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 **1**

2 Methodology **1**

2.1 Data cleaning 1

2.2 Features and identifiers 1

2.3 Naive Bayes 2

2.4 Training and testing the model 2

3 Data **2**

4 Results **3**

5 Interpretation **5**

6 Error analysis **5**

7 Concluding remarks and future work **5**

7.1 Future work 6

8 Disclosure Statement **6**

9 Appendix **6**

9.1 State 6

9.2 Gender 7

1 Introduction

This report presents an analysis performed on a data set of 5011 New year's resolution tweets containing text material from many different authors along with their demographic information and New Year's resolution category. By applying Natural Language Processing (NLP) and 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 exploratory data analysis to gain insight on our data. This illustrated how feature distributions like 'author state' were heavily skewed.[1]

2.1 Data cleaning

Before starting the analysis of our data set we cleaned the data: Removing all unknown characters, splitting contractions into separate words, stemming all words, ignoring stop words from the English dictionary and manually removing words that, in the context of the date set at hand, can be considered stop words.

2.2 Features and identifiers

The location-based feature was split into three sub-features: location/city, state, and region. On a city-level there were too few tweets associated with each city. On a regional level the distribution was fairly balanced, which is desirable. However, this was less prone to interesting findings due to a lack of specificity. Therefore, we settled on analysing the location feature on a state-level.

To create our identifiers we created vocabularies including data on the five most frequent features in a feature set. Similarly, to avoid words that were rarely mentioned and thereby not indicative of a feature, we chose a minimum of 20 mentions per word. This ensures a more even representation of the groups.

We had two methods for generating vocabularies: One prioritises words mostly unique to one of the features but rarely present in other features. The other picked the most common words in each category.

Formula for method 1:

$$Word\ value = \frac{(Word\ count\ in\ category)}{(Word\ count\ in\ total) * (Tweet\ count\ in\ category)}$$

We vectorized the tweets using SKlearn to make a boolean matrix of whether the word was in a tweet; The columns representing each word in the vocabulary, and rows representing each tweet.

[1] See Figure 1 in section 3

This array of boolean values is what the model uses to predict a feature.

In addition to gender and location we implemented a vocabulary for the use of profanity.[2]

2.3 Naive Bayes

As a model for classifying features from tweet content we have used the SKlearn Naive Bayes (NB) MultinomialNB[3] function. This allows us to use NB on multiple categories, instead of just having a binary relationship.

The model assigns, to each word in our vocabulary, the probability of being in a category.

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

Each x_i is a boolean value 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. The product of these probabilities and the probability of the category is the score, of which the model picks the highest as its prediction.

Our model assumes that the presence of one word does not influence the presence of another, i.e. they are independent.

2.4 Training and testing the model

For training and testing, we used the k-folds method, while also shuffling the data to prevent ordered data bias. 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 risk of over-fitting and an unrepresentative test-set. We could do this, because the run time was not too bad for our fairly small data set.

3 Data

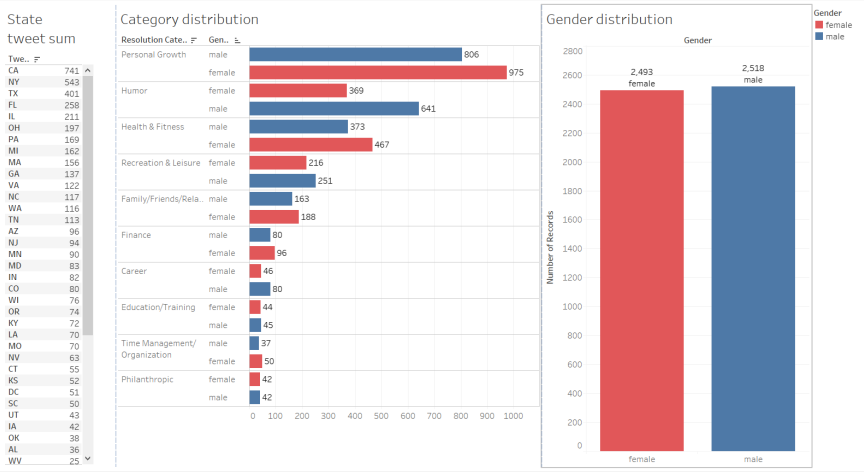
The data set was provided by 'crowdflower'.[4] It contains New Year's resolution tweets from the 2014/2015 New Year. Notable features of the data set include gender, location (city, state and region), New year's resolution topic and timestamps. To compensate for the lack of the recommended third demographic feature (age), we interpret the resolution topic as characterising the individuals' interests. This allows for rich and interesting interpretations.

[2] This did not prove useful due to the time restraints.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[4] https://data.world/crowdflower/2015-new-years-resolutions

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



4 Results

The following is the results for precision and recall of the three investigated categories, along with predictions based on profanity. Elaboration follows in section 5. Further confusion matrices and strong identifiers can be found in the appendix. Applying bi-grams only worsened results, so we have excluded this from the report.

Table 1: Classification report on given models and variables.

Accuracy		
Model	Variables	Accuracy
k-fold sklearn	Words/Gender	57.3%
k-fold sklearn	Words/Categories	50.8%
k-fold sklearn	Words/States	33.7%
NB in nltk (train/dev/test)	Profanity/Gender	49.0%
NB in nltk (train/dev/test)	Profanity/Categories	36.5%
NB in nltk (train/dev/test)	Profanity/States	18.4%

Table 2: Precision and recall for specific categories using the k-fold model.

Precision and Recall		
Category	Precision	Recall
Family/Friends/Relationships	71%	7%
Health & fitness	72%	54%
Humor	40%	17%
Personal Growth	48%	88%
Recreation & Leisure	53%	9%

Figure 2: Top 10 strongest identifiers

```
Family/Friends/Relationships: like stop time friend meet love family make friends people
Health & Fitness: healthy workout day lose weight smoking drink stop gym eat
Humor: gonna bitch going eat just ûï people like make stop
Personal Growth: love time let sta people better like life make stop
Recreation & Leisure: happy want sta travel mor meet going play read watch
```

Figure 3: Confusion matrix, prediction of resolution category. A strong colour in the decreasing diagonal is favorable

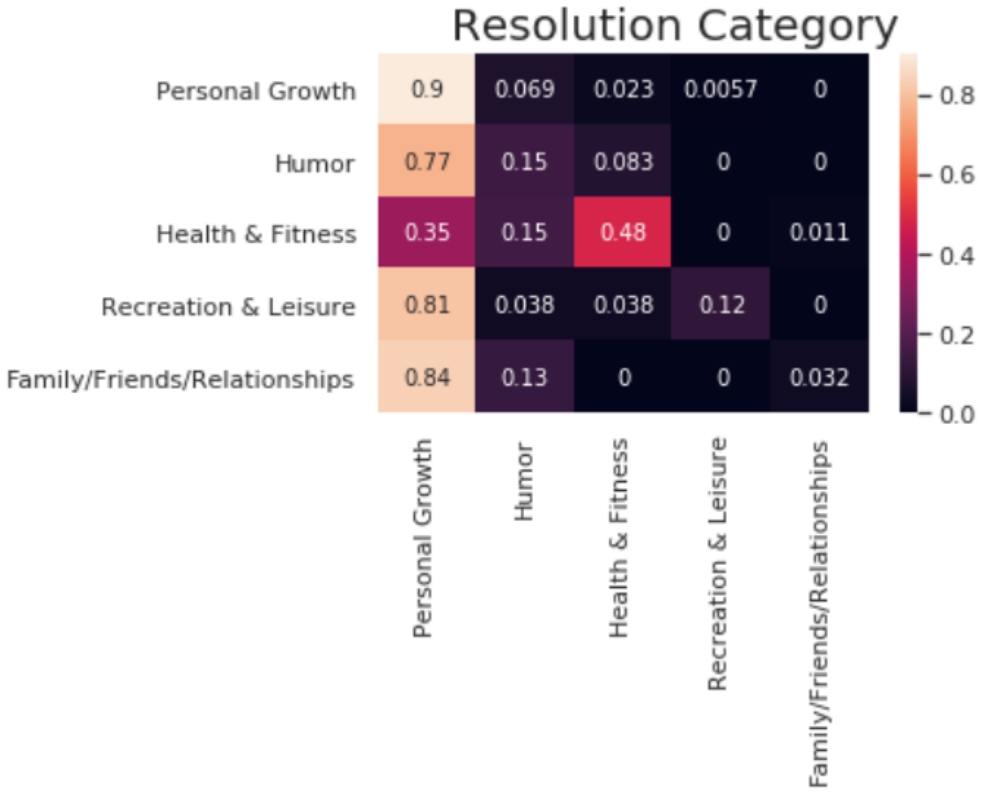


Table 3: Precision and recall for states.

Precision and recall		
State	Precision	Recall
California	34%	86%
Florida	0%	0%
Illinois	0%	0%
New York	28%	14%
Texas	28%	2%

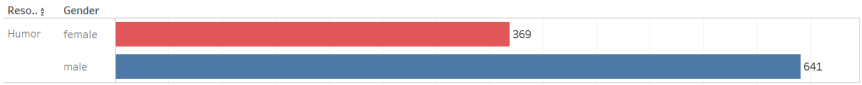
5 Interpretation

Generally the calculated accuracy for our models are quite low. Our confusion matrix shows a weak diagonal; most of the time, the model predicts the tweet to belong to the largest category, therefore the bright left column (Personal growth is the largest category).

The strong identifiers in Figure 2 have decent connections to their topics. In the first category, 'meet', 'friend', 'family' and more, we too would associate with relationships. Other categories show similarly good associations. Our model manages to correctly find the significant words for each category. For state (appendix), these associations are less clear.

We didn't find any significant cross-dimensional relationships between our major categories. The only exception that may be worth mentioning is the gender distribution for the category 'Humor'.[5] There was about twice as many males as females in the category. This emphasizes that being male is notably more indicative of writing in the context of this topic, humor. In general our data didn't contain as many cross-dimensional tendencies as we expected.

Figure 4: Gender Distribution of total tweets under the category "Humor" (Made in Tableau



6 Error analysis

Our feature-set distribution is imbalanced (except gender), thus it can not analyse a tweet from a random feature. To allow for this, we should include an even amount of tweets from each feature e.g. by over- or under sampling the data set. Since our data is very limited, this was not feasible. Preferably we would collect more data.

The reliability of Naive-Bayes is questionable since it assumes that all the predictors are independent, and this is highly unlikely in our case.

When we fit and transform in our code, we do it on the whole data set-therefore also the test set. This could bias the weight of out model this is an error.

7 Concluding remarks and future work

We ended up with a Naive-Bayes model that predicts demographic features from tweets. Its prediction are biased towards tweets from the largest categories, but still it had some success on categories smaller than this. The model is quite flawed, and has some bias in our results (small/uneven data,

[5] See, section 5, Figure 4: (Gender Distribution for the Category "Humor").

biased classifiers). Therefore our results are not reliable.

7.1 Future work

We found it interesting to see if hashtags could be an identifier for gender and/or resolution category. Due to time constraints we unfortunately weren't able to implement this.

The correlation between alcohol consumption on New Year's Eve and a possible increase in profanity could be researched by comparing the timestamp of the tweets and the usage of profanity.

To begin with the Twitter date set was very dirty. For slight improvements we could do even more cleaning and pre-processing of our data in order to improve the results.

8 Disclosure Statement

Most of the work has been a team effort where we worked together at ITU.

9 Appendix

9.1 State

Figure 5: Confusion matrix, prediction of state

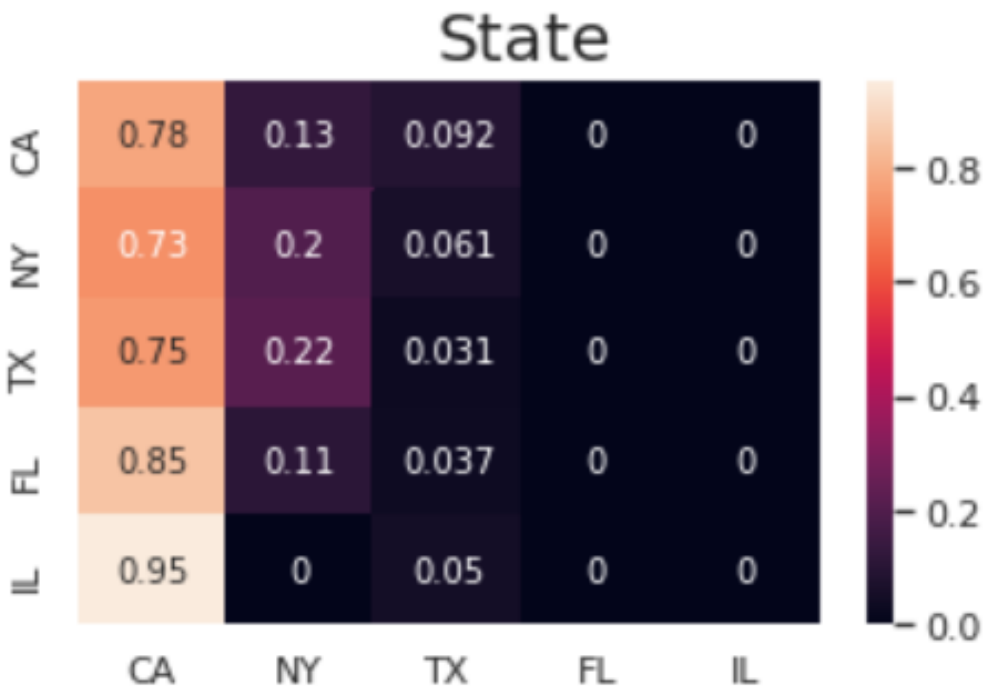


Figure 6: Informative feature, state, unigrams

```
CA: ûï going just love time want sta like make stop
FL: poss want happy money sta time day people stop make
IL: drink let fuck love day stop time happy like make
NY: like just time life good happy people make eat stop
TX: let happy time try sta like just better make stop
```

9.2 Gender

Figure 7: Confucion matrix, prediction of gender

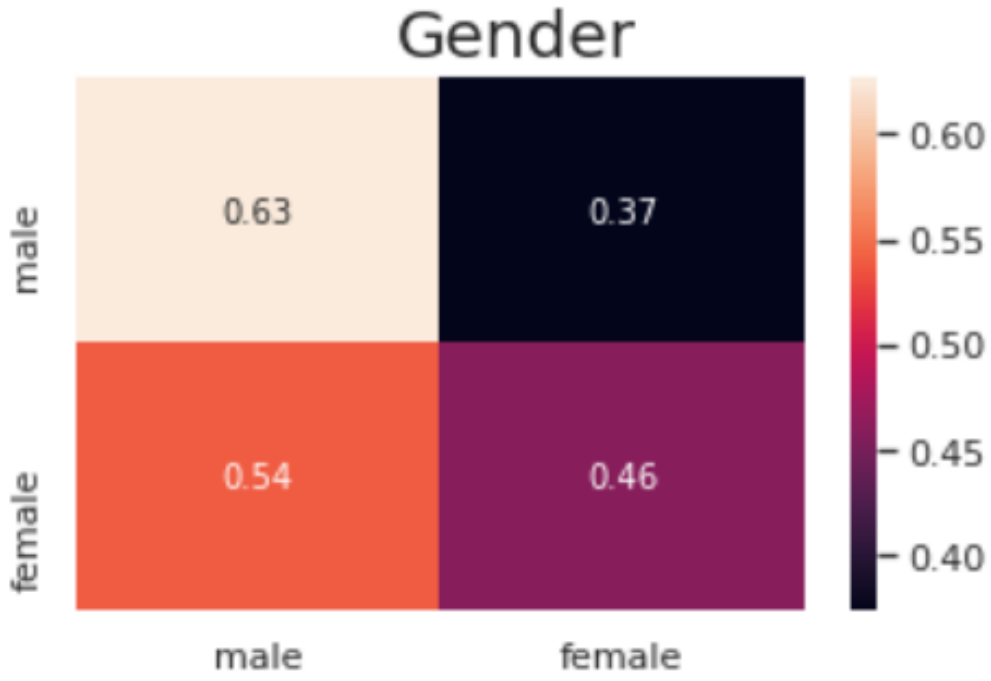


Figure 8: Informative feature, gender, unigrams

```
female: people day just better going time sta like make stop
```