DEMOGRAPHIC ANALYSIS OF BREAST CANCER AWARENESS

INTRODUCTION

This project performs a demographic analysis of Breast Cancer Awareness for US twitter users by their gender and location. This project provides an idea about the awareness for breast cancer in the US population. Awareness here implies: sharing one's own/friends'/family's experience, information about causes of the disease, prevention and symptoms, participation in awareness programs, and donation to breast cancer foundations.

To carry out this analysis I propose following hypothesis:

1. Females are more likely to share their experiences
2. Tweets from less aware region are likely to be low and imply more incidences of breast cancer in that region

DATA

To infer and analyze gender of the twitter users for their awareness towards breast cancer, I collected ~66K tweets using Twitter's Search API. The data was collected using 'breast cancer' keywords. I used the tweet text and description field for tokenization, to infer the gender. US census names and manual labels were used for training sets of the gender identification experiments. For location inference user reported location was used to identify which State that user belongs to. The reported location was tokenized and compared with the list of cities from US and corresponding state was reported in the result.

METHODS

Data Preprocessing:

About ~66K tweets were collected as raw data and processed for their relevance. Out of these I used the unique and English language tweets (~38K) to find their relevance towards the topic. To determine the relevance of tweets I used a manually labelled training set of 500 tweets labelled as: (1) for relevant and(0) for irrelevant. Using Logistic Regression model I trained the data with 5-fold cross validation accuracy of 85%. Using this model for predicting the relevance of tweets I identified about ~30K tweets as relevant and used these for further processing.

To perform the gender and location analysis, I filtered the tweets based on the location field from user's profile by ignoring tweets from users who did not report any location and then restricting those who reported, to the locations, that had places which could be identified as US locations (US locations are maintained in the 'ListOfUSCities' file). This preprocessing brought down the raw data to (~14K) tweets used for the actual experimentation.

Further analysis is done using these ~14K tweets (let's call it experiment_data). I performed gender identification of gender using manually trained data and US census trained data.Different combinations of parameter for identifying features from user description and tweet text strings were considered for gender analysis and the best setting with ngram (1,2) with binary set to True was used. Below are few accuracy measures of the logistic regression model experimented with different settings:
CountVectorizer(tokenizer = stringtokens,min_df=1, max_df=1., binary=True, ngram_range=(1,1)) accuracy= .68
CountVectorizer(tokenizer = stringtokens,min_df=1,max_df=1., binary=False, ngram_range=(1,1)) acuracy =.66
CountVectorizer(tokenizer = stringtokens,min_df=1,max_df=1., binary=False, ngram_range=(1,2)) accuracy= .695
CountVectorizer(tokenizer = stringtokens,min_df=1, max_df=1., binary=True, ngram_range=(1,2)) accuracy= .715

EXPERIMENTS

Gender Inference and Analysis:

1. Using US census data for training

a) Training 200 tweets

With US census data I labelled 200 tweets into 3 categories as (-1) unknown, (0) males and (1) females. Then using a Logistic Regression model I trained this data with 5-fold cross validation accuracy of 71.5% and used it to predict the genders of the experiment_data. This experiment showed the difference in male and female awareness as assumed in hypotheses 1. Also, the results show unknown gender category as the highest aware among these three classified categories. This is because the model is trained on census data which can only identify males and females and keep the rest as unknown. There are many research

organizations and Breast Cancer Foundations working on the awareness programs to whom the census trained data might have categorized as unknown gender.

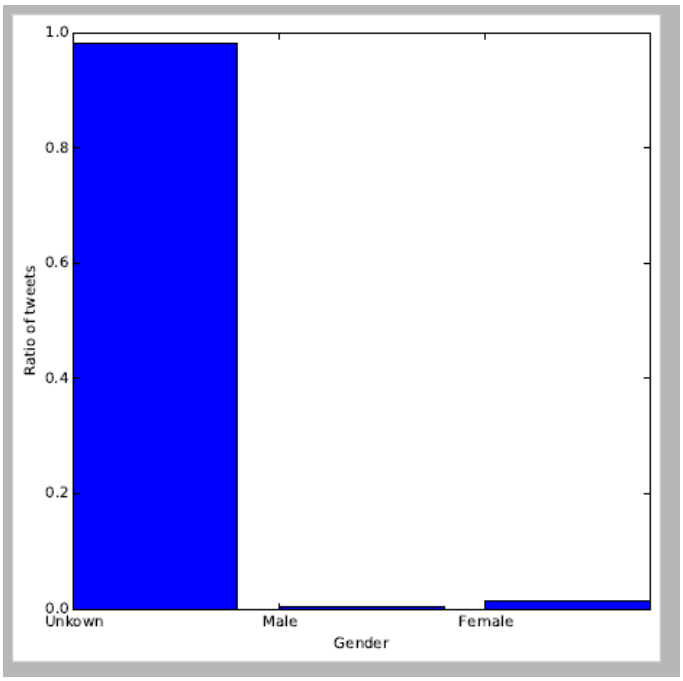
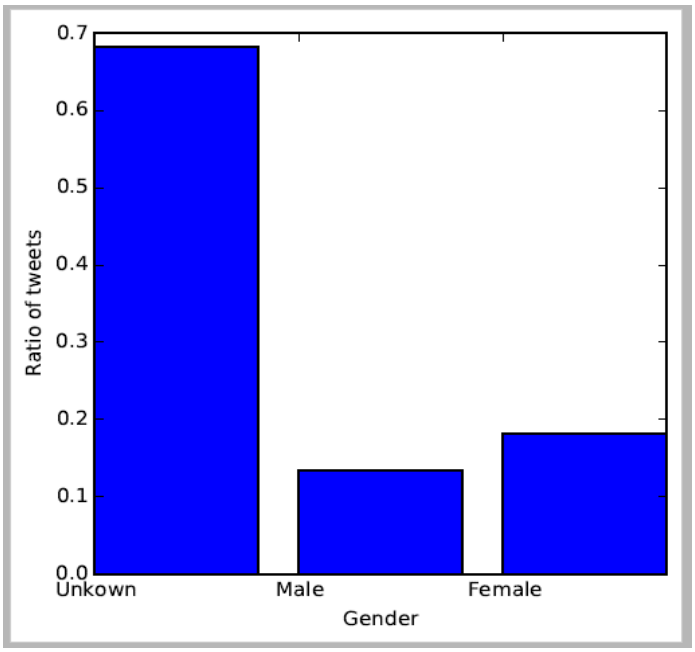


Fig:1 Messages classified by gender in three categories

b) Training complete data

This experiment is done so as to verify if the accuracy can be increased with larger training set. In this experiment I trained the model on complete data and predicted the genders with 82.5% accuracy. This provides an example for overfitting data and explains why we need to identify the correct size of training data to be considered when dealing with large size data for experiments



.

Fig:2 Gender classification with census data for complete data as training set

2. Using manually labelled data for training

In this experiment I manually labelled data set of 200 tweets using user description and tweet text in 4 categories as: (-1) unknown, (0) males, (1) females and (2) organizations. I trained this model using Logistic Regression and predicted gender for experiment_data. It was found that it's not actually the unknown gender but the organizations who tweet more about breast cancer. This increased per category count of each gender but the accuracy of this model was observed to be 43.5% which is less than experiment 1. The reason for lower accuracy might be because the manually labelled data introduced a new category i.e. organizations. The increase in counts for male and female may be because the census data removed the ambiguous names that fall into both categories of male and female, and few among such names were marked to either one of the categories while manually labelling the data thus providing more samples for respective categories.

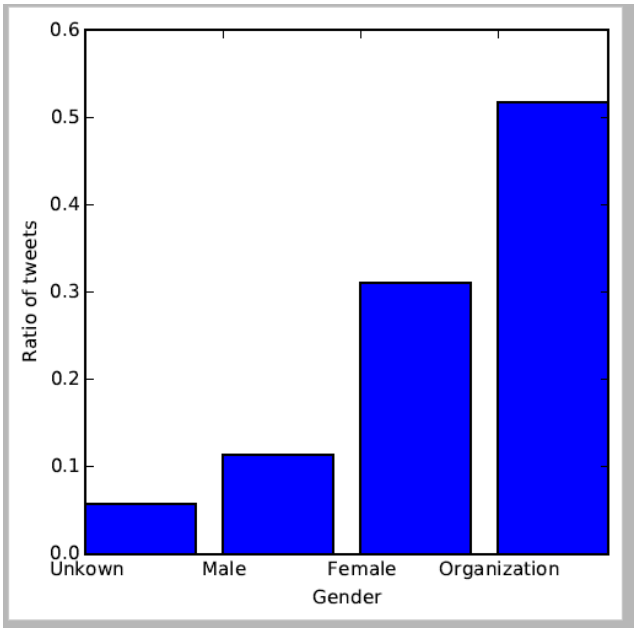


Fig:3 Gender classification with manually labelled data in four categories

Location Inference and Analysis:

The location analysis was done using the user provided locations. I processed the user location tokens for identifying if any of those matches with the US places list and then reporting the State of this location. For example if a user reported 'Chicago USA' as location on the profile then 'IL' was reported as the location. Results from the location analysis shows 10 least aware locations as: VT,DC,PR,WY,RI, DE,WV,NH,SD,TN which can be correlated (a few exceptions) to 'Rates of Getting Breast Cancer' published by Centre for Disease Control and Prevention in fig:5

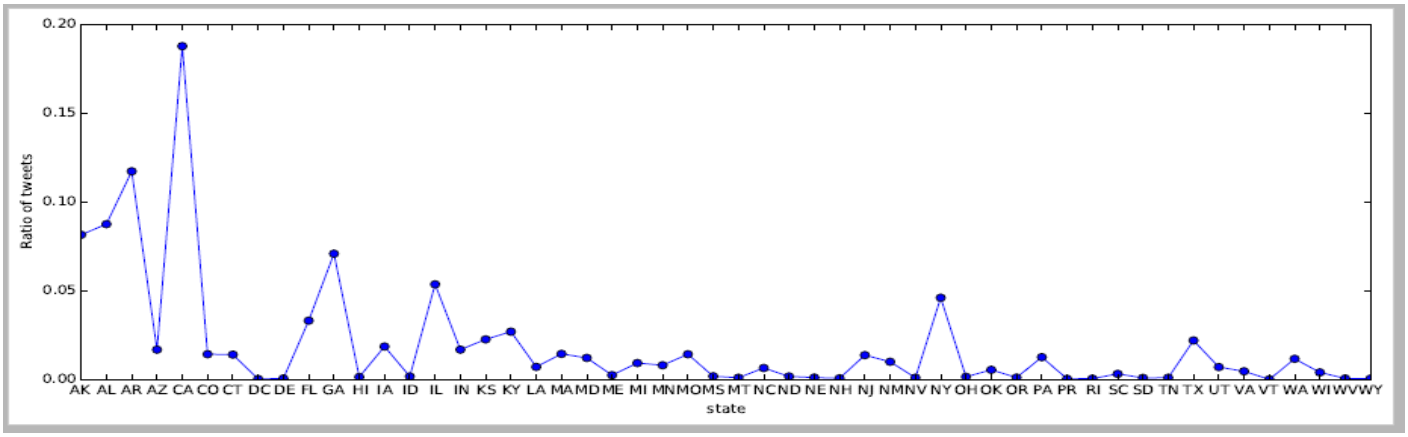


Fig:4 Awareness proportion of users reported by states

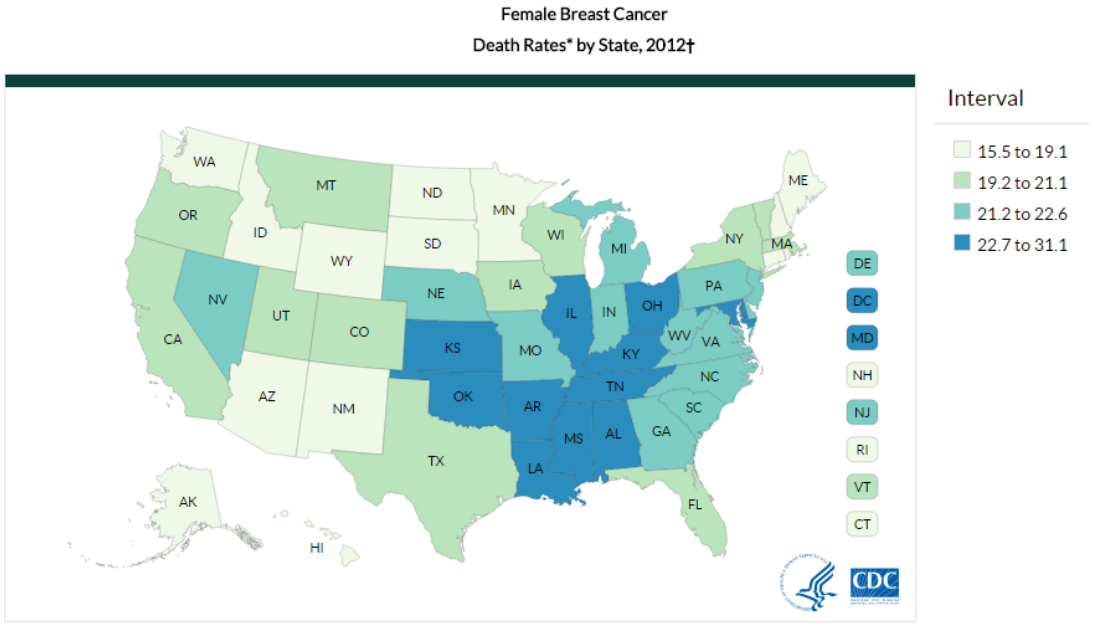


Fig:5 CDC reported breast cancer incidences by State

RELATED WORK

Many projects in demographic analysis have utilized US census names for gender inference. This project implemented two approaches for inferring gender, one using census names and second by labelling data manually. Though training manually labelled data could not beat the accuracy of census trained data it could identify organizations as a different category which actually was one of the highest contributing category towards the awareness cause than individuals.

CONCLUSION AND FUTURE WORK

It reflects from the gender analysis that organizations are the ones who talk more about awareness causes. In spite of many initiatives taken by such organizations male and female counts remain low, for discussions on awareness programs, compared to the responses that, could be observed if one analyzes Justin Bieber's latest album release. Female count as assumed in hypotheses 1 appeared greater than male count. Results from CDC also proved the hypotheses with some approximation. Maintaining a list of organizations and using it with male and female names list for training data might have given more insights for the presenting the correct counts of respective categories and can be considered for future implementations. Location inference could help in presenting more appropriate data if all users report their location on profile.

References

http://www.cdc.gov/cancer/breast/statistics/state.htm
Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, Jeremy Rodrigue, 2012, A Demographic Analysis of Online Sentiment during Hurricane Irene.
Elaine Cristina Resende and Aron Culotta, 2015, A demographic and sentiment analysis of e-cigarette messages on Twitter.