INTEGRATING MULTIPLE SOURCES OF CUSTOMER FEEDBACK DATA

Office of Publications and Special Studies
Office of Technology and Survey Processing

CIVIC DIGITAL FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

DANIEL ZHAO
Yale College ‘21
Statistics & Data Science / Global Affairs

SCOPE

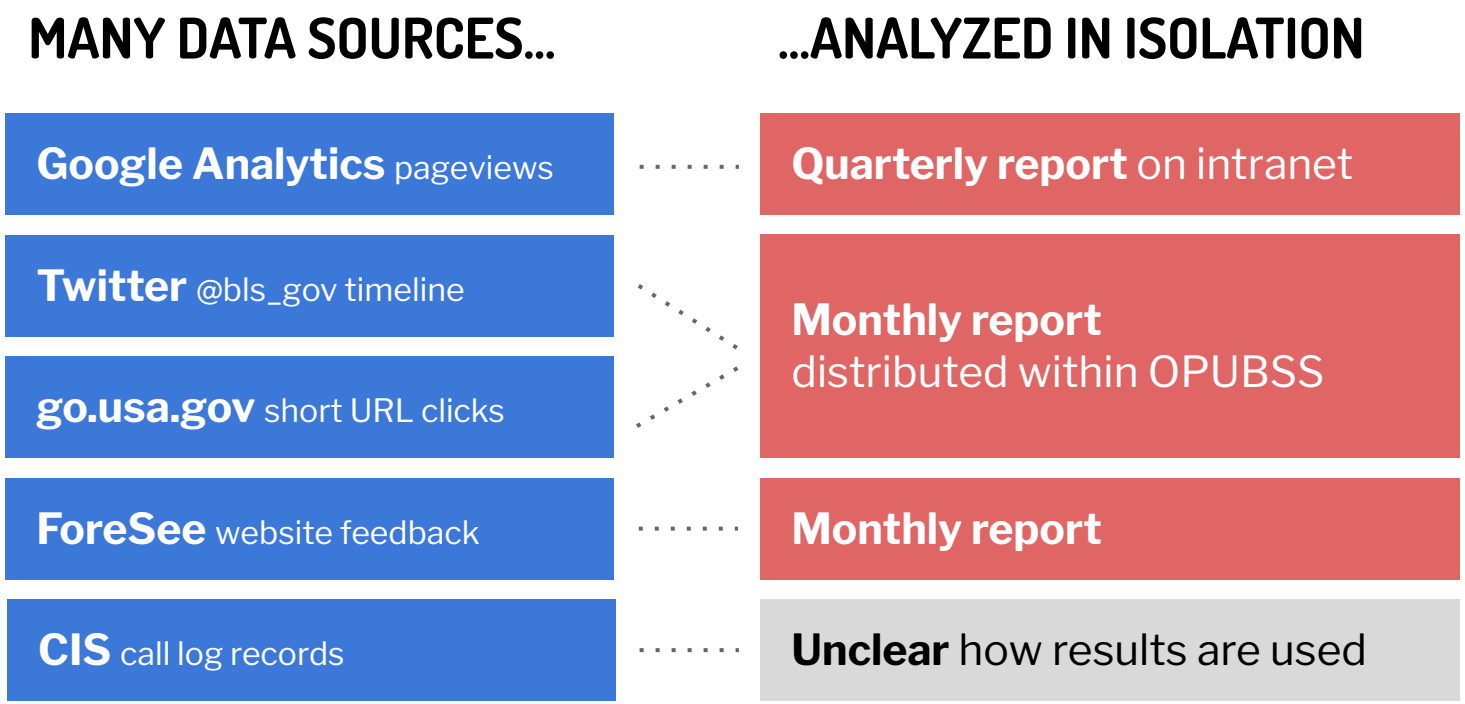
Customer **feedback** **data**

users of BLS data

anything that tells BLS what data is being used + how

numbers, text

FROM SILOED REPORTING...



...TO A UNIFIED FRAMEWORK

DATA SOURCES

1. **Google Analytics**
   Pageviews, users
2. **CIS**
   Internal call log records
3. **Twitter**
   Timeline, mentions
4. **go.usa.gov**
   Short URL clicks
5. **ForeSee** or new tool?
   Website feedback
6. **Meltwater**
   News aggregator

PROCESSING

1. **Extracting webpage info**
   Web scraping (BeautifulSoup)
2. **Predicting BLS program**
   Rule-based, spaCy
3. **Text & tweet cleaning**
   stringr, regex
4. **Sentiment analysis**
   To be implemented
5. **Prediction & classification**
   Logistic regression

MODULES

Methods to merge all data sources and extract relevant metrics
(varies by use case)

REPORTING

SAMPLE PIPELINE

Which BLS publications are most-viewed?

DATA SOURCES

1. **Google Analytics**
   Pageviews, users
2. **CIS**
   Internal call log records
3. **Twitter**
   Timeline, mentions
4. **go.usa.gov**
   Short URL clicks
5. **ForeSee** or new tool?
   Website feedback
6. **Meltwater**
   News aggregator

PROCESSING

1. **Extracting webpage info**
   Web scraping (BeautifulSoup)
2. **Predicting BLS program**
   Rule-based, spaCy
3. **Text & tweet cleaning**
   stringr, regex
4. **Sentiment analysis**
   To be implemented
5. **Prediction & classification**
   Logistic regression

MODULES

Link webpages that appear across the data sources

REPORTING

SAMPLE PIPELINE

Which BLS products are people talking about, and how are they talking about it?

DATA SOURCES

1. **Google Analytics**
   Pageviews, users
2. **CIS**
   Internal call log records
3. **Twitter**
   Timeline, mentions
4. **go.usa.gov**
   Short URL clicks
5. **ForeSee** or new tool?
   Website feedback
6. **Meltwater**
   News aggregator

PROCESSING

1. **Extracting webpage info**
   Web scraping (BeautifulSoup)
2. **Predicting BLS program**
   Rule-based, spaCy
3. **Text & tweet cleaning**
   stringr, regex
4. **Sentiment analysis**
   To be implemented
5. **Prediction & classification**
   Logistic regression

MODULES

Unsupervised similarity analysis using cosine or jaccard similaity

REPORTING

SAMPLE PIPELINE

Which BLS products are people looking for on the website, and is it easy to access?

DATA SOURCES

1. **Google Analytics**
   Pageviews, users
2. **CIS**
   Internal call log records
3. **Twitter**
   Timeline, mentions
4. **go.usa.gov**
   Short URL clicks
5. **ForeSee** or new tool?
   Website feedback
6. **Meltwater**
   News aggregator

PROCESSING

1. **Extracting webpage info**
   Web scraping (BeautifulSoup)
2. **Predicting BLS program**
   Rule-based, spaCy
3. **Text & tweet cleaning**
   stringr, regex
4. **Sentiment analysis**
   To be implemented
5. **Prediction & classification**
   Logistic regression

MODULES

Predicting topic of customer comments, then performing unsupervised clustering

REPORTING

BACKEND DETAILS

R Shiny

R packages: `httr`, `RJDBC`, `rtweet`, `googleAnalyticsR`
Python packages: `spaCy`, `requests`, more to come

`reticulate`

Allows calling Python NLP packages from R Shiny. Handles R-to-Python type conversions on-the-fly (and vice versa).

Easy authentication

Browser-based popup login via **`rtweet`** and **`googleAnalyticsR`**. Other data sources may require simple API key.

Serverless

Run client-side using R portable, without a need for dedicated R Shiny server. Via Stephen York.

CODE STRUCTURE

Modular function design enables the same processing functions to act on data from all different data sources

---

`fetch_*()`

Takes authentication info
Returns dataframe

`extract_*()`

Takes raw text
Returns processed text

module functions

Merge data from sources
Specific to each pipeline

`auth_*()` for authentication

`str_*()` for additional string processing, to supplement `stringr`

EXAMPLE FUNCTIONS:

MOST FUNCTIONS WORK ON MOST DATA SOURCES

`auth_*()`	`fetch_*()`	`extract_*()`
`auth_cis()`	`fetch_cis_inquiries()`	`extract_page_title()`
`auth_ga()`	`fetch_ga_top_opub()`	`extract_program_office()`
`auth_gousa()`	`fetch_gousa_links()`	`extract_clean_title()`
`...`	`fetch_twitter_timeline()`	`extract_publication_type()`
	`...`	`...`

Potential future implementations

`auth_foresee()`	`fetch_foresee_feedback()`	`extract_trendline()`
`auth_meltwater()`	`fetch_mw_feed()`	`...`
`auth_upubs()`	`fetch_upubs_views()`	

GOALS ACCOMPLISHED

Bringing together separate data sources positions stakeholders to unlock more value out of data that BLS already pays for

Platform for future dashboarding

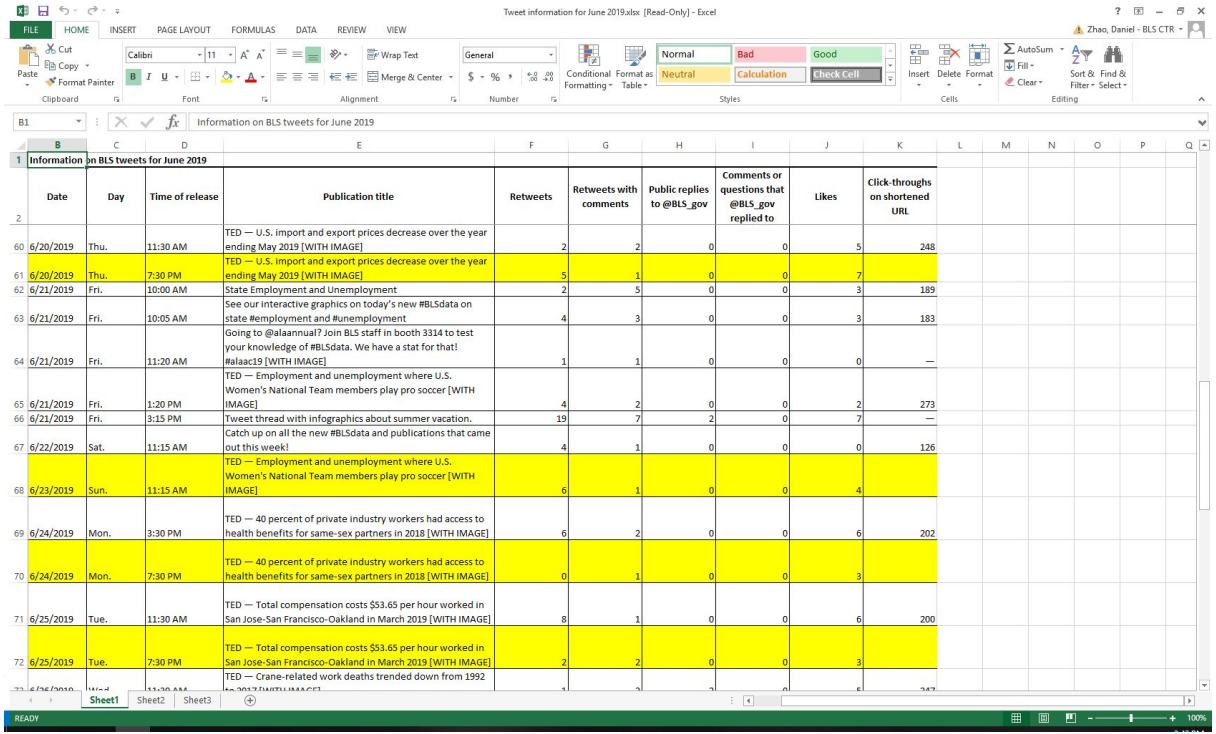
Potential to automate existing quarterly GA report, if internal R Shiny server is set up

Use case for R Shiny and Python integration

Supports BLS-wide push for R. Proof-of-concept for R Shiny dashboards that leverage Python's NLP capabilities.

Supports push for data science

Immense potential for NLP to unlock new insights and inform communications decisionmaking

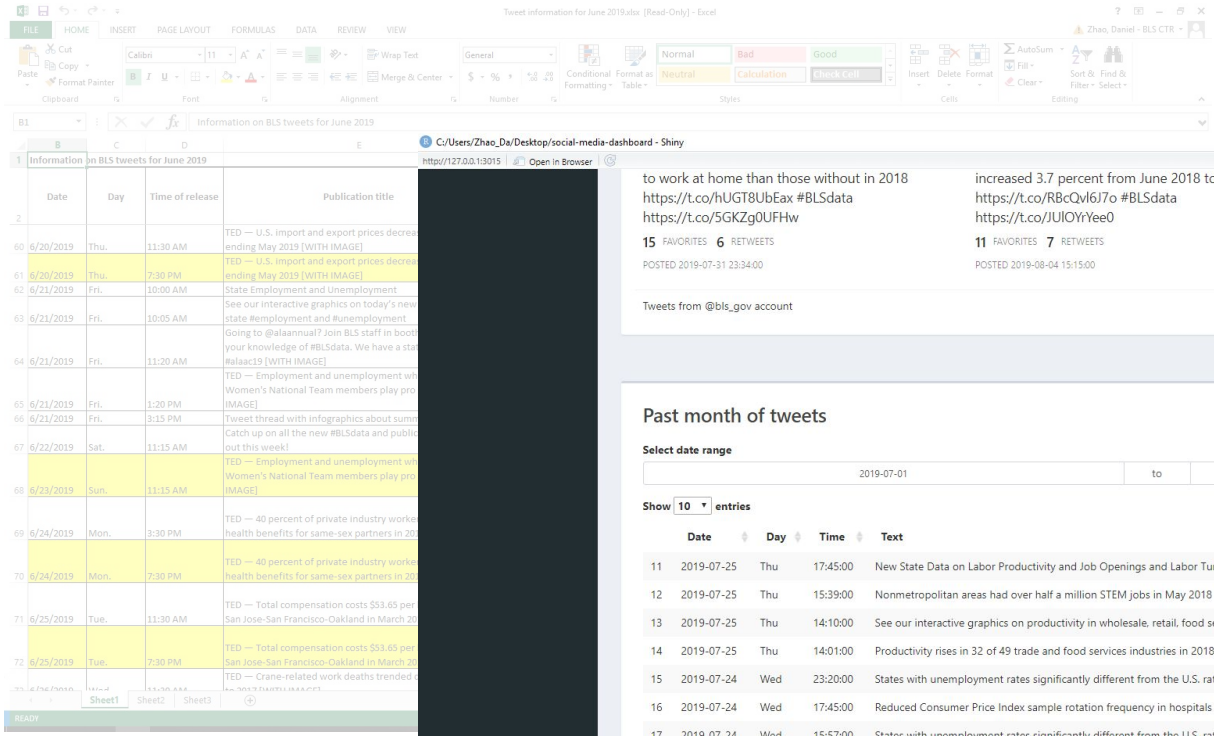


Tweet information for June 2019.xlsx [Read-Only] - Excel

Information on BLS tweets for June 2019

Date	Day	Time of release	Publication title	Retweets	Retweets with comments	Public replies to @BLS_gov	Comments or questions that @BLS_gov replied to	Likes	Click-throughs on shortened URL
6/20/2019	Thu.	11:30 AM	TED — U.S. import and export prices decrease over the year ending May 2019 [WITH IMAGE]	2	2	0	0	5	248
6/20/2019	Thu.	7:30 PM	TED — U.S. import and export prices decrease over the year ending May 2019 [WITH IMAGE]	5	1	0	0	7	
6/21/2019	Fri.	10:00 AM	State Employment and Unemployment	2	5	0	0	3	189
6/21/2019	Fri.	10:05 AM	See our interactive graphics on today's new #BLSdata on state #employment and #unemployment	4	3	0	0	3	183
6/21/2019	Fri.	11:20 AM	Going to @alaannual? Join BLS staff in booth 3314 to test your knowledge of #BLSdata. We have a stat for that! #alaac19 [WITH IMAGE]	1	1	0	0	0	—
6/21/2019	Fri.	1:20 PM	TED — Employment and unemployment where U.S. Women's National Team members play pro soccer [WITH IMAGE]	4	2	0	0	2	273
6/21/2019	Fri.	3:15 PM	Tweet thread with infographics about summer vacation.	19	7	2	0	7	—
6/22/2019	Sat.	11:15 AM	Catch up on all the new #BLSdata and publications that came out this week!	4	1	0	0	0	126
6/23/2019	Sun.	11:15 AM	TED — Employment and unemployment where U.S. Women's National Team members play pro soccer [WITH IMAGE]	6	1	0	0	4	
6/24/2019	Mon.	3:30 PM	TED — 40 percent of private industry workers had access to health benefits for same-sex partners in 2018 [WITH IMAGE]	6	2	0	0	6	202
6/24/2019	Mon.	7:30 PM	TED — 40 percent of private industry workers had access to health benefits for same-sex partners in 2018 [WITH IMAGE]	0	1	0	0	3	
6/25/2019	Tue.	11:30 AM	TED — Total compensation costs $53.65 per hour worked in San Jose-San Francisco-Oakland in March 2019 [WITH IMAGE]	8	1	0	0	6	200
6/25/2019	Tue.	7:30 PM	TED — Total compensation costs $53.65 per hour worked in San Jose-San Francisco-Oakland in March 2019 [WITH IMAGE]	2	2	0	0	3	
			TED — Crane-related work deaths trended down from 1992						

EXAMPLE: AUTOMATING REPORTING



Past month of tweets

Select date range

2019-07-01 to 2019-07-31

Show 10 entries Search:

	Date	Day	Time	Text	Retweets	Favorites	URL clicks
11	2019-07-25	Thu	17:45:00	New State Data on Labor Productivity and Job Openings and Labor Turnover	18	13	347
12	2019-07-25	Thu	15:39:00	Nonmetropolitan areas had over half a million STEM jobs in May 2018	1	2	275
13	2019-07-25	Thu	14:10:00	See our interactive graphics on productivity in wholesale, retail, food services & drinking places	1	3	461
14	2019-07-25	Thu	14:01:00	Productivity rises in 32 of 49 trade and food services industries in 2018	2	1	2852
15	2019-07-24	Wed	23:20:00	States with unemployment rates significantly different from the U.S. rate in June 2019	6	8	Repost
16	2019-07-24	Wed	17:45:00	Reduced Consumer Price Index sample rotation frequency in hospitals and household utilities	3	4	155
17	2019-07-24	Wed	15:57:00	States with unemployment rates significantly different from the U.S. rate in June 2019	9	9	384
18	2019-07-24	Wed	14:05:00	See our interactive graphics on today's new #BLSdata on gross job gains and losses	3	3	985
19	2019-07-24	Wed	14:01:00	Gross job gains 7.7 million and gross job losses 6.9 million in the 4th quarter of 2018	7	12	10709
20	2019-07-23	Tue	23:10:00	Nevada had largest percentage increase in employment over the year ended June 2019	3	8	Repost

Showing 11 to 20 of 79 entries

Previous 1 2 3 4 5 ... 8 Next

SPECIAL THANKS TO...

CODING IT FORWARD

Rachel Dodell
Chris Kuang
Hillary McLauchlin

MENTORS

Christian Moscardi Census Bureau
Jared Rodman Uber

CIVIC DIGITAL FELLOWSHIP

BUREAU OF LABOR STATISTICS

Wes Chou Office of Technology & Survey Processing
Amrit Kohli Division of Enterprise Web Systems
Emily Liddell Office of Publications & Special Studies
Jay Meisenheimer Office of Publications & Special Studies
Michael Levi Office of Publications & Special Studies
Anna Grace Rutledge Office of Emp. & Unemp. Statistics

DATA

Karen Krein OCOMM
Cody Parkinson OPUBSS

TECHNICAL

Brandon Kopp OSMR
Stephen York OPLC

CONTACT INFORMATION

DANIEL ZHAO

daniel@danielzhao.com
(973) 919-6559