social data collaboratory

Social Media and Public Opinion Research: A Road Map for Rigor, Transparency & Replicability

Sherry Emery, MBA, Ph.D.

AAPOR Webinar
January 24, 2019

NORC at the UNIVERSITY of CHICAGO

Outline

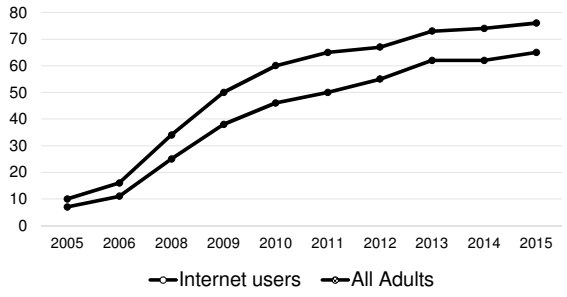
- Why social media data? How does this relate to Public Opinion Research?
- How social media data are **collected, filtered**, and **reported** can vary widely
- How to collect the data?
  - ➜ Use "search filter"
- How good are your data?
  - ➜ Assess the quality of search filter
- How to report about the data?
  - ➜ Reporting standard

NORC at the UNIVERSITY of CHICAGO

Why Social Media

- Sources of observing health attitude, intention, and behavior
- The real world for youth and young adults[1]
- Data dimensions to consider
  - Amount
  - Content
  - Source
  - Diffusion & network

% of all American adults and internet-using adults who use at least one SNS[2]



1. 90% of US 18-29 year olds use social media; 2. Source: Pew Research Center surveys

3

NORC at the UNIVERSITY of CHICAGO

Social Media Data 101

- Social media data Rule 1
  - Analysis is "easy"
  - Data collection and management represent at least 90% of the work
- Social media data Rule 2
  - How you collect (and report) the data WILL influence inferences/conclusions

4

NORC at the UNIVERSITY of CHICAGO

About Rule 1:
Anatomy of a Tweet (json format) (for example)


```
"id":"tag:search.twitter.com,2005:472724530994356224",
"objectType":"activity",
"actor":
    {
        "objectType":"person",
        "id":"id:twitter.com:22866991",
        "link":"http://www.twitter.com/joyoushealth",
        "displayName":"Joy McCarthy",
        "postedTime":"2009-03-05T02:32:24.000Z",
        "image":"https://pbs.twimg.com/profile_images/3588771321/0ffc5055c0dda79e7c32c86dea3edb83_normal.png",
        "summary":"Holistic Nutritionist CNP/RNCP, Author of JOYOUS HEALTH: Eat & Live Well Without Dieting. Natural Health expert for the Morning Show & Steven and Chris.",
        "links":
            [
                {
                    "href":"http://www.joyoushealth.ca","rel":"me"
                }
            ],
        "friendsCount":5828,
        "followersCount":18290,
        "listedCount":707,
        "statusesCount":32438,
        "twitterTimeZone":"Central Time (US & Canada)",
        "verified":false,
        "utcOffset":"-18000",
        "preferredUsername":"joyoushealth",
        "languages":
            [
                "en"
            ],
        "location":
            {
                "objectType":"place",
                "displayName":"Toronto, Canada"
            },
        "favoritesCount":324
    },
"verb":"post",
"postedTime":"2014-05-31T13:01:25.000Z",
"generator":
    {
        "displayName":"HootSuite",
        "link":"http://www.hootsuite.com"
    },
"provider":
    {
        "objectType":"service",
        "displayName":"Twitter",
        "link":"http://www.twitter.com"
    },
"link":"http://twitter.com/joyoushealth/statuses/472724530994356224",
"body":"Today is World No Tobacco Day! Check out these tips for breaking bad habits, whatever they may be, from @bitewithlove http://t.co/87vItPLtUl",
```


at the UNIVERSITY of CHICAGO

More About Rule 1: Example of SDC Workflow

Operationalize	Data Collection	Data Coding	Classification	Reporting
• Data viability • Key search rule generation • Refine search rule	• Estimate/download • Data ingestion • Language filter	• Exploration • Sampling • Human labeling • Reliability	• Classifier training • Classifier testing/tuning • Classifier validation	• Data export • Results exploration • Visualization

NORC at the UNIVERSITY of CHICAGO

How to Collect Social Media Data?

- Use search filter
- **Search Filter =** Keyword + Search Rule

  Keyword selection is not simple
  - Language and culture vary and change
  - Different language norms, tech constraints, and social functions across platforms

  Search rules for more focused search

Stryker, Wray, Hornik, & Yanovitzky. Journalism & Mass Communication Quarterly 2006 Jun 01;83(2):413-430.

7

NORC at the UNIVERSITY of CHICAGO

Not All Data Are Good Data!

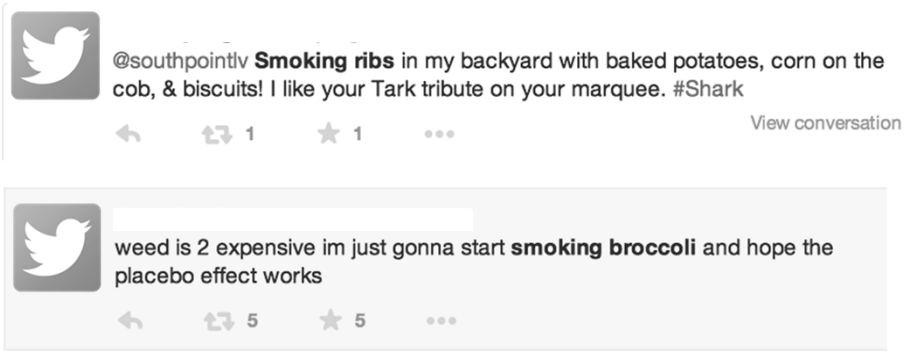
- "SMOKING" is an important keyword in tobacco research

8

NORC at the UNIVERSITY of CHICAGO

Not All Data Are Good Data!

- "SMOKING" is an important keyword in tobacco research



Need filter out irrelevant contents!
Otherwise biased inference

9

NORC at the UNIVERSITY of CHICAGO

Not All Data Are Good Data!

- Smoking cigarettes vs. marijuana ⇨ Different sentiment
- Quality of search filter ⇨ Validity of inference

- Mixture of good and bad data in massive quantity
- Use search filter to filter out irrelevant contents
- The search filter affects the amount and content of data

10

NORC at the UNIVERSITY of CHICAGO

Search Filter Development

NORC at the UNIVERSITY of CHICAGO

Search Filter Development

1. Build a list of **search keywords** (Stryker et al. 2006)
   - **I. Generate a list of candidate keywords** based on expert knowledge, systematic search of topic-related language, and other resources.
   - **II. Screen the keywords** by examining relevance and frequency.
   - **III. Discard keywords** that return posts with high proportion of irrelevant contents or relatively low frequency.
   - **IV. Add and screen new keywords** when new relevant terms and phrases emerge.

     Repeat II to IV until no more new relevant terms
2. Integrate keywords with **search rules**

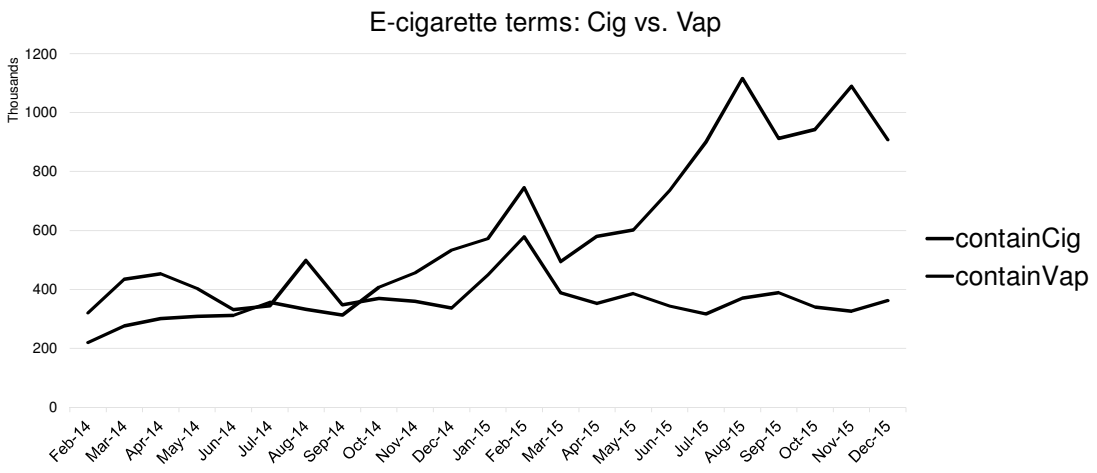
   e.g., “atomizer” NOT “perfume”

Stryker, Wray, Hornik, & Yanovitzky. Journalism & Mass Communication Quarterly 2006 Jun 01;83(2):413-430.

12

NORC at the UNIVERSITY of CHICAGO

Language changes! Stream vs. Historic



13

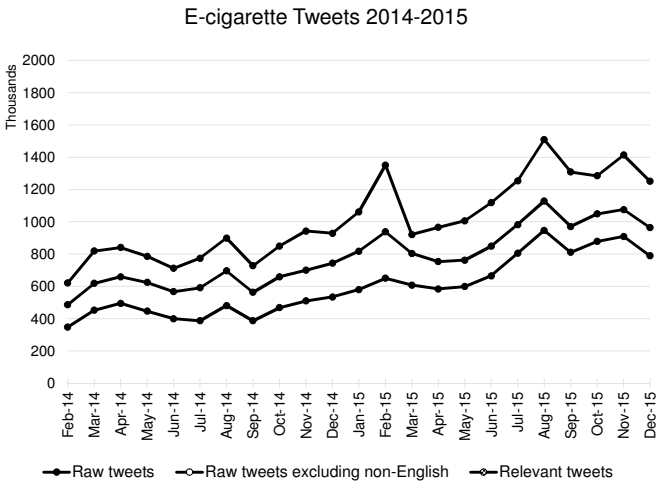
Language (English) Filter

- Language filter affects amount and content of the data
- Metadata: Twitter as an example
  - Actor's language: User's default language (if user provides)
  - Lang[3]: Machine-detected language[1] of the tweet text.
  - Gnip's language value[3]: Gnip's language detection. Language detection 1.0.
  - Twitter_lang: Language detection 2.0
- Machine learning algorithm
  - Language detection libraries of python
  - e.g. langid (Lui and Baldwin 2012), langdetect, pycld2 (Dick Sites[2])

1. BCP 47 language identifier ; 2. CLD: compact language detection; 3. Not provided from Gnip 2.0

14

Language (English) Filter



- Example:
  - E-cigarette tweets
  - Filter on actor's language, lang, Gnip's language value – English if 50% or more of available language fields indicates English
- Different filters
  - Change amount and content
  - Affect classifier training
- Report
  - whether language filter is used
  - how it is carried out.

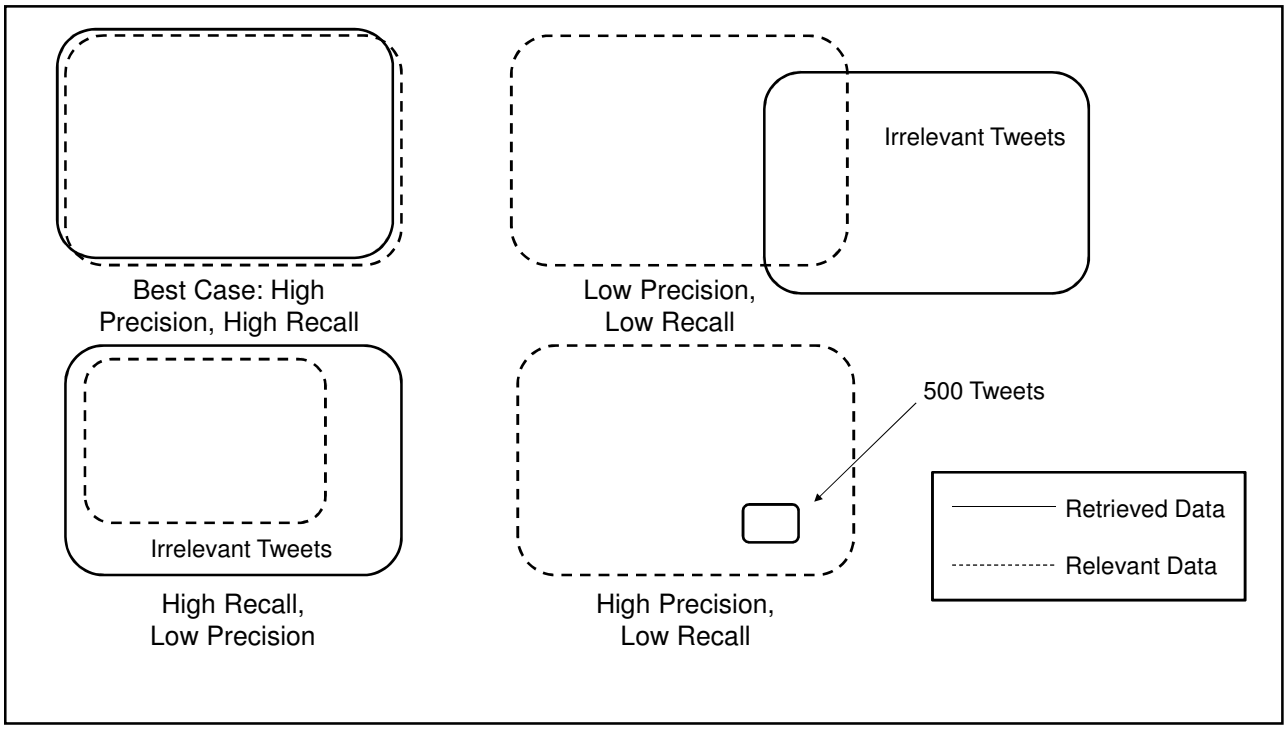
15

NORC at the UNIVERSITY of CHICAGO

Search Filter Assessment

NORC at the UNIVERSITY of CHICAGO

Garbage In, Garbage Out



Retrieval Data Quality Measures

- **Recall = a/(a+c)**
  How much of the relevant messages is retrieved?
- **Precision = a/(a+b)**
  How much of the retrieved messages is relevant?
- F-Score
- Specificity = d/(b+d)
- Negative predictive value = d/(c+d)

	Human Coding	
Search Filter	Coded Relevant	Coded Not Relevant
Retrieved	a (TP)	b (FP)
Not Retrieved	c (FN)	d (TN)

19

NORC at the UNIVERSITY of CHICAGO

Retrieval Data Quality Measures

$$Recall = \frac{(precision)P(retrieved)}{(precision)P(retrieved) + P(relevant|unretr)[1 - P(retrieved)]}$$

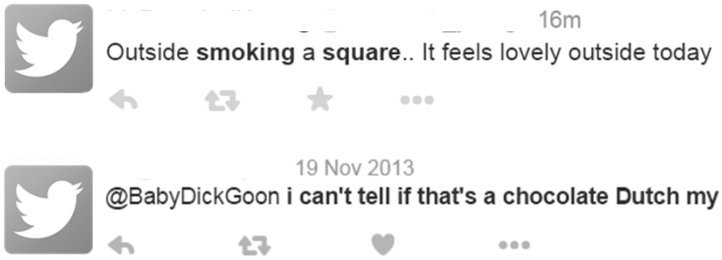
Retrieval recall and precision ≠ **Classifier** recall and precision

Trade-off between recall and precision

20

NORC at the UNIVERSITY of CHICAGO

Humans Can Make Errors



21

NORC at the UNIVERSITY of CHICAGO

Humans Can Make Errors

22

NORC at the UNIVERSITY of CHICAGO

Humans Can Make Errors: Subject Matter Expertise is Necessary

Human coding may not be a gold standard

- Ambiguous language
- Short messages
- Creative terms, unknown acronyms, slang, and colloquial
- Misspelling
- Fatigue

- Human coding has <100% recall, <100% specificity
  - ➜ Biased assessment of search filter quality (Staquet et al. 1981)

Staquet et al. Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis 1981;34(12):599-610.

23

NORC at the UNIVERSITY of CHICAGO

Concrete Examples: E-Cigarette Messages on Twitter

NORC at the UNIVERSITY of CHICAGO

E-Cigarette Search Filter: depends on your RQ

Category	Keywords and Rules
Variations and alternative terms of e-cigarettes	ecig(s), "e cig(s)", e-cig(s), ecigarette(s), e-cigarette(s), ehokah, e-hookah, ejuice(s), e-juice(s), eliquid(s), e-liquid(s), esmokes, e-smoke(s), lavatube(s), smokestik(s)
E-cigarette device parts	cartomizer(s), atomizer(s), NOT perfume
Specific brand of e-cigarettes	@blucig, from:blucig, blu cig, blu cigarette, njoy cig, njoy cigarette, "green smoke" "south beach smoke", eversmoke, "Joye 510", joye510, joyetech, logicecig, logicecigs, smartsmoker, "v2 cig(s), v2cig(s), zerocig(s)
Behavior	vaper(s), vaping

25

NORC at the UNIVERSITY of CHICAGO

Data Collection Experiment: How You Get Data Matters

- Tweets posted from Jan 15 – Jun 15, 2015 via 3 APIs
- Consistent keywords across the APIs
- Keywords for the three topics

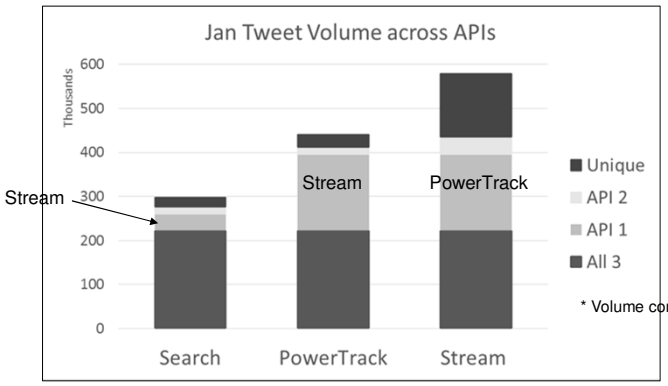
Tobacco	E-Cigarettes	Anti-Smoking
cig hookah(s) tobacco shisha rello(s) cigarillo(s) skoal snus Marlboros	ecig vaper(s) Vaping eliquid(s) e-liquid(s) cartomizer	@drfriedencdc smokefree secondhand smoke quitline(s) #quitnow

26

NORC at the UNIVERSITY of CHICAGO

Tweet Volume

- The tweets largely overlapped between the 3 APIs.
- But, each API retrieved unique tweets too.
- Unique tweets may result in different research conclusion.



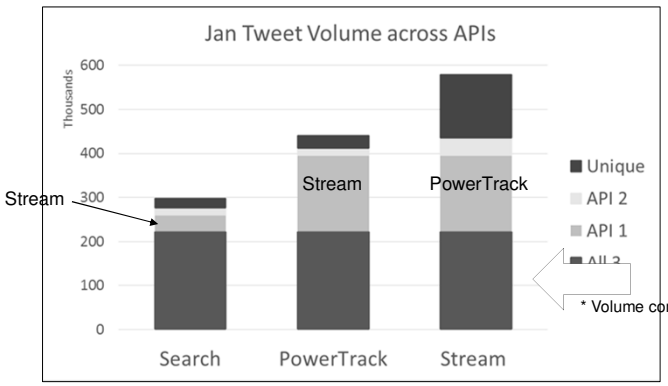
* Volume combining all three topics

27

NORC at the UNIVERSITY of CHICAGO

Tweet Volume

- The tweets largely overlapped between the 3 APIs.
- But, each API retrieved unique tweets too.
- Unique tweets may result in different research conclusion.



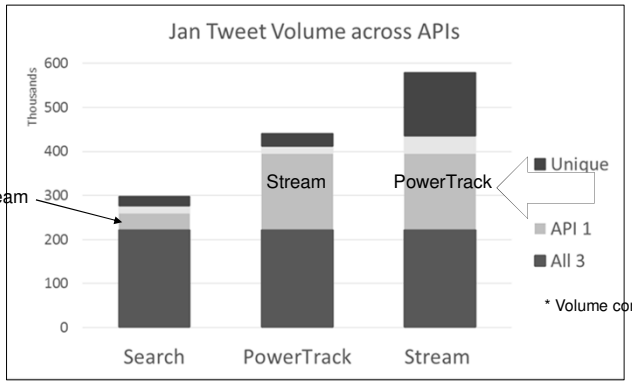
* Volume combining all three topics

28

NORC at the UNIVERSITY of CHICAGO

Tweet Volume

- The tweets largely overlapped between the 3 APIs.
- But, each API retrieved unique tweets too.
- Unique tweets may result in different research conclusion.

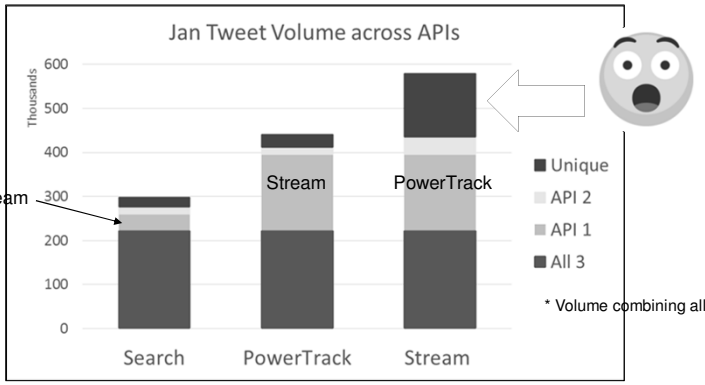


* Volume combining all three topics

29

Tweet Volume

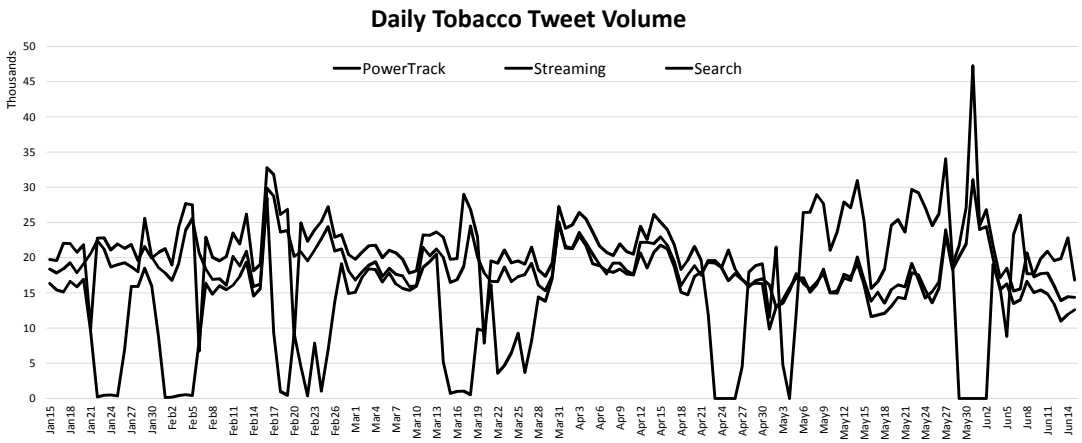
- The tweets largely overlapped between the 3 APIs.
- But, each API retrieved unique tweets too.
- Unique tweets may result in different research conclusion.



* Volume combining all three topics

30

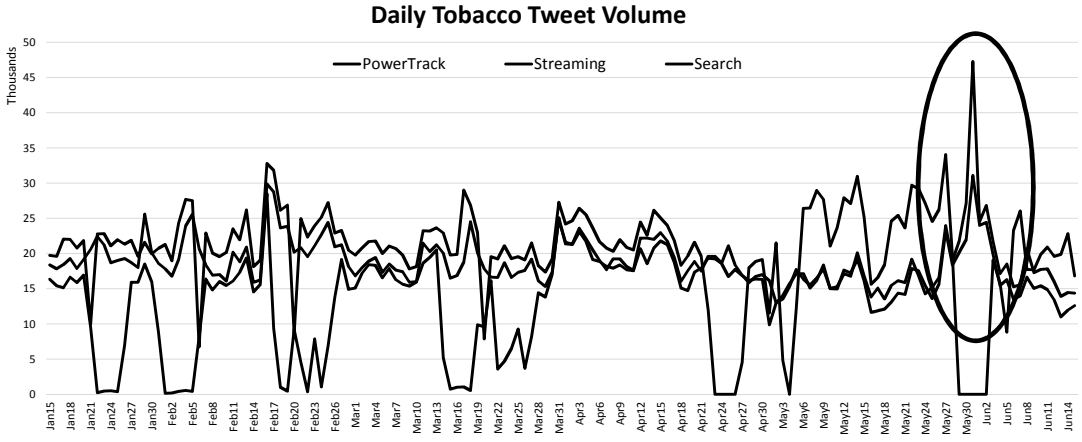
Tobacco Volume



NORC at the UNIVERSITY of CHICAGO

31

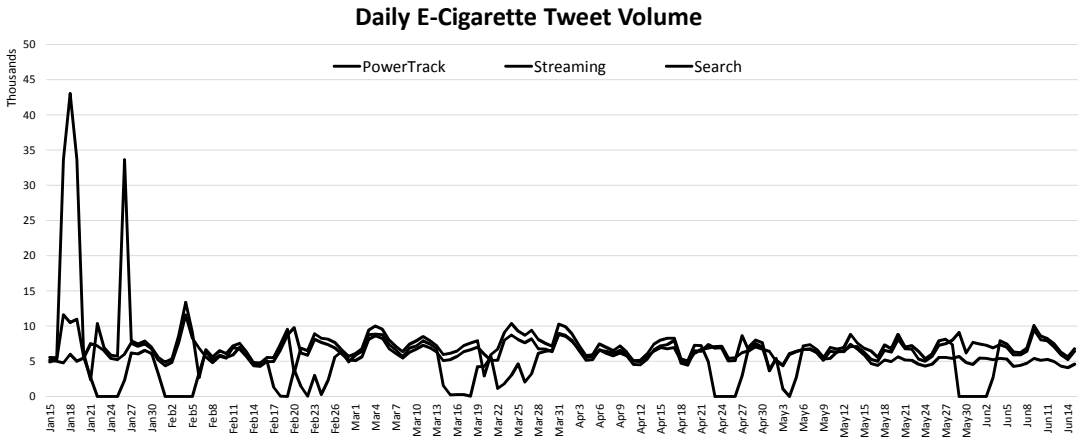
Tobacco Volume



NORC at the UNIVERSITY of CHICAGO

32

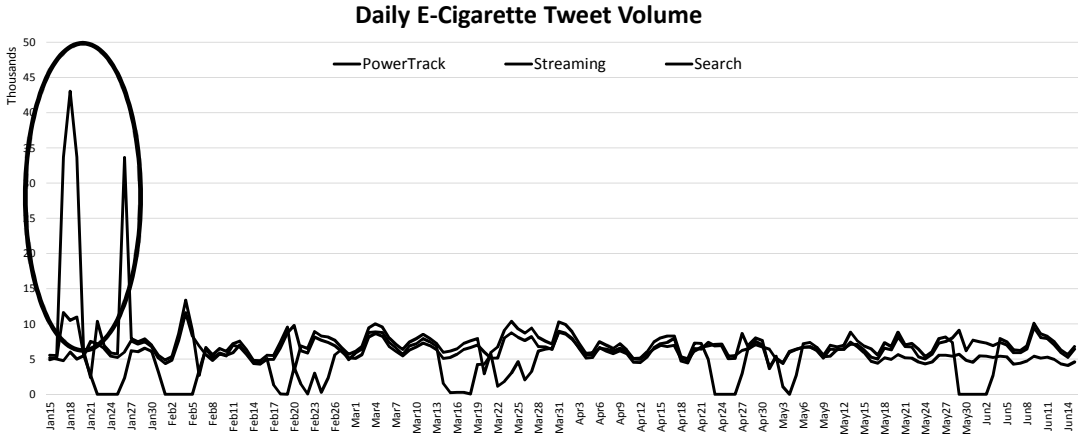
E-cigarette



NORC at the UNIVERSITY of CHICAGO

33

E-cigarette



What are the **sources** and **content** of these tweets?

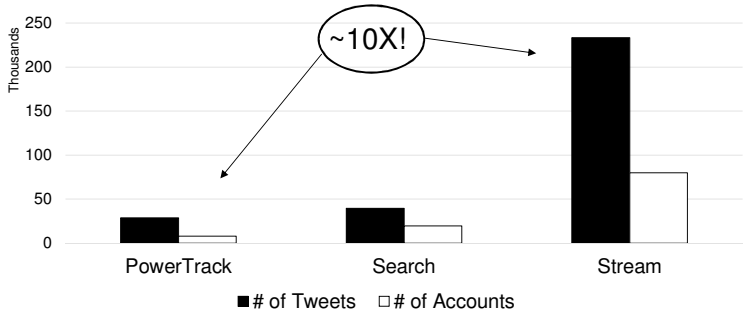
NORC at the UNIVERSITY of CHICAGO

34

E-Cigarette Tweets & Accounts

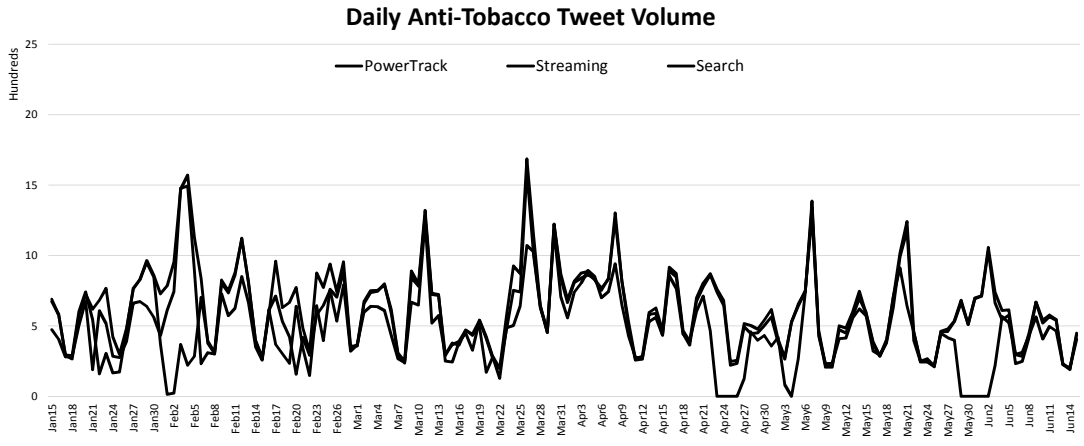
- Jan e-cigarette tweets

	PowerTrack	Search	Stream
# of Tweets	28,785	39,569	233,372
# of Accounts	8,130	19,568	80,037
Avg # of Tweets / Account	3	2	3



35

Anti-smoking



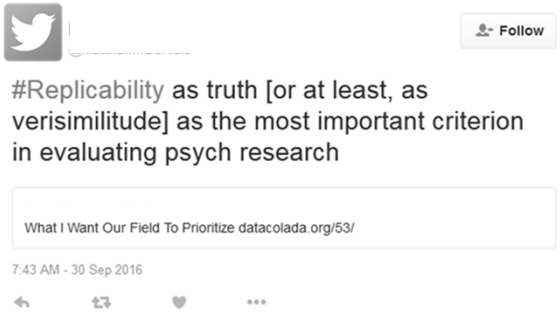
36

Reporting Standard For Social Media Data Use

NORC at the UNIVERSITY of CHICAGO

Minimum Disclosure

- **Data**
- **Development of search filter**
- **Assessment of search filter**



38

NORC at the UNIVERSITY of CHICAGO

Minimum Disclosure

- **Data**
  - Scope of the study
    - Definition of e-cigarette posts
  - Platform, time frame
    - Twitter, Oct 1-Oct 31 2015
  - Source or method used to access data
    - Twitter Streaming API
- **Development of search filter**
  - Keywords generation and refinement
    - Criteria to drop or add keywords
    - Precision and frequency of keywords.
  - List of final keywords and search rules
    - Acceptable signal-to-noise ratio
    - Research topic determines the definition of “noise”

NORC at the UNIVERSITY of CHICAGO

Minimum Disclosure

- **Assessment of search filter**
  - Assumption about human coding
    - Human coding as gold standard
  - Sampling frame and size for human coding
    - Proportionate stratified sampling, oversample of certain keywords, etc.
  - Quality measures
    - Inter-coder reliability
    - Retrieval precision & recall
  - Classifier training, if used to retrieve relevant data
    - Classifier precision & recall

NORC at the UNIVERSITY of CHICAGO

Preferred Disclosure

- Source code
- Model equations
- Coding/labeling instructions manual
- Ethical concerns/need for IRB review
- Data decay assessment

NORC at the UNIVERSITY of CHICAGO

Summary & Discussion

- Social media are valuable and alternative or complementary data sources for public opinion and behavioral research
- Collecting social media data that are both **precise** and **accurate** is critical to reaching correct research conclusions
- Need a standard of reporting social media data collection, filtering and quality, so that quality of data retrieved and analyses may be compared across different studies
- Our method to develop search filter and assess its quality can be adapted to other text-based social media data
- Future research: semi-automation of keyword selection

NORC at the UNIVERSITY of CHICAGO

Thank You!