Using automatic clustering to identify themes from vape related content on TikTok

Tanvi Anand[1], Srijith Radhakrishnan[3], Nikhil C Mohan[3], Juhan Lee[2], Rachel Ouellette[2], Dhiraj Murthy[1], Grace Kong[2]

[1]Computational Media Lab, UT Austin, [2]Department of Psychiatry, Yale School of Medicine, [3]Manipal Institute of Technology

The University of Texas at Austin
Communication Studies
Moody College of Communication

The University of Texas at Austin
Cockrell School of Engineering

Yale University
School of Medicine

Introduction

- TikTok is a frequent source of youth exposure to e-cigarette content
- Manual coding of TikTok content requires many human labor hours
- Machine learning techniques, such as image clustering, can facilitate the distillation of e-cigarette content on TikTok into common themes

Data Collection

- Scraped 16 vape-related words (e.g., “e-cigarette”, “e-liquids”) and 15 hashtags (e.g., #vape, #vapelife)
- N=812 (non-English videos and videos not available removed from 1510 collected videos)
- We took one screenshot per video that best reflected the video's e-cigarette content.

OPTICS Clustering Algorithm

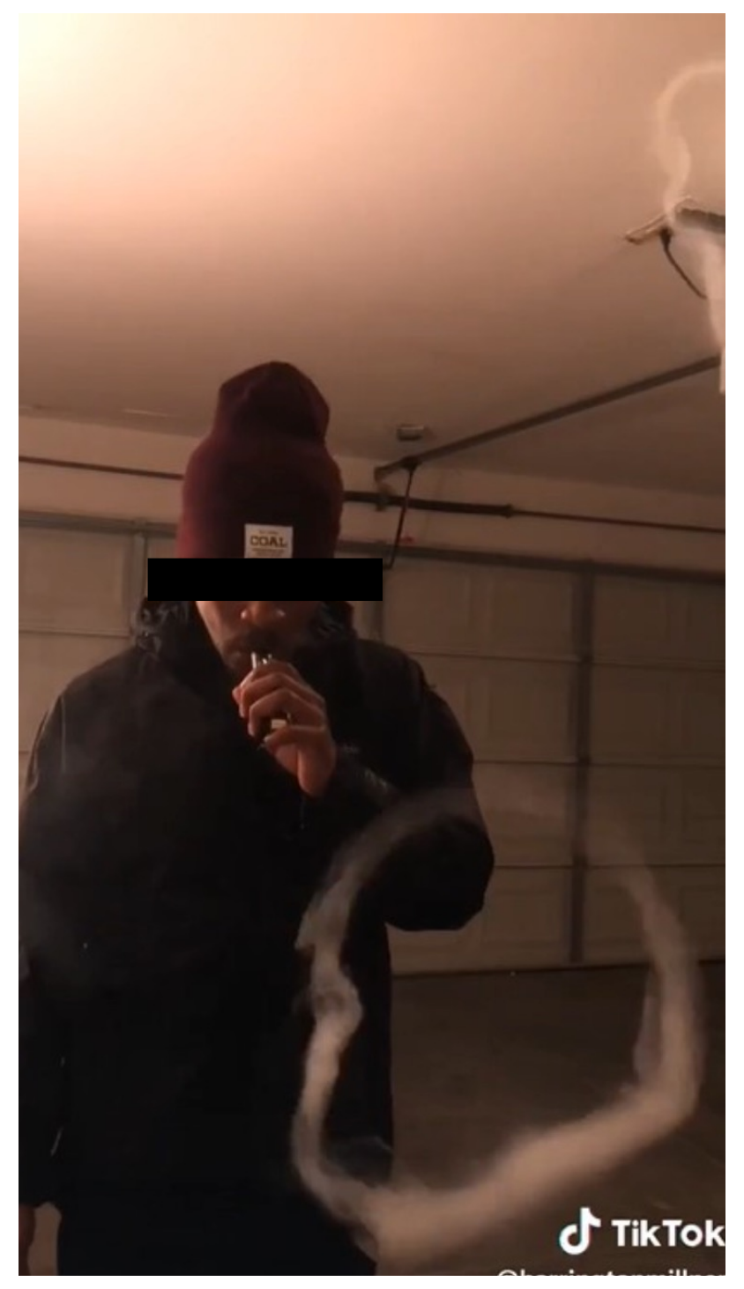
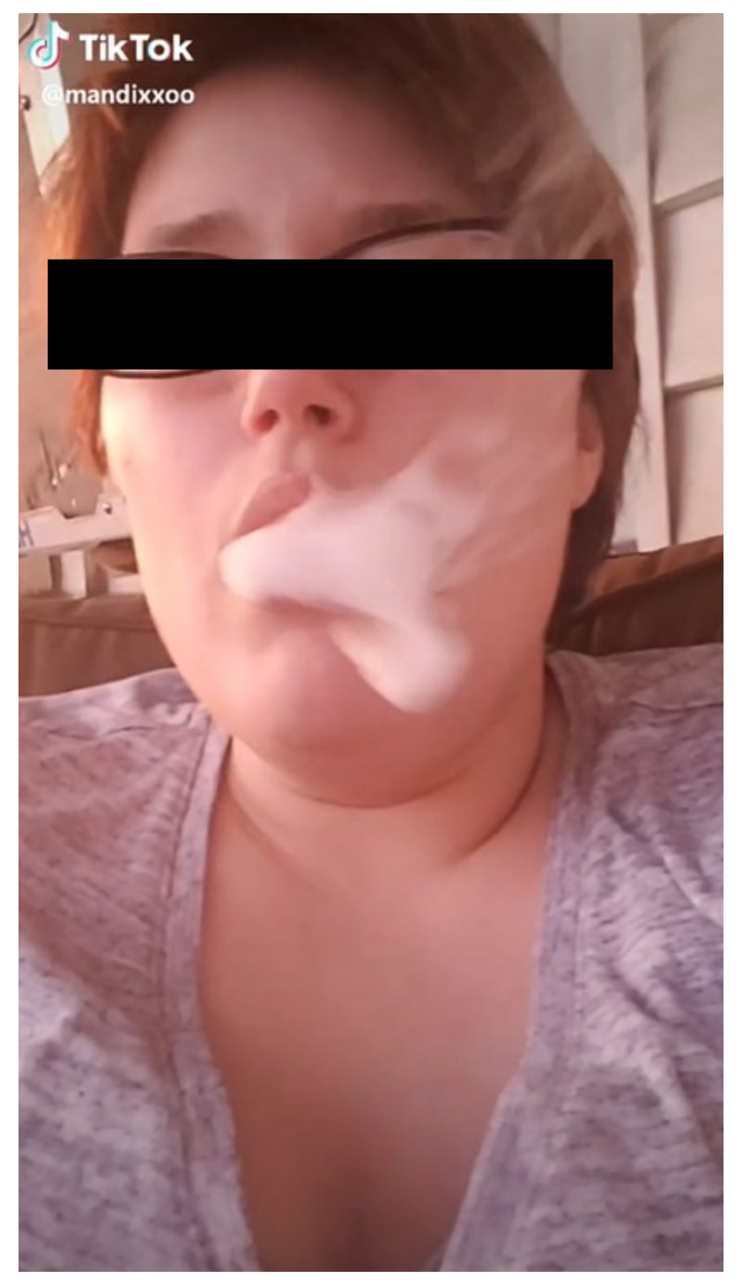
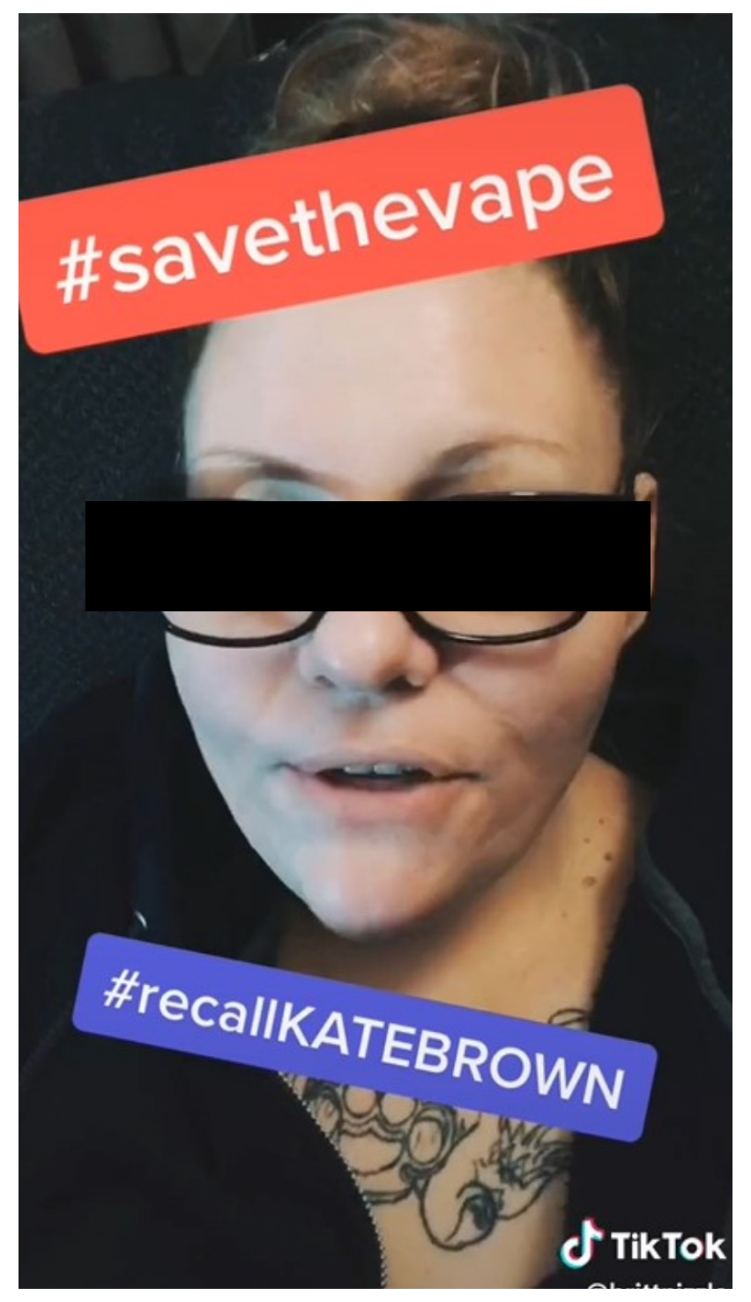
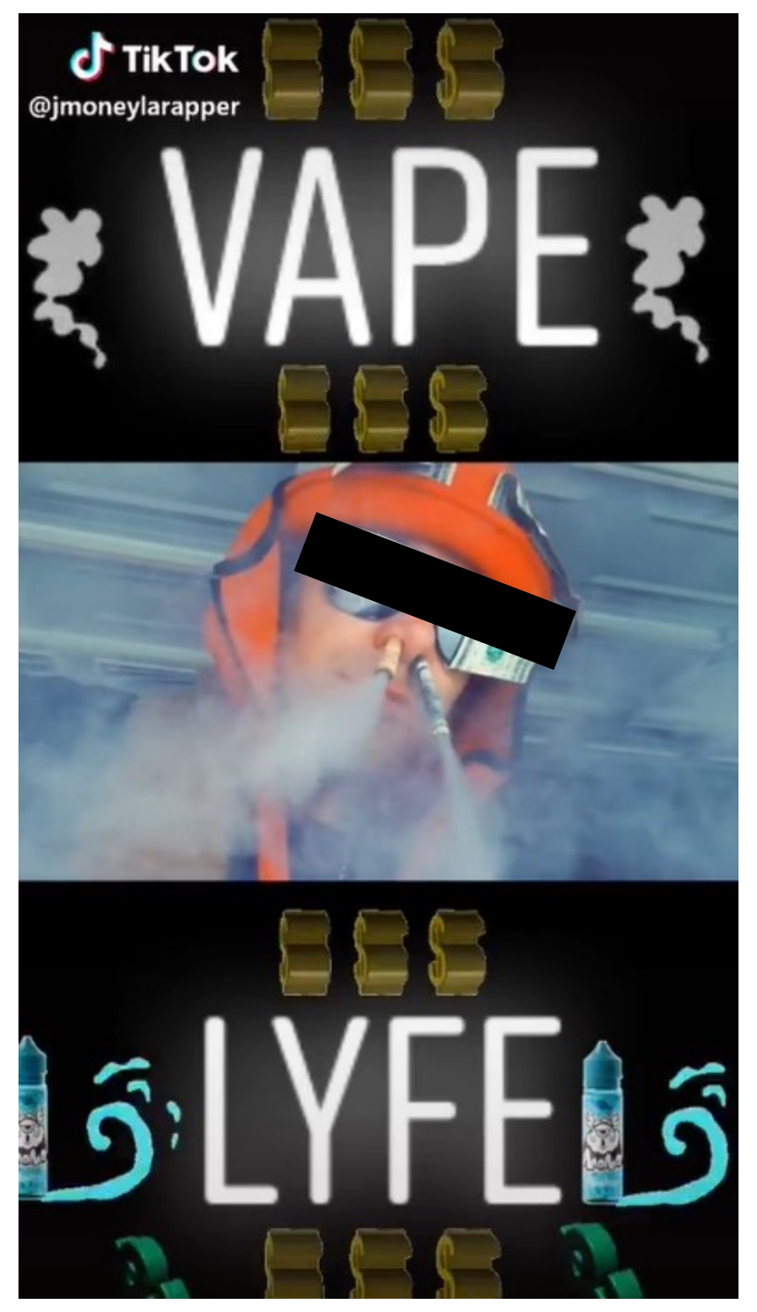
- We use “OPTICS”: Ordering Points To Identify Cluster Structure
- OPTICS is an unsupervised clustering algorithm
- It identifies groups of data points that are “similar” to each other based on certain features
- Features are obtained from images in our data
- No manual labelling of data is required while inputting into the model

Methods

- The OPTICS clustering model successfully grouped images into 20 clusters based on visual mathematical similarities.
- We decided the final number of clusters by examining using grid search for optimal parameters
- We manually scanned resulting clusters for coherence among images.

Results: Image Clusters and Model Results

- We found optimal parameters for OPTICS clustering to be epsilon=10e-10 and xi=0.01
- Combined with qualitative analysis, the research team analyzed the clusters and identified 7 overarching vaping related themes from the 20 clusters.

7 themes:						
1. Visible Vaping Devices	2. Visible Smoke Rings	3. Small Clouds of Vapor	4. Videos edited to have a vape related text overlay	5. Vapor covering faces	6. Videos with both text and vapor	7. Other (e.g.: people talking about vapes in the video but no visible vapes present)
						

Blackened to preserve privacy.

Conclusion

- Our model successfully clustered images into 7 interpretable themes.
- Clustering can be used to effectively interpret vape-related image data into common themes at scale.
- This method can be used to identify vape related content on social media and may be useful to identify trends across time.

Future Work

- This methodology can be extended to other photo/video based social media platforms (e.g., Twitter, Instagram) to identify vape and other tobacco related content
- Further work is needed to evaluate the capacity of machine learning to monitor e-cigarette content on social media to inform tobacco regulatory science

Acknowledgements

Funding: National Institutes of Health
Grant #: R01DA049878
(PI: Kong, G)

Contact

- **Tanvi Anand:**
tanviaanand@utexas.edu