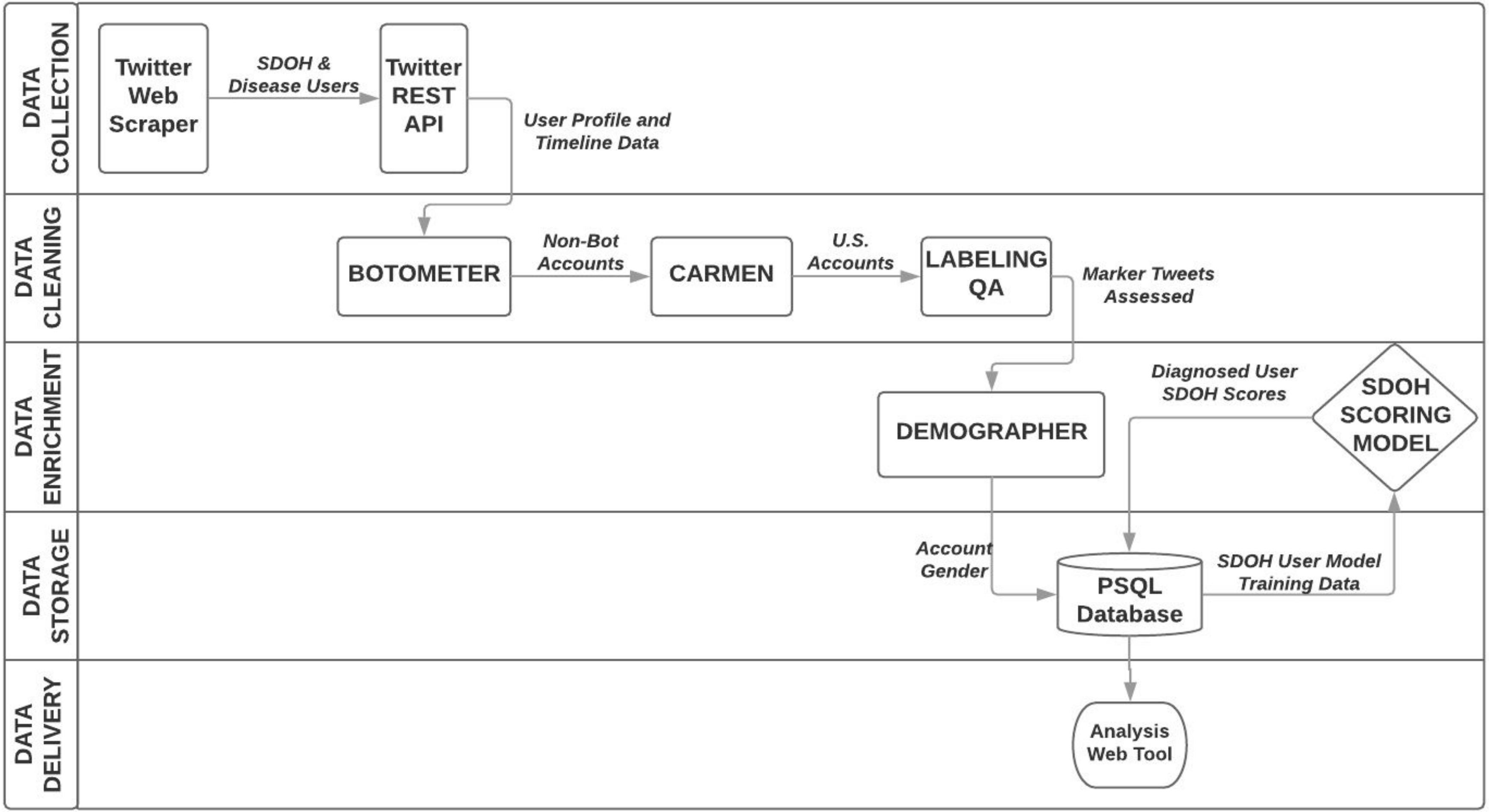
Social Media Determinants of Health

Week 8 Update

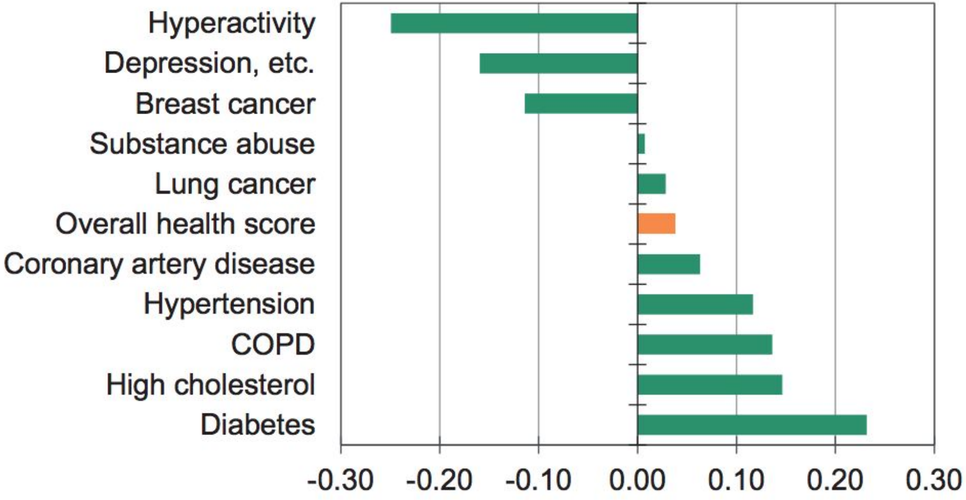
Pipeline Overview



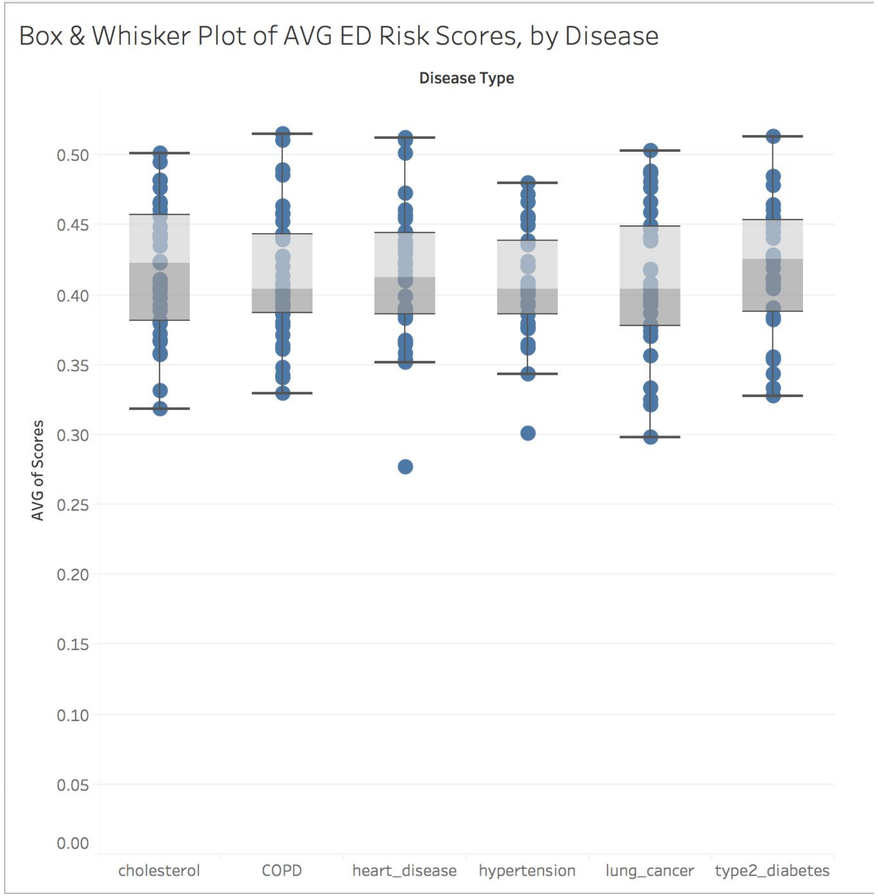
BCBS Report Benchmark

Chart E2: Education Has Mixed Effects

Effect of % population with college degree on condition z-score



Sources: BCBS, Moody's Analytics


Box & Whisker Plot of AVG ED Risk Scores, by Disease

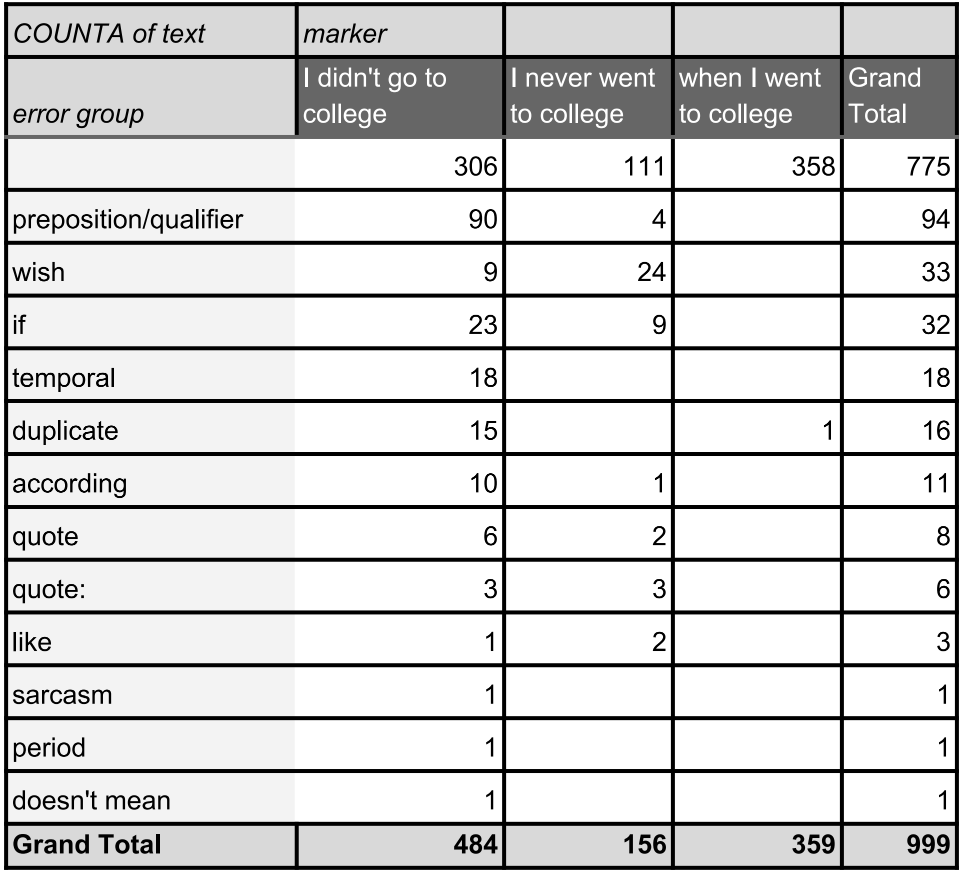
Data Cleaning

- 1000 “marker” tweets examined.

- “No college” marked tweets
  - Many turns of phrases
  - 417 accurately marked
  - 177 labels reversed programmatically
  - 32 removed programmatically
  - 14 ambiguous context
  - After cleaning, 97% accurate labels

- “College” tweets well-marked

- Final Label Ratio of 431:535



COUNTA of text	*marker*			
error group	I didn't go to college	I never went to college	when I went to college	Grand Total
	306	111	358	775
preposition/qualifier	90	4		94
wish	9	24		33
if	23	9		32
temporal	18			18
duplicate	15		1	16
according	10	1		11
quote	6	2		8
quote:	3	3		6
like	1	2		3
sarcasm	1			1
period	1			1
doesn't mean	1			1
Grand Total	**484**	**156**	**359**	**999**

Data Loading

Disease_subject_user
marker_tweet_id
handle
marker_tweet
search_phrase
disease_population

Disease_subject_user_profile_detail
handle
latitude
longitude
gender
follower_count
favorites_count
friends_count
bot_likelihood

Disease_subject_user_tweet_history
tweet_id
handle
tweet_text
tweet_datetime

Sdoh_model_user_profile_detail
handle
latitude
longitude
gender
follower_count
favorites_count
friends_count
bot_likelihood

Sdoh_model_user_tweet_history
tweet_id
handle
tweet_text
tweet_datetime

Sdoh_model_user
marker_tweet_id
label
handle
marker_tweet
search_phrase
sdoh_model

Data Loading

Tweets were scraped from 2014-2018 with these 14 phrases:

- *“I didn’t go to college”*
- *“When I went to college”*
- *“I never went to college”*
- *“my lung cancer”*
- *“I have lung cancer”*
- *“I was diagnosed with lung cancer”*
- *“I was diagnosed with COPD”*
- *“I have COPD”*
- *“my COPD”*
- *“I have high cholesterol”*
- *“my high cholesterol”*
- *“I have type 2 diabetes”*
- *“my type 2 diabetes”*
- *“I was diagnosed with type 2 diabetes”*

Data Loading - Method

For each “marker” tweet, we pulled:

- Profile json
- 100 recent tweets from the marker tweet (just the max amt if less than 100 tweets available)
- Label profile json with 3 things:
  - Location (carmen library)
  - Gender (demographer library)
  - Bot Likelihood (botometer)

Quick Stats:

Data pulled from twitter: **>10GB**

~160k disease subject tweets + ~560k sdoh tweets **= 720,000 tweets**

~950 disease subject profiles + ~6,500 education user profiles = **7,500 profiles**

Education Model Improvement

Accuracy before: **0.6** => After: **0.86** (in parallel with i.e not including re-labeling, demographer)

Feature Engineering

- Text Sentiment Extraction
  - Polarity Score
  - Subjectivity Score
- Text Cleaning
  - remove symbols/links etc from the tweet text
- Next Step:
  - external dictionaries e.g Slangs, sentence compositions
  - demographic Info
- Balancing the positive labels VS negative labels
  - 100000 tweets from + labels
  - 100000 tweets from - labels

Future Roadmap

Week 8: End-to-End pipeline complete, data cleaning, model refinement

Week 9: All education and disease-subject data loaded, cleaned; model refined

Week 10: Presentation 2: Front-end web tool prototype with education model

Week 11: Implement at least 2 other SDOH models (housing, employment)

Week 12: Further work on model implementation, web tool development

Week 13: Final analytics front-end built and incorporated into web tool

Week 14: Final Presentation