Categorizing E-Cigarette-related Tweets using BERT Topic

Murthy, D.,[1] Keshari, S.,[1] Arora, S.[2], Yang, Q.[3], Loukas, A.[2], Harrell, M. B.[4], Hébert, E.[4], T., Schwartz, S. J.[2], Wilkinson, A. V.[4]

1. Computational Media Lab, University of Texas at Austin, Austin, TX
2. University of Texas at Austin, Austin, TX
3. Texas Christian University, Fort Worth, TX
4. UT Health Houston, School of Public Health, Michael & Susan Dell Center for Healthy Living; Austin, TX

SIGNIFICANCE

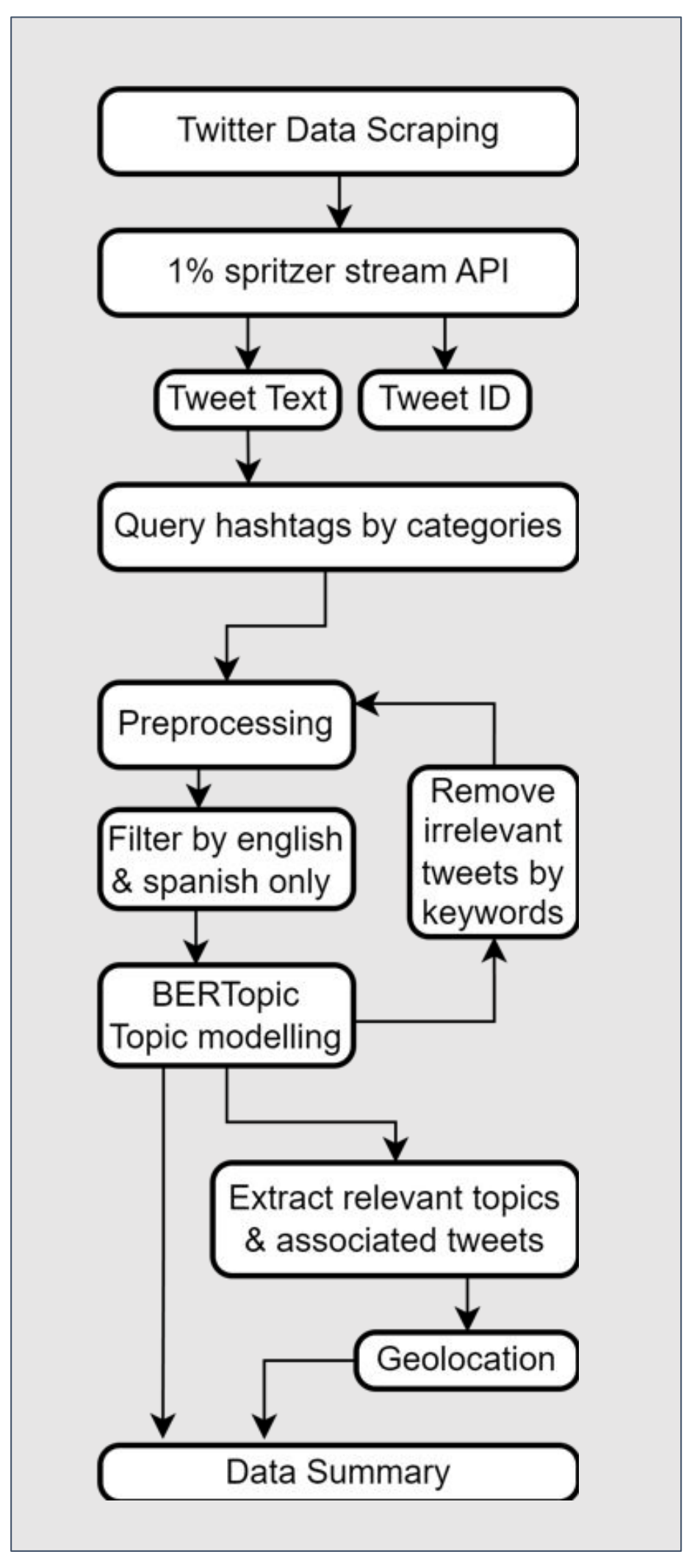
- Social media platforms have become a popular forum for discussing e-cigarettes
- Several studies have found a significant increase in e-cigarette and vaping discussions
- Topic modeling remains underutilized as a method for analyzing e-cigarette related messages on social media
- This study focused on ascertaining the extent to which themes and topics in e-cigarette-related tweets can be rendered using machine learning and topic modeling

DATA

154,281 (121,000 unique) tweets from 98,634 unique individuals collected from November 2022-February 2023 using a custom developed Python script deployed on a Oracle Cloud virtual machine.

METHOD

- Topic modeling (BERTopic) was used to derive vape-related tweet clusters.
- Experts in tobacco control removed irrelevant topics through an iterative process.
- Automated geoparsing methods were used to infer the location of tweets.
- Python langdetect library was used to filter out non-English and Spanish language tweets.



Twitter Data Scraping

1% spritzer stream API

Tweet Text

Tweet ID

Query hashtags by categories

Preprocessing

Filter by english & spanish only

Remove irrelevant tweets by keywords

BERTopic Topic modelling

Extract relevant topics & associated tweets

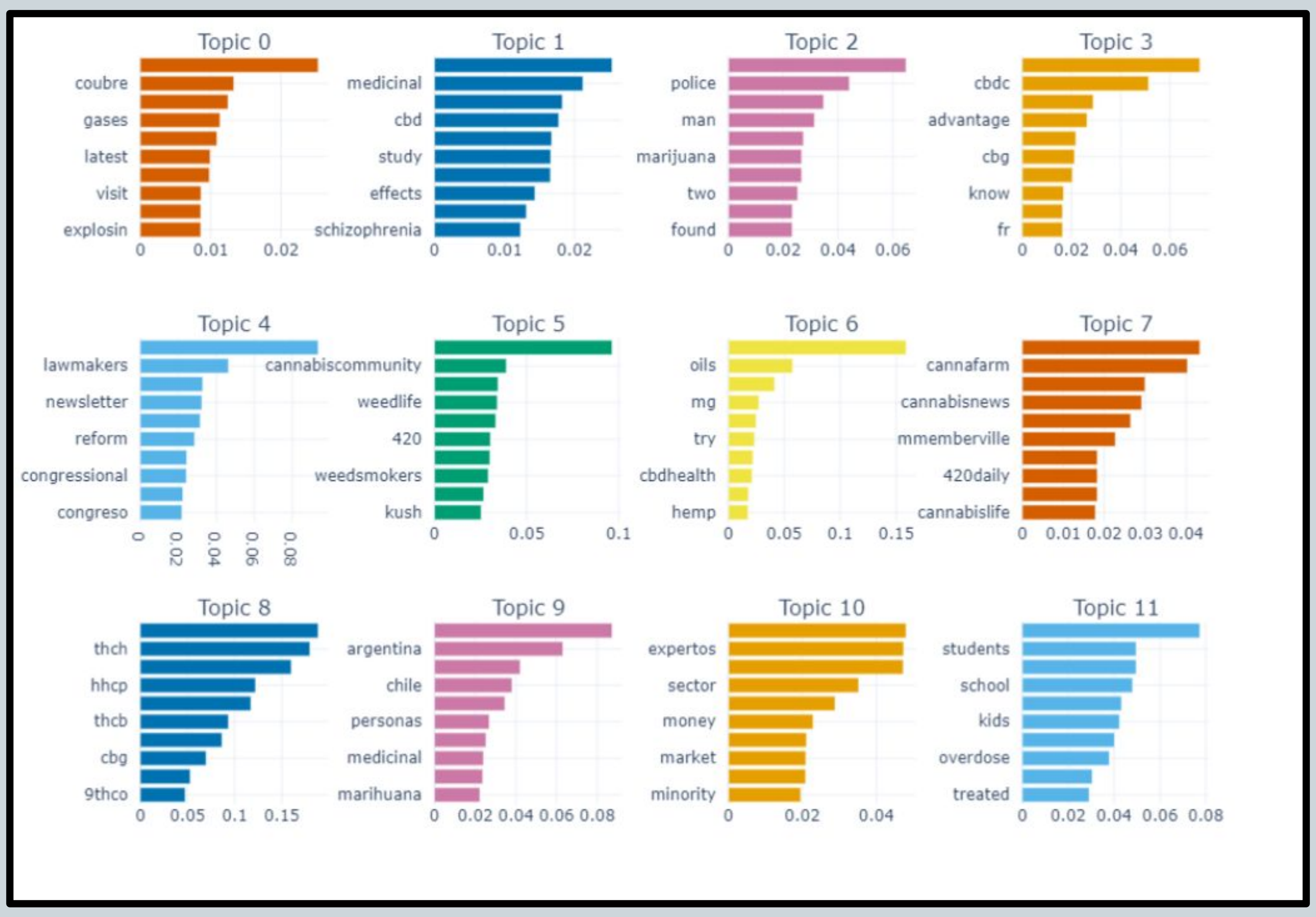
Geolocation

Data Summary

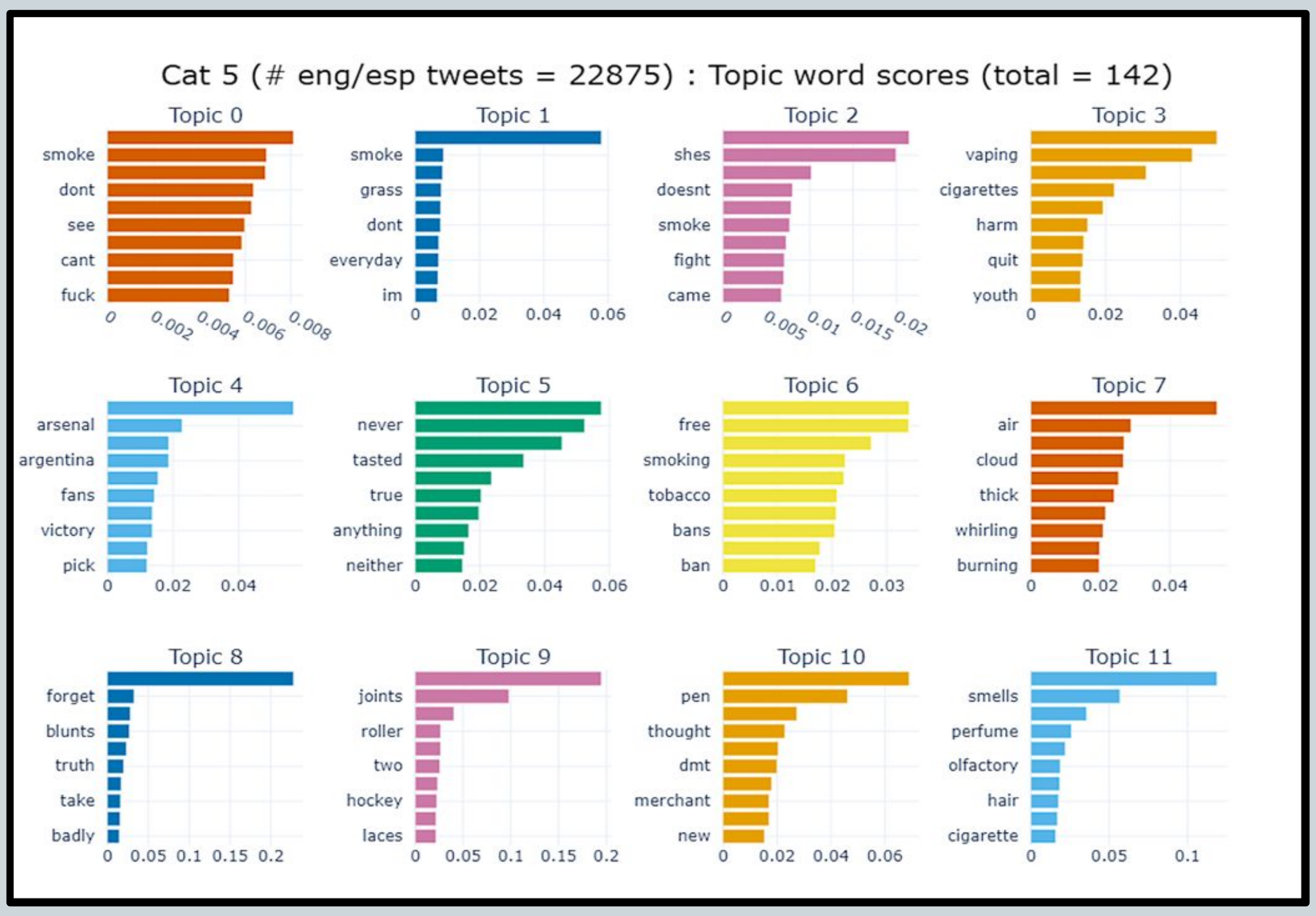
Computational topic modeling successfully identified 90 relevant topics (e.g., regarding brands, flavors, and regulation) across 6 thematic categories

RESULTS

- Clustered >100,000 e-cigarette-related tweets in English and Spanish into a total of 90 relevant topics (e.g., regarding brands, flavors, and regulation) across 6 broad thematic categories.
- Correlation and inter-topic map analysis indicated most topics were unique (correlation value < .5) and did not overlap.
- Inter-topic maps visually confirmed that topics within each category (n=6 categories) were mutually exclusive and unique.
- United States had the highest number of tweets related to vaping; however, most tweets (95%) could not be geolocated.



Topic modeling results of top 12 topics in category 2, with e-cigarette-related topics framed with red boxes; values represent c-TF–IDF scores per keyword



Topic modeling results of top 12 topics in category 5, with e-cigarette-related topics framed with red boxes; values represent c-TF–IDF scores per keyword

CONCLUSION

- This study highlights the need for continued monitoring of e-cigarette-related discussions on social media platforms.
- Social Media can have significant implications for public health messaging.
- Our study provides valuable insights into the topics and themes that are most discussed in e-cigarette-related tweets on Twitter and demonstrates the potential of BERTopic as a tool for analyzing large-scale social media data.

This research was funded by the National Institutes of Health under award number R01MD017280 to AL and AVW.

TEXAS
The University of Texas at Austin

UTHealth® Houston
School of Public Health

MICHAEL & SUSAN DELL
CENTER *for* HEALTHY LIVING

Project
VAMoS