The Anatomy of a Restaurant Review

By Shivayogi Biradar, Sharyu Deshmukh, Jianchao Yang
DS5500, Fall 2018 --- Prof. Cody Dunne, Northeastern University

Motivation and Data

Summary of Data:
The dataset comes from AI Challenger, an AI programming competition.
https://challenger.ai/competition/fsauor2018

The original data is in Chinese. We used Google Translate API to translate some sample reviews to English, just for the visualization purpose.
The English translations are available in the project repository
https://github.com/ktmud/fsauor2018

The dataset incorporates two layers. The first layer is the coarse-grained evaluation object such as 'location' and the second layer is the fine-grained emotion sentiment score such as 'traffic convenience' in location. We predict the sentiment score for the second layer. For e.g. the sentiment of review-type 'traffic convenience' would be predicted but not the 'location' overall.

Data Preprocessing

There are in total 20 sentiment elements in 6 categories, each labeled with -2 (not mentioned), -1 (negative), 0 (neutral), and 1 (positive). Therefore, there are in total 21 useful columns in the dataset, with the first column being text of the reviews and other columns the sentiment labels.

The review content column is free text data and not directly digestible by visualizations. We built feature matrices to convert it to categorical data type (words, phrases) and numerical (TF-IDF Vectorization) data in order to visualize it. All other columns (for sentiment labels) are **ordinal**. The dataset has no missing values.

Data Analysis

We visualized the distribution of all second-level label values through a pairplot of histograms that checked the distribution of the sentiment labels in each aspect. It is observed that most aspects were not mentioned by the majority of the reviews. The distribution/frequency of labels are encoded as the height of the bars and multiple aspects are organized into small multiples. This visualization is efficient in quickly understanding the prevalence of sentiment topics, and get a general feeling of how the correct predictions should look like.

Next we visualized a Correlation plot of all the numeric variables which helps to understand the relationship between variables. In our case it is interesting to note that there is high correlation (0.44) between dish taste and overall experience which is intuitive as well because we primarily go to restaurant to enjoy a great meal of which taste is a major component.
We can also see high correlation amongst variables related to environment which may suggest that we can drop a couple of them in variable selection if there is no significant loss of information in the model.

Task Analysis

Index #	“Domain” Task	Query Task (Low-Level)	Search Task (Mid-Level)	Analyze Task (High Level
1	Glance through topics in a single review	Summarize	Locate	Discover
2	Compare predicted and actual sentiment labels	Compare	Browse	Discover
3	Summary of global distribution of labels in selected dataset	Summarize	Lookup	Discover

Domain Task 1: This visualization is a grouped barchart where users can see a sentiment score on various topics within the reviews such as dish taste, easy to find location, price range, service quality etc. The four sentiments are represented by four colored segments in the bar chart and by mouse over each segment we can see the sentiment and the predicted probability respectively.
This is a quick glance of the whole review to help them digest a review more quickly. At the analytic task level, we are **summarizing** the information buried in text as colored bars. The 20 topics are known and users need to **locate** them in the visualization, if they exist. Finally, at the high-level, we are basically helping users **discover** the mentioned topics from the text.
With such high-level summarization, this feature would be useful to anyone who wants to read customer reviews more efficiently. They can be either restaurant owners or customers looking for restaurants.

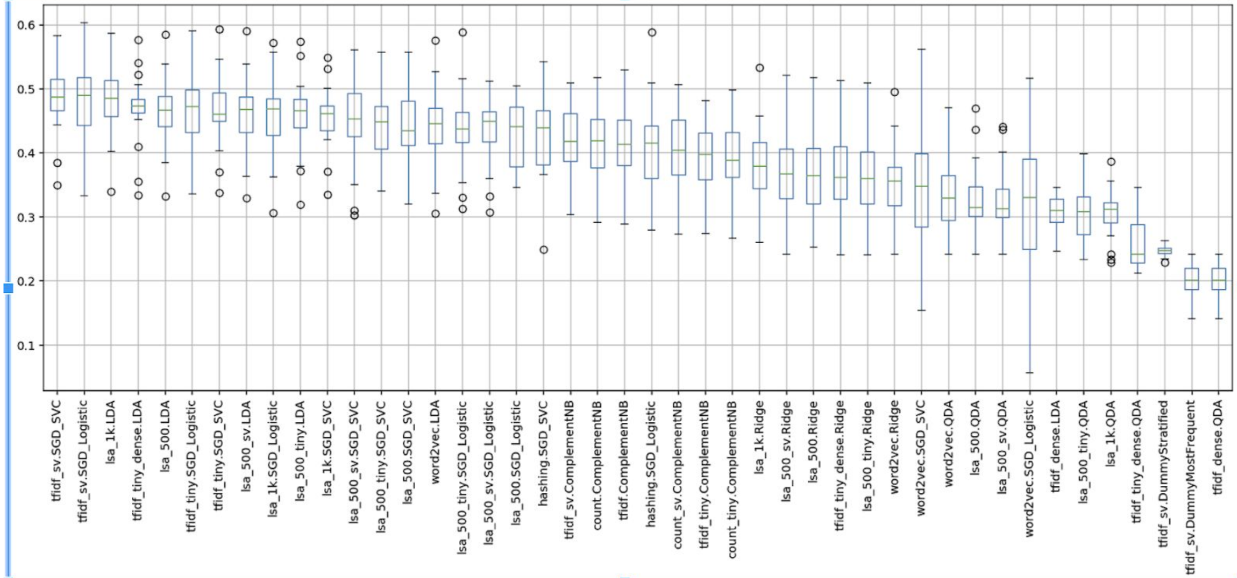
Domain Task 2: In this task we are visualizing the **comparison** of predicted and actual labels. By mouse over each segment we can see the sentiment and the predicted probability respectively. It helps users to **browse** for which topics the predicted labels were similar to actual labels and hence helps them **discover** the meaningful patterns

Domain Task 3: This visualization shows the global distribution of predicted probabilities for each particular topic. By mouse over each segment we can see the sentiment and the predicted probability respectively. This helps in **summarizing** information globally and shows the overall distribution by **looking up** the topics in the entire reviews text and **discovering** topics.

Model Description

We used Google Translate to translate the reviews for training dataset from chinese to english. We also use original Chinese reviews for data analysis and prediction in Chinese language.

We tried various featuring engineering models for our dataset in which TF-IDF with LDA, Logistic Regression, and Support Vector Classifier performed the best.



Our final app uses the following 7 feature models from the various feature models we tried based on their accuracy and ease of loading:

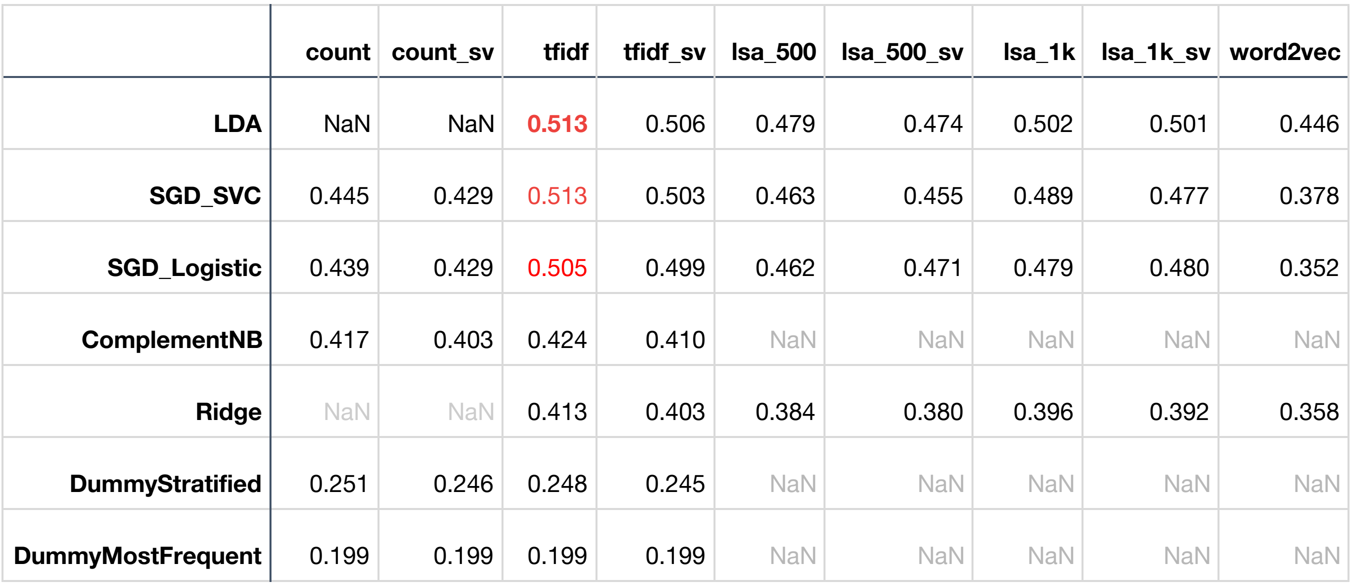
1. Word Count(4k): simple term count with 4,000 most frequent terms.
2. Word Count(2k): simple term count with 2,000 most frequent terms
3. TFIDF(4k)- SVD(500): TF-IDF vectors based on word count, then passed to SVD for dimension reduction.
4. TFIDF(4k)- SVD(500): same as above but with different vocabulary and SVD component size.
5. TFIDF(2k)- SVD(500): same as above but with different vocabulary and SVD component size.
6. TFIDF(2k)- SVD(500): same as above but with different vocabulary and SVD component size.
7. Word2V(400): word vectors of 4,00 dimensions. Feature matrix for each review is simply the average of the Word2Vec vectors of all words in the document.

We have limited the word features in the production for faster loading of predictions.

Our Predictive Models used in the app are :

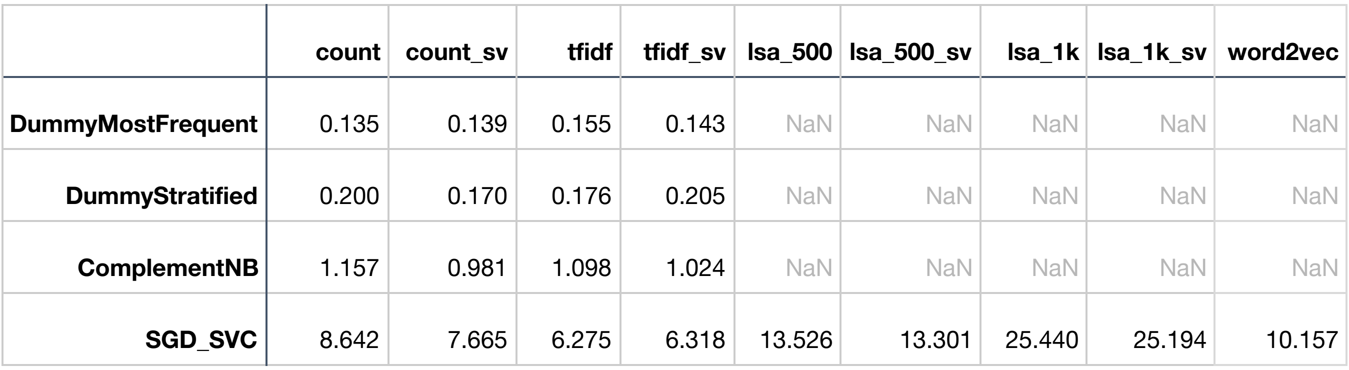
1. Complement NB
2. Logistic Regression with SGD
3. Linear SVC with SGD
4. Ridge Classifier

Model Performance : **F1 Score**



	count	count_sv	tfidf	tfidf_sv	lsa_500	lsa_500_sv	lsa_1k	lsa_1k_sv	word2vec
LDA	NaN	NaN	**0.513**	0.506	0.479	0.474	0.502	0.501	0.446
SGD_SVC	0.445	0.429	0.513	0.503	0.463	0.455	0.489	0.477	0.378
SGD_Logistic	0.439	0.429	0.505	0.499	0.462	0.471	0.479	0.480	0.352
ComplementNB	0.417	0.403	0.424	0.410	NaN	NaN	NaN	NaN	NaN
Ridge	NaN	NaN	0.413	0.403	0.384	0.380	0.396	0.392	0.358
DummyStratified	0.251	0.246	0.248	0.245	NaN	NaN	NaN	NaN	NaN
DummyMostFrequent	0.199	0.199	0.199	0.199	NaN	NaN	NaN	NaN	NaN

Model Performance : **Training Speed (Seconds)**



	count	count_sv	tfidf	tfidf_sv	lsa_500	lsa_500_sv	lsa_1k	lsa_1k_sv	word2vec
DummyMostFrequent	0.135	0.139	0.155	0.143	NaN	NaN	NaN	NaN	NaN
DummyStratified	0.200	0.170	0.176	0.205	NaN	NaN	NaN	NaN	NaN
ComplementNB	1.157	0.981	1.098	1.024	NaN	NaN	NaN	NaN	NaN
SGD_SVC	8.642	7.665	6.275	6.318	13.526	13.301	25.440	25.194	10.157

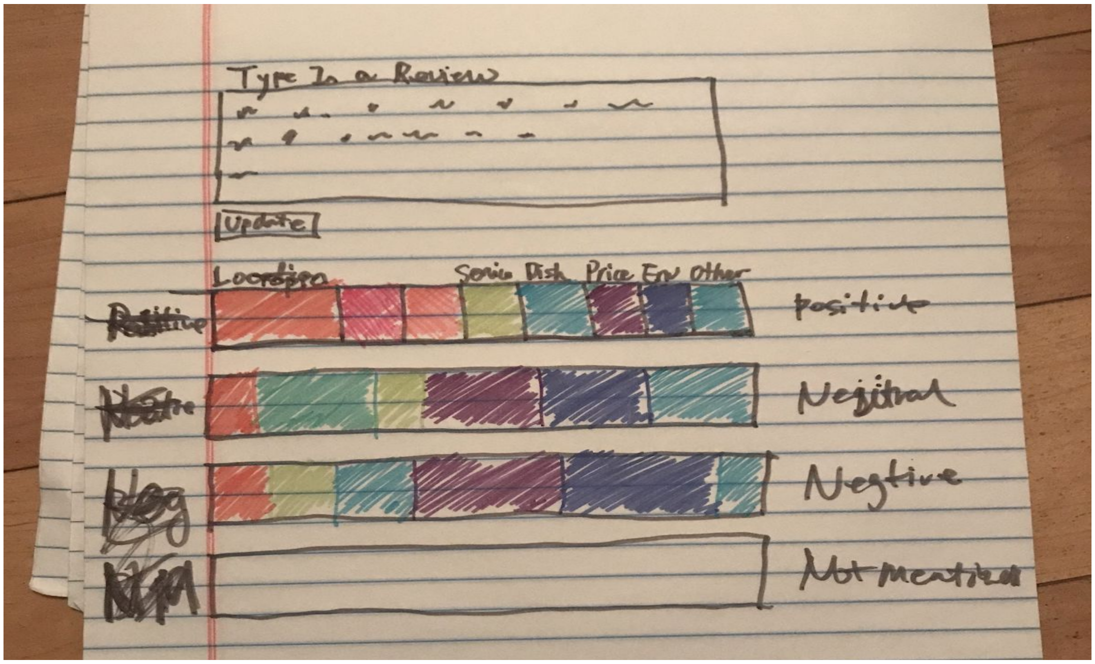
SGD_Logistic	8.628	7.395	7.003	5.827	15.801	15.361	29.475	27.155	52.316
LDA	NaN	NaN	NaN	NaN	58.567	54.765	148.312	143.328	30.537
Ridge	NaN	NaN	NaN	NaN	69.216	65.362	166.688	174.574	39.619

As described above the complement naive bayes model has relatively great accuracy and takes less time to train, which makes it a perfect candidate for baseline.

Design Process

After brainstorming for ideas and designs, we zeroed upon the following three design sketches for our visualization.

Visualization #1

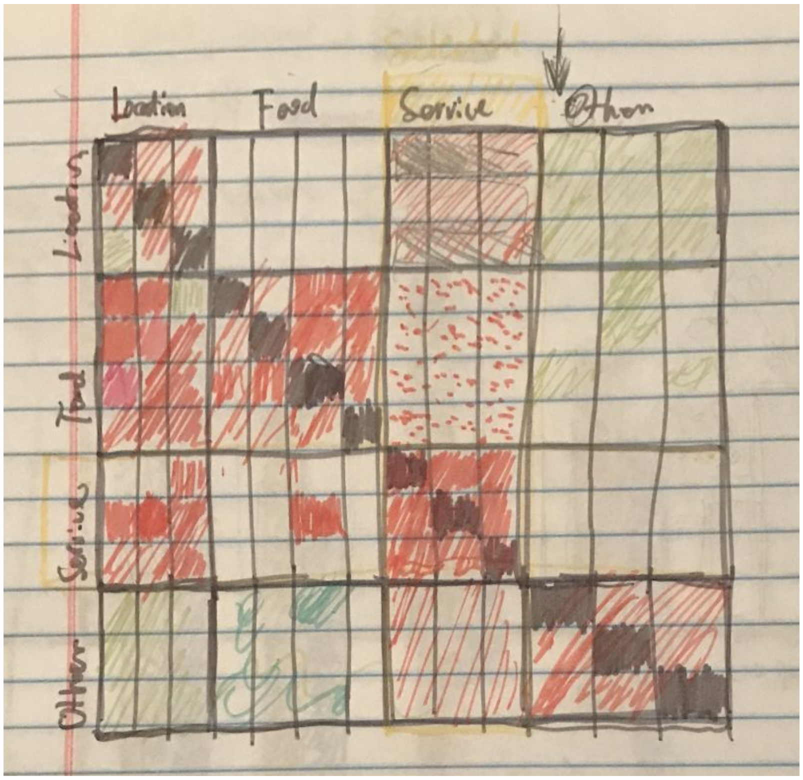


This visualization tries to fulfill our Task #1: “Glance of topics in a single review.”

Each sentiment label (positive, negative, neutral, not mentioned) will have its own horizontal bar. Color of the segments in a bar represents a first-level topic. The length of the segments represent the relative probability that the topic is mentioned in the given sentiment label AND the given review. The overall absolute probability for a sentiment label will be encoded as opacity of the bars. E.g. if a review has 99% positive in all the topics, the “Positive” bar will be brighter than other bars. For training data with true labels, everything is predicted with 100%/0% probability, then all segments will have equal size, but either 100% or 0% opacity.

The visualization is able to encode all sentiment labels and topic categories, as well as predicted probability in one graph, which is useful for Data Scientist to check the confidence of our models, as well as overall distribution of topics in one review.

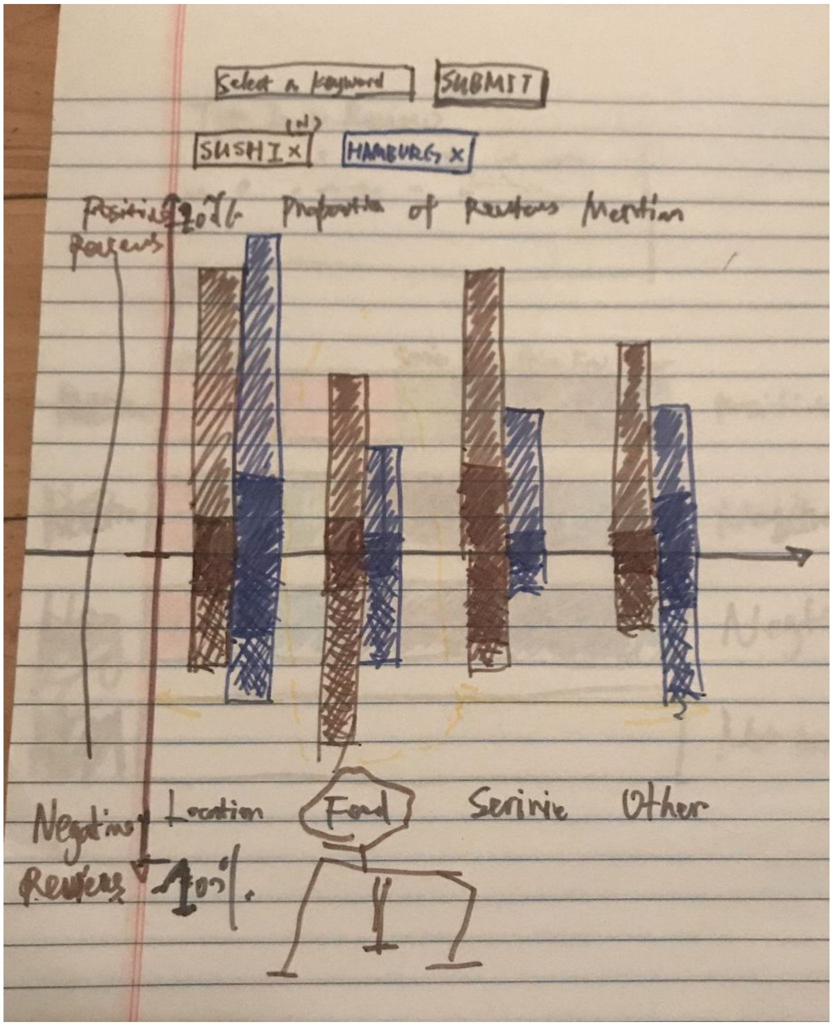
Visualization #2



This visualization allows users to browse the correlation between multiple sentiment categories at both levels. (VDA P53) They can even check the correlation between a first-level topic to a second-level topic. In order to do this, we will need to add 6 additional columns represent how the first-level topics are mentioned---we use the most frequent label in their subtopics.

By allowing users to toggle between two different level of topics, we are able to encode the 26x26 correlation matrix into a 6x6 or 20x20 heatmap. The grouping of first-level topics by bold separation lines is a clean and intuitive representation of the hierarchy of the topics (VDA P351).

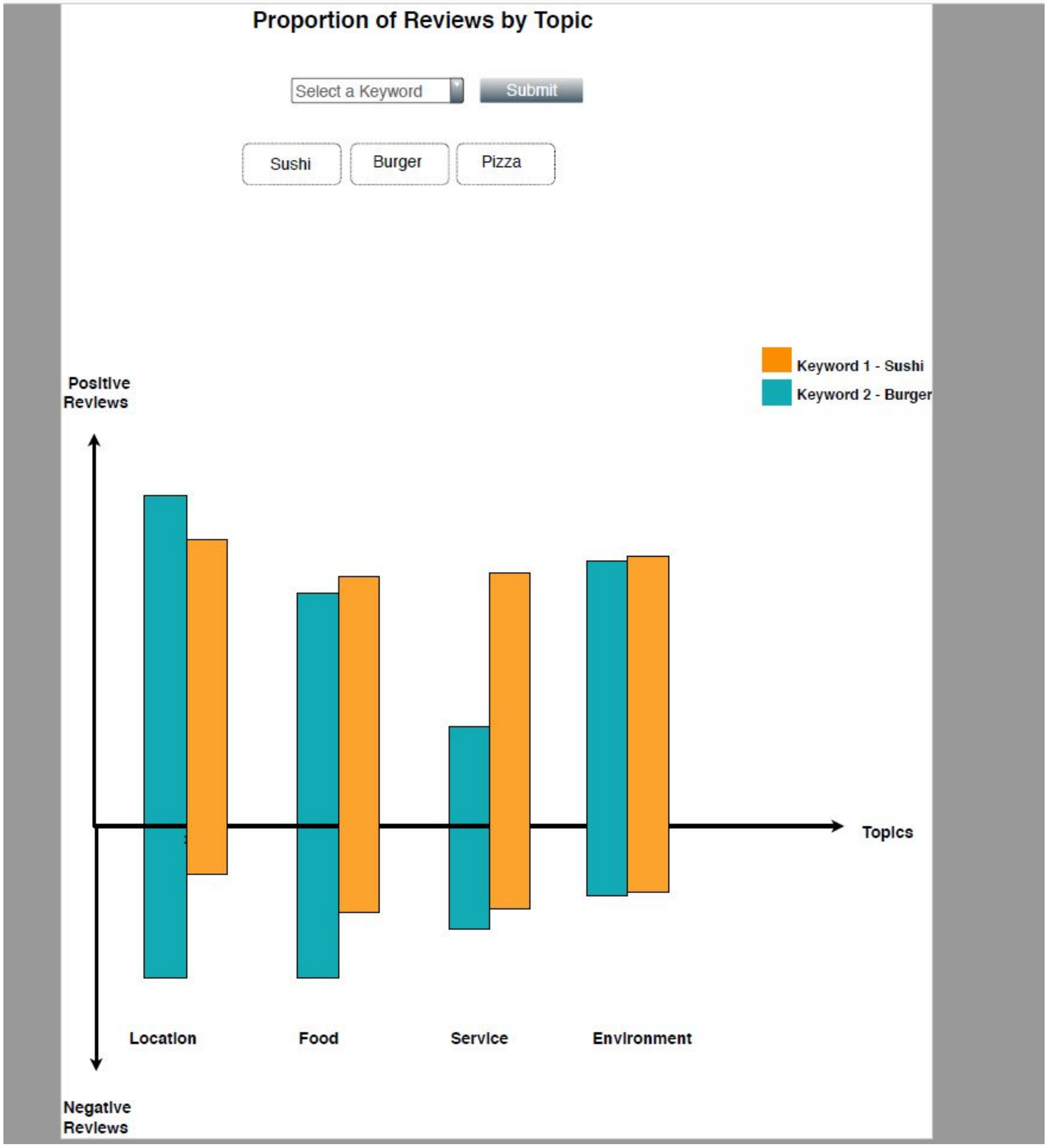
Visualization #3



This visualization fulfills the task “summary of review sentiments for segments of reviews”. The grouped bar chart is a common and direct choice for comparing groups of values on a categorical variable (VDA, P281). Since the end user are consumers of our prediction insights (restaurant owners, account managers), the underlying probabilities do not matter anymore. Instead, we show proportion of reviews that mentioned a specific topic in either positive or neutral way. Neutral mentions is the center segment that is centered around zero. “Not mentioned” will not take space in the bars.

This visualization is able to capture all the sentiment labels we have and is easy to consume for end users.

Based on these visualizations we developed one digital design sketch as follows.



Final Visualization

Final Visualization has been built on HTML, CSS, D3 and Flask (Python).
Models have been trained in python and have been deployed in production via dokku.
Packages used have been mentioned in requirements.txt.

For final visualization and UI walkthrough please follow along the following video tutorial:

Final Visualization walkthrough:
https://www.youtube.com/watch?v=cmAa0jb-3Iw&feature=youtu.be

We have mostly used bars in our visualization to represent quantitative and categorical variables. The colors used are mostly red-for negative, blue- for neutral and green- for positive connotations. The color scale used in the visualization performs good on color blind scale. Finally we have created an app with the following functionalities :

1) Find and Filter reviews by keywords
2) Give an idea of sentence polarity on the review by marking red and green colors for negative and positive sentiments respectively
3) Give a choice to the user to select feature processing and base classifier
4) Provide a user an idea about accuracy of the current baseline models

   Scope For future:
   1) Improve accuracy of existing models
   2) Add a correlation plot included in the design stage
   3) Modify sentence polarity to include sentiment topic tags.

Evaluation Plan

Our visualization tool provides the functionality to users to select the dataset. Since one of the target users for this tool are the restaurant owners who would like to check how their business is performing, data exploration is not supported by this tool. It focuses more on the Knowledge Discovery. It delivers a visualization that displays topic-wise

1. predicted probabilities for each sentiment label
2. actual and predicted labels
3. global distribution of labels

The Data Exploration is supported by the initial visualizations developed to understand data. Visualizations such as

- distribution of label values using histogram pair-plot to check the distribution of the sentiment labels in each aspect
- word cloud to determine the most common words in the reviews to help Data Scientist understand which words are likely to be more important in making sentiment predictions
- correlation plot help understand how strong is the relationship between two variables.

The other target user for this tool are the data analysts who would feed the review text in the tool by choosing a dataset, provide a keyword, choose a feature processing algorithm and base classifier. The visualization thus obtained is analyzed and conveyed effectively to the restaurant owners. This leads to Hypothesis Generation by analysts about the results obtained for different aspects and topics in restaurant reviews. It ultimately contributes in Decision Making for the restaurant owners in terms of how to improve their service in different aspects. (lecture15-validation and evaluation, slide 30)

For visual encodings/interaction idioms the goal is to see how effectively the message that you are trying to convey through your visualization is delivered. To reach this goal interviews can be conducted with the target users and check if the Visual encoding/interaction idiom requirement is fulfilled i.e. if the way you are representing it isn't effective for them or doesn't work at all. This would be the case of Wrong Idiom (VAD pg.75). Particular qualitative and quantitative results can be taken from the user by performing experiments. Quantitative results can be taken by checking what accuracies user is receiving when using the tool and can be compared with the actual accuracies obtained during development and testing phase. Qualitative results can be obtained by checking how effective and user-friendly is the tool in terms of usage.
In early phase physical samples (paper prototypes) can be used to get feedback on the color encodings, visual encodings and marks channels used so that there is no trivial repetitive improvements done in the development of the actual tool. Field study is another effective way to conduct the experiment and work towards the goal of effectively communicating through visualizations. In field experiment users are investigated in actual settings. This gives a clear

idea of how effective the tool would be once it is deployed and made available to the real world which ultimately is what our goal is.