp.github.io/stanza/sentiment.html

Sentiment analysis per topic



The sentiments are overall negative across the topics, but the plot shows some variation between topics in polarity of responses

Example 1: Emotional-social needs



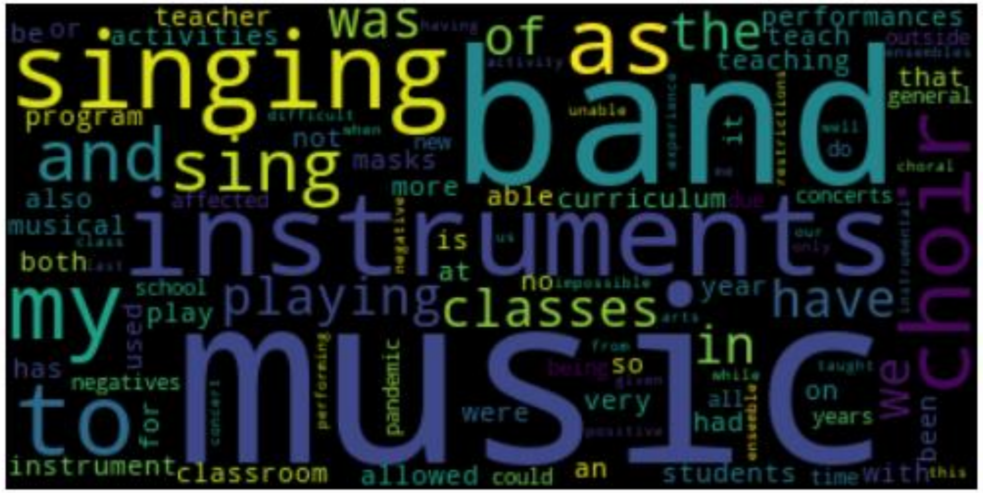
Total Response : 30

Negative polarity: 10%

Example response (paraphrased):

- I am more perceptive of my student's emotions, and I provide my students with emotional support. We have several activities that provide resources to our students to help handle their emotional and mental health needs.

Example 2: Effect on teaching music



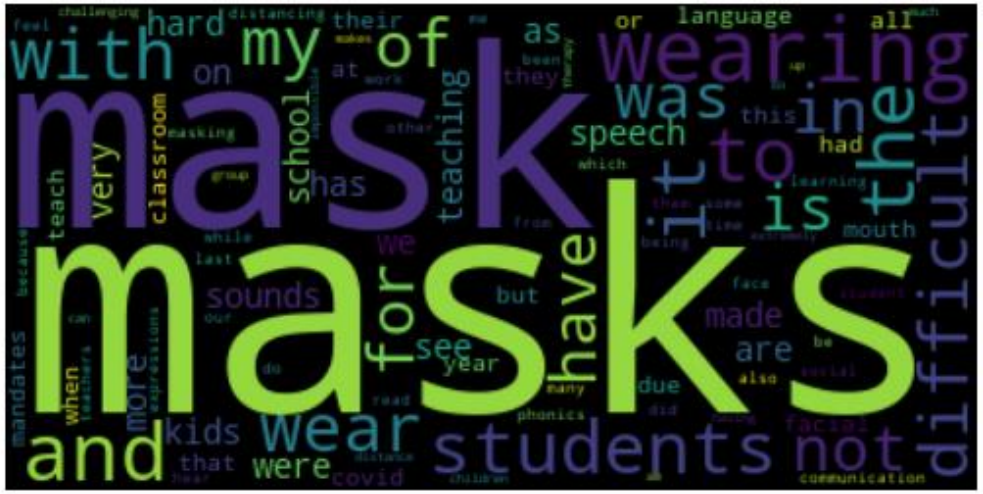
Total Response : 90

Negative polarity: 70%

Example responses (paraphrased):

- It was extremely difficult to teach music during the pandemic because we couldn’t be indoors and had to wear masks.
- I used singing and playing music to excite students and bring them joy. It was a therapeutic experience for my students.

Example 3: Effect of masks



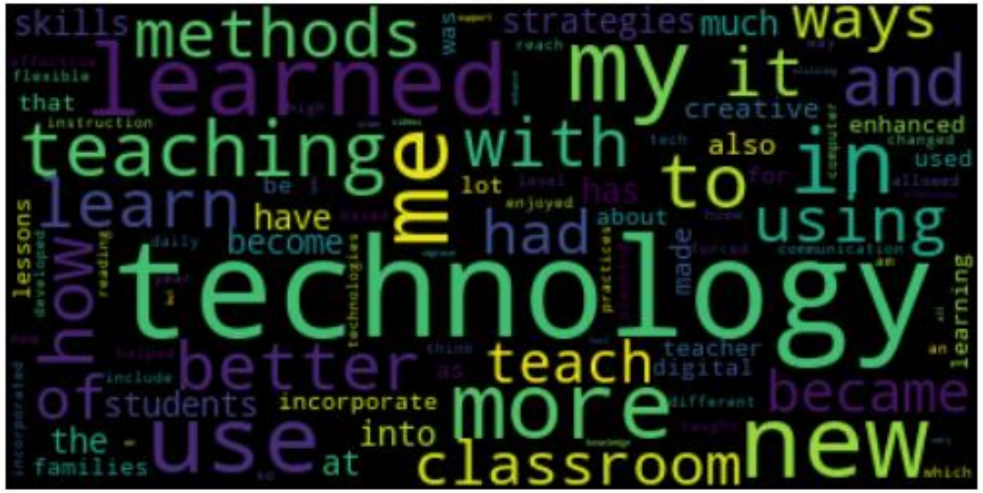
Total Response : 350

Negative polarity : 80%

Example responses (paraphrased):

- I couldn’t hear students through masks that made it hard to teach phonics. Students reading and writing abilities were adversely affected.
- It was hard to teach while being required to wear masks.

Example 4: Improving knowledge of technology



Total Response : 150

Negative polarity: 30%

Example responses (paraphrased):

- I have many new technical teaching abilities. Using technology in my class has improved my teaching.
- I now know how to use technology in my teaching.

Conclusions

Limitations

- Bigger clusters were more heterogenous in response
- Topic modeling still requires some manual interpretation

Highlights

- Combining the sentiment analysis and topic modeling can further automate the response exploration
- Topic analysis revealed themes such as emotional and social needs, difficulty using mask, integrating technology to teaching, and effect on teaching music
- While sentiment was largely negative, there were positive topics like learning how to use technology in the classroom.

Acknowledgements

- Louis Avenilla (U.S. Census)
- Patrick Campanello (U.S. Census)
- Ugochukwu Etudo (Brite Group / U.S. Census)
- Haley Hunter-Zinck (U.S. Census)
- Maura Spiegelman (NCES)

QUESTIONS?

arezou.koohi@census.gov