Proof of concept:
Natural language filtering of U.S. VAERS data

MIKE RIGHTMIRE

CANDIDATE
LAB INFORMATICS DATA SCIENTIST
AGILENT, WALDBRONN

Agilent Technologies

Introduction

The U.S. VAERS (Vaccine Adverse Event Reporting System) was established in 1990, under the management of the U.S. Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA) to collect and analyze vaccine adverse event data with the hopes of identifying unpredicted and unforeseen side effects once the vaccines are available to the community.

Currently there are more than 397,000 VAERS reports for 2021, with 31,000+ of these classified as serious adverse events in need of investigation. With limited resources, triaging for investigation becomes a challenging task for both government and private researchers.

This project is a proof-of-concept intended to allow for the discovery and filtering of individual adverse event reports based on specified criteria.

https://vaers.hhs.gov/about.html

Table of Content (ToC)

Overview

The data

- Both mandatory and voluntary reports of adverse events (AEs) may be submitted by anyone.
  - Consumers can freely report an adverse event to VAERS.
  - Healthcare professionals are *required* to report certain adverse events.
  - Vaccine manufacturers are required to report *all* adverse events that come to their attention.
- Much of the critical VAERS data is reported in free-form text fields.
- Only serious AEs are validated or investigated.
- VAERS data is available to the public via their website, https://vaers.hhs.gov/data.html

https://vaers.hhs.gov/about.html

Overview

Challenges

- Much of the critical VAERS data exists only as unconstrained text fields.
- This text includes reports from both medical professionals, and laypeople – making syntax and vernacular unpredictable.
- The quality of the descriptions vary greatly, ranging from a single word to long, detailed patient histories.
  - Average word count: 77
  - Minimum word count: 1
  - Maximum word count: 2996
- Misspellings, entry errors, idioms, and grammatical issues are commonplace.

Overview

Solutions

- A *supervised* Natural Language Processing (NLP) pipeline using Python NLTK, and SKLearn.
- TF-IDF logistic regression.
- Adjustable parameters for iteration, percent cutoff, and majority selection.

Success criteria

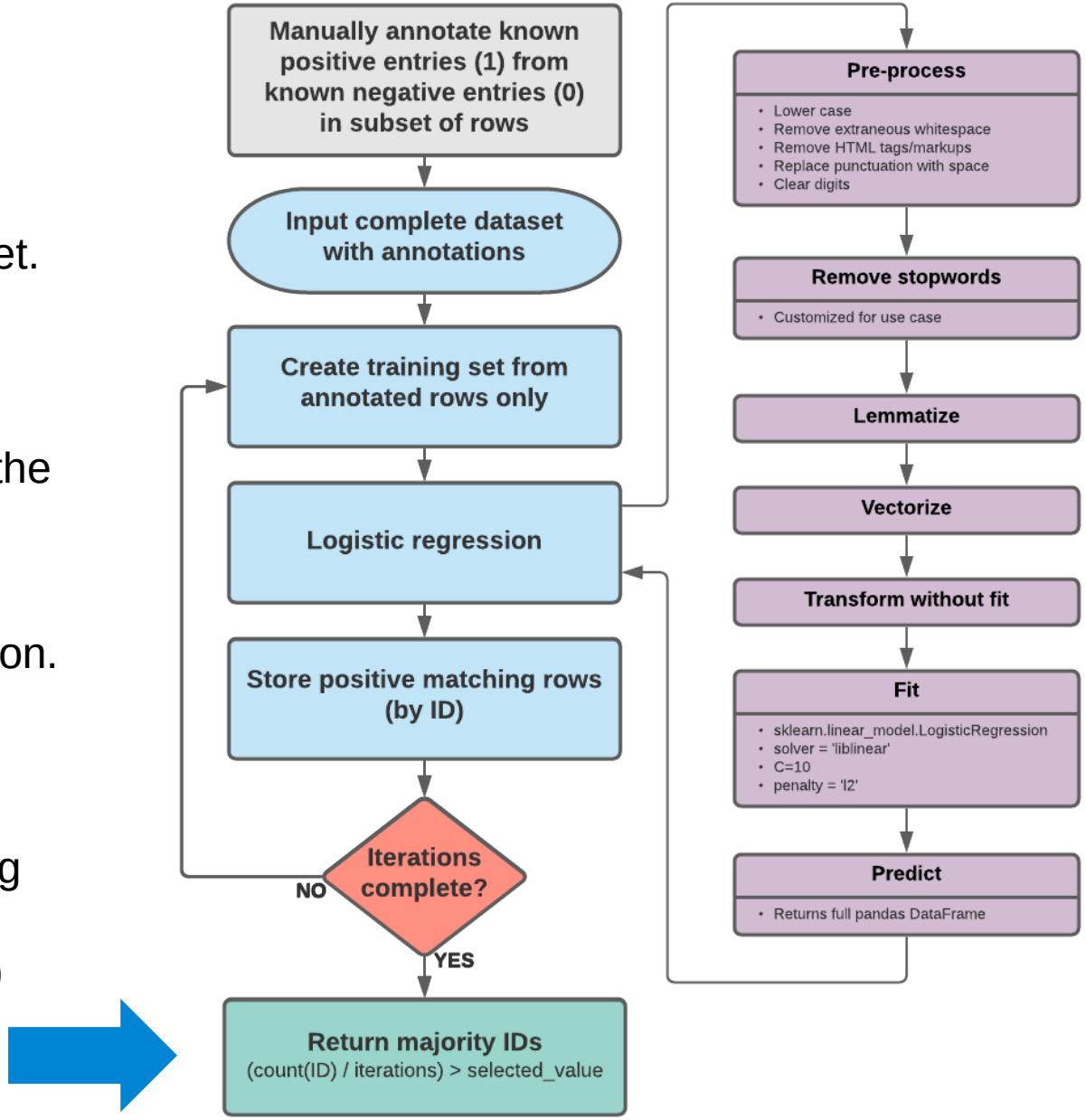
- Manual annotation of the data for training is <10% of data set.
- Fewer than 5% false positives.

https://vaers.hhs.gov/about.html

Pipeline

General flow

1. Annotate the training set.
2. Train based on the annotated rows.
3. Logistic regression on the complete dataset.
4. Capture the output of each training/test session.
5. Reiterate.
6. Run an election on all stored results, capturing only those which pass majority (default: 50%.)



ToC

Methods

How many deaths reported involved confirmed Sars-CoV-2 infections?

- The complete 2021 VAERS data was downloaded on the first week of July.
- Data was filtered to include only feature "DIED" == "Y"
- The columns "*SYMPTOM_TEXT*" and "*LAB_DATA* “ were selected as features.
- The columns "*SYMPTOM_TEXT_TARGET*" and "*LAB_DATA_TARGET*" were added and annotated with a 1 (Positive) or a 0 (Negative) to indicate the commensurate column was Covid(+) or Covid(-)/Unknown.
- The feature columns were annotated individually, without referencing the other column for confirmation. I.e. If the column was unclear, the training was marked as “0” based on the unclear column alone.

ToC

Methods

How many deaths reported involved confirmed Sars-CoV-2 infections?

Examples

Positive (1)	Negative (0)
After vaccination, **patient tested positive for COVID-19**. Patient was very ill and had numerous chronic health issues prior to vaccination. Facility had a number of patients who had already tested positive for COVID-19. Vaccination continued in an effort to prevent this patient from contracting the virus or to mitigate his risk. This was unsuccessful and patient died.	Feb 8 **states she had a cold**. Feb 9 added stomach ache and nausea. Feb 9 visited urgent care facility for exam and Covid-19 test. **Rapid test results were negative**. Appeared tired but fine. Told to go home and rest. Feb 10 at 9:00 am found dead on the floor in pool of blood and aspirated. Excessive blood in toilet, pooled on floor and hallway rug.
Vaccine 12/30/2020 Screening PCR done 12/31/2020 Symptoms 1/1/2021 **COVID test result came back positive** 1/2/2021 Deceased 1/4/2021	**Patient death within 60 days of receiving the COVID vaccine series**
Patient **had been diagnosed with COVID-19** on Dec. 11th, 2020. Symptoms were thought to have started on 12/5/2020. Received Moderna vaccine on 12/23. Unexpected death on 1/8/2021. Resuscitation attempts unsuccessful	Spoke to RN at ER. Resident had fever of 102.9 upon arrival to the ER. He coded on the way back from his CT scan. He was **diagnosed with right sided pnuemonia**, respiratory failure and sepsis. Resident just passed away in ER and was not even admitted. Family was there with him
Patient developed **Covid pneumonia** dx 1/15/21, patient expired	**respitory colase**
+ COVID 19	**COVID 19**

ToC

Training and testing

How many reported deaths involved confirmed Sars-CoV-2 infections?

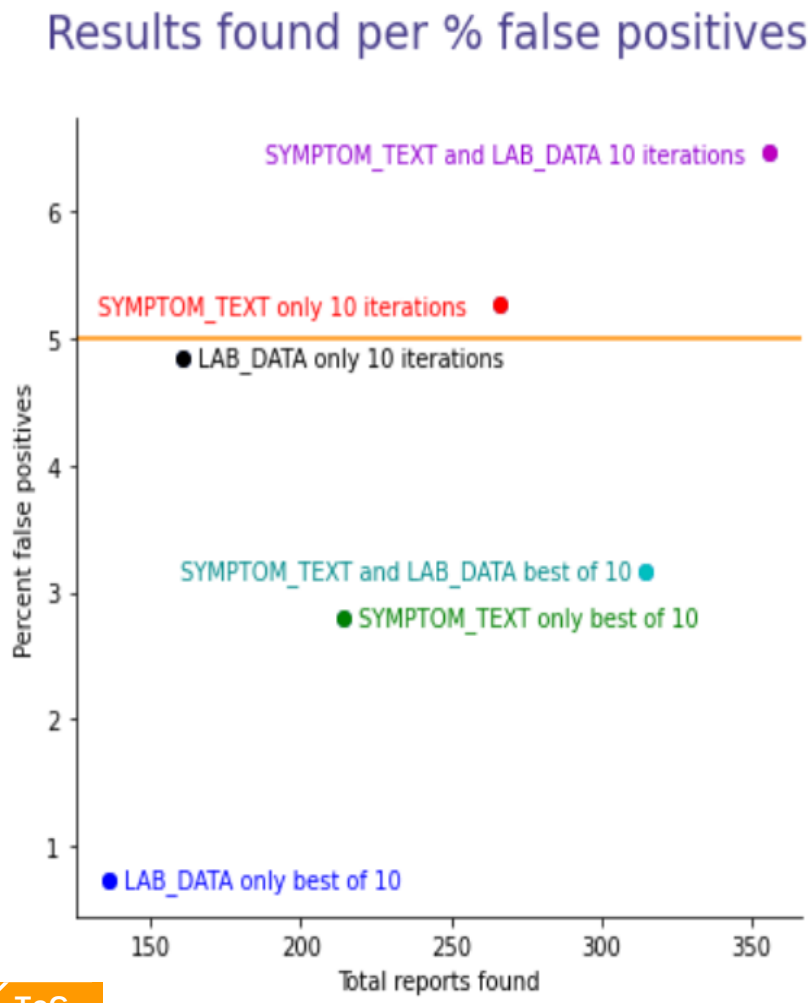
- Training and test were run against the selected columns individually and together, using both the "*10 iterations with election*" and "*best of 10"* methods.
  - The *majority* parameter for the elections was set to > 0.5 (simple majority.)
  - The percent cutoff (*perc_cutoff*) parameter was at 80 (80% confidence in datapoint.)

- Re-training (training adjustment) used to mark commonly misidentified phrases.
  - "*Patient was hospitalized x 3 and died within 60 days of receiving a COVID vaccine series*"
  - *"Patient death within 60 days of receiving the COVID vaccine series*"
  - *"Sars PCR 2/2/21"* (versus *"Pos Sars PCR 2/2/21.")*
  - *"COVID 19"* (versus "*COVID 19 positive.")*

Feature	Initial Positive	Initial Negative	Adjusted Positive	Adjusted Negative
SYMPTOM_TEXT	101	101	101	108
LAB_DATA	105	104	105	111

Training and testing

How many reported deaths involved confirmed Sars-CoV-2 infections?



- *LAB_DATA* (only) iterations vs. best of:
  - 8/161 vs. 1/136
  - 18.38% more hits.
  - 575.78% greater false positives rate.
- *SYMPTOM_TEXT* (only) iterations vs. best of:
  - 14/266 vs. 6/214
  - 24.3% more hits.
  - 87.72% greater false positives rate.
- *SYMPTOM_TEXT* and *LAB_DATA* iterations vs. best of:
  - 23/356 vs. 10/315
  - 13.02% more hits.
  - 103.51% greater false positives rate.

ToC

Conclusions

How many reported deaths involved confirmed Sars-CoV-2 infections?

- Best overall score, *SYMPTOM_TEXT* and *LAB_DATA* best of 10.
- *LAB_DATA* (only) provided the most accurate search, likely due to information density.
- *SYMPTOM_TEXT* combined with *LAB_DATA* provided better results than either column alone.
- Different methods (iteration vs. best-of) still contain value for differing use cases. I.e. Early to late filtering.

ToC

Summary

- A pipeline was implemented using natural language scoring on two open text features of the VAERS data.
- Supervised training to answer a specific question required the manual selection of ~8% of entire data set.
- Training was implemented in two rounds: an initial training, and a remedial training to remove commonly misidentified phrases.
- Each feature was scored independently, and results tabulated based on each individual feature, as well as for the features' combined efforts.
- The model successfully identified the expected volume of data points, within the established margins of error.

References…

Title
Jupyter Notebook
Data set annotated for supervised learning
VAERS User Guide/Legend for data set
Download current VAERS data
NLP Tutorial for Text Classification in Python

THANK YOU