Individual Report on Project (#36): Rating Classification of Yelp Reviews

Name: Vaibhav Behl

Email address: vbehl@usc.edu

Rest of my group: Farhan Khwaja, Nagarjun Pola, Yao Fan

Name of the Bitbucket repository for your group's project: csci544-group36

1. Project Overview

The goal of our project was to predict ratings of user reviews found on Yelp. The ratings were on a scale from 1 to 5, where lower rating meant a negative review and higher rating corresponded to a positive review. So this was a multi-class classification problem. But, this problem can also be seen as a sentiment classification, since the score predicted by our system directly corresponds to the sentiment associated with the review. To help in identifying the sentiment of the review and thus help in predicting the correct star rating, we used some advanced features like selective bigrams for positive and negative words which gave us a good increase in accuracy thus validating our hypothesis that there is a positive correlation between review sentiment and its rating. I'll discuss these advanced features later on in more detail.

To start with, we used the Yelp Academic dataset to get user reviews for local businesses. The dataset was highly skewed because there were more 5-star reviews compared to 1-star reviews. To deal with this we selected our dataset to have equal number of reviews for each class. During pre-processing we performed the standard text cleaning by removing special characters and punctuations, and also applied porter-stemming. We noticed here that using a stop word remover decreased the accuracy since it would remove words like 'very', which would transform reviews from 'very good restaurant' to 'good restaurant'. We also used two different techniques to generate our features, namely- Word2Vec and Bag of words (n-gram).Both of these techniques had different data workflow, but we evaluated them on the same set of algorithms, which were:

- Naïve Bayes
- Logistic Regression (stochastic)
- SVM (stochastic)
- Random Forests

All these algorithms were from Scikit-learn packages. For the bag of words model, we used four different features for these algorithms- Unigram TF (term frequency), Unigram TF with rare words removed, Unigram TF-IDF with rare words removed and Advanced Features (selective bigrams). The first three features here are self-descriptive. The Advanced Feature here consisted of making two bigrams (with left and right word) of all the words which were found in a list of pre-defined positive and negative words, and also adding a normalized count of positive and negative words found in each review. These four features were evaluated in the bag of words model over the four models we built using the classification algorithms.

Finally, we evaluated the results on two different corpora- One consisted of only restaurant reviews, while the other one didn't make such a distinction. We did this to check whether the domain of a review has any effect on the classification accuracy. One key problem we faced while evaluating the results was that the F-score was not such a good measure of classification accuracy. This was because it equally penalized in case when our algorithm predicted 4 for a rating of 5, and when it predicted 1 for the same review with rating 5. Clearly in both cases there is a big difference, but the F-score will not be able to measure that. To deal with this I suggested that we use MSE (mean squared error) to measure the accuracy of our classifier. This worked well because it penalized highly when the predictions were very wrong compared to when the prediction was closer to the original class(example, for predicted:4 and original:5, MSE is 1, but for predicted:1 and original:5, MSE is 16). In the end, our evaluation consisted of plotting MSE over different features for the four models we used. We found that Naïve Bayes using Advanced features performed the best for both the corpus (restaurant and mixed).

2. My Project Tasks

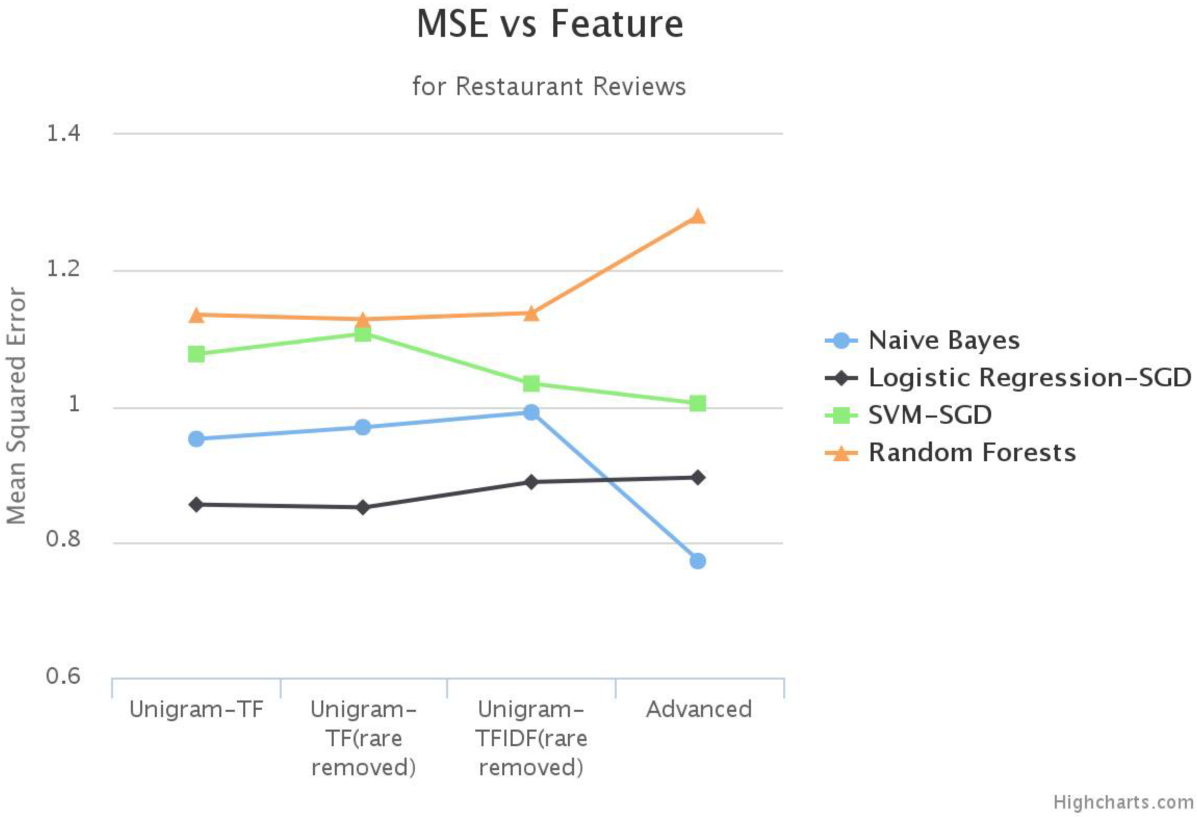
We split the work of this project into two parts: one to generate results starting with Bag of Words (n-gram) model and other for Word2Vec model. I and Yao Fan worked on the Bag of Words model approach. I worked mostly on data pre-processing and model generation, while Yao Fan did all the evaluations. Apart from this, we both worked on the 'advanced' feature generation, which was one of the four features we used.

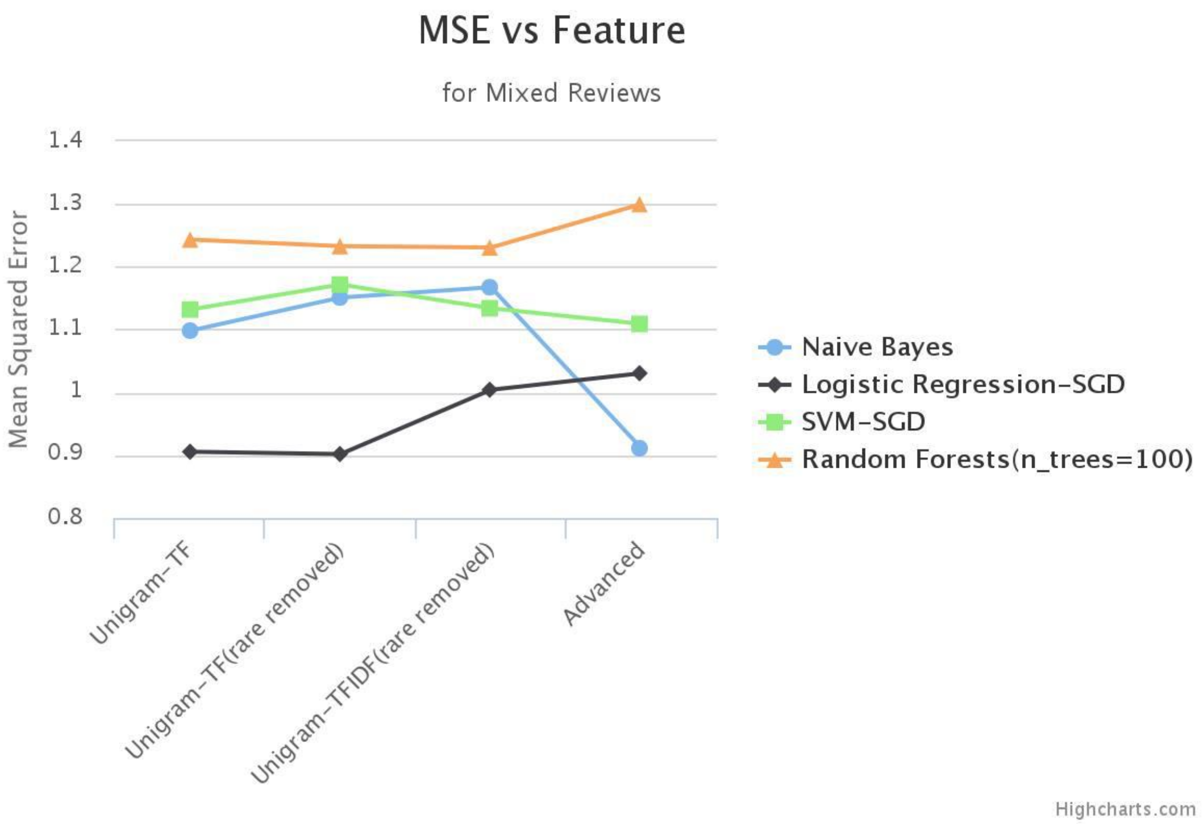
Following are the major steps involved for the Bag of Words model approach:

- Data Pre-processing: This stage was mostly handled by me. In this stage I used as input the Json data for reviews and businesses, from Yelp. All our features are represented in the standard SVMLight format which were converted to python's sparse matrix format before being passed to algorithms. So all the Json data was first converted to CSV format and then to SVMLight format. In between these conversions we performed the other pre-processing steps like:
    - Lower-casing all words
    - Removing all non-alphabet characters (special characters and punctuations)
    - Removing words with length less than or equal to two (this was used as a replacement for NLTK's stop-word removal)
    - Porter-Stemming
- Feature Generation: While converting from Json to CSV format, we also converted to the specific feature format we wanted it in. There were four different feature format, and all of them used the previous one as input, i.e. were generated iteratively.
    - Feature1: Unigram TF (term frequency) - This was the simplest feature which consisted of frequency of each word represented in the format (word_index:frequency_count). Along with this, I also generated a vocab dictionary file which would help when generating the later features.
    - Feature2: Unigram TF with low frequency words removed- For this feature vector we used the dictionary file from previous step and removed all words in the feature file whose frequency was below a certain threshold. After this step, we regenerated the feature file and its corresponding new vocab dictionary file.
    - Feature3: Unigram TF-IDF with low frequency words removed- For this features we used the scikit-learn's TF-IDF transformer.

- Feature4: Advanced: This feature consisted of making two bigrams, one for left word and other for the right, for all the words that were present in a pre-defined positive and negative words list[2]. Having a vocabulary dictionary file helped in identifying the actual word from their integer index and then creating two new bigrams, which were subsequently added to the vocab dictionary file as well.

- Evaluation: During the evaluation we used MSE (mean squared error) instead of the F-score because they were better suited for this purpose (I talked about this in detail in the project overview section). All our results were evaluated using a 5-fold CV on the whole labelled dataset. Below are two graphs which describe the progression of MSE for different models over different feature. First graph is for Restaurant only reviews and second one is for mixed (all) reviews.





- Observations: It can be inferred from the graphs that logistic regression performs the best for the first three features, but naïve bayes does the best when using advanced features. This is true for both the graphs. It can also be seen that in general the range of MSE values is less for Restaurant corpus compared to the mixed corpus. The results for Word2Vec are not in this graph, but their performance was poorer than naïve bayes using advanced features.
  Note: To see all numerical results please refer to the file (all_features_n_gram.txt) in the repository.

Apart from this I also wrote the python script to filter out restaurant only reviews from general user reviews based on the business category associated with the business id for that review. This script was also used by other two guys while they were working on their Word2Vec implementation. Also the initial version of the text-cleaning script that I used was written by Farhan, but it was modified later to suit the project needs.

3. Project Resources

- Our dataset was obtained from the Yelp Dataset Challenge (http://www.yelp.com/dataset_challenge) Which consists of 1.6 million reviews by 366k users of 61k different business categories.
- Word2Vec libraries from Gensim package (https://radimrehurek.com/gensim/).
- Porter Stemmer from NLTK (http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.porter)
- Scikit-learn for the Machine Learning Algorithms (http://scikit-learn.org/stable/index.html)

4. References

1. Okanohara, D., & Tsujii, J. I. (2005). Assigning polarity scores to reviews using machine learning techniques. In Natural Language Processing–IJCNLP 2005 (pp. 314-325). Springer Berlin Heidelberg.
2. Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342-351). ACM.

5. Review 1

The group number was: 2

What was the goal of the group's project?

The title of the group's project was- Automated Short Answer Grading. Their main goal was to build machine learning models that will allow them to grade short answers. Their motivation behind this project was that this system will be fair to the students, will be cost efficient and give them timely feedback.

How did the group attempt to accomplish the goal?

Their outline of this system was that they used graded short answers out of which they extracted NLP features to train their model, which will be later used for predictions. They used Hewlett dataset which contained short answers written by students and each answer was labelled by human graders (used as base-truth).

Their pipeline started with the standard pre-processing which consisted of stop word removal, POS tagging, Porter Stemmer, etc. Next during feature engineering, they used features through which they wanted to extract sentence quality and content fluency. To do this they used various features like essay length, TFIDF, Bag of words, n-gram dictionaries, etc. Next they used Feature Selection to select the best features. After that they listed their list of algorithms used some of which were- KNN, Naïve Bayes, SVM, etc. For selecting the hyper-parameters of their final models they used 5-fold CV.

How did the group evaluate their work and what was the result?

To evaluate their systems, they used 'kappa' score to measure the degree to similarity between their predictions and true value (human reported score). They also showed the effect of number of features on an algorithms accuracy. Their final results showed that there wasn't one single algorithm (model) which performed the best for all test sets, though Deep Learning was close to the best answer in most cases.

6. Review 2

The group number was: 10

What was the goal of the group's project?

The title of this project was- Movie Genre identification based on plot summary. Their goal was to identify movie genre from a small plot summary. Based on this summary the program would report a genre from a list of six. So this was basically a multi-class classification problem.

How did the group attempt to accomplish the goal?

To start with, this group selected a dataset containing many movie plot summaries along with metadata like actor's name. They used CMU's movie summary corpus for this project. They used this dataset for training and used features like TF, POS tagging, etc. They also assigned genre-based weights to all actors based on the number of movies of a particular genre he had acted in. Finally, they used multi-class SVM on this dataset to build a model which will be later used for predictions.

How did the group evaluate their work and what was the result?

To evaluate their work, they used precision/recall/Accuracy and also reported a confusion matrix. The confusion matrix showed that the Comedy genre was the easiest to identify. During presentations, they also showed a little demo (in slides) which consisted of a web-portal in which you would have to enter a plot summary and it will show you the predicted genre. It was reported that adding an actors name to the plot summary greatly increased the accuracy of that plot summary, hence showing that stereotyping actors was working to a certain extent.

7. Review 3

The group number was: 11

What was the goal of the group's project?

The title of the project was- Disease Diagnosis Hotline. Their goal was to build a medical diagnosis system which diagnosed diseases based on the interaction with the user through a chat-based system. This kind of project will help in remote diagnosis and save unnecessary doctor's visits.

How did the group attempt to accomplish the goal?

Their system was a chat engine where a user interacted with a bot which asked him various questions and according to user replies tried to diagnose diseases. This system consisted of two main sub-tasks, mainly: Dialog Management and Information Extraction. The Natural Language Generation was handled using OpenDial with hand-crafted dialogs. They achieved Natural Language Understanding using Information Extraction, where they used techniques like sentiment analysis and synonym extraction.

How did the group evaluate their work and what was the result?

To evaluate the performance, the group first built a baseline system in which they just used regular expression to extract symptoms from the user messages. In the next iteration they built an advanced system using synonym extraction and polarity identification to improve diagnosis. They calculated the final results by comparing the performance between baseline and advanced system, and using user's feedback. They did not show any numeric results in the slides since they were still working on it. But, they did mention that when they real people used this system, their reviews were mixed. They attributed this to the limitations of the system, key ones being that dialogs are hand-crafted and that this is highly dependent on the domain so a large amount of data is needed to make a good diagnosis system.

8. Review 4

The group number was: 40

What was the goal of the group's project?

The title of the project was – Cuisine Classification based on Ingredients. Their main goal here was to make a system that can identify the cuisine given the ingredients needed for a recipe. So this was basically a multi-class classification problem.

How did the group attempt to accomplish the goal?

Firstly, this group collected labelled data having ingredients and the corresponding cuisine. In their data Italian was the most common cuisine. They also had highly skewed classes in their dataset, as was visible in the histogram, but they did not handle it and directly passed it onto the classification algorithm.

Then they did pre-processing which mostly consisted of Lemmatization, removing special characters and using TF-IFD as a feature to represent the ingredients in each review. After this they applied various multi-class classification algorithm like linear SVC's and Naïve Bayes.

How did the group evaluate their work and what was the result?

They evaluated their system on two metrics- Accuracy of classification and time taken. For both these metrics Linear SVC's performed the best. They also noted that pre-processing decreased the execution time. I think if they had handled skewed classes than their accuracy would have increased even further.

9. Review 5

The group number was: 19

What was the goal of the group's project?

The title of the project was- Wrong Preposition Detection System. The goal of this project was to design a system that would predict preposition for a sentence, and if the predicted proposition was different compared to the original one, than it will prompt the correct preposition. Their main motivation for this project was that there are a lot of non-native English speakers these days, and systems like this will aid them greatly.

How did the group attempt to accomplish the goal?

They choose the top 25 prepositions and converted this into a multi-class classification algorithm. For data, they used two sources- one was from news articles sources like CNN, NY Times, etc. which are much more curated compared to the other source which was from Wikipedia. They used a mixture of manual data scrapping and Wikimedia api to get data. After that they did pre-processing and feature extraction. They used features like- Tokens and POS tags in a 2-word window around the preposition, Head Verb, Head Noun, etc. They also used a combination of POS tags to create extra features. After that, they various machine learning classification algorithms like SVM and Decision Trees.

How did the group evaluate their work and what was the result?

For evaluation they used weighted accuracy. SVM did the best out of all the algorithms they used and it was used in the backed system that they design which would ask user to input a sentence and tell whether there is any mistake in it. They also acknowledge that since they are using so many classes, it's difficult to achieve high accuracy.

10. Review 6

The group number was: 20

What was the goal of the group's project?

The title of the project was- An Approach to Automatic Trending Tweet Summarization. Their main goal was text summarization such that the processed text still retains a significant portion of information from the original text.

How did the group attempt to accomplish the goal?

To accomplish their goal they used an extraction-based text summarization technique. In this approach they used LESK algorithm, which finds the meaning of a word depending upon the context it was used in. They also used WordNet which is a lexical database for English words. Their data sources were twitter api and other online articles which they found.

Their pipeline started with cleaning task which included removing punctuations, links and special characters. Next, they removed stop words and also tokenized sentences into a word and POS tag pair. During the summarization step, they iterated over all the text data, and assigned weights to them depending upon their importance, which was measured by the overlap between the text and its meaning (generated by LESK). They selected the sentence with top four weights as their output.

How did the group evaluate their work and what was the result?

For evaluation they compared their summary with that of a human. They used two metrics, namely- Cosine Similarity and Semantic Similarity. They evaluated their system on various topics and reported the average metric and the max metric. Based on the results they inferred that text summarization depends a lot on the quality of text. For example, the summarization quality of formal sources like news will be higher compared to opinions and sarcastic comments. They also noted that more data would generally give better results.

11. Review 7

The group number was: 30

What was the goal of the group's project?

The title of the project was- Text Classification Based on CAMEO event codes. Their goal was to classify news articles based on CAMEO event codes. CAMEO is a framework for coding event data which is typically used by news broadcasters.

How did the group attempt to accomplish the goal?

They started this task by reducing the number of CAMEO codes from 300 to 26, after which they identified important words in each category. Their pre-processing consisted of using NLTK for stemming, tokenization and removing punctuations. They calculated TF-IDF scores for each category and unlabeled data. After that, they used cosine similarity index and selected documents with highest similarity for each category.

How did the group evaluate their work and what was the result?

Their initial accuracy was low because of some many categories they had to classify on and the fact that labelled data was short compared to unlabeled news data. They increased this accuracy by manually selecting 100 documents for each category, they also labelled files for test data. After this they applied multinomial naïve bayes and achieved an accuracy between 60-70 percent.

12. Review 8

The group number was: 21

What was the goal of the group's project?

The title of the project was- Feature Specific Sentiment Analysis. The goal of the project was to analyze the sentiments associated with a feature of a product which is being reviewed by someone. Example, for a smart-phone there could be features like responsiveness, battery and camera. Their goal was to predict the sentiment associated with each of these features.

How did the group attempt to accomplish the goal?

For this task they used data from flipkart.com and amazon.com for a particular product. Firstly, they did pre-processing on this data, which consisted of sentence segmentation, sentence correction and tagging POS. The two main tasks after this were identifying features and predicting sentiment for that feature. For feature identification, they restricted their domain to cell phones and found the important features associated with them using manual inspection. Then for predicting sentiments, sentences associated with a feature were combined and a positive and negative sentiment score was given. To achieve this they used tools like- Stanford core NLP and algorithms like- CRF and maximum entropy.

How did the group evaluate their work and what was the result?

In Results, they reported the positive and negative sentiment score of features associated with that brand of cell phone (Moto G 2nd gen). For evaluation they compared their accuracy with that of a benchmark. They also reported that their trained CRF and Maximum Entropy had better accuracy than Stanford's NLP.

13. Review 9

The group number was: 18

What was the goal of the group's project?

The title of the group’s project was- Restaurant Review Analysis. The goal was to generate a summary review about a restaurant given many user reviews. This summary review was based on a fixed number of categories which were chosen by them. Their program gave a numeric output for each of these category.

How did the group attempt to accomplish the goal?

To accomplish this goal they used restaurant review data from Yelp. They cleaned and pre-processed this data by using spacy.io to extract Noun and Noun Phrases. After that they selected four categories on which they would generate their final output. These categories were- food, ambience, service and price. They used NLTK-Vader to generate sentiment for a review after which they gave rating for each category.

How did the group evaluate their work and what was the result?

They evaluated their work by comparing the textual reviews and the rating generated. They also compared between the user rating and the rating generated by their program. They gave their ratings on a scale of -1 to 1, where negative rating meant negative sentiment and positive rating meant positive sentiment.

14. Review 10

The group number was: 39

What was the goal of the group's project?

The title of the group’s project was- Text Summarization and Multi-class Perceptron for Email Classification. They had two goals in their project, one was to classify emails into particular categories and other was summarize an email.

How did the group attempt to accomplish the goal?

For the classification of emails into categories they used multi-class perceptron. They used the Enron Email data as their dataset. They started by pre-processing these emails into different categories and also split this data into train and test. After that they trained the perceptron by giving weights to the vocabulary for each category.

For the second part of text summarization, they wanted to generate a summary text for each email. For doing this, they parsed each documents into sentences, and assigned ranks to them, after which they summarized using a subset of ranked sentences. They used a graph based ranking model for this step.

How did the group evaluate their work and what was the result?

For their first task of perceptron classification, they achieved an average F-score of 0.75. For their second task of text summarization the average F-score was about 0.70.