

Political Analysis of Social Media Data

Keyword Expansion

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- Keyword expansion
- Video lectures & exercises

Keywords in document selection

- How can we select a set of social media posts for analysis from a large unstructured set?
- How can we rapidly code social media posts to a specific topic or set of topics?

Computer-Assisted Keyword and Document Set Discovery from Unstructured Text

Gary King Harvard University

Patrick Lam Thresher

Margaret E. Roberts University of California, San Diego

Abstract: *The (unheralded) first step in many applications of automated text analysis involves selecting keywords to choose documents from a large text corpus for further study. Although all substantive results depend on this choice, researchers usually pick keywords in ad hoc ways that are far from optimal and usually biased. Most seem to think that keyword selection is easy, since they do Google searches every day, but we demonstrate that humans perform exceedingly poorly at this basic task. We offer a better approach, one that also can help with following conversations where participants rapidly innovate language to evade authorities, seek political advantage, or express creativity; generic web searching; eDiscovery; look-alike modeling; industry and intelligence analysis; and sentiment and topic analysis. We develop a computer-assisted (as opposed to fully automated or human-only) statistical approach that suggests keywords from available text without needing structured data as inputs. This framing poses the statistical problem in a new way, which leads to a widely applicable algorithm. Our specific approach is based on training classifiers, extracting information from (rather than correcting) their mistakes, and summarizing results with easy-to-understand Boolean search strings. We illustrate how the technique works with analyses of English texts about the Boston Marathon bombings, Chinese social media posts designed to evade censorship, and others.*

Many applications require keyword selection

- The words that identify topics can over time (conversational drift)
 - “gay marriage” → “marriage equality”
 - “pro-choice” → “reproductive rights”
 - “late-term abortion” → “partial-birth abortion”
- Governments that censor posts lead to creative word use to avoid the censors
 - “Ai Weiwei” (in Chinese) → “AWW”
- Statistical analysis often requires only a subset of documents
 - Keywords often used to select these documents

General benefits to keyword selection

- Intuitive
- Fast
- Can be improved simply with more effort
- Can augment statistical approach to correct mistakes
- Especially useful for rare events or topics
 - Among millions of posts, cannot easily manually code posts on a topic if there are very few posts about it

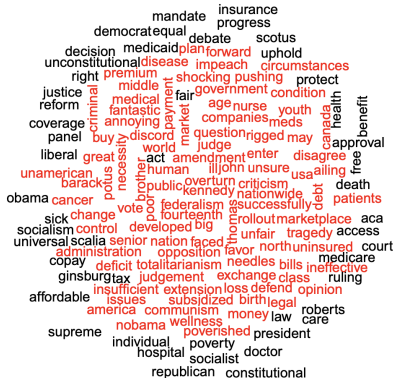
Humans are bad at keyword selection

- Everyone can come up with a few keywords
- But are unreliable
 - Few people would select the same keywords
- Experiment to show this...

Experiment

- Ask 43 undergraduates the following:
 - We have 10,000 twitter posts, each containing the word “healthcare,” from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obamacare.
- Ask students to do the same for an analogous task regarding the Boston Marathon bombings

Little overlap between 43 sets of keywords



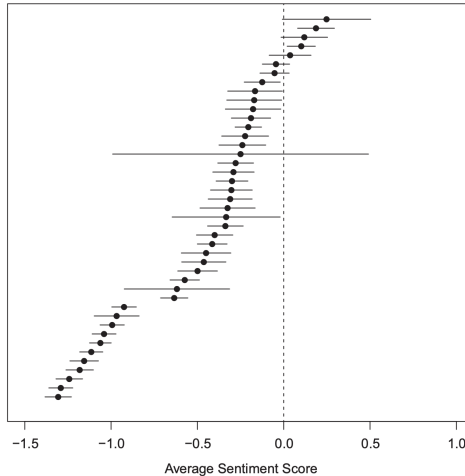
A Obamacare



B Boston Bombings

Note: Word clouds of keywords were selected by human users; those selected by one and only one respondent are in red (or gray if printed in black and white). The position of each word within the cloud is arbitrary.

Sentiment analysis of 43 sets are very different



Note: Each document set was selected by a different keyword list, with point estimates (as dots) and 95% confidence intervals (horizontal lines) shown.

So what do we do?

- Use the fact that people are bad at thinking of keywords, but are excellent at determining whether a keyword is good if they see it
- Use a statistical keyword discovery keyword technique to prompt a researcher with candidate keywords
- Iteratively build a keyword list

Keyword expansion

- **Reference Set**

- Documents classified as being examples of our topic, person, event, organization etc. of interest

- **Search Set**

- Documents which aren't classified
- Some may be from the topic, person, event, organization etc. of interest, whereas other will not be

- **Target set**

- Documents in the Search Set that are from the topic, person, event, organization etc. of interest
- Our goal is to discover these documents and add them to the reference set

Example

- All tweets from Members of Congress from 2019-2020
- Want to find all tweets related to COVID-19
- Can define the **reference set** from the start as any tweet that contains the keyword “covid-19”
- All other tweets (i.e. those without the keyword “covid-19”) are then in the **search set**
- Goal is to find a **target set** within the search set that are tweets about COVID-19, but just don't happen to include the keyword “covid-19”

Methodology

- Find documents in the search set that look like those in the reference set
- Are words in the reference set that distinguish them from all the documents in the search set
- A tweet in the reference cite might read:
 - “I support the emergency relief funding bill to combat the **COVID-19 pandemic**, and urge my colleagues to vote in favor **#covid19**”
 - This tweet is in the reference set because it includes the term “COVID-19”
 - But *co-occurring terms* (e.g. “pandemic”, “#covid19”) will be predictive of tweets that are in the reference set

TABLE 1 The Keyword Algorithm

-
1. Define a reference set R and search set S .
 2. Using a diverse set of classifiers, partition all documents in S into two groups: T and $S \setminus T$, as follows:
 - (a) Define a training set by drawing a random sample from R and S .
 - (b) Fit one or more classifiers to the training set using as the outcome whether each document is in R or S .
 - (c) Use parameters from classifiers fit to the training set to estimate the predicted probability of R membership for each document in S . (Of course, every document *is* in S , and so the prediction mistakes can be highly informative.)
 - (d) Aggregate predicted probabilities or classifications into a single score (indicating probability of membership in T) for each document in S .
 - (e) Partition S into T and $S \setminus T$ based on the score for each document and a user-chosen threshold.
 3. Find keywords that best classify documents into either T or $S \setminus T$, as follows:
 - (a) Generate a set of potential keywords by mining S for all words that occur above a chosen frequency threshold, K_S .
 - (b) Decide whether each keyword $k \in K_S$ characterizes T or $S \setminus T$ better, by comparing the proportion of documents containing k in T with the proportion of documents containing k in $S \setminus T$.
 - (c) Rank keywords characterizing T by a statistical likelihood score that measures how well the keyword discriminates T from $S \setminus T$. Do the analogous ranking for keywords characterizing $S \setminus T$.
 4. Present keywords in two lists to the user, to iterate and choose words of interest or for use in building a document retrieval query.
 5. If sufficient computational power is available, rerun Steps 1–4 every time the user makes a measurable decision, such as adding a keyword to Q_T to improve the lists of keywords to consider.

Step 1: Define a reference set R and search set S

- Select a minimal set of keywords that will classify a topic, event, person, organization, etc. of interest
- e.g. You have a dataset of tweets from Members of Congress, and want to find all posts related to the COVID-19 pandemic
- Define as your **reference set** any tweet that contains the keyword “covid-19”
- Your **search set** contains all tweets without the keyword “covid-19”

Step 2: Use supervised learning model to partition the search set into a target set (T) and non-target set ($S \setminus T$)

- Combine reference set and random sample of the search set
 - Could use *all* data in the reference and search sets, but that might take a very long time to fit a model
- Fit a machine learning model with the text as data to predict membership in the reference set
- Use the fitted model to predict membership in the reference set for *all* data in the search set
- Based on the predictions, separate the search set into a target set (T) and a non-target set ($S \setminus T$)
- Why? Documents in the target set will be those that look like those in the reference set

Step 3: Find keywords that best classify documents into either T or $S \setminus T$

- Find all keywords in the search set
- Compare proportion of each keyword in the target T set and not-target set $S \setminus T$ to determine which set it fits best with
- Rank the keywords by a likelihood measure to determine which keywords best discriminate documents in T from documents in $S \setminus T$

Step 4: Examine keywords in two lists to discover new ones to expand your original keyword list

- Look at, say, the top 25 keywords that best discriminate T from documents in $S \setminus T$
- Examine these keywords in the documents to understand their use
- Add relevant keywords to the list that defines the reference set

Example keywords lists

TABLE 3 Top 25 Keywords in the Boston Bombings Validation Example

Target Keywords	Nontarget Keywords
peopl, thought, prayforboston, prayer, fbi, affect, arrest, cnn, pray, video, obama, made, bomb, bostonmarathon, heart, injur, attack, releas, victim, terrorist, sad, news, sick, rip, investig	marathon, celtic, game, miami, weekend heat, tsarnaev, new, play, red watertown, open, back, sox, job mom, tonight, win, fan, monday bruin, reaction, liam, tomorrow, payn

Note: The validation example is from the target T and nontarget $S \setminus T$ search set lists produced by a single noniterative run of the algorithm, without human input.

Step 5: Iterate


- Re-run Steps 1 thru 4 with the new set of keywords to see if you can discover more

To reiterate

- **Step 1:** Use keywords to classify posts into a reference set and a search set
- **Step 2:** With text from each social media post, use a supervised machine learning model to predict which posts in the search set look like they belong to the reference set
- **Step 3:** Measure how well the words in the predicted target and not-target sets discriminate between the two sets
- **Step 4:** Examine the top keywords to discover candidate keywords to add to the original list
- **Step 5:** Iterate

Examples of applications

Political Knowledge and Misinformation in the Era of Social Media: Evidence From the 2015 UK Election

Kevin Munger^{1*} , Patrick J. Egan², Jonathan Nagler³, Jonathan Ronen⁴ and Joshua Tucker⁵

¹Department of Political Science and Social Data Analytics, Penn State University, State College, PA, USA, ²Department of Politics, New York University, USA, ³Department of Politics, Center of Data Science, and Social Media and Political Participation Laboratory, New York University, USA, ⁴Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany and ⁵Department of Politics, Department of Russian and Slavic Studies, Center of Data Science, and Social Media and Political Participation Laboratory, New York University, USA

*Corresponding author. E-mail: kmm7999@psu.edu

(Received 29 July 2019; revised 20 December 2019; accepted 24 March 2020; first published online 22 December 2020)

Abstract

Does social media educate voters, or mislead them? This study measures changes in political knowledge among a panel of voters surveyed during the 2015 UK general election campaign while monitoring the political information to which they were exposed on the Twitter social media platform. The study's panel design permits identification of the effect of information exposure on changes in political knowledge. Twitter use led to higher levels of knowledge about politics and public affairs, as information from news media improved knowledge of politically relevant facts, and messages sent by political parties increased knowledge of party platforms. But in a troubling demonstration of campaigns' ability to manipulate knowledge, messages from the parties also shifted voters' assessments of the economy and immigration in directions favorable to the parties' platforms, leaving some voters with beliefs further from the truth at the end of the campaign than they were at its beginning.

What is the effect of exposure to social media posts about a given topic on knowledge about that issues?

- Survey-linked social media data
- Know which users were exposed to which tweets
- Observe knowledge before and after exposure (panel data)

Topic measurement problem:

- How do we know whether a tweet is about a specific topic?
 - Ties to the EU
 - Immigration
 - Islamic State
 - The economy
- Need to measure topics of tweets on social media

Final keyword list

Table 1. Top terms pertaining to the topics under study

Ties to the EU	Immigration	ISIS	Economy
brexit	immigration	isis	cuts
no2eu	detention	jihad	benefits
betteroffout	uncontrolled	kobane	budget
eureferendum	ukip	islam	welfare
eu	obama	iraq	vat
euref	farage	syria	osborne
grexit	policy	fundamentalist	tax
scoxit	controls	iraqi	tory
stayineu	reform	mosul	disabled
flexcit	immigrants	kurds	tories
referendum	illegal	kurdish	spending
ciuriak	eu	quran	austerity
yestoeu	labour	ypg	cut
ivotedukip	yarl	raqqa	reform
nothankeu	mug	palmyra	benefit
nox	bbcqt	islamic	ids
spexit	mass	twitterkurds	nhs
nunelected	bordersecurity	fighters	ifs
efta	nigel	ramadi	labour
frexit	ncustoms	muslim	disability
uk	time4atimelimit	kobani	budget2015
scaremongers	Noamnesty	beheading	Healthh
annually	Debate	bb4sp	Cameron
irexit	Immigrant	beheadings	Reforms
britty	leadersdebate	peshmerga	Government

Note: examples of the terms we found to tend to co-occur with our anchor terms, allowing us to identify the terms that comprise the topics of interest. These are the top twenty-five terms per topic.

Results

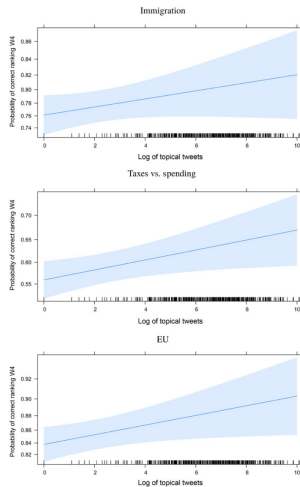


Figure 3. Effect of topical tweets on probability of correctly identifying relative party placement by issue in Wave 4
 Note: plots of the estimated effects of topical tweets received on the probability that the subject correctly ranked the four parties on that topic in Wave 4 of the survey.
 Source: Table 8 in Appendix C (page 6).

POLITICAL SCIENCE

How the news media activate public expression and influence national agendas

Gary King,^{1*} Benjamin Schneer,² Ariel White³

We demonstrate that exposure to the news media causes Americans to take public stands on specific issues, join national policy conversations, and express themselves publicly—all key components of democratic politics—more often than they would otherwise. After recruiting 48 mostly small media outlets, we chose groups of these outlets to write and publish articles on subjects we approved, on dates we randomly assigned. We estimated the causal effect on proximal measures, such as website pageviews and Twitter discussion of the articles' specific subjects, and distal ones, such as national Twitter conversation in broad policy areas. Our intervention increased discussion in each broad policy area by ~62.7% (relative to a day's volume), accounting for 13,166 additional posts over the treatment week, with similar effects across population subgroups.

What is the effect of exposure to news media on a topic on discussion about that issue?

- Large-scale field experiment
- Coordinate with news media to time the publishing of stories about a given topic
- Need to measure the amount of topic discussion on social media

Keyword expansion

(b) Algorithmic Approach

- i. Apply the keyword generation procedure described in King, Lam, and Roberts (54) to generate additional potential keywords. In brief, when adding a new set of posts, run this algorithm on a random sample of posts drawn from the existing body of posts pooled with the newly discovered posts. The model is trained based on which posts are in the existing set of posts and which set are in the newly added set. Then fit the model on the remaining newly added posts to identify posts that are similar to those from the existing set of posts. Extract new, relevant keywords from these newly identified “on-topic” posts based on how frequently the keywords are used in the newly identified posts. The main feature of the algorithm is that it learns from, rather than correcting, mistakes in classifying posts as relevant and mines keywords from those mistakes.
- (c) After generating lists of candidate keywords, we then selected a final set of keywords based on the pool of keywords generated from the methods described above.

Results

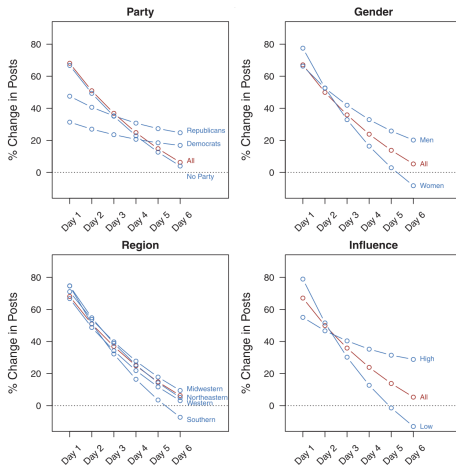


Fig. 3. Causal effect of the news media on the percent change in social media posts by political party, gender, region, and influence on Twitter. Axes are defined as in Fig. 2A.

Do Violent Protests Affect Expressions of Party Identity? Evidence from the Capitol Insurrection

GREGORY EADY *University of Copenhagen, Denmark*

FREDERIK HJORTH *University of Copenhagen, Denmark*



PETER THISTED DINESEN *University College London, United Kingdom, and University of Copenhagen, Denmark*

The insurrection at the United States Capitol on January 6, 2021, was the most dramatic contemporary manifestation of deep political polarization in the United States. Recent research shows that violent protests shape political behavior and attachments, but several questions remain unanswered. Using day-level panel data from a large sample of US social media users to track changes in the identities expressed in their Twitter biographies, we show that the Capitol insurrection caused a large-scale decrease in outward expressions of identification with the Republican Party and Donald Trump, with no indication of reidentification in the weeks that followed. This finding suggests that there are limits to party loyalty: a violent attack on democratic institutions sets boundaries on partisanship, even among avowed partisans. Furthermore, the finding that political violence can deflect copartisans carries the potential positive democratic implication that those who encourage or associate themselves with such violence pay a political cost.

What is the effect of the Capitol Insurrection on expressions of Republican partisanship?

- Difference-in-differences analysis of social media bios
- Estimate difference in removal of partisan terms from a bio among Republicans relative to Democrats
- Need to measure explicit expressions of partisanship

Table A1: Keyword target list based on republican seed word

	Feature	Likelihood	p	n_target	n_reference
1	!	189.61	0.00	210.00	230.00
2		125.85	0.00	84.00	49.00
3		81.77	0.00	38.00	10.00
4	#maga	70.38	0.00	34.00	10.00
5	.	67.88	0.00	790.00	2432.00
6	trump	58.54	0.00	28.00	8.00
7	conserv	58.39	0.00	25.00	5.00
8	love	54.93	0.00	74.00	95.00
9	god	52.95	0.00	40.00	28.00
10	,	40.38	0.00	629.00	2025.00
11	#resist	32.39	0.00	25.00	18.00
12	maga	31.12	0.00	11.00	0.00
13	america	29.80	0.00	16.00	5.00
14	#kag	29.64	0.00	12.00	1.00
15	#trump2020	29.16	0.00	14.00	3.00
16	wife	26.59	0.00	35.00	44.00
17	dog	26.44	0.00	26.00	25.00
18	vote	26.22	0.00	12.00	2.00
19	christian	23.73	0.00	17.00	11.00
20	mother	21.86	0.00	26.00	30.00

Results

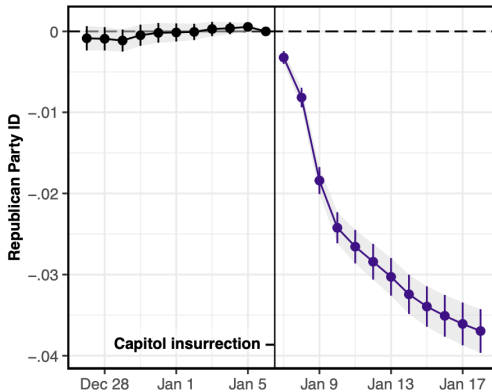


Figure 2: Event study estimates (with 95% CIs). Data were collected each morning, and thus observations on January 6 (before vertical line) are pre-insurrection. Standard errors are clustered at the user level.

Conclusions

- Basic idea is keyword expansion is quite straightforward
 - Learn from the “mistakes” of a supervised learning model
- Many steps, but once one has the basic idea, it's very easy to understand
- Shown to provide high precision compared to keyword selection by researchers off the top of their head