Correlating Poll Results with Twitter Sentiment on Obamacare

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

ACA
AFFORDABLE
CARE ACT

twitter

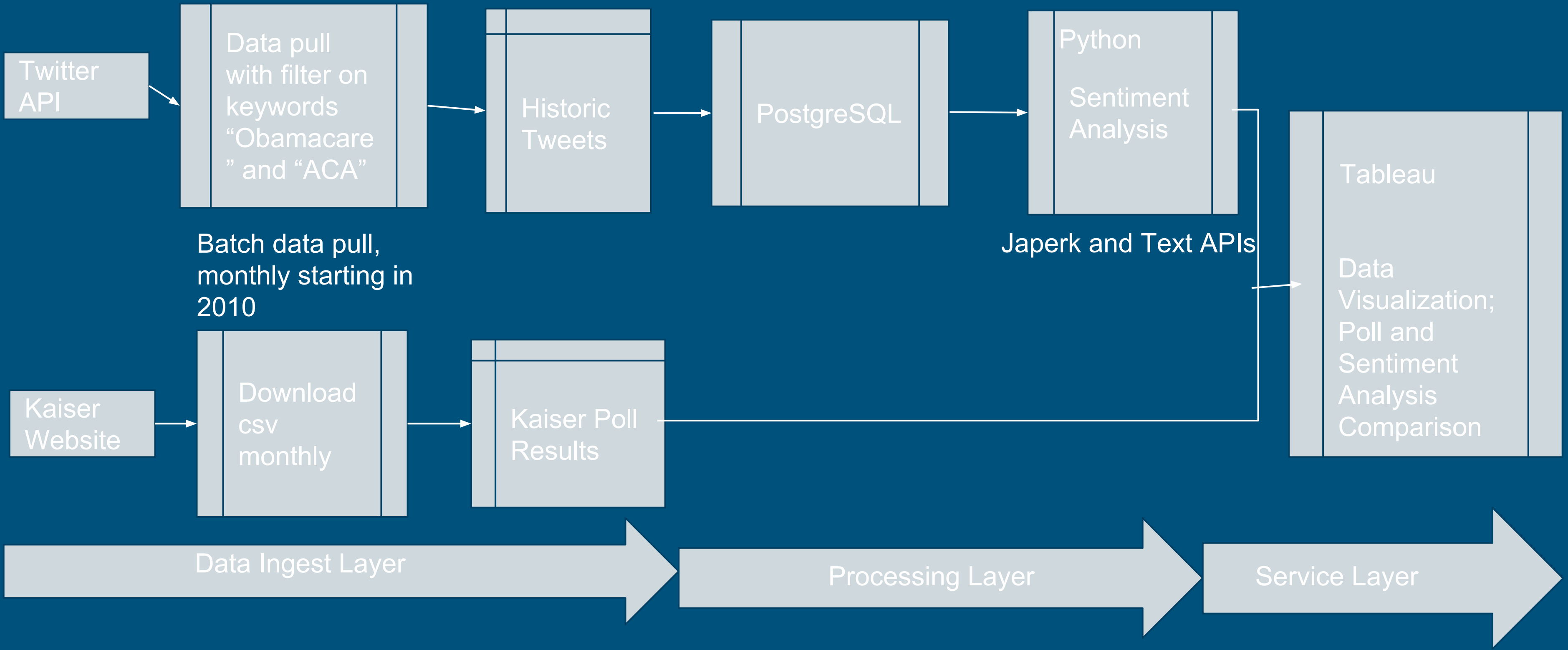
W205

Tingwen Bao
Jay Cordes
Alex Jamar
Michelle Liu

Overview of the problem

- Develop a data infrastructure to hold both ACA poll results from Kaiser and Twitter posts
- Analyze the correlation between the result of public polls and sentiment on Twitter
- Twitter claims that the real-time, public orientation of its social network makes it a reliable barometer of the public's constantly changing moods and interests

Overall Architecture



Twitter API

Data pull with filter on keywords “Obamacare” and “ACA”

Historic Tweets

PostgreSQL

Python

Sentiment Analysis

Batch data pull, monthly starting in 2010

Japerk and Text APIs

Tableau

Data Visualization; Poll and Sentiment Analysis Comparison

Kaiser Website

Download csv monthly

Kaiser Poll Results

Data Ingest Layer

Processing Layer

Service Layer

Acquisition of Twitter Data

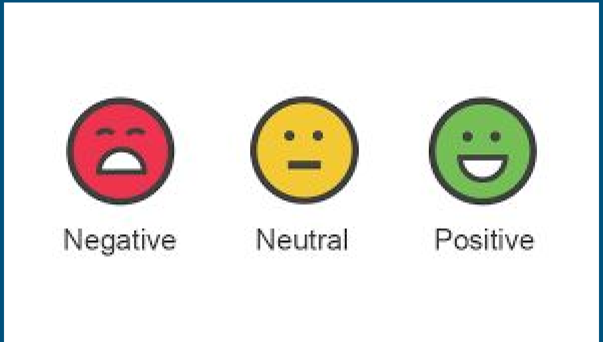
- Adopted Python library, GetOldTweets
  - Bypass the time constraint limitations of Twitter API
  - Search for tweets in English about Obamacare or ACA since its introduction in 2010
  - Collects 1000 tweets per day every 7 days.
- Ran on an Amazon Web Services Elastic Cloud Compute virtual machine
- Stored about 400,000 Tweets

The Database

- Created a python script to import tweets tab-delimited file into PostgreSQL
- Used command-line argument with filename for flexibility
- Added surrogate primary key (bigserial data type is auto-incrementing) and indexes
- Exceptions handled gracefully and script keeps importing
- Skipped transactions and kept everything simple and maintainable
- PostgreSQL is very easy to use and facilitated later sentiment scoring / analysis steps

Sentiment Analysis

- Randomly selected 10% of the tweets for analysis
- Used 2 sentiment analysis APIs
  - Sent POST requests containing the Tweet text to the APIs
  - Received JSON objects in return with either “positive,” “negative,” or “neutral” labels
  - Mapped positive → 1, negative → -1, neutral → 0
- Updated Tweets database with the results



Evaluation & Visualization

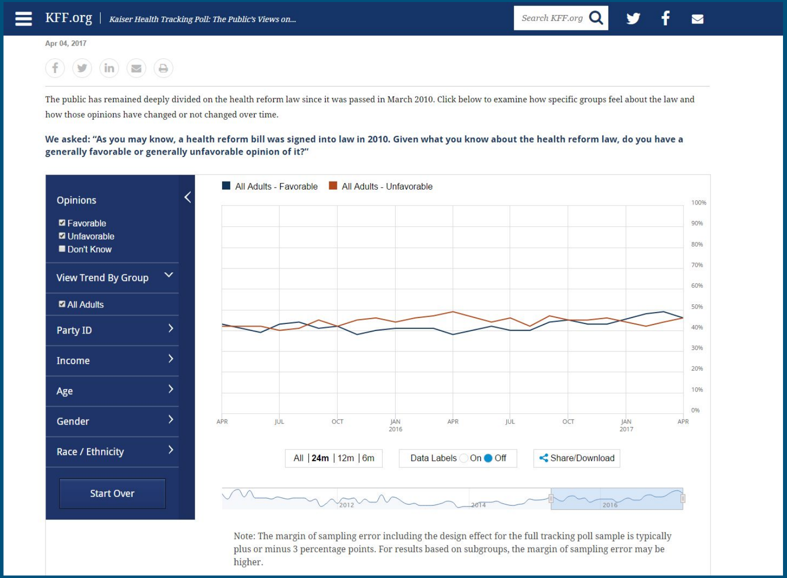
Public Poll Data: Kaiser Health Tracking Poll

Why Tableau:

1. Can easily connect both databases and flat files
2. Relative strong visualization with interactive features
3. Various ways to share/access

Drawbacks:

1. Not strong at stats
2. Could become slow after scale up



Apr 04, 2017

The public has remained deeply divided on the health reform law since it was passed in March 2010. Click below to examine how specific groups feel about the law and how those opinions have changed or not changed over time.

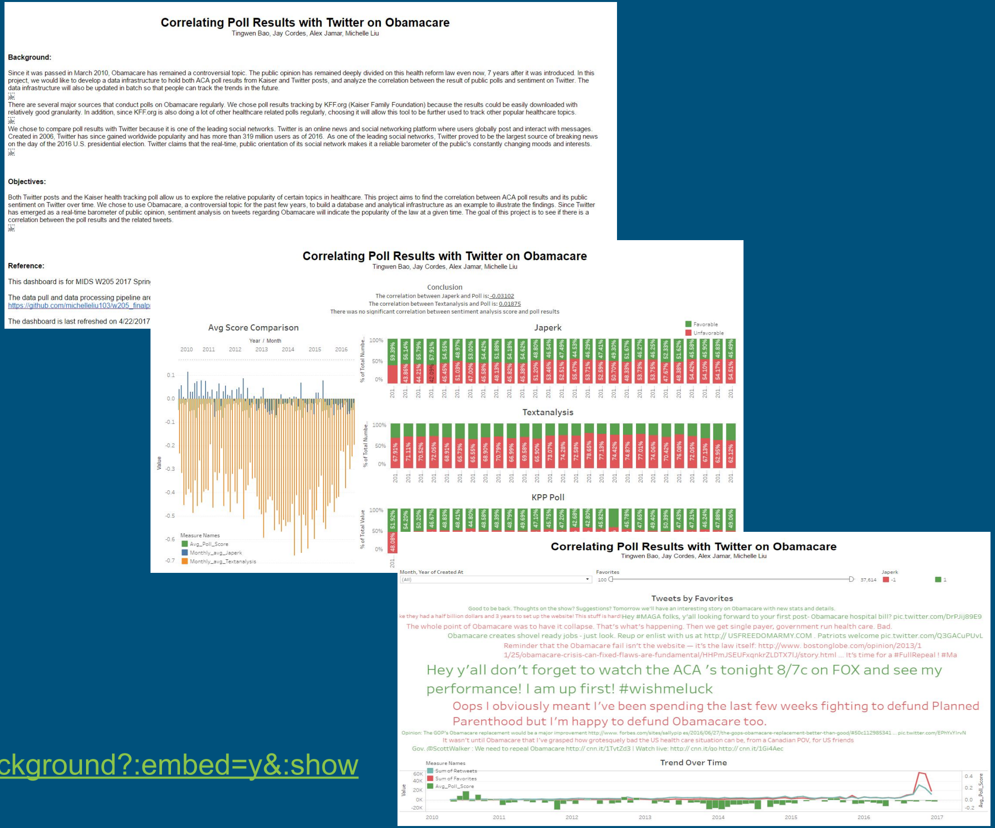
We asked: "As you may know, a health reform bill was signed into law in 2010. Given what you know about the health reform law, do you have a generally favorable or generally unfavorable opinion of it?"

Note: The margin of sampling error including the design effect for the full tracking poll sample is typically plus or minus 3 percentage points. For results based on subgroups, the margin of sampling error may be higher.

Tableau Dashboard

To evaluate and visualize the result, we create a dashboard with 3 tabs:

- Background: a general introduction of the background and objective of this project
- Comparison: visuals used to compare the 2 different sentiment analysis method and understand its correlation with public poll data
- Explore Tweets: word cloud of top tweets by favorites and its trend by time with interactive filters



Correlating Poll Results with Twitter on Obamacare

Tingwen Bao, Jay Cordes, Alex Jamar, Michelle Liu

Background:

Since it was passed in March 2010, Obamacare has remained a controversial topic. The public opinion has remained deeply divided on this health reform law even now, 7 years after it was introduced. In this project, we would like to develop a data infrastructure to hold both ACA poll results from Kaiser and Twitter posts, and analyze the correlation between the result of public polls and sentiment on Twitter. The data infrastructure will also be updated in batch so that people can track the trends in the future.

There are several major sources that conduct polls on Obamacare regularly. We chose poll results tracking by KFF.org (Kaiser Family Foundation) because the results could be easily downloaded with relatively good granularity. In addition, since KFF.org is also doing a lot of other healthcare related polls regularly, choosing it will allow this tool to be further used to track other popular healthcare topics.

We chose to compare poll results with Twitter because it is one of the leading social networks. Twitter is an online news and social networking platform where users globally post and interact with messages. Created in 2006, Twitter has since gained worldwide popularity and has more than 319 million users as of 2016. As one of the leading social networks, Twitter proved to be the largest source of breaking news on the day of the 2016 U.S. presidential election. Twitter claims that the real-time, public orientation of its social network makes it a reliable barometer of the public's constantly changing moods and interests.

Objectives:

Both Twitter posts and the Kaiser health tracking poll allow us to explore the relative popularity of certain topics in healthcare. This project aims to find the correlation between ACA poll results and its public sentiment on Twitter over time. We chose to use Obamacare, a controversial topic for the past few years, to build a database and analytical infrastructure as an example to illustrate the findings. Since Twitter has emerged as a real-time barometer of public opinion, sentiment analysis on tweets regarding Obamacare will indicate the popularity of the law at a given time. The goal of this project is to see if there is a correlation between the poll results and the related tweets.

Reference:

This dashboard is for MIDS W205 2017 Sprin

The data pull and data processing pipeline are
https://github.com/michelleliu103/w205_finalp

The dashboard is last refreshed on 4/22/2017

Correlating Poll Results with Twitter on Obamacare

Tingwen Bao, Jay Cordes, Alex Jamar, Michelle Liu

Conclusion

The correlation between Japerk and Poll is: -0.03102

The correlation between Textanalysis and Poll is: 0.01875

There was no significant correlation between sentiment analysis score and poll results

Correlating Poll Results with Twitter on Obamacare

Tingwen Bao, Jay Cordes, Alex Jamar, Michelle Liu

Month, Year of Created At: (All)

Favorites: 100 – 37,614

Japerk: 🟥 -1 🟩 1

Tweets by Favorites

Good to be back. Thoughts on the show? Suggestions? Tomorrow we'll have an interesting story on Obamacare with new stats and details.

ke they had a half billion dollars and 3 years to set up the website! This stuff is hard!

Hey #MAGA folks, y'all looking forward to your first post- Obamacare hospital bill? pic.twitter.com/DrPJij89E9

The whole point of Obamacare was to have it collapse. That's what's happening. Then we get single payer, government run health care. Bad.

Obamacare creates shovel ready jobs - just look. Reup or enlist with us at http:// USFREEDOMARMY.COM . Patriots welcome pic.twitter.com/Q3GACuPUvL

Reminder that the Obamacare fail isn't the website — it's the law itself: http://www. bostonglobe.com/opinion/2013/1 1/25/obamacare-crisis-can-fixed-flaws-are-fundamental/HHPmJSEUFxqnkrZLDTX7lJ/story.html ... It's time for a #FullRepeal ! #Ma

Hey y'all don't forget to watch the ACA 's tonight 8/7c on FOX and see my performance! I am up first! #wishmeluck

Oops I obviously meant I've been spending the last few weeks fighting to defund Planned Parenthood but I'm happy to defund Obamacare too.

Opinion: The GOP's Obamacare replacement would be a major improvement http://www. forbes.com/sites/sallypip es/2016/06/27/the-gops-obamacare-replacement-better-than-good/#50c112985341 ... pic.twitter.com/EPhYxYIrvN

It wasn't until Obamacare that I've grasped how grotesquely bad the US health care situation can be, from a Canadian POV, for US friends

Gov. @ScottWalker : We need to repeal Obamacare http:// cnn.it/1TytZd3 | Watch live: http:// cnn.it/go http:// cnn.it/1Gi4Aec

https://us-east-1.online.tableau.com/t/tingwen/views/Obamacare/Background?:embed=y&:showShareOptions=true&:display_count=no&:showVizHome=no

Results

- Using different sentiment analysis method yields different conclusion of Obamacare related tweets
  - Japerk’s score indicates that the positive and negative tweets of Obamacare are close to a tie
  - Textanalysis’s score indicates that there are about twice as many negative tweets as positive.

- KPP poll shows public has a nearly even spread on their opinion on Obamacare.
  - More positive at the start and then more negative, and most recently, almost even.

- No significant correlation between sentiment analysis score and poll results

- Tweets with most favorites and retweets happened in 2016 rather than when it was first introduced.
  - Right before election when people mentioned Obamacare to express their political stands.

Future Work

Roadmap for improving the solution with increased usage and increasing data size: improving the ways in acquiring the data, storing and processing it more efficiently, and overall, scaling up our solution.

- Instead of pulling the same number of tweets every week, grab all related tweets or a proportional number of them
- Default the sentiment scores to null instead of 0 and/or score all of the tweets in our database
- Automate and import the polling data into a PostgreSQL table as well.
- Create aggregation table to reduce the processing time of Tableau.
- Strip out all of the tab characters in the tweets before writing it to CSV.

Conclusion

- Text sentiment analysis is nowhere near perfect
- We did not find correlation between poll results and two types of Twitter sentiment scores
- The 3Vs were important challenges for our project: long term needs of the solution will have to deal with the sheer *volume* of the data coming in from Twitter, and to increase the *velocity* of doing sentiment analysis and processing of the data.
- Our program could provide a useful test for any new sentiment analysis approaches that are developed by validating them against polling data.
- If our app finds a strong correlation between Twitter sentiment scores and polling data, we would have provided evidence for the ability to do real-time polling on any subject.

References

http://kff.org/interactive/kaiser-health-tracking-poll-the-publics-views-on-the-aca/#?aRange=twoYear

https://github.com/Jefferson-Henrique/GetOldTweets-python

https://market.mashape.com/textanalysis/sentiment-analysis

https://market.mashape.com/japerk/text-processing