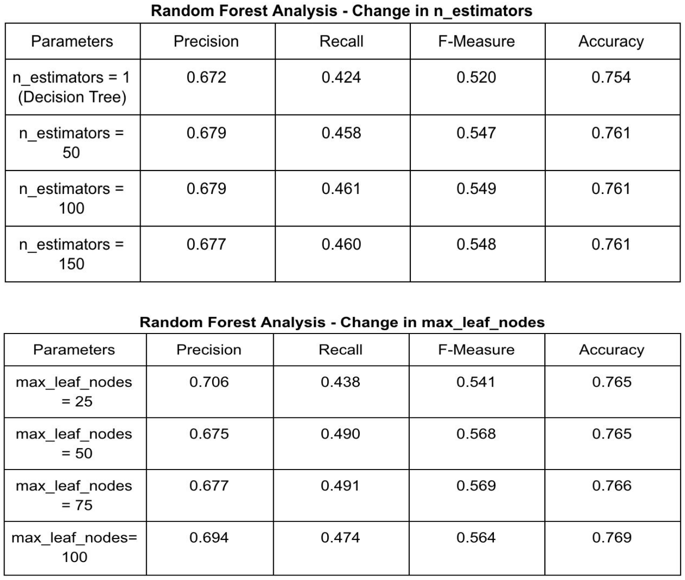
Recommendations from Yelp Dataset

RESULTS:

- **Topical Authority**: We have trained the model using Random Forest and Support Vector Machine.
    - Random Forest: We have used 5-fold cross-validation and were able to obtain an accuracy of 0.92. However, the dataset is very skewed i.e., the positive examples are very less when compared to negative examples. Hence we can not consider accuracy as a correct measure of correctness. For skewed data, F-measure would be a much appropriate measure for determining the correctness of the results.F−measure=2*(precision*recall)(precision+recall)We got the following results:(1) precision=0.46(2) recall=0.08(3) F-measure = 0.14The above results were obtained when we have used the complete dataset. The recall value obtained based on the above-mentioned features is very less.So, we have trained the model using a subset of the dataset which represents the complete data set. We have reduced the skewness considerably in the subset which we have chosen. Figure 1 depicts the change in the F-Measure and accuracy as the number of support vector machines (n_estimators) are varied. Also when n_estimator is set to 150, the second table in Figure 1shows the change in accuracy and F-measure as max_leaf_node(Maximum number of leaf nodes in the trees grown). The best results that we were able to get was:(1) precision=0.69(2) recall=0.47(3) F-measure = 0.564(4) Accuracy = 0.7699.1.2SVM.The accuracy and F-measure obtained using SVM are more or less similar to the accuracy and F-measure obtained using RandomForest. Figure 2 depicts the F-Measure and Accuracy for different tuning parameters. We parameters which we have tried tuning are Gamma (Kernel Coefficient), C (penalty parameter)and Tol (tolerance for stopping criterion).



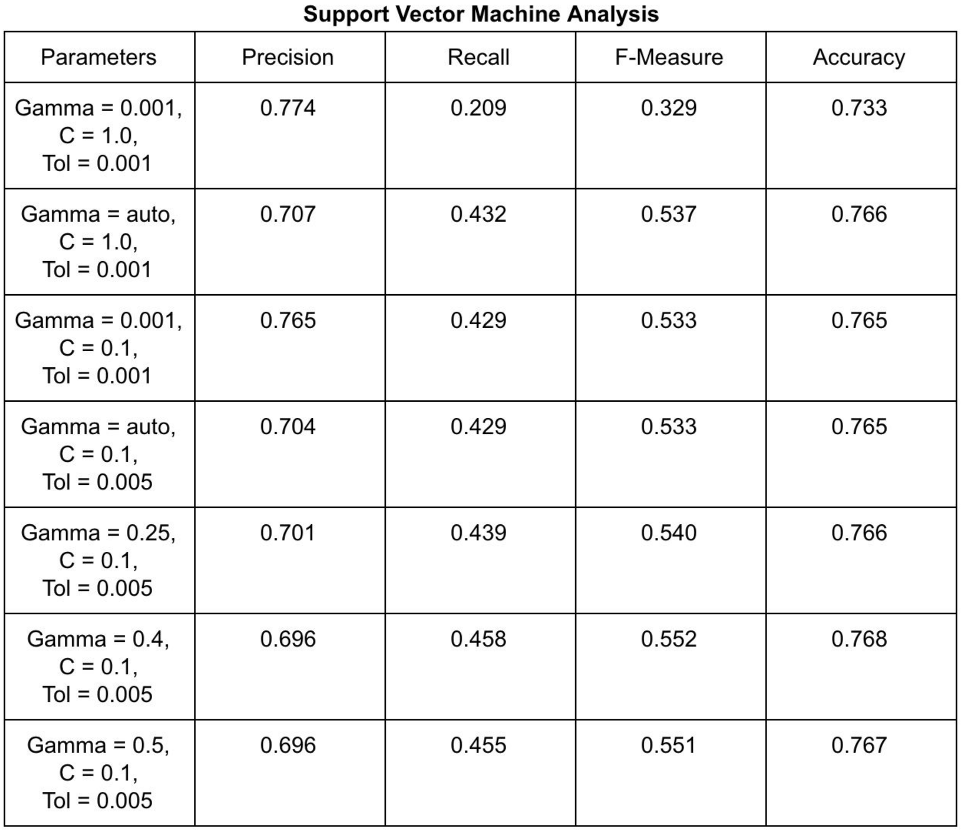
Random Forest Analysis - Change in n_estimators

Parameters	Precision	Recall	F-Measure	Accuracy
n_estimators = 1 (Decision Tree)	0.672	0.424	0.520	0.754
n_estimators = 50	0.679	0.458	0.547	0.761
n_estimators = 100	0.679	0.461	0.549	0.761
n_estimators = 150	0.677	0.460	0.548	0.761

Random Forest Analysis - Change in max_leaf_nodes

Parameters	Precision	Recall	F-Measure	Accuracy
max_leaf_nodes = 25	0.706	0.438	0.541	0.765
max_leaf_nodes = 50	0.675	0.490	0.568	0.765
max_leaf_nodes = 75	0.677	0.491	0.569	0.766
max_leaf_nodes= 100	0.694	0.474	0.564	0.769

Figure 1: Random Forest Evaluation and Results



Support Vector Machine Analysis

Parameters	Precision	Recall	F-Measure	Accuracy

Gamma = 0.001,
C = 1.0,
Tol = 0.001
	0.774	0.209	0.329	0.733

Gamma = auto,
C = 1.0,
Tol = 0.001
	0.707	0.432	0.537	0.766

Gamma = 0.001,
C = 0.1,
Tol = 0.001
	0.765	0.429	0.533	0.765

Gamma = auto,
C = 0.1,
Tol = 0.005
	0.704	0.429	0.533	0.765

Gamma = 0.25,
C = 0.1,
Tol = 0.005
	0.701	0.439	0.540	0.766

Gamma = 0.4,
C = 0.1,
Tol = 0.005
	0.696	0.458	0.552	0.768

Gamma = 0.5,
C = 0.1,
Tol = 0.005
	0.696	0.455	0.551	0.767

Figure 2: Support Vector Machine Results

○ Location Authority: As we have used Gaussian Mixture Model, we clustered each of the user's rated restaurants into two clusters. After clustering them, we choose the cluster which is denser and find the centroid of it. As we can see in Figure 3, they are two figures of different users, in the first figure, the density of the clusters are 45 and 57, the centroid is 35.2250973639,-80.8411199818 which is the green-colored point. Similarly, for the second figure, the densities are 58and 167, the centroid taken is 36.1125319434,-115.168108273 which is the orange point. All the other restaurants that are rated are represented using blue point. Hence one of either the green or the orange (depending on the density of the cluster) is taken as the reference location of the user. Finally, we take in the input of the location and compute the distance of the input location with all the users. We then select the ones which have the least distance and give them as the result

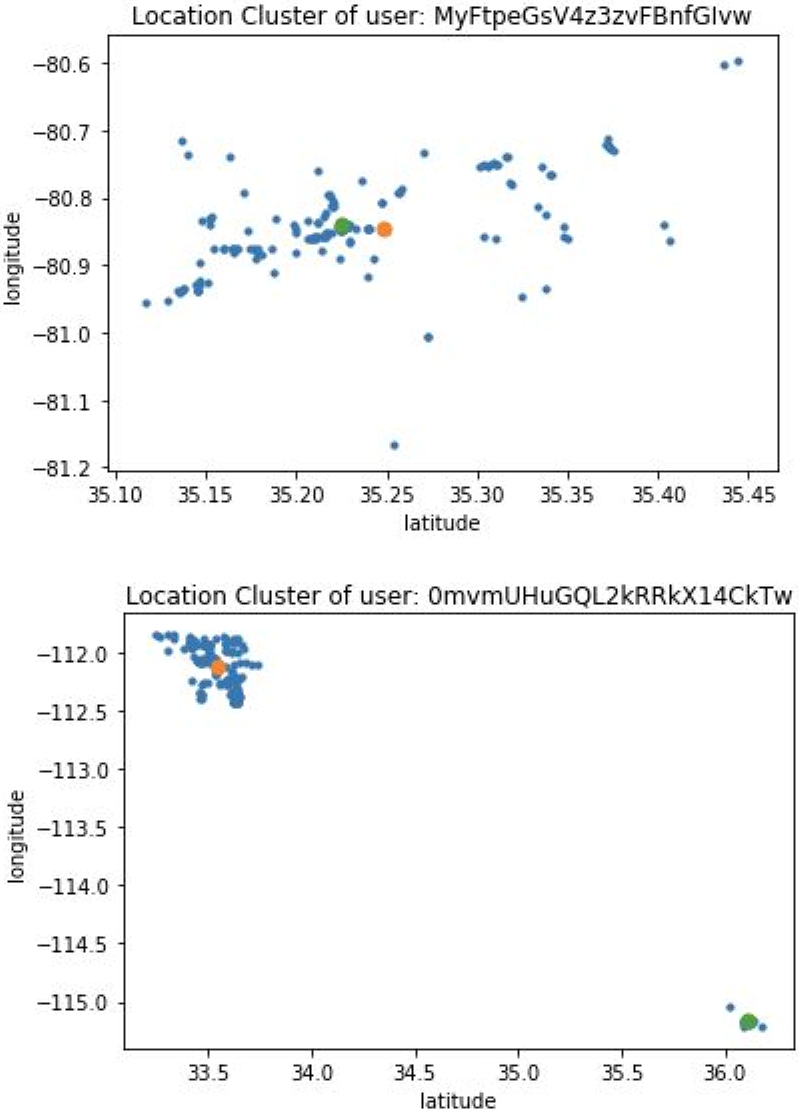
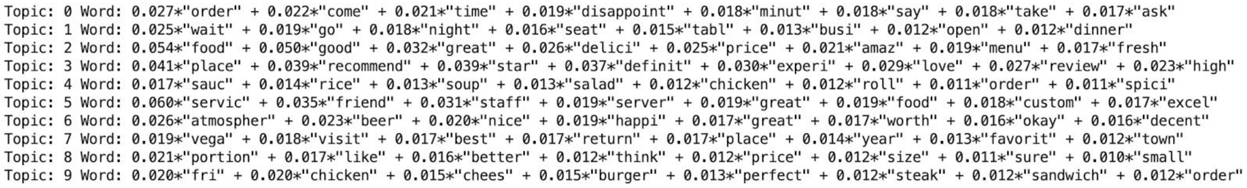


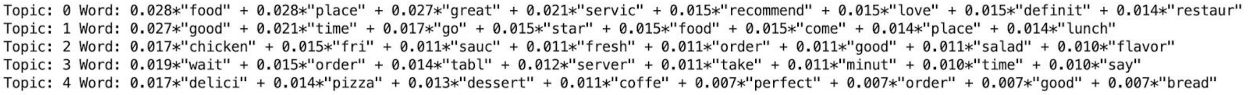
Figure 3: Gaussian Mixture cluster

- **Rating key attributes of a restaurant:** In Figure 4, we performed LDA with 10 topics and assigned each topic to a certain attribute. For example, topics 4,7 and 9 correspond to Food attribute. Topics 3 and 6 correspond to Service attribute. Similarly, in Figure 5, we performed LDA with 5 topics. In Figure 6, stars attribute to represent the average ratings of the restaurant obtained from the dataset, whereas ratings for attributes like Food, Service, Value For Money and Ambience are averaged for each restaurant with the values we got from the sentiment analyzer.



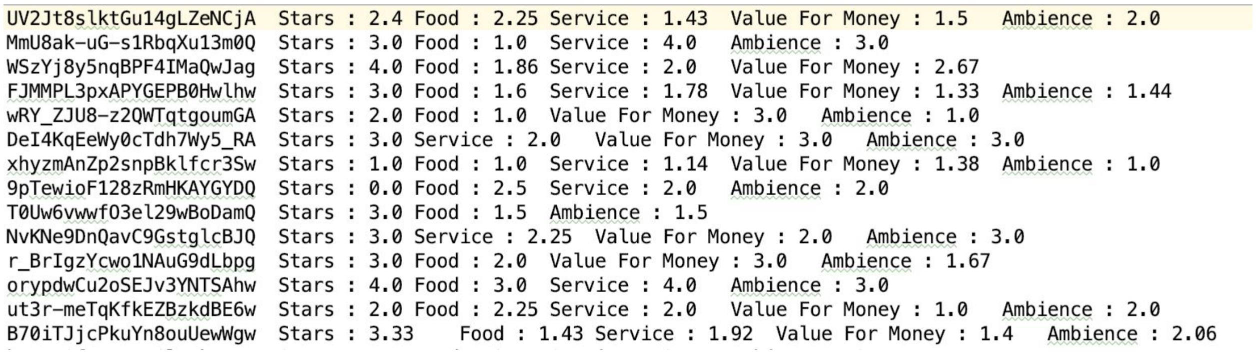
```
Topic: 0 Word: 0.027*"order" + 0.022*"come" + 0.021*"time" + 0.019*"disappoint" + 0.018*"minut" + 0.018*"say" + 0.018*"take" + 0.017*"ask"
Topic: 1 Word: 0.025*"wait" + 0.019*"go" + 0.018*"night" + 0.016*"seat" + 0.015*"tabl" + 0.013*"busi" + 0.012*"open" + 0.012*"dinner"
Topic: 2 Word: 0.054*"food" + 0.050*"good" + 0.032*"great" + 0.026*"delici" + 0.025*"price" + 0.021*"amaz" + 0.019*"menu" + 0.017*"fresh"
Topic: 3 Word: 0.041*"place" + 0.039*"recommend" + 0.039*"star" + 0.037*"definit" + 0.030*"experi" + 0.029*"love" + 0.027*"review" + 0.023*"high"
Topic: 4 Word: 0.017*"sauc" + 0.014*"rice" + 0.013*"soup" + 0.013*"salad" + 0.012*"chicken" + 0.012*"roll" + 0.011*"order" + 0.011*"spici"
Topic: 5 Word: 0.060*"servic" + 0.035*"friend" + 0.031*"staff" + 0.019*"server" + 0.019*"great" + 0.019*"food" + 0.018*"custom" + 0.017*"excel"
Topic: 6 Word: 0.026*"atmospher" + 0.023*"beer" + 0.020*"nice" + 0.019*"happi" + 0.017*"great" + 0.017*"worth" + 0.016*"okay" + 0.016*"decent"
Topic: 7 Word: 0.019*"vega" + 0.018*"visit" + 0.017*"best" + 0.017*"return" + 0.017*"place" + 0.014*"year" + 0.013*"favorit" + 0.012*"town"
Topic: 8 Word: 0.021*"portion" + 0.017*"like" + 0.016*"better" + 0.012*"think" + 0.012*"price" + 0.012*"size" + 0.011*"sure" + 0.010*"small"
Topic: 9 Word: 0.020*"fri" + 0.020*"chicken" + 0.015*"chees" + 0.015*"burger" + 0.013*"perfect" + 0.012*"steak" + 0.012*"sandwich" + 0.012*"order"
```

Figure 4: LDA using TF-IDF model for number of topics as 10.



```
Topic: 0 Word: 0.028*"food" + 0.028*"place" + 0.027*"great" + 0.021*"servic" + 0.015*"recommend" + 0.015*"love" + 0.015*"definit" + 0.014*"restaur"
Topic: 1 Word: 0.027*"good" + 0.021*"time" + 0.017*"go" + 0.015*"star" + 0.015*"food" + 0.015*"come" + 0.014*"place" + 0.014*"lunch"
Topic: 2 Word: 0.017*"chicken" + 0.015*"fri" + 0.011*"sauc" + 0.011*"fresh" + 0.011*"order" + 0.011*"good" + 0.011*"salad" + 0.010*"flavor"
Topic: 3 Word: 0.019*"wait" + 0.015*"order" + 0.014*"tabl" + 0.012*"server" + 0.011*"take" + 0.011*"minut" + 0.010*"time" + 0.010*"say"
Topic: 4 Word: 0.017*"delici" + 0.014*"pizza" + 0.013*"dessert" + 0.011*"coffe" + 0.007*"perfect" + 0.007*"order" + 0.007*"good" + 0.007*"bread"
```

Figure 5: LDA using TF-IDF model for number of topics as 5.



```
UV2Jt8slktGu14gLZeNCjA  Stars : 2.4 Food : 2.25 Service : 1.43  Value For Money : 1.5   Ambience : 2.0
MmU8ak-uG-s1RbqXu13m0Q  Stars : 3.0 Food : 1.0  Service : 4.0   Ambience : 3.0
WSzYj8y5nqBPF4IMaQwJag  Stars : 4.0 Food : 1.86 Service : 2.0   Value For Money : 2.67
FJMMPL3pxAPYGEPB0Hwlhw  Stars : 3.0 Food : 1.6  Service : 1.78  Value For Money : 1.33  Ambience : 1.44
wRY_ZJU8-z2QWTqtgoumGA  Stars : 2.0 Food : 1.0  Value For Money : 3.0   Ambience : 1.0
DeI4KqEeWy0cTdh7Wy5_RA  Stars : 3.0 Service : 2.0   Value For Money : 3.0   Ambience : 3.0
xhyzmAnZp2snpBklfcr3Sw  Stars : 1.0 Food : 1.0  Service : 1.14  Value For Money : 1.38  Ambience : 1.0
9pTewioF128zRmHKAYGYDQ  Stars : 0.0 Food : 2.5  Service : 2.0   Ambience : 2.0
T0Uw6vwwfO3el29wBoDamQ  Stars : 3.0 Food : 1.5  Ambience : 1.5
NvKNe9DnQavC9GstglcBJQ  Stars : 3.0 Service : 2.25  Value For Money : 2.0   Ambience : 3.0
r_BrIgzYcwo1NAuG9dLbpg  Stars : 3.0 Food : 2.0  Value For Money : 3.0   Ambience : 1.67
orypdwCu2oSEJv3YNTSAhw  Stars : 4.0 Food : 3.0  Service : 4.0   Ambience : 3.0
ut3r-meTqKfkEZBzkdBE6w  Stars : 2.0 Food : 2.25 Service : 2.0   Value For Money : 1.0   Ambience : 2.0
B70iTJjcPkuYn8ouUewWgw  Stars : 3.33    Food : 1.43 Service : 1.92  Value For Money : 1.4   Ambience : 2.06
```

Figure 6: Attribute ratings and actual ratings (Stars) for restaurants.