Demographics and Sociolinguistics

Astrid Machholm, astma@itu.dk
Christoffer Ebbe Sommerlund, csom@itu.dk
Gustav Bakhauge, guba@itu.dk
Kasper Thorhauge Grønbek, 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 new years resolution tweets containing both text material from many different authors along with their demographic information and new year resolution category. By applying Natural Language Processing (NLP) through statistical calculations, we can determine which words in a tweet are indicative of these 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. This illustrated how feature distributions like 'author state' were heavily skewed.

2.1 Data cleaning

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.

2.2 Features and identifiers

NEEDS REWORDING/REVIEW The feature regarding geographic was split into three sub-features: location (city), state, and region. On a city-level there were too few tweets associated with each city, thus this feature was not usable. On a regional level the distribution was approximately which is desirable for getting the best test results. However, we felt that this was less interesting to work with due to the lack of specificity in subjects and thereby interesting findings. Therefore, we settled on analysing the location feature on a state-level, creating a mask that chooses the five states with the highest tweet-representation. NEEDS REWORDING/REVIEW

To create our identifies we created overall vocabularies for each feature consisting of sub-vocabularies including data on the five most frequent features in a feature set. Similarly, to avoid words that were rarely mentioned and thereby not actually indicative of a feature, we chose a minimum of 20 mentions per word. Using the entire text as input for our model is highly specific. Instead we created a binary check for each tweet, to check if select words (vocabulary) were present in this tweet. This array of boolean values is what the model used to predict a feature.

INSERT FORMULA!

For building our vocabulary, we wanted groups to be evenly represented. We had 2 methods for this, build from different philosophies. One prioritised words mostly unique to one of the features (meaning rarely present in other features), and another picked the most common words in each category. We vectorized the tweets using SKlearn to make a boolean matrix. The columns representing each word in the vocabulary, and rows representing each tweet.

In addition to gender and location we also implemented a vocabulary for the use of profanity.The festive circumstance surrounding New Year's Eve is often related to alcohol consumption and thus in many cases an increase in profanity. Therefore, we wanted to see if profanity use was in correlation to the timestamp of the tweet.

2.3 Niave Bayes

NAIVE BAYES: THIS IS COPY PASTED The assumption made here is that the predictors/features are independent, i.e. presence of one particular feature does not affect the other. Hence it is called naive. LATER: Another assumption made here is that all the predictors have an equal effect on the outcome. END COPY PASTE

The probability is calculated with the Naive Bayes formula:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Each x_i is a boolean of whether the word is present in the tweet. The result is the probability of having a feature, given the words identified in the tweet. Note: When the model is trained, the denominator is constant, so it can be omitted.

2.4 Training and testing the model

For training and testing, we used the k-folds method. K-folds allowed us to get the most out of our limited data by maximising our train data, while also testing our model on all of our data, minimising over-fitting or an unrepresentative test-set. We could do this, because the run time was not too bad for our fairly small data set (5000 tweets when analysing for gender. Less for other categories)

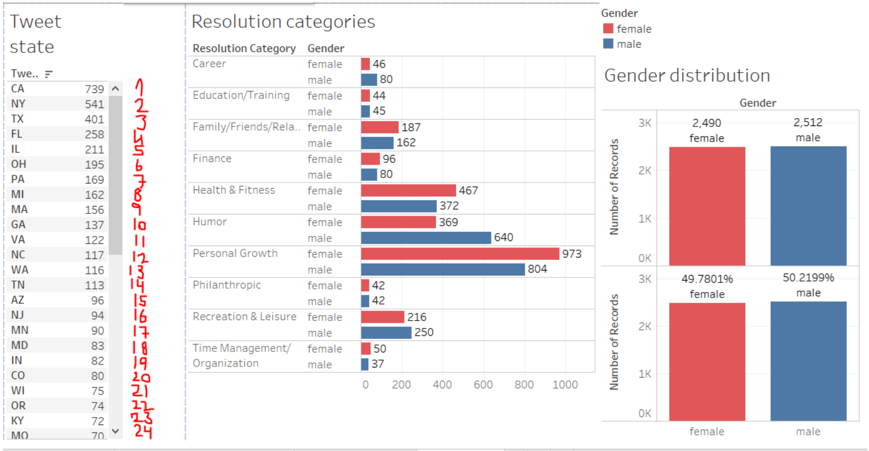
Preparation of data for specific model or general (vocabulary, vectorization, n-grams)

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

3 Data

The data set was located using Googles dataset search engine. It was provided by data.world/crowdflower/. It contains tweets about new year resolutions from the dates 31/12/2014 and 01/01/2015. Notable features of the data set include gender, location (city, state and region), new year resolution topic and timestamps. To compensate for the lack of the recommended third demographic feature (age),

Figure 1: Tweet distribution for each set of features (Made in Tableau)



Tweet state

Twe..		
CA	739	1
NY	541	2
TX	401	3
FL	258	4
IL	211	5
OH	195	6
PA	169	7
MI	162	8
MA	156	9
GA	137	10
VA	122	11
NC	117	12
WA	116	13
TN	113	14
AZ	96	15
NJ	94	16
MN	90	17
MD	83	18
IN	82	19
CO	80	20
WI	75	21
OR	74	22
KY	72	23
MO	70	24

Resolution categories

Resolution Category	Gender	
Career	female	46
	male	80
Education/Training	female	44
	male	45
Family/Friends/Rela..	female	187
	male	162
Finance	female	96
	male	80
Health & Fitness	female	467
	male	372
Humor	female	369
	male	640
Personal Growth	female	973
	male	804
Philanthropic	female	42
	male	42
Recreation & Leisure	female	216
	male	250
Time Management/ Organization	female	50
	male	37

we interpret the resolution topic that describe the individuals' interests, as an alternative. This allows for rich and interesting interpretations.

4 Results

Accuracy

NB in nltk (train/dev/test) : 12.6%

k-fold (sklearn) : 57.3%

NB in sklearn (train/dev/test) : 48%

5 Interpretation

6 Error analysis

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

Since our feature-set distribution is imbalanced, 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. The best way to make up for this imbalance is to collect more data but since this project Under and over sampling (remove some or copy some at random from either the largest or smallest categories - or a combination). Because the project is so small we can just leave it as is and argue for the outcome.

7 Concluding remarks and future work

8 Disclosure statement