Demographics and Sociolinguistics

Astrid Machholm, astma@itu.dk
Christoffer Ebbe Sommerlund, csom@itu.dk
Gustav Bakhauge, guba@itu.dk
Kasper Grønbæk, kgro@itu.dk
Sarah Dueholm Ramezanpour, sarr@itu.dk

28th of February 2020

Contents

1 Introduction

This report present an analysis performed on a data set of around 5000 tweets containing both text material from many different authors along with their demographic information like gender and US state. By applying Natural Language Processing (NLP) through statistical calculations, we can determine which words are indicative of demographic features.

2 Methodology

In the process of settling on a data set, we made some initial plots to get an overview of the distribution of features. Some feature distributions like 'state' were heavily skewed, so we decided to only include data on the 5 most frequent features. Similarly, we avoided rarely mentioned words (we chose min of 20).

Before starting the actual analysis of our chosen data set we cleaned the data. Removing all unknown characters, splitting contractions into separate words, removing stop words, and stemming all words to decrease the amount of unique words in the total vocabulary.

Thereafter, we shuffle all the data by rows whilst keeping the same column order. This ensures that the classifiers we create are not biased towards a specific ordering of the tweets.

The data set is split into k-folds. We chose this form of slitting our data for train and test because our data set is fairly small (around 5000 tweets when analysing for gender. Less for location).

get the most data to test on. Because our data set is fairly small it does not take too long to run the program each time and thus k-folds are

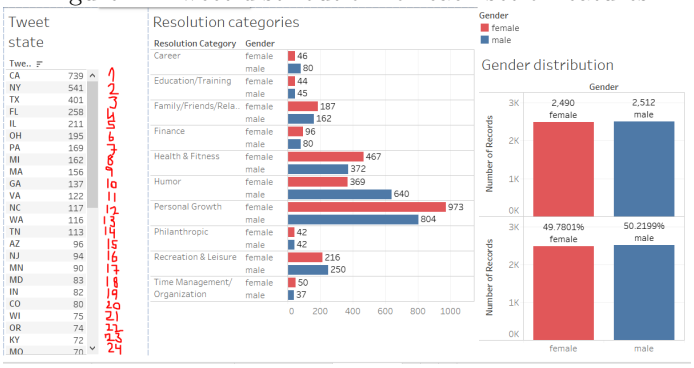
Preparation of data for specific model or general (vocabulary, vectorization, n-grams)

Choice of model (nltk or SKlearn or something else) The math behind it all

3 Data

Our data set contains tweets about new year resolutions from the dates 31/12/2014 and 01/01/2015. The data set contains information about some demographic categories. The authors of the tweets are males and female and from all over USA. Their location is split into three categories (city, state, and region) from which we have chosen [the most frequent states/region]. Furthermore, the content of the tweets are divided into topics and overall categories of the new year

Figure 1: Tweet distribution for each set of features



resolutions. The data set does not contain information about age or any other third demographic feature, thus we have chosen to use

4 Results

5 Interpretation

6 Error analysis

Small data set, even smaller when eleminating small features Too much different in size between largest and smallest state representation in tweets.

Since our featureset distribution is uneven, it would not be applicable to a tweet from a random feature. The estimated accuracy is only representative on data with around the same relative distribution. If we wanted the model to classify a tweet from a random feature, we could include an even amount of tweets from each feature. Since our data is already very limited, this approach was not feasible.

7 Concluding remarks and future work

8 Disclosure statement