Demographic Analysis of Breast Cancer Awareness

Mayuri Kadam

A20353234

Problem

- What ratio of population is aware of breast cancer and what region they belong to?
- Aware here implies: Share experience, talk about research, causes, prevention, participate/contribute to programs and events

Approach

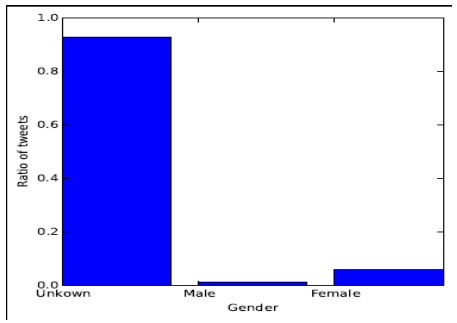
- Collected (~66K) tweets having mention of breast cancer and related keys.
- Classified tweets as relevant and irrelevant using a manually labelled(~500) training set.
- Excluded tweets mentioning sales and deals by retails celebrating Breast Cancer Month, and those that did not talk anything directly about the subject
- Filtered relevant tweets based on : Language(English) and Region(US)
- Tokenized tweet text and user description
- Performed experiments by training manually labelled data and training census labelled data.
- Identified genders as categories: Female, Male, Organization and Unknown
- Identified the States that user belong to by tokenizing the user provided locations.

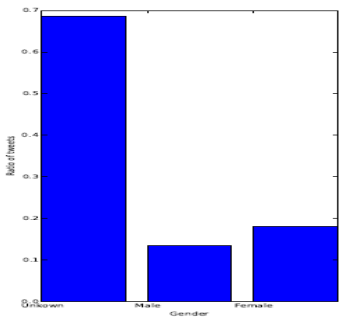
Data

- Raw : Tweets
- Fields : Description, Text, Location, Language
- Other : US Census names, US Cities and States
- Size:
- Raw ~66K
- Relevant ~39K
- US and English ~14K

Results

- Gender Analysis:
- Accuracy=.660 Accuracy=.810 Accuracy=.400





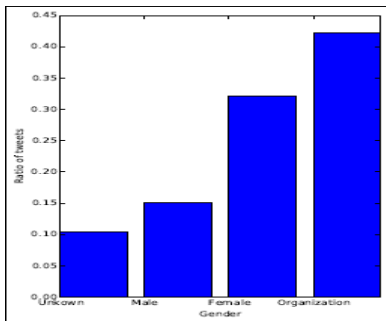
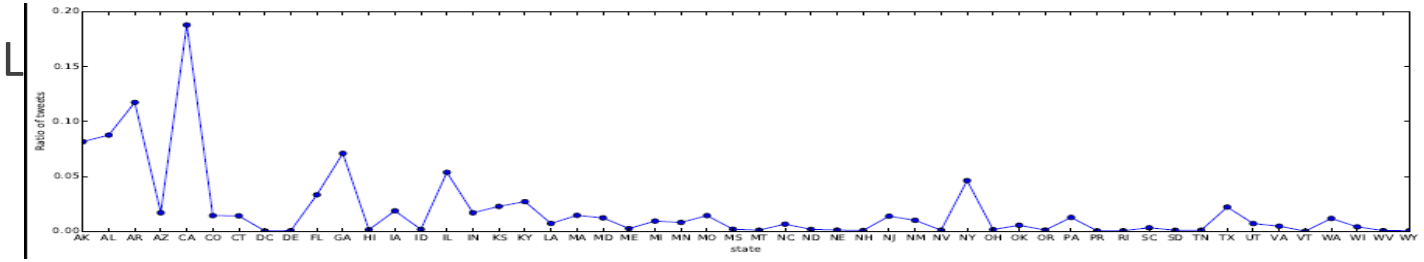


fig:1 fig:2 fig:3



Conclusion

Organizations are more likely to tweet about awareness programs than individuals.

Difference in male and female gender distribution is observed as assumed.

Spike in tweets from state AR (a less populated) might be the result of Breast Cancer foundation organization working on the awareness program.

Gender analysis using census names tend to produce more accuracy in this experiment, since it could classify males and females very well separating rest as unknown.

Manually labelled data divided the training set into four categories which could have caused less samples per category, yielding less accuracy than training with three categories using census data.

Insight: Individual responses to such awareness causes tends to be low in spite of many organization promoting the awareness