First Year Project BSFIYEP1KU

Mini Project 1

Identifying the girl next door
- a study in natural language processing

Alan Ispani - alai@itu.dk

Christian Rønsholt - roen@itu.dk

Frederik Høngaard - frph@itu.dk

Mikkel Petersen - mikpe@itu.dk

Virginija Juozapaityte - virj@itu.dk

IT University of Copenhagen

Spring 2020

Introduction

This project is an exercise in natural language processing (NLP). Specifically, we shall analyse a large data set containing biography-essays and labels scraped from the dating site OkCupid to establish whether certain text/language is a strong identifier for a given label.

Methodology

This analysis seeks to expose whether certain language in a given text is an identifier for some class.

To do so, we have applied the `nltk`-library with the Naïve-Bayes classifier on a number of different essays. Naïve Bayes works by returning the conditional probability of some event - here that a text can be classified given presence of a specific token. Formally:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Applied in our analysis:

$$P(\text{label} \mid \text{n-gram in text}) = \frac{P(\text{n-gram in text} \mid \text{label}) \cdot P(\text{label})}{P(\text{n-gram in text})}$$

The `NaiveBayesClassifier` is applied on the data by firstly creating a shuffled feature set, i.e. we classify a certain essay as a tuple in the format (features,label), then splitting the feature set into a train-, dev- and test-set in a 80/10/10 ratio. The `nltk`-classifier is then trained on the train-data, fine tuned/developed on our dev-set and subsequently applied on the test-data - which is how the code is delivered in `code.ipynb`.

The evaluation of our classifiers has gone beyond measuring accuracy by introducing precision, recall and F1-score. These evaluation metrics give a deeper insight to the performance of our model and prevent incorrectly lauding the performance of a model on an imbalanced data set.

Data

The data is the *"OkCupid Profile Data for Intro Stats and Data Science Courses"* by Kim and Espededo-Land[1]

The data set is a `.csv`-file that contains 10 essays by each of $\sim 60,000$ users of the OkCupid dating service in San Francisco bay area. To each user upwards of 10 labels are given. Among these are sex, age and ethnicity, which will be the focus of our analysis. Location was **not chosen** as it is too specific (only within a 25 mile area). Hence, the ethnicity label was explored instead.

[1] Albert Y. Kim and Adriana Escobedo-Land, Journal of Statistics Education July 2015, Volume 23, Number 2; Retrieved from https://github.com/r-spark/okcupid

Pre-processing of the raw data consisted of two steps; the first step was to clean all the text of html-tags and punctuation so that words could be splitted on white-space.

Step 2 of pre-processing was balancing the distributions of labels. Males make up more than 60% of the raw data. Age and ethnicity distributions are similar for males and females alike, hence we removed approximately $11,000$ male entries at random resulting in an evenly distributed data set with respect to sex.

Results

The results of our analysis Naïve-Bayes and evaluation of our model are presented in Table 1.

The three classes, sex, age and ethnicity, are presented and divided into applicable subgroups. Sex is divided in male - female; age is split into ≤ 30 and > 30; and ethnicity is divided by white and non-white. Our model is analyzed on both essay0 (self-summary) and essay4 (interests). Precision, recall and F1-score are all calculated to give a full and extensive evaluation of the performance of our accuracy.

Moreover, Table 2 presents the strong identifiers for a label with a likelihood ratio for the given n-gram.

Table 1: Summary of results

Label	Text	N-gram	Accuracy	Precison	Recall	F1-score
Sex (m/f)	about me	Uni	62.1%	0.61 / 0.64	0.72 / 0.51	0.66 / 0.57
Sex (m/f)	about me	Tri	54%	0.53 / 0.57	0.82 / 0.24	0.65 / 0.35
Age (> 30 / ≤ 30)	about me	Uni	63.5%	0.65 / 0.63	0.51 / 0.75	0.57 / 0.68
Age (> 30 / ≤ 30)	about me	Bi	62.9%	0.62 / 0.65	0.49 / 0.76	0.56 / 0.68
Ethnicity (non-white / white)	about me	Uni	56.4%	0.42 / 0.71	0.64 / 0.52	0.52 / 0.60
Ethnicity (non-white / white)	about me	Bi	62.9%	0.48 / 0.66	0.24 / 0.85	0.31 / 0.74

Table 2: "Fun" and strong identifiers

Label	Given N-gram	Ratio	
Sex	easi go guy	78.5 : 1	male : female
Sex	girl next door	26.3 : 1	female : male
Sex	tri new recip	12.3 : 1	female : male
Age	sensual	5.9 : 1	$> 30 : \leq 30$
Age	haha	5.6 : 1	$\leq 30 : > 30$
Age	adventur time	7.6 : 1	$\leq 30 : > 30$
Age	harry potter	6.2 : 1	$\leq 30 : > 30$

Interpretation

The model is a relatively good predictor - acc. between $> 62\%$ for the best fitting models. It is best at identifying gender. This is most likely because this group is the simplest i.e. there are only two different labels for the class. This also meant it was easily balanced out.

Conversely, it performs worst when trying to label ethnicity. We attribute this fact to the composition of the ethnicity labels; around half the group is labelled white, the other half has ~ 200 different labels that we categorize as non-whites.

The strongest identifiers are typically bi- and trigrams, although the model is more accurate analysing unigrams. An explanation could be that certain trigrams are more distinct, yet less frequent. It makes sense that "easi go guy" and "girl next door" are distinct and that they would not appear as frequently in all of the essays as the n-grams that are shared by both groups.

Error Analysis

One needs to consider the nature of this data i.e. scraping a specific dating service. The groups are not perfectly balanced - e.g. median age is 30, ethnicity consists of half whites and half distributed between hundreds of subgroups.

As relates to the analysis method a shortcoming might be that by applying Naïve-Bayes we assume independence among the variables in the data set which probably is far from the actual case.

It would be a possibility to do a multi-class analysis instead of binary classification. Splitting of the "non-whites" group into more distinctive groups e.g. Asians, Hispanics, African-Americans etc. could feasibly raise the accuracy predicting non-white labels.

Concluding remarks and future work

Our model applies the 2000 most frequent n-grams from the essays in order to predict whether the presence of a certain n-gram identifies the label of the essay. However, frequent tokens might not be that independent between labels.

Instead we could look into identifying a label by maybe less frequent, yet more distinct language usage. With more computing power and time we could do a larger selection of n-grams, then do several iterations with the weakest identifiers removed for each step.

In relations to ethnicity, essay4 (interest) might be better than essay0 (self-summary), as it might generate stronger identifiers for ethnicity. As different cultures have different interests e.g. books, that are unique for their heritage. E.g. the qur'an has a great likelihood to be a strong identifier for middle eastern ethnicity.