Mining and Analyzing Data from Social Media for Pragmatic Research

- Bethany Kwan, PhD, MSPH; Jenna Reno, PhD
- University of Colorado Anschutz Medical Campus

COPRH Con

Colorado Pragmatic Research in Health Conference

ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

Learning Objectives

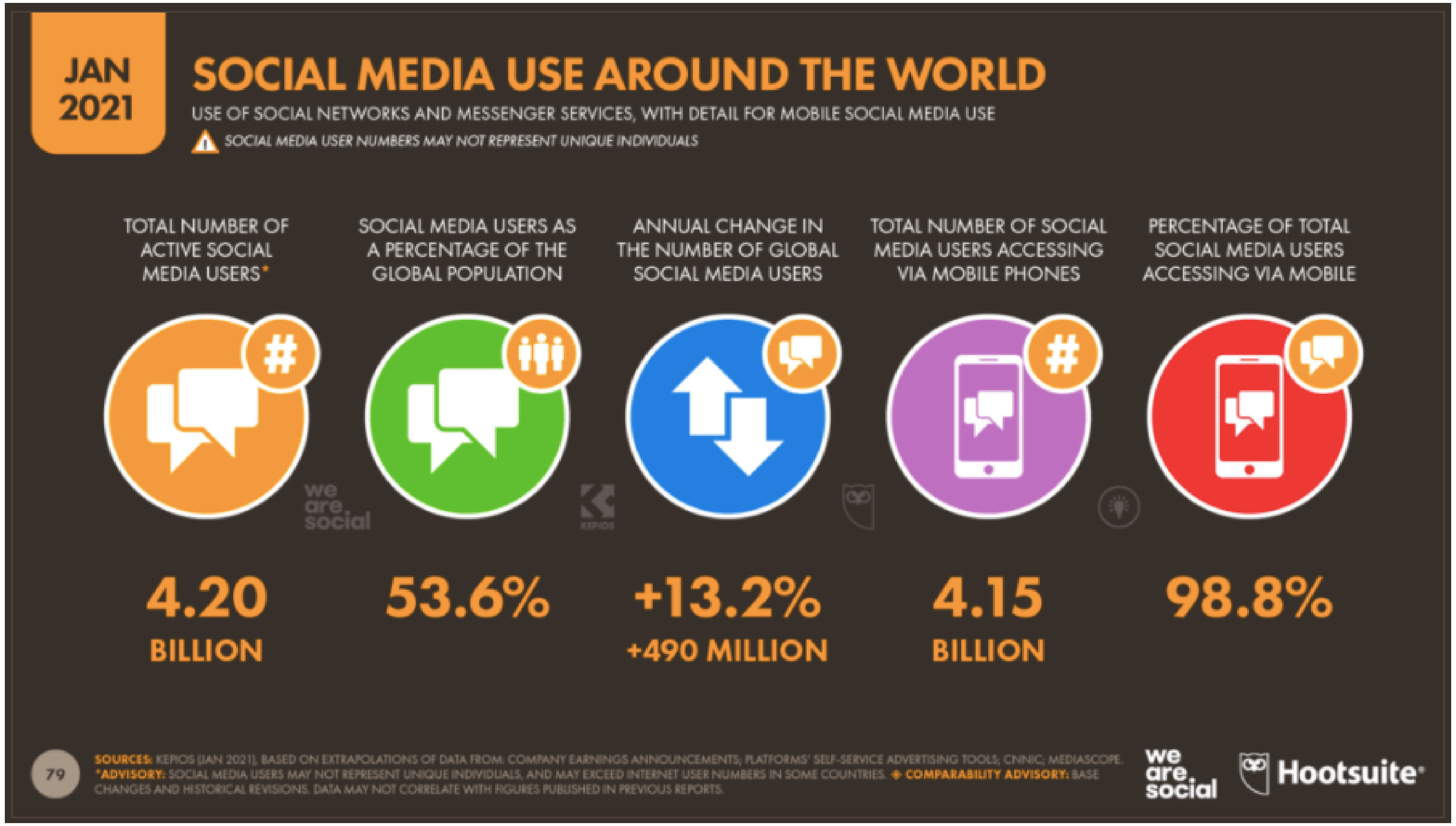
1. Identify audiences and potential uses of social media in pragmatic research
2. Identify approaches to mining data from social media and the web for research
3. Describe quantitative and qualitative analysis methods appropriate for social media data

Conflict of Interest statement

- Neither Dr. Kwan nor Dr. Reno have a financial or any other conflict of interest related to this talk.
- We assert that we have no financial interest in any of the social platforms or third party vendors for data mining mentioned in this talk.

Social Media Platforms and Audiences

- Facebook
- Twitter
- Instagram
- Snapchat
- Reddit
- YouTube
- LinkedIn
- Blogs
- Patient forums “PatientsLikeMe”
- Other online communities



JAN 2021

SOCIAL MEDIA USE AROUND THE WORLD

USE OF SOCIAL NETWORKS AND MESSENGER SERVICES, WITH DETAIL FOR MOBILE SOCIAL MEDIA USE

⚠ SOCIAL MEDIA USER NUMBERS MAY NOT REPRESENT UNIQUE INDIVIDUALS

TOTAL NUMBER OF ACTIVE SOCIAL MEDIA USERS*	SOCIAL MEDIA USERS AS A PERCENTAGE OF THE GLOBAL POPULATION	ANNUAL CHANGE IN THE NUMBER OF GLOBAL SOCIAL MEDIA USERS	TOTAL NUMBER OF SOCIAL MEDIA USERS ACCESSING VIA MOBILE PHONES	PERCENTAGE OF TOTAL SOCIAL MEDIA USERS ACCESSING VIA MOBILE
4.20 BILLION	53.6%	+13.2% +490 MILLION	4.15 BILLION	98.8%

79

SOURCES: KEPIOS (JAN 2021), BASED ON EXTRAPOLATIONS OF DATA FROM: COMPANY EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; CNNIC; MEDIASCOPE. *ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS, AND MAY EXCEED INTERNET USER NUMBERS IN SOME COUNTRIES. ◈ COMPARABILITY ADVISORY: BASE CHANGES AND HISTORICAL REVISIONS. DATA MAY NOT CORRELATE WITH FIGURES PUBLISHED IN PREVIOUS REPORTS.

we are social

Hootsuite®

Use of Social Media in Pragmatic Research

- Implementation and conduct of research
  - Stakeholder and community engagement or “citizen science”
  - Dissemination and messaging channels
  - Recruitment and outreach
- Source of data for research
  - “Secondary use”
  - Communication research
  - Network analysis
  - Ethnographic research
  - Public health surveillance
  - Patient-generated health outcomes data

Taylor J, Pagliari C. Mining social media data: how are research sponsors and researchers addressing the ethical challenges?. Research Ethics. 2018 Apr;14(2):1-39.

Benefits of Social Media for Research

- Real-time data
- Reaching large numbers of people
- Reducing costs
- Understanding what information people seek and engage with
- Identifying trending topics, ideas
- Environmental scans
- Can engage users via tools they're already familiar with
- Built in networking/ability to share/re-share
- Data mining tools forgo the need for transcribing
- NLP can provide be applied to content analysis of large datasets

Taylor J, Pagliari C. Mining social media data: how are research sponsors and researchers addressing the ethical challenges?. Research Ethics. 2018 Apr;14(2):1-39.

Challenges with using social media data for research

- Inequitable access
- Selection bias
- Data accessibility
- Non-standard data (data quality and formatting concerns)
- Non-traditional sampling
- Ethical considerations

Guidelines for Online Research and an Ethical Framework for Text Mining

- https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_Guideline-for-online-research.pdf
- Ethical considerations
  - Are data private or public?
  - Consent—should users be asked?
  - Anonymity
  - Weighing harms against benefits
  - Research for public benefit
  - Legal concerns and site terms and conditions
  - Governance of data, annotations, algorithms, and linkage
  - When is ethical approval needed?

Taylor J, Pagliari C. Mining social media data: how are research sponsors and researchers addressing the ethical challenges?. Research Ethics. 2018 Apr;14(2):1-39.

Ford E, Shepherd S, Jones K, Hassan L. Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. Frontiers in Digital Health. 2021 Jan 26;2:62.

Case Example: Use of Twitter for Stakeholder Engagement in Research on Palliative Care for People with Brain Tumors

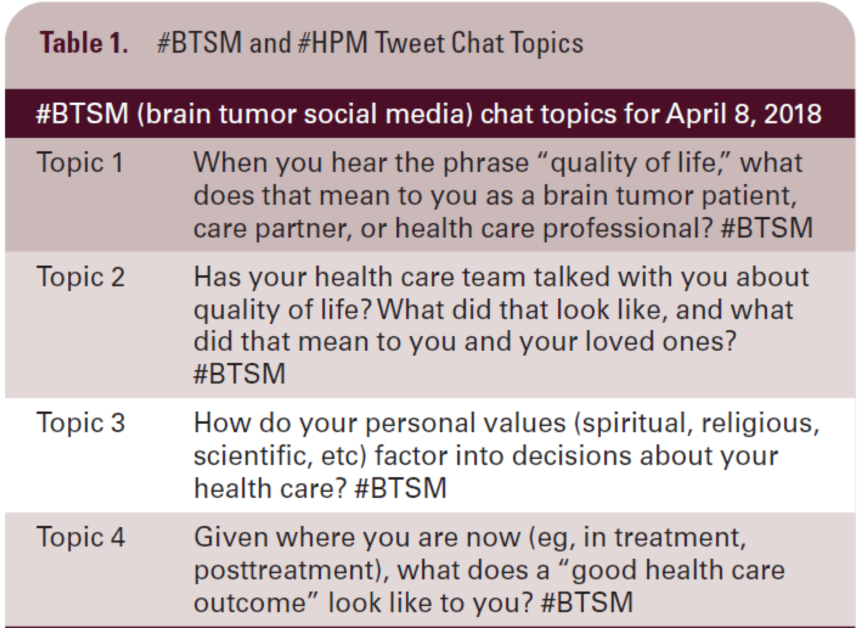


Table 1. #BTSM and #HPM Tweet Chat Topics

#BTSM (brain tumor social media) chat topics for April 8, 2018	
Topic 1	When you hear the phrase “quality of life,” what does that mean to you as a brain tumor patient, care partner, or health care professional? #BTSM
Topic 2	Has your health care team talked with you about quality of life? What did that look like, and what did that mean to you and your loved ones? #BTSM
Topic 3	How do your personal values (spiritual, religious, scientific, etc) factor into decisions about your health care? #BTSM
Topic 4	Given where you are now (eg, in treatment, posttreatment), what does a “good health care outcome” look like to you? #BTSM

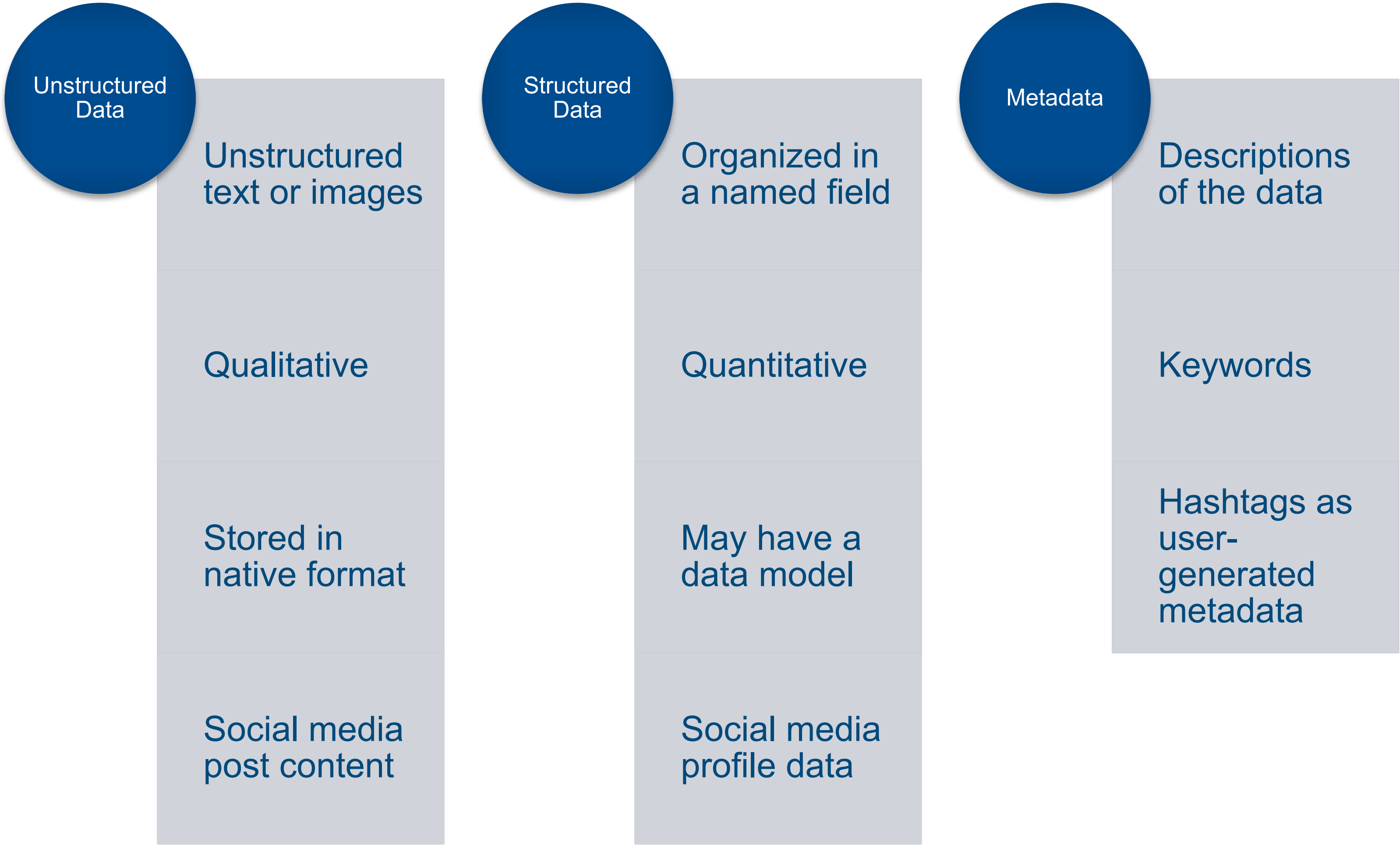
Theme 2: There is need to address quality of life in the context of healthcare, decision making about treatment, and support for care partners	
The healthcare system needs to provide better support for care partners	“So important… to turn your head and look at the suffering caregiver sitting next to the patient and ask How are you doing? Also ask caregiver lens on how the patient is doing because there may be forgetfulness or minimization. The tired caregiver knows what's going on.” – care partner
Patients and care partners appreciate when providers discuss quality of life early, but not immediately at diagnosis, and wish it was emphasized more	“I don't recall the phrase of ‘QOL’ came up specifically. But the side effects were clearly discussed pre-treatment. Having gone thru treatment, now I realized if #QOL was more emphasized to the patient, the process would have gone better.” – patient

Salmi L, Lum HD, Hayden A, Reblin M, Otis-Green S, Venechuk G, Morris MA, Griff M, Kwan BM. Stakeholder engagement in research on quality of life and palliative care for brain tumors: a qualitative analysis of# BTSM and# HPM tweet chats. Neuro-oncology practice. 2020 Dec;7(6):676-84.

Activity #1: Know Your Audience

- Pick one of the “uses of social media for pragmatic research”
- State a specific hypothetical or real example of how you might use social media in one or more ways in your research
- Who is your audience?
  - Adopters, influencers, potential saboteurs
- Where might you find this audience on social media?
- How do they use social media?
- Who are the influencers on social media?
- How might you partner with existing online communities?
- How will you conduct your research ethically and equitably?

Social Media as a Data Source



Unstructured Data

- Unstructured text or images
- Qualitative
- Stored in native format
- Social media post content

Structured Data

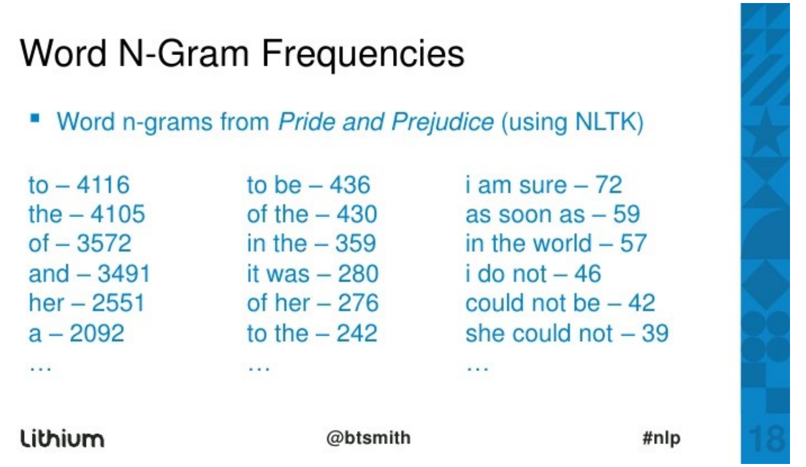
- Organized in a named field
- Quantitative
- May have a data model
- Social media profile data

Metadata

- Descriptions of the data
- Keywords
- Hashtags as user-generated metadata

Mining Data from Social Media: Mining techniques

- Manual approaches
- Connection via an Application Programming Interface (API)
  - Free on Twitter – search and download tweets (but limited to 1% of tweets)
- Third party vendors
  - Licensed with the platform for broader access (can be expensive)
  - Symplur: https://www.symplur.com/products/signals/
  - Social listening tools
- Data preprocessing methods
  - Named entity recognition and normalization (automated and manual)
    - Named entity recognition: identification of entities such as drugs, diseases, and medical events
    - Normalization: Mapping to predefined categories or standard medical ontologies
    - Dictionary lookup: Matching text strings to medical lexicons
  - Text mining techniques (extracting features of free-text for further analysis)
    - N-gram, word embedding, sentence-dependency-based parse tree, Latent Dirichlet Allocation (LDA) topic modeling
  - Fuzzy adaptive resonance theory network based Information Retrieval (FIR) scheme



Word N-Gram Frequencies

- Word n-grams from *Pride and Prejudice* (using NLTK)

to – 4116
the – 4105
of – 3572
and – 3491
her – 2551
a – 2092
…

to be – 436
of the – 430
in the – 359
it was – 280
of her – 276
to the – 242
…

i am sure – 72
as soon as – 59
in the world – 57
i do not – 46
could not be – 42
she could not – 39
…

Lithium @btsmith #nlp 18

Ru B, Yao L. A literature review of social media-based data mining for health outcomes research. Social Web and Health Research. 2019:1-4.

Chen LS, Lin ZC, Chang JR. FIR: an effective scheme for extracting useful metadata from social media. Journal of medical systems. 2015 Nov;39(11):1-4.

Social Listening

Crowdtangle:

- a public insights tool from Facebook that makes it easy to follow, analyze, and report on what's happening with public content on social media. (free to researchers)

Agora Pulse:

- synchronises your social media accounts around the clock, offers unlimited reports and graphics of performance analytics, retains all your account data, compares your page with others on key metrics.

Hootsuite:

- a social media listening tool with specific search terms in real- time. Can be used to monitor mentions of your brand, products, or relevant keywords you are interested in. Also handy to track all of your social media accounts in one dashboard.

Iconosquare:

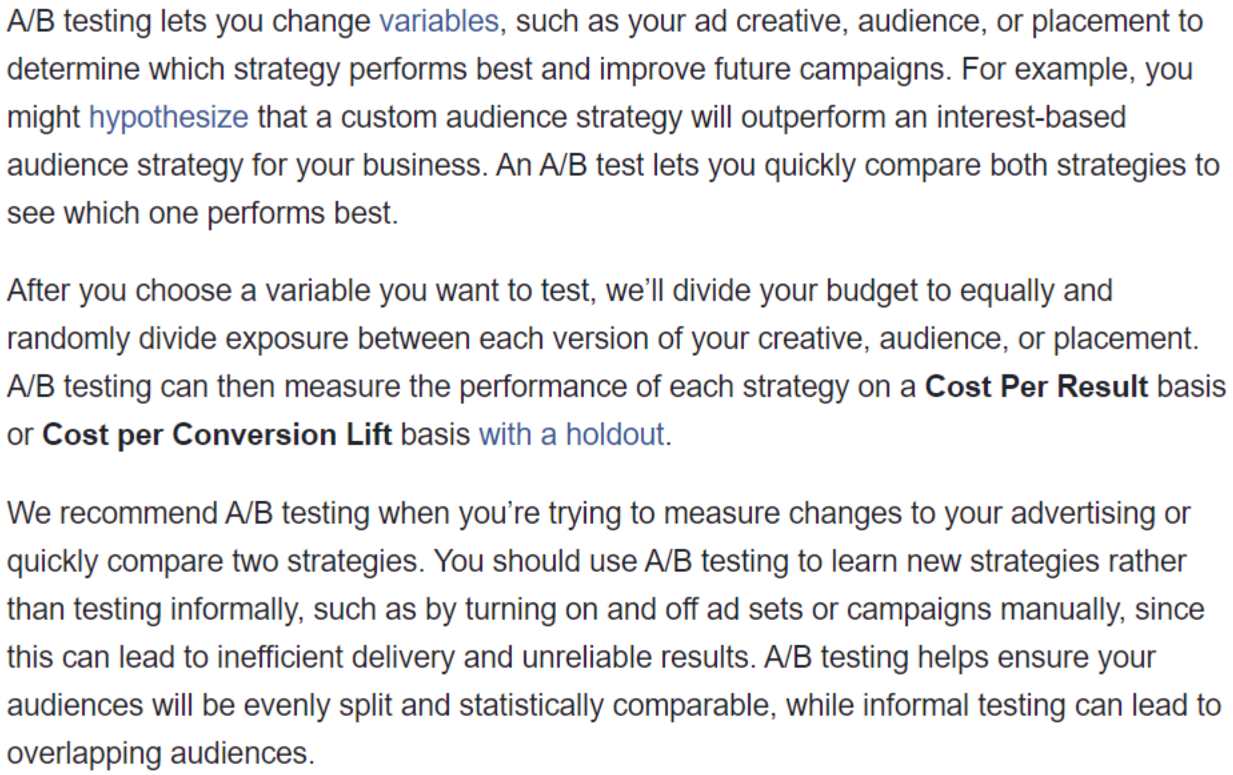
- allows effective management of conversations and your social media accounts. Also facilitates communication planning.

Sprout Social:

- a popular and user-friendly social media management software – contains tools such as social performance reporting, advanced social analytics, social monitoring and listening tools, and advanced social listening (at the moment does not include visual networks such as YouTube).

Analysis of Data from Social Media

- Qualitative content analysis
- Term frequencies
- Network analysis
- Natural Language Processing
- Supervised and unsupervised machine learning
- Hypothesis testing
  - Facebook message testing



A/B testing lets you change variables, such as your ad creative, audience, or placement to determine which strategy performs best and improve future campaigns. For example, you might hypothesize that a custom audience strategy will outperform an interest-based audience strategy for your business. An A/B test lets you quickly compare both strategies to see which one performs best.

After you choose a variable you want to test, we'll divide your budget to equally and randomly divide exposure between each version of your creative, audience, or placement. A/B testing can then measure the performance of each strategy on a **Cost Per Result** basis or **Cost per Conversion Lift** basis with a holdout.

We recommend A/B testing when you're trying to measure changes to your advertising or quickly compare two strategies. You should use A/B testing to learn new strategies rather than testing informally, such as by turning on and off ad sets or campaigns manually, since this can lead to inefficient delivery and unreliable results. A/B testing helps ensure your audiences will be evenly split and statistically comparable, while informal testing can lead to overlapping audiences.

Ru B, Yao L. A literature review of social media-based data mining for health outcomes research. Social Web and Health Research. 2019:1-4.

Who do you need on your team?

Qualitative and content analysis

- Qualitative methodologist

Data mining, network analysis, machine learning

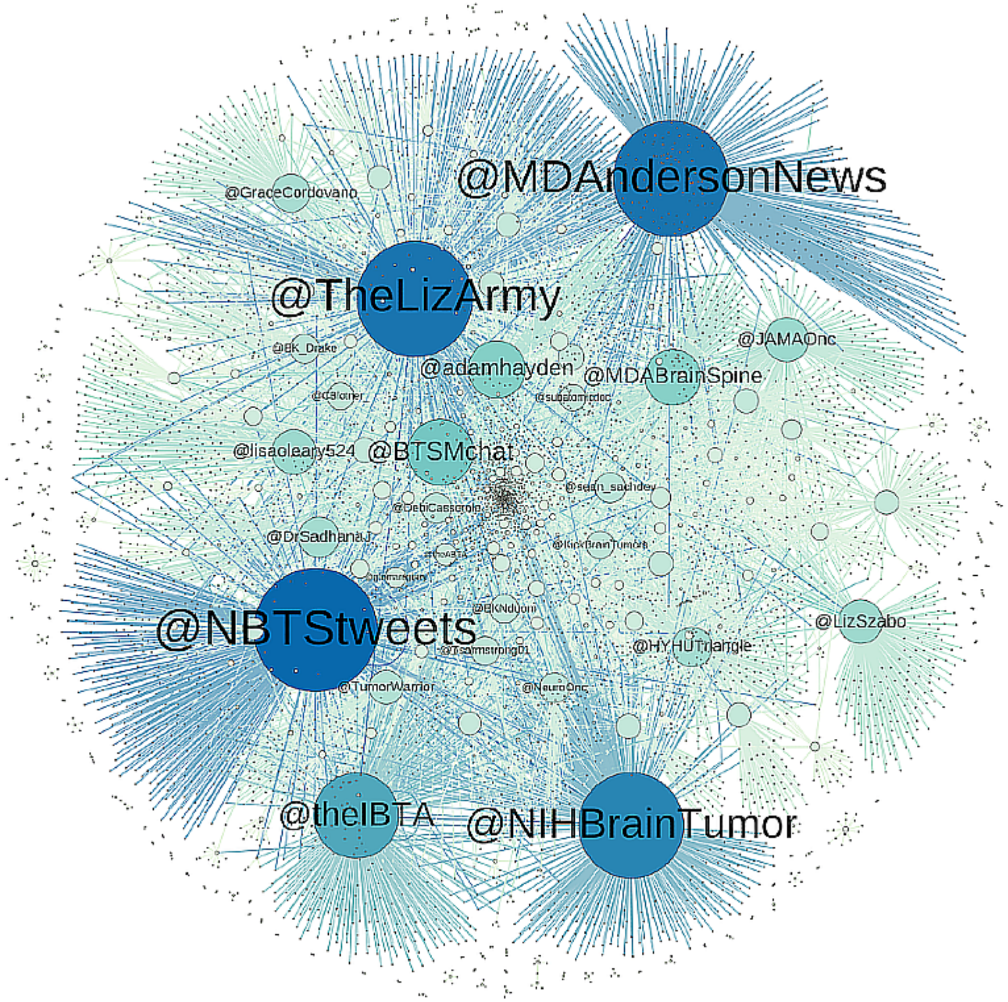
- Biostatistician

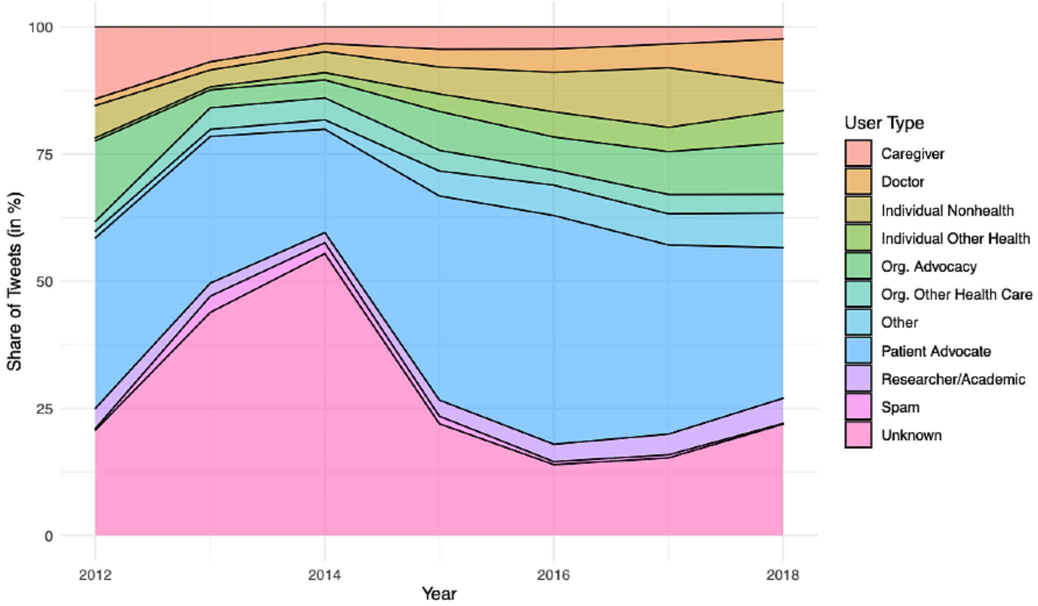
Social media engagement experts

Patient and community stakeholder representatives

- Familiar with/part of social media communities

Case example: Network analysis of #BTSM community





Feliciano J, Salmi L, Blotner C, Hayden A, Nduom E, Kwan B, Katz M, Claus E
Brain Tumor Discussions on Twitter (#BTSM): Social Network Analysis
J Med Internet Res 2020;22(10):e22005
URL: https://www.jmir.org/2020/10/e22005
DOI: 10.2196/22005

Activity #2: Social media data mining and analysis plan

- Consider the audience, social media platform, and research topic you considered in Activity #1.
- What data types might be available from that social media platform?
  - Text data
  - Structural data
  - Metadata
  - Other?
- How might you mine that data?
  - Manual
  - APi/third party
- How might you analyze that data?
- Who do you need on your team?

Thank you!

- Bethany Kwan, PhD, MSPH bethany.kwan@cuanschutz.edu @BethanyKwan
- Jenna Reno, PhD jenna.reno@cuanschutz.edu @RenJenPhD