Make Public Perception Data Work for You: Collecting and Analyzing Social Media Data

SEPTEMBER 29, 2022

Annice Kim, PhD, MPH
Robert Chew, MS
Jamie Guillory, PhD, MS, MBA
Social media data overview and what research questions we can answer with these data

Data science methods to gain insights on public health audience segments

Recommendations for incorporating social media monitoring into your existing surveillance process
The New York Times

‘Very Harmful’ Lack of Data Blunts U.S. Response to Outbreaks

HealthAffairs

Social Media Listening
Social listening is the process of collecting data from social platforms and forums on a chosen topic...the collected data is then analyzed to find trends and useful insights.

Brandwatch
Why Health Agencies Use Social Media Data

Communication
Surveillance
Policies/Regulations
What Health Agencies Want to Know

- Gauge public response to health education campaigns.
- Identify and address misinformation.
- Identify influencers who will amplify messages and detractors who undermine messages.
What Health Agencies Want to Know

- Monitor emerging products and health issues.
- Identify emerging consumer perceptions and misperceptions about health.
- Monitor emerging consumer health behaviors.
What Health Agencies Want to Know

- Understand public support/opposition for policies.
- Identify potential impact of regulatory policies and unintended consequences.
- Monitor marketplace for emerging products and misleading claims.
- Monitor compliance and enforcement of regulations.
Social Media Listening – Example Research Questions

- How are brands marketing their new synthetic nicotine products? What are consumers’ perceptions and reactions to these products?
- What are consumers’ perceptions and anticipated behavioral responses to the upcoming ban on menthol cigarettes?
- How are consumers using vaping products as a cigarette smoking cessation tool?
- What are North Carolina residents discussing on social media regarding travel and travel safety amid COVID-19?
- What questions do consumers have about staying safe during the COVID-19 pandemic?
Benefits/Opportunities

Near Real-Time Data Collection

Massive Volume

Users’ Own Voices
Limitations

Representation
Some Data Limited
Access
Social Media Platforms & Data Access

Restricted

Open

Best Access
Social Media Listening Process

Set-up

1. Identify research question, time frame & platforms for analysis.

Listen

2. Develop query to identify conversations on social media & forums.
3. Review results & refine query to remove off-topic posts.

Analyze

4. Categorize posts into key themes.
5. Qualitatively code posts & summarize results.

Report

Key Metrics

- Volume of conversations about your issue
- Key spikes in volume trends and what is driving these peaks
- Top topics related to your issue
- Emerging keywords/themes
- Illustrative sample posts
**Surveillance of Emerging Products**

**Case Study: Synthetic Nicotine – Conversation Volume**

**October 7, 2020:** Twitter volume peaked in early October 2020 when Kenya’s health minister banned LYFT synthetic nicotine pouches.

**September 6, 2021:** Increase around September 9 PMTA deadline for brands to submit application to stay on the market under FDA regulation.
Surveillance of Emerging Products
Case Study: Synthetic Nicotine – Top Topics

<table>
<thead>
<tr>
<th>TOPIC NAME</th>
<th>MENTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthetic nicotine</td>
<td>1957</td>
</tr>
<tr>
<td>tag us in your story</td>
<td>259</td>
</tr>
<tr>
<td>synthetic e-fuels</td>
<td>122</td>
</tr>
<tr>
<td>quit smoking</td>
<td>352</td>
</tr>
<tr>
<td>derived from tobacco</td>
<td>149</td>
</tr>
<tr>
<td>long cut</td>
<td>211</td>
</tr>
<tr>
<td>pure nicotine</td>
<td>99</td>
</tr>
<tr>
<td>zyn pouches</td>
<td>1071</td>
</tr>
<tr>
<td>tobacco products</td>
<td>218</td>
</tr>
<tr>
<td>Zyn nicotine pouches</td>
<td>332</td>
</tr>
</tbody>
</table>

Top phrases centered around the following themes:

- **Synthetic Nicotine Products/Brands**
- **Tobacco/Smoking Cessation**
- **Promotional Posts**
- **Comparisons with Tobacco**
Surveillance of Emerging Products

Key Takeaway: Consumers discuss using synthetic nicotine pouches and other similar products to quit smoking and/or vaping.

- User is asking for recommendations on how to quit smoking, and another user replies “zyn tobacco free nicotine pouches.”

- User attributes success quitting smoking to tobacco free nicotine pouches.

- User says tobacco free nicotine pouches are similar to dip pouches. They also describe them as helpful with managing nicotine withdrawal.
April 26, 2021: Twitter and Reddit volume peaked because of the FDA announcement to ban menthol cigarettes and flavored cigars. Volume peaked around 115K total mentions.
Monitoring Potential Impact of Policy

Case Study: Federal Menthol Ban – Top Topics

Top phrases centered around the following themes:

- Race/Ethnicity
- Menthol Cigarette/Flavored Cigar Ban
- Menthol/Flavored Tobacco
- Government
Consumer suggests the menthol ban undermines Black people’s autonomy and will result in criminalizing more Black people.

User critiques decision to make menthol cigarettes illegal. Suggests punishing tobacco companies or providing universal healthcare would better address the issue.

User interprets menthol ban as banning a flavor of cigarettes because Black people like it.

The Biden admin seems to think that black people are children to be managed and coddled because they’re unable to make decisions for themselves like smoking menthol cigarettes. All this will do is criminalize more black people.

“Smoking menthols kills Black people so instead of punishing the Tobacco companies for pushing this poison, or enacting a universal healthcare system that will save lives, we will simply make them illegal 😞”

saying you’re banning menthols because they disproportionately impact the African-American community is a wild way of saying you’re banning a flavor of cigarettes because black people like it
Data Science Methods to Gain Insights on Public Health Audiences
Social media listening platforms can cover many relevant use cases. However, additional user information is often important for public health surveillance and evaluation.
How would you approach this problem?

Traditional Approach: Qualitative (Deductive) Coding

Steps
- Determine categories/codes of interest.
- Take a sample of users.
- Collect and read information from each user’s profile, posts, and comments.
- Assign codes to each user, based on best professional judgement.

Strengths
- Conceptually, easy to understand and explain

Limitations
- Manually taxing and challenging to scale/repeat

Alternative Approach: Supervised Machine Learning

Steps
- Create labeled data set, following same steps from deductive coding.
- Split data into training and test sets.
- Develop model to predict categories of interest.
- Evaluate trained model’s performance.

Strengths
- Easy to scale to users outside your coded sample and to repeat periodically over time

Limitations
- More challenging to develop and diagnose
Supervised Learning for Audience Segmentation

- Developed classification models to assign users into predicted age groups.
  - Twitter (Morgan-Lopez et al., 2017)
  - Reddit (Chew et al., 2021)

- Developed classification models to assign “User Types” to accounts.
  - Marketer, Individual, News Media, Public Health Agencies, Advocates, Bots
  - Twitter (Kim et al., 2017)
Reddit Age Category Model: Overview

- Manually labeled age groups by searching for user age self-reports
  - 1,325 Adolescents (13–20)
  - 831 Adults (21–54)

- Features
  - Derived from metadata and up to last 100 posts and comments

- Modeling
  - Compared several supervised learning algorithms
    - Logistic Regression, KNN, SVMs, RF, Gradient Boosted Trees
  - Conducted feature selection to reduce from ~1500 to ~15 variables.
  - Reported 5-fold cross-validated metrics, on training set, as well as test set performance.

### Variable group

<table>
<thead>
<tr>
<th>Metadata used</th>
<th>Example</th>
<th># variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary Statistics</td>
<td>All</td>
<td>Median post score</td>
</tr>
<tr>
<td>Subreddit Frequencies</td>
<td>Posts and Comments</td>
<td>Frequency user posted to “Teenagers” subreddit</td>
</tr>
<tr>
<td>Emoji Frequencies</td>
<td>Comments</td>
<td>Frequency of “😞” used by user</td>
</tr>
<tr>
<td>Literary Characteristics</td>
<td>Posts and Comments</td>
<td>Average Flesch Reading Ease score</td>
</tr>
<tr>
<td>Patterns in Posting</td>
<td>Posts and Comments</td>
<td>Percentage of user's posts that were videos</td>
</tr>
<tr>
<td>Term Usage</td>
<td>Comments</td>
<td>TF-IDF score for the term “school”</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reddit Age Category Model: Variable Interpretation

**Adolescents (13–20)**
- Created account more recently
- More likely to post in popular subreddits
- More likely to post to r/teenager
- More likely to use the term “school”

**Adults (21–54)**
- Wrote more sentences per comment
- Higher Coleman Liau Index (readability)
- More likely to post in r/news
- More likely to use the term “home”

*Chew et al., 2021*
Who is posting on social media?
Use Case 1: Cigarette Conversations Among Predicted Youth vs. Adults

Are there differences in themes about cigarettes between youths and adults on Reddit and Twitter?

- Identified conversations about smoking.
- Applied age prediction algorithms to 17,707 Twitter and 23,364 Reddit users posting about cigarettes.
  - Guillory et al., 2022
- Conducted SML on final samples:
  - 5,598 youth and 6,347 adults (Twitter)
  - 5,417 youth and 6,131 adults (Reddit)

Adults discuss smoking cessation more than youth on both Reddit and Twitter.

Youth on Reddit discuss health consequences of smoking; Adults do not.
In the late 2010s, JUUL was the most popular ENDS in US, accounting for nearly 75% of sales.

Among high school students who had ever used e-cigarettes and had tried 1 of 4 leading brands, JUUL was the top brand used (64.2%).

Social media and JUUL’s advertising strategy of using young adult models and influencers may have influenced JUUL’s popularity.

JUUL claimed it did not target youth and voluntarily initiated plan to address youth access, appeal, and use of JUUL products.
Are kids being exposed to social media posts by JUUL?

- Examined 11,861 accounts following JUUL’s Twitter account in April 2018.
  - Kim et al., 2019
- Predicted age of JUUL followers based on metadata and sample of tweets.
  - ~80% < 21 years old
  - ~45% < 18 years old
Recommendations for Incorporating Social Media Monitoring in Your Work
Do you need to monitor social media?

- Do you need continuous surveillance on a health topic?
- Are you monitoring an emerging issue that is rapidly evolving?
- Do you need early indicators to help inform other research or response efforts?
- Is there currently no gold standard method for capturing these data?
- Is your target audience of interest on social media?
Social Media Monitoring is an Iterative Process

1. What do you want to know?
2. Where can we look?
3. Are these conversations happening online?
4. What insights can we extract?
5. Are results helpful?
Clarify Research Questions:
Know what you can / cannot answer with social media data

- **Very specific questions:**
  - “Did our campaign influence vaccine hesitant adults to change their attitudes and get vaccinated?”

- **Difficult to answer because:**
  - Specific audience
  - Behavioral outcomes
  - Causal changes

- **Instead, ask these types of questions:**
  - What are emerging trends in health topic x?
  - How is the public responding to policy x?
  - What questions do they have about x?

- **Easier to answer because:**
  - Open-ended, exploratory
  - Identifying emerging issues and trends
  - Early indicators/signals that can inform further research and/or triangulation with other data
Triangulate Social Media Insights with other Data Sources
Evolving Social Media Ecosystem – Where is the signal?

- What data are accessible?
- How will you get these data?
- What is your budget and resources?
Social Media Analysis Maturity

1. Personal Ad Hoc Monitoring
2. Social Media Listening
3. Qualitative Coding
4. Custom Computational Methods
5. Custom Dashboard/Triangulation With Other Data

Value

Difficulty
expectation  reality
Need for Paradigm Shift

What we value in science
- Rigor
- Validity
- Control, experimentation, manipulation
- Generalizability

What we need
- Adaptive, nimble, lean research
- Timeliness
- End user mindset
- Faster dissemination process
Balance Tradeoffs

Signal Detection
Timeliness
Scalability Frequency

Rigor
Deep Dive
Customization

Inform communication
Inform other data collection
Inform prevention development

Publication
Evaluate outcome
Enforce policies/regulations

Outputs
Knowledge of social media data is essential for public health
Continually upskill and bring others along on the journey
Work with Us!

RTI’s MULTIDISCIPLINARY EXPERTISE
SOCIAL MEDIA ANALYSIS

SUBJECTS
- Infectious Disease
- Communicable Diseases
- Emerging Drugs
- Disease Prevention

DATA SCIENCE
- Computational Text Analysis
- Natural Language Processing
- Data Mining
- Predictable Modeling
- Social Network Analysis
- Data Visualization Tools

COMMUNICATION SCIENCE
- Public Discourse Analysis
- Message Testing and Evaluation
- Digital Strategy
- Risk Perceptions / Misperceptions
- Content Analysis
- Message Development

UNDERSTAND PROBLEM AND CLIENT NEEDS
RIGOROUS METHOD AND RAPID, SCALABLE ANALYSIS
SYNTHESIS AND ACTIONABLE INSIGHTS
Thank you

Annice Kim | akim@rti.org
Robert Chew | rchew@rti.org
Jamie Guillory | jamieguillory.contractor@rti.org

RTI International is a trade name of Research Triangle Institute. RTI and the RTI logo are U.S. registered trademarks of Research Triangle Institute.