Department of Computer Engineering

CS 464 - Introduction to Machine Learning

Spring 2016

Progress Report

Foursquare Meal Classification

as Liked or Disliked Based on Comments

Group 17

Burak Eserol

Fikret Kaya

Işıtan Yıldız

Mert Gür

Tables of Contents

List of Tables, Equations and Figures

Problem Description

Our project will cluster certain menu items from a restaurant, and will give a percentage for each menu item how it is liked and disliked by the commenters.

Introduction

Foursquare is a social media platform where people find different places to eat, shop, and visit. People generally rely on other people's comment about the places to decide which place they want to go. When we consider the number of users in Foursquare, there are huge number of comments. According to Foursquare chief executive Dennis Crowley, there are more than 60 million registered users and more than 50 million monthly active users using Foursquare [1]. For the places where people can eat and drink, people comment about whether they liked the food or not. However, everybody has different understanding of taste and it is very hard for users to read all of the comments before they decide where they want to eat.

Our aim in this project is to classify certain menu items from a specific restaurant as liked or disliked based on people's comments. Thus people will be able to see which menu item is most liked or not without need to read all of the comments about that restaurant.

Dataset

We decided to use Turkish Foursquare comments for the restaurants. We generated the dataset from 10 different well known restaurants until now. The reason behind selecting Turkish comments is to understand them easily and label them accurately. Name of the restaurants and number of comments generated until now are as follows:

Restaurant	Number of Comments
Zigana Pide	432
Yıldız Aspava	618
Quick China	243
PizzaİIforno	256
Özsüt	171
McDonalds	124
Mado	939
Liva Bistro	467
CookShop	582
Burger King	178

Table 1 : Restaurant Names and Number of Comments in Dataset

Work Done So Far

Labelling

We labelled all of the comments mentioned above. Labelling consists of the food names that are liked or disliked.

Preprocessing

Preprocessing of the comments in dataset contains several steps:

- Convert all of the words into lower case words.
- Change all Turkish Characters into respective English Characters.
- Remove all punctuation and symbols from comments like "!, /, %, & etc..".
- Stemming words. Replace words that have same meaning but different morphemes with a same word.
  This step is very important in terms of clustering the words. Turkish is a highly agglutinative language and Turkish words have many grammatical morphemes or endings that determines the meaning of the word. Therefore, we group up words with different ending but same meaning. For example, words "Guzeldi", "Guzeller", "Guzelmiş" have different endings, yet they all have the same meaning "Guzel".
- Replace words that have typo with the word that has minimum distance with. We detect typos if we can not find the word in the dictionary, we assume that word has typo. We use Levenshtein Edit Distance to measure the distances between words.
  For the last two steps, we use a Turkish Dictionary dataset[2] that contains approximately 1.300.000 words in it.

Meal Detection

In meal detection part, first the meals as labeled in the training labels are put into separate arrays for each restaurant considered. This step consists of extracting all food labels and adding only non-duplicate entries into the menu arrays.

Next, the arrays containing menu items are extended such that each array also includes each item it previously had in a subarray, so that these subarrays can be extended later to include different but equivalent names for the same menu item. This is done such that the index of the subarray to be extended is equal to length of the main array over 2 (integer division) plus the index of the original name of the menu item for said subarray.

Afterwards, each comment and label pair in the training data is searched such that, in each comment any word chain consisting of the same number of words as one of the menu item names in the label for that particular comment is compared to all alternative names of stated menu item in the respective subarray according to levenstein edit distance. The chain of words with the lowest edit distance over length value is found, and if this value is below a certain threshold, the word chain is removed from the comment, respective menu item is removed from the label, and unless the word chain is already in the respective subarray, it is added to that subarray. This process loops until no change happens in any of the subarrays. Currently, we have a recall of approximately 0.52526 and a F-Measure of approximately 0.68875.

To calculate Recall, Precision and F-Measure values, we used below formulas as shown in the class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Equation 1 : Recall Formula

$$\text{Precision} = \frac{TP}{TP + FP}$$

Equation 2 : Precision Formula

$$\text{F} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

Equation 3 : F-Measure Formula

Threshold	Recall
0.1	0.10450
0.2	0.27405
0.3	0.42907
0.4	0.47889
0.5	0.52526
0.6	0.58201
0.7	0.65398
0.8	0.69827
0.9	0.79239
1	0.83253

Table 2 : Threshold and Recall Values

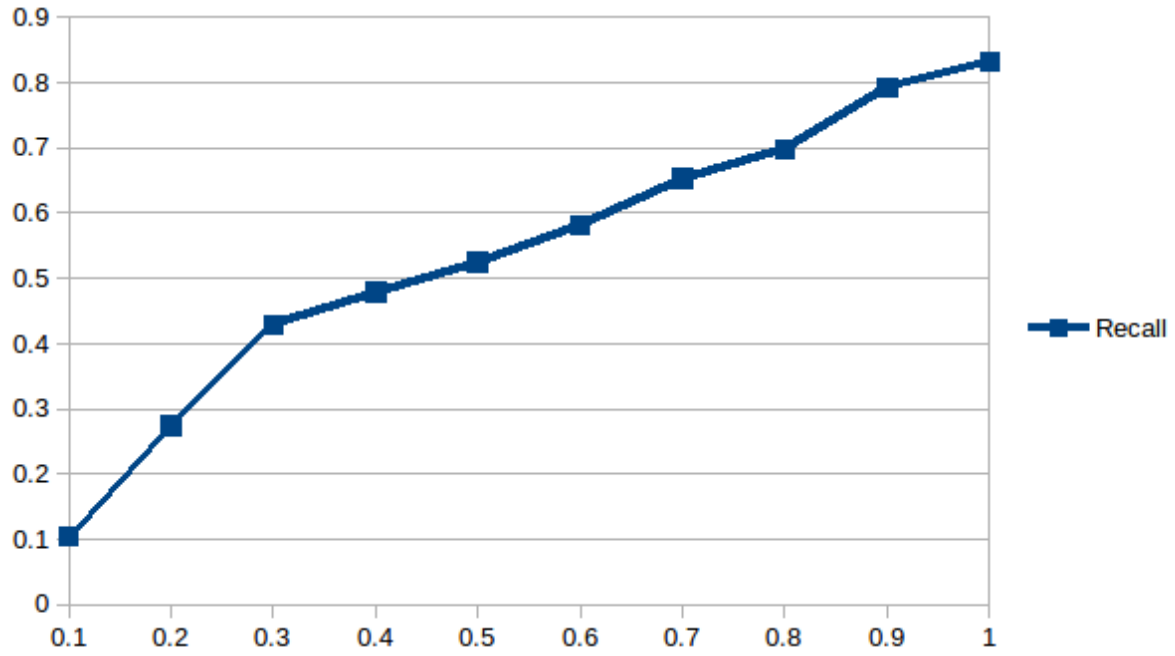

Figure 1 : Threshold vs Recall Graph

Remaining Work to be Done

- Using a binomial bag of words model on comments in the training data to generate positive and negative probabilities for each word.
- Selection of words to be used as features according to mutual information.
- Checking the words not selected for use in classification to generate more alternative names for menu items.
- Using the trained model on test data and checking the statistical measurements for our classification.

Division of Work

We divided our project into 3 parts; preprocessing on text, meal detection on those preprocessed texts and classification. Up to the present, we have finished preprocessing and meal detection. Our team consists of 4 members and all members involved in the research process; however, the implementation of divided parts have been performed in pairs as follows:

- Preprocessing of dataset
    - Burak Eserol

    - Mert Gür
- Meal detection on preprocessed dataset
    - Fikret Kaya
    - Işıtan Yıldız

The remaining works are divided into 4 portions and each portion will be performed by group members as follows:

- Generating positive and negative probabilities for each word, using binomial bag of words model : Fikret Kaya
- Selection of words to be used as features according to mutual information : Mert Gür
- Checking the words not selected for use in classification to generate more alternative names for menu items : Işıtan Yıldız
- Using the trained model on test data and checking the statistical measurements for our classification : Burak Eserol

References

[1] Dennis Crowley, VentureBeat's 2015 GrowthBeat Conference, August 18, 2015.
[2] Turkish Dictionary Dataset. github.com/hrzafer/resha-turkish-stemmer/blob/master/src/main/resources/generated.dict