Project 5 – Advanced Data Mining Applications

CS548 / BCB503 / CS583 Knowledge Discovery and Data Mining - Fall 2019

Prof. Carolina Ruiz

Student: Bhoomi Kalpesh Patel

Description of the particular problem within the selected data mining topic to be addressed in this project	/15
Description of the approach used in this project to tackle the above problem. *All data mining techniques you use in this project for pre-processing, mining and evaluation must have been covered in class during this semester*	/25
Description of the dataset selected	/15
Appropriateness of the dataset selected with respect to this topic/problem	/10
Guiding questions	/10
Preprocessing	/10
Experiments: • Sufficient & coherent (most experiments must be performed in Python)	/25
• Objectives, Data, Additional Pre/Post-processing	/20
• Presentation of results	/20
• Analysis of results	/30
Overall discussion, comparisons, and conclusions	/20
TOTAL	/200

Total Written Report: _____/200 = _____/100

Class Presentation: _____/100

Class participation during project presentation: _____/100

Do not exceed the given page limits for this written report

Topic: Text Mining

1. **Description of the particular problem within the selected data mining topic to be addressed in this project:**

For an E-Commerce company, the products they sell and the reviews they obtain from customers by selling their items play a major role in deciding the fate of the company. By analyzing the reviews obtained, this company can easily target to increase the sales and further predict the future sales.

2. **Description of the approach used in this project to tackle the above problem:**

To tackle the above problem, we use different approaches such as Sentiment Analysis to identify percentage of positive, negative or neutral reviews. Further, analyzing each word of the reviews by clustering can assist in identifying words that are very much similar and thus help to extract what are aspects of the products that matter to the customers. We can also identify what kind of products the customers would like. Finally, by applying anomaly detection to the reviews, we can get rid of unnecessary data and extract only the necessary patterns.

3. **Dataset Name:** Women's E-Commerce Clothing Reviews
4. **Where found:** Kaggle
5. **Dataset Description:** This dataset consists of 23k rows and has 10 attributes. Each row represents a customer review with respect to a product that belongs to a class and department. Therefore, the dataset has following 10 features:
   Clothing ID; which is unique for each piece, Age; age of the customer writing the review, Title; title of the review, Review Text; the review written by the customer for the product, Rating; It ranges from 1 to 5 where 1 and 2 means Negative reviews, 3 means Neutral and 4 and 5 means positive. Recommended IND; a binary value which states if a person would recommend the product or not, Positive Feedback count; no. of positive reviews by the customer, Division Name, Department Name and Class Name represent the categorical name of the product.

6. **Initial data preprocessing, if any:**

There were missing values in the column, title and review text. These missing values were represented as NA. These values were converted by using the following: df = df.fillna("").

7. **Three Guiding Questions about the dataset domain:**
   1. Using the reviews related to Department of Tops and Dresses can we predict the sentiments of the user?
   2. Given the reviews of Blouses, skirts and Sweaters, can we get any insights that will be useful to the organization?
   3. Based on the reviews for tops and bottoms, are there any words that are typo, use slang or are not related to the product?

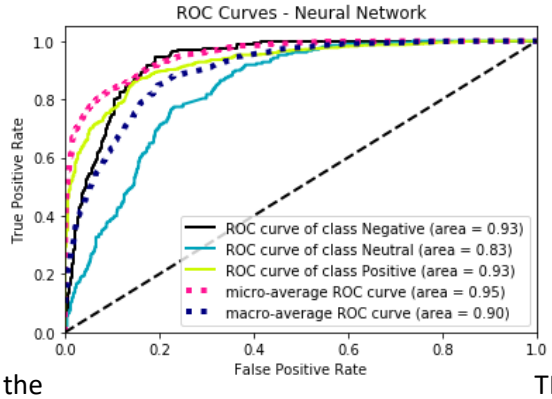
Summary of Experiments. *At most 2 page.*								
GQ	**Python function**	**Pre-process**	**Mining Technique**	**Parameters used**	**Results**	**Time taken**	**Evaluation**	**Observations about experiment Observations about visualization Interpretation results**
1	sklearn.ensemble.RandomForest, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Random Forest	n_estimators=20, criterion='entropy', random_state=0	Accuracy = 0.78, f1score= 0.78 Precision= 0.78 Recall=0.78	1 min 4 secs	Using confusion matrix, it was observed that the highest precision, recall and f1 score was obtained for Positive reviews	With the ROC curve, it was observed that ROC of Positive reviews was 0.87, the highest of all, followed by ROC of Negative reviews = 0.86 and ROC of Neutral reviews = 0.75
1	sklearn.ensemble.RandomForest, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Random Forest	n_estimators=40, criterion='gini', random_state=0, min_samples_leaf=10	Accuracy = 0.76, f1score= 0.76, Precision= 0.76, Recall=0.76	1 min 39 secs	Using confusion matrix, it was observed that the precision for Negative/Neutral reviews was 1.0 whereas for the Positive reviews it was 0.78	By plotting ROC curve for this model, it was observed that ROC Curve for positive class was 0.89, while that of Negative was 0.91 and of Neutral was 0.80
1	sklearn.neural_network.MLPClassifier, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Neural Networks	hidden_layer_sizes=(10, 10), activation='relu', solver='sgd', alpha=0.0001, learning_rate='constant'	Accuracy = 0.81, f1score= 0.81 Precision= 0.81 Recall=0.81	14 mins 42 secs	Using confusion matrix, it was observed that precision, recall and f-measure was highest for the Positive class whereas lowest for the Neutral class	By plotting ROC Curve for the MLPClassifier, it was observed that ROC Curve for Negative and Positive reviews was the same i.e 0.93 whereas for the neutral reviews, the AUC was 0.83
1	sklearn.neural_network.MLPClassifier, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Neural Networks	hidden_layer_sizes=(10, 5,2), activation= 'tanh', solver= 'sgd', learning_rate='adaptive'	Accuracy = 0.80, f1score= 0.80, Precision= 0.80 Recall=0.80	14 mins 29 secs	With confusion matrix, it was observed that the positive reviews precision, recall and f-measure was less compared to the above model	On plotting ROC Curve, observed that micro-average ROC curve was 0.95 whereas for macro-average it was 0.89. Again, ROC Curve for positive and negative reviews was the same
2	sklearn.cluster.KMeans, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences,	K-Means Clustering	n_clusters= 3, init='k-means++', max_iter= 200, n_init= 1	SSE = 330.21 # iterations = 12	2 mins 55 secs	Silhouette score obtained was 0.16. This implied that though the clusters were very close, the	On visualizing, it was observed that 3 clusters were obtained. Of these two were overlapping clusters. But when the same clusters where project

		Applying TFIDF					words within each cluster were very similar	on a 3 dimension, the clusters were at ease to interpret
2	sklearn.cluster.KMeans, gensim.models.Word2Vec	Removing punctuations, stopwords from sentences, Converting to vector of words	K-Means Clustering	n_clusters= 3, init='k-means++', max_iter= 200, n_init= 1	SSE = 2583.88 # iterations = 20	29 secs	Silhouette score obtained was 0.33. This was very better than just applying TFIDF, stating the cluster formation was better comparatively	On visualizing, it was observed that there was better separation between the clusters. The outliers were detected as the 3rd clusters whereas, majority of the points were part of 2 compact clusters
2	sklearn.cluster.KMeans, gensim.models.Word2Vec	Removing punctuations, stopwords from sentences, Converting to vector of words	K-Means Clustering	n_clusters=k, init='random', max_iter =100, n_init =1, algorithm='elkan'	SSE = 2583.88 # iterations = 20	40 secs	adjusted_rand_score= -0.012 normalized_mutual_info_score= 0.043 adjusted_mutual_info_score= 0.03	Using random as initialization for kmeans, did not affect the cluster formation, thus concluding that with word embedding the model was able to capture the distances between the words.
3	scipy.stats, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Zscore (Gaussian Distribution)	threshold = 3	No. of outliers obtained = 10143 No. of non outliers = 182897	52 secs	To evaluate this model, used IQR. Computing 1st and 3rd Quartile, the words that did not lie in IQR = Q3 – Q1 were outliers.	Using scatter plot, it was observed that there were few data points which were not included in the boxes. Even, with scatter plot, outliers were on the left side.
3	Sklearn.neighbors.LocalOutlierFactor, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Local Outlier Factor (Density based)	n_neighbors =10, algorithm=kd_tree, contamination=0.05	% of outliers = 4	9 mins 18 secs	Using negative_outlier_factor_, we could identify the words marked as outliers by converting them in the range [-1,1]	On visualizing, it was observed that even on dense regions there are lot of outliers. This is because some of the reviews used slang words which led to identifying them as outliers
3	Sklearn.neighbors.LocalOutlierFactor, sklearn.feature_extraction.text.TfidfVectorizer	Removing punctuations, stopwords from sentences, Applying TFIDF	Local Outlier Factor (Density based)	n_neighbors =30, algorithm='auto',contamination =0.1	% of outliers = 8	8 mins 56 secs	Using negative_outlier_factor_, we could identify the words marked as outliers by converting them in the range [-1,1]	On visualizing, it was observed that with the increase in the contamination parameters, more reviews were classified as outliers as compared to above.

Analysis of Results: (at most 2 page) 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.

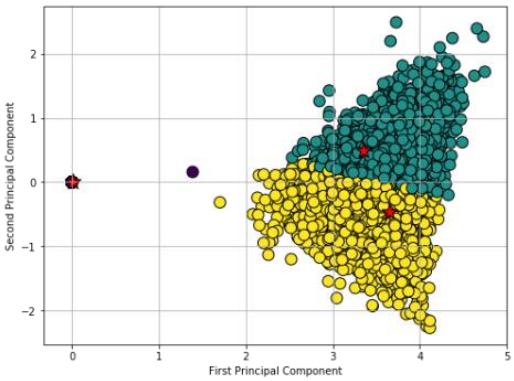
1. Varying the parameters played a very important role in the results and the observations obtained.
   Guiding Question 1: In random forests, n_estimators and criterion were the most crucial parameters. As the value of n_estimators increased from 20 to 40, the time taken to build the model also increased. With entropy as the splitting method, the accuracy and confusion matrix obtained was better as compared to gini criterion.
   Guiding Question 1: In Neural Networks, with increase in hidden layers, the time taken to build the model increased. As we changed the value of activation function from relu to tanh, the accuracy and the area under curve for each review class dipped.
   On comparing Random Forests with Neural Networks to perform sentiment analysis, it was observed that Neural Networks provided the best results.
   Guiding Question 2: Preprocessing played a very important role in the formation of the clusters. Using TFID, when the initialization method was changed from k-means++ to random, the cluster formation was widely vague. Whereas, on using a pretrained GloVE model, changing the initialization method did not change the formation of clusters.
   Guiding Question 3: Using Gaussian distribution, it was observed that the outliers were correctly detected using mean and standard deviation. With Local Outlier Factor, changing the algorithm was a major factor for detecting the outliers. Using algorithm as kd_tree instead of auto, the optimization increased. By increasing the contamination value from 0.1 to 0.15, the no. of outliers increased drastically.
2. Guiding Question 1: From the point of view of the domain, just by reading the reviews of the users for Tops and Dresses, it was observed that majority of them were positive. On further analysis, it was also observed that people whose reviews were positive, tend to recommend the product to other users. Moreover, 40% of people with neutral reviews also tend to recommend the product. In the experiments, the models were correctly able to identify the three classes with positive class being of the highest percentage, thus concluding, that majority of the costumers were happy with these products.
   Guiding Question 2: Initial visualization showed there was just one big cluster with some outliers. But on performing cluster analysis, it was observed that words which are similar were clustered together. For example, on clustering with n = 3, few words from clusters were as follows: Cluster1: dress, jean, skirt, Cluster2: size, look, wear, color and Cluster 3: love, perfect, comfort. Further using pre-trained model, the clusters obtained were far better than applying TFIDF wherein with 3 clusters, the 3rd cluster detected outliers.
   Guiding Question 3: Just by looking at the reviews for tops and bottoms, it was observed that some of the customers whose reviews were positive were negative used slang language. For example, one of the slang used were waaaay, reallllllly, etc. With using Gaussian Distribution and Local Outlier Factor, the model was correctly able to identify these words and outliers. Some of the output obtained using these experiments were aaaaaaand, andnotpayforship, etc.

3. Guiding Question 1: The best model obtained for sentiments analysis was using Neural Networks.

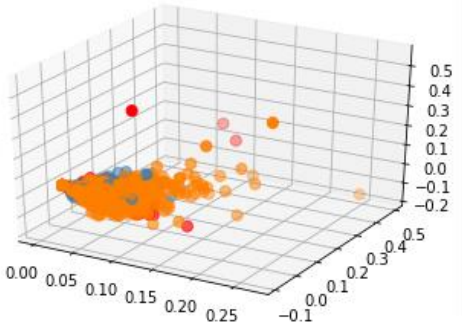


For AUC with Random Forest, the ROC Curve for Positive reviews was 0.87, that of negative reviews was 0.86 and that of neutral reviews was 0.75. Whereas when ROC Curves were plotted for Neural Networks, it was observed that the model was able to learn from the patterns from the reviews of the users.

Guiding Question 2: Cluster analysis obtained using pre-trained GloVE model, formed different cluster as compared to the TF-IDF. It was observed that 3 overlapping clusters are formed. But on further analysis, it was observed that these clusters had similar words in within the cluster. The figure to the left represents the cluster obtained using GloVE model. It was observed that, with 3 clusters, 2 clusters have majority of the data, whereas the 3rd cluster detected outliers. Even after using init='random', the Glove Model formed the same clusters thus stating that, the pre-trained model was less sensitive to the data instances. By comparing two clusters formed by TFIDF and GloVE model and using relative indices such as adjusted_rand_score, normalized_mutual_info_score and adjusted_mutual_info_score. A negative score was obtained for adjusted_rand_score which implied that the cluster formations were completely different.



Guiding Question 3: Using Local Outlier Factor, when the output was projected on a 3-dimensional space, the anomalies could be easily observed. Since, LOF uses KNN which is based on density, the points which are far away from that neighbors are marked as anomalies. As seen in the graph, the red points are the anomalies obtained. When the graph was plotted for n_neighbors as 30 and contamination to 0.1, the percentage of outlier increased. Further even with ZScore, the values obtained for outliers followed a Gaussian distribution wherein, the points with threshold greater than 3 were marked as outliers. Finally, by using BoxPlot of Gaussian Distribution, we could identify the outliers with respect to the other words and also with respect to the outfits.



Summary of what you learned in this project:

In this class, we distinctively learned models such as Decision Trees, Neural Networks, Clustering, Anomaly detection, etc. With this project, by implementing Random Forests and Neural Networks to text data, I understood how these models are used for text classification. Further, by implementing text clustering using K-Means and GloVE model, I understood how clustering can be extended to apply on text data. By using GloVE model, I understood how pre-trained vectors affect clustering of text data. Finally, I learnt how to perform anomaly detection not just by using statistical methods but also using nearest neighbors.