Finding the best dish of a restaurant using NLP Techniques

CSCI 6509
Deepak Munjal
B00748375

P1 – Project Statement

1. Project Title:

Finding the best dish of a restaurant using NLP techniques.

2. Names of the member(s) of the group:

Individual Project

3. Problem statement:

There are many restaurants and each have n number of dishes. User reviews vary with different dishes, restaurants, personal choices etc. Finding best dish of a restaurant based on the user reviews is an important problem in this regard. This is not only important for customers but for owners too. Customers find this information interesting because this allows them to know the name of the best dish and get an idea about the restaurant that they plan to visit. And similarly it is important for the restaurant owners too because it allows them to know view of general public about their restaurant and dishes & can help them grow their business. This can be considered as an important business problem and this project tries to provide a solution to it by analyzing user reviews. For that, we will be scraping reviews from online resource(s) and these reviews will be analyzed further to see if it says about the dish positively, negatively or are neutral and then best dish of a particular restaurant can be found based on the analysis.

4. List of possible approaches with citations to relevant work:

The first task is to scrap user reviews from online resource. In my case, I will be using www.yelp.ca[3] to scrap user reviews. This will be done in Python. I planned to extract user reviews for twenty-ish restaurants of Halifax and then use this data for the project. All reviews will be manually labeled in beginning for the dish name(s) mentioned in any particular review along with its class i.e. positive, negative or neutral depending on the review. Once this task is done we will divide the data into training and test data, most probably in 70%-30%. Then for sentiment analysis[6] of the reviews, we will use various learning models with this data i.e. Support Vector Machine[5], Logistic Regression and Random Forest[4]. The features for these models can be either simple words of the review directly or it can be score of various words written in the user review or can also use words to vectors approach in this. These are the various possible approaches, one of these will be used while working on that stage of the project. Classes will include Positive, Negative and Neutral[1]. The training data will be used to train the model. Out of the learning models discussed above, the one with the best accuracy will be used for this project. The accuracy will be calculated by the confusion matrix for all the learning models. Using this whole approach, we will try to find best five dishes of a particular restaurant. One issue with this approach is that we are assuming that the names of the dishes mentioned in the user reviews will be exactly same as the one on the actual Restaurant Menu. But since this is not the scenario in most cases. Eg. some users may write the dish name as “chips” instead of the actual name as per restaurant menu “Potato Fries” etc[1]. So this project will include one more task as well, which is Dish Name Mapping[1]. This task is independent of the learning model task. In this, matching of the dish name, written by the user in review, will be done with the dish name on the restaurant menu. This matching maybe of various types - Exact match, Partial match, Substring match, Fuzzy match etc[1]. In Exact match, the name of the dish mentioned in the user review matches exactly with the one mentioned in the restaurant’s menu. Similarly in the Substring match, some part of the dish name mentioned in the user review matches with the one mentioned in the menu of the restaurant, in order. In Partial match, at least half of the name matches and similarly in fuzzy match, it handles if there are any typos in the dish name written in the user review and hence it doesn’t match partially or as substring but as fuzzy match. So using these approaches, we will first find the dish names mentioned in the user review[7] and then will do sentiment analysis of that review, trying to find if the review talks about the dish in a positive, neutral or negative sense and how strong the expression[1] of positivity, negativity or neutrality is. Depending on these, a score will be assigned to each dish and then total score will be calculated for each dish for all the reviews, which will decide the best dishes of a particular restaurant.

5. Project plan for the rest of the term:

Task	Schedule
Scraping of data from Yelp website	Currently working on till 15th March
Manual labelling of the reviews	16th March to 17th March
Dish Name Mapping task	18th March to 22nd March
Designing learning model and evaluation	23rd March to 29th March
Testing the project	30th March to 31st March
Creating Presentation (P)	01st April to 03rd April
Working on Report (R)	04th April to 08th April
Review of the whole project	09th April
Presentation of the slides	10th April
Report Submission	10th April

6. List of references:

[1]"Gong, A., & Lu, J. Picking Out Good Dishes from Yelp.". [Online]. Available: http://nlp.stanford.edu/courses/cs224n/2015/reports/4.pdf

[2]V. Keselj, "CSCI 4152/6509 - Course Project", Web.cs.dal.ca, 2017. [Online]. Available: https://web.cs.dal.ca/~vlado/csci6509/project.html

[3]"Restaurants in Halifax - Yelp", Yelp, 2017. [Online]. Available: https://www.yelp.ca/c/halifax/restaurants

[4]"Random forest", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest

[5]"Support Vector Machines". [Online]. Available: http://cs229.stanford.edu/notes/cs229-notes3.pdf

[6]"Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).", 2004. [Online]. Available: http://dmlab.uos.ac.kr/html/lecture/Textmining(2007-2)/Sentiment%20analysis%20using%20support%20vector%20machines%20with%20diverse%20information%20sources-2004.pdf

[7]"Tsubiks, O. (2012). Mining Consumer Trends from Online Reviews: An Approach for Market Research.", 2012. [Online]. Available: https://dalspace.library.dal.ca/bitstream/handle/10222/15401/Tsubiks%2c%20Olga%2c%20MEC%2c%20CS%2c%20Aug%202012.pdf