Unsupervised vs Supervised: Sentiment Analysis on Hotel Review Ratings

Introduction.

We wanted to investigate the polarity of hotel reviews from Tripadvisor. Maika and Liz are both travel enthusiasts and wanted to use a dataset that was related to a travel experience. Since learning about clustering in class, we compared the performance of an unsupervised clustering method, k-means clustering, versus a supervised clustering method, Gaussian Naive Bayes clustering. We used embeddings of the review text as high dimensional features for assigning the hotel reviews into two clusters: the 1 label and cluster was assigned to positive reviews and the 0 label and cluster was assigned to negative reviews.

Experimental Setup.

We used a dataset of hotel reviews extracted from Tripadvisor, which has users' review comments and ratings of hotels. The following describes the setup:

1. Cleaning and conversion of the dataset

   We defined and excluded non-qualifying examples such as nondescript reviews that had just '#NAME' as the text. The original dataset had ~130k examples. We also removed duplicate examples, which resulted in ~150 non-qualifying examples. In the end, we had to limit our dataset size to 10k examples due to token constraints enforced when obtaining the word embeddings from the OpenAI API. However, we still maintained the original proportion of the number of positive reviews vs the number of negative reviews.

   We then converted review ratings to binary labels. The original examples had ratings on a scale from 1 to 5, and we converted the labels of examples with 1 or 2 ratings to "0," and those of examples with 4 or 5 ratings to "1." We removed examples with a rating of 3, which consisted of ~12% of the cleaned dataset (~16k examples). After the conversion, the label 0 indicated "bad rating" and 1 indicated "good rating."

   As the final phase of this step, we used the OpenAI API to obtain Ada 2 embeddings of each review. Specifically, we converted the textual comments of each review to numerical vectors. After the cleaning and conversion, each example in our data included a review comment of vector type as features and a binary rating of 0 or 1 as labels

2. Implementation of unsupervised learning

We implemented the k-means clustering to create 2 clusters. After building the model with the scikit-learn library, we built the k-means clustering by coding from scratch. Here is an overview of our model: (1)initialize the centroids randomly, (2) assign each data point to the nearest centroid, (3)recompute centroids as the mean of all data points assigned to each centroid, (4)repeat steps 2 and 3 until convergence (i.e., when centroids do not change significantly).

3. Implementation of supervised learning
   We implemented the Gaussian Naive Bayes (GNB) classifier to train and classify our dataset. After building the model with the scikit-learn library, we built the GNB classifier from scratch. Here is an overview of our model: (1)calculate the mean and variance for each feature for each label, (2)use these parameters to estimate the probability density of each feature given a label, (3) apply Bayes' theorem to compute the posterior probability for each label, (4) classify each example into the label with the highest posterior probability.

4. Evaluation
   We ran both models 20 times each and calculated average scores. For unsupervised learning, we calculated the silhouette score and the accuracy by comparing the actual label with the assigned cluster. For interpretation of the silhouette score, 1 means that the clusters are dense and well separated, whereas 0 means clusters are overlapping. For supervised learning, we calculated the accuracy and f1-score. We used a t-test to decide whether the accuracy comparison between the two models was statistically significant. We could not test other scores (silhouette and f1) as these could not be compared between models.

Results.

For unsupervised learning, the average silhouette score was close to 0 (0.058), and the average accuracy was very low (0.49) (Table 2). For supervised learning, the accuracy was close to 100% (0.98) (Table 1). The t-test on the accuracies had a p-value of 8.0606e-51, which is less than 0.05 and significant. Therefore, the supervised clustering was significantly more accurate than unsupervised clustering. The 3D visualization of clusters in Figure 2 shows that the supervised model cluster matches the actual clustering via ground truth labels much more than the unsupervised model cluster and its visualization in Figure 1.

Conclusions.

Our results indicated that for this particular problem of predicting review ratings, supervised learning had better assignment of clusters to actual polarity.

Appendix.

Table 1. Results on Supervised Learning

Cluster	F1-score (avg)	Accuracy
0	0.83929	0.98199
1	0.99046	0.98199

Table 2. Results on Unsupervised Learning

Silhouette Score (avg)	Accuracy
0.05847	0.49433

Table 3. T-test between mean accuracy of Supervised v. Unsupervised

P-value
8.0606e-51

Figure 1. Clustering from k-means (Unsupervised Learning) versus Actual Label

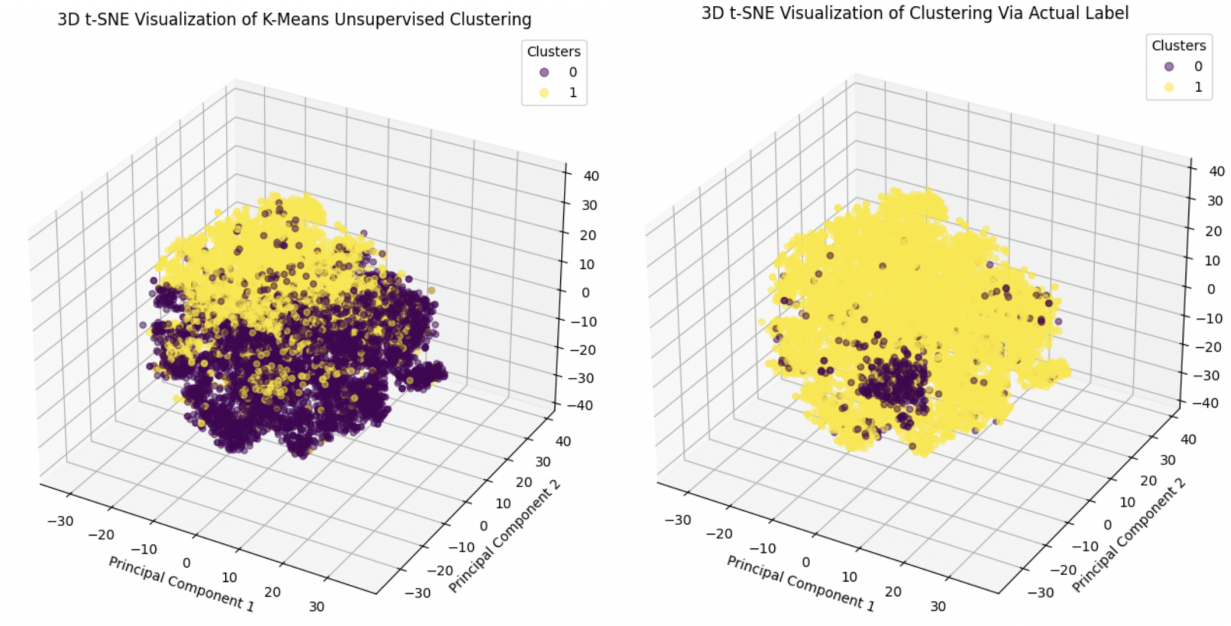
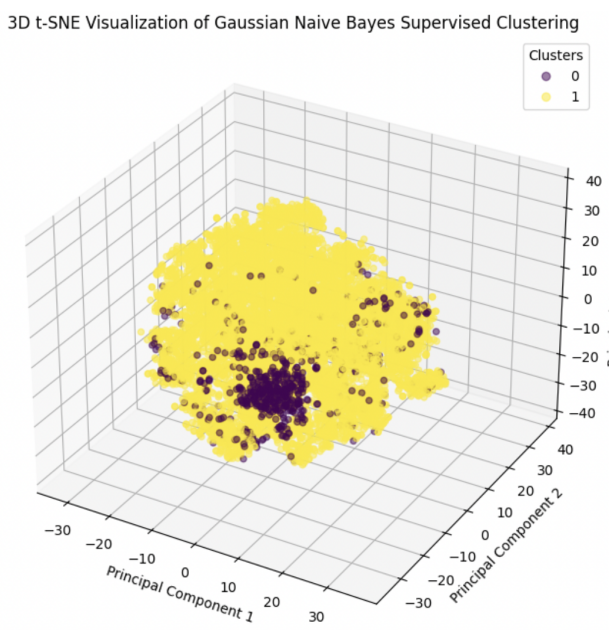
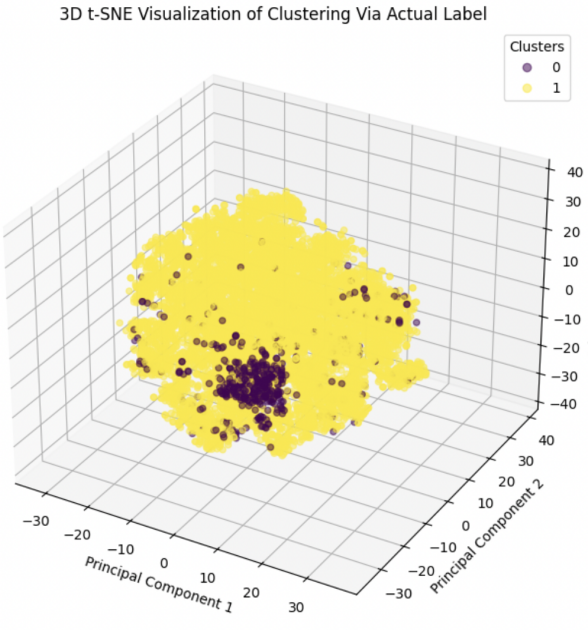


Figure 2. Clustering from Gaussian Naive Bayes (Unsupervised Learning)


3D t-SNE Visualization of Gaussian Naive Bayes Supervised Clustering
Clusters
0
1
Principal Component 1
Principal Component 2


3D t-SNE Visualization of Clustering Via Actual Label
Clusters
0
1
Principal Component 1
Principal Component 2