CLASSIFICATION OF E-COMMERCE DATA & SPAM REVIEW DETECTION SYSTEM

By
Khushboo Mantri
Samiksha Gaikwad

INTRODUCTION:

E-Commerce website advertisements attract billions of customers which ends up in billions of orders. Most of their advertisements are of the similar products, so there has to be a way to find out the similar products and cluster them, so that these clusters can be used to find out customers interests and give them suggestions. The challenge is to group same products by considering product level equivalency (PLE). Products need to be clustered together if they have the same conditions even if they belong to different listings in a single group. Product Level Equivalency considers manufacturer specifications which means it doesn't consider details like condition, price, shipping cost, location but features like brand, color, size, style are needed to be considered.

Fake reviews, fake comments, fake blogs, fake social networking postings, deceptive messages are identified by opinion spam detection. The review-centric sites such as yelp can be considered while detecting fake review detection. The supervised techniques consider distinct features generated from the reviews as well as the behavior of the reviewer. A publically available large scale and generated dataset has been considered provided by the yelp reviews which are classified using few well known supervised classifiers which bifurcate the reviews as true or deceptive by considering various features of the data. The work proposes and evaluates some additional new features that can be suitable to classify genuine and fake reviews.

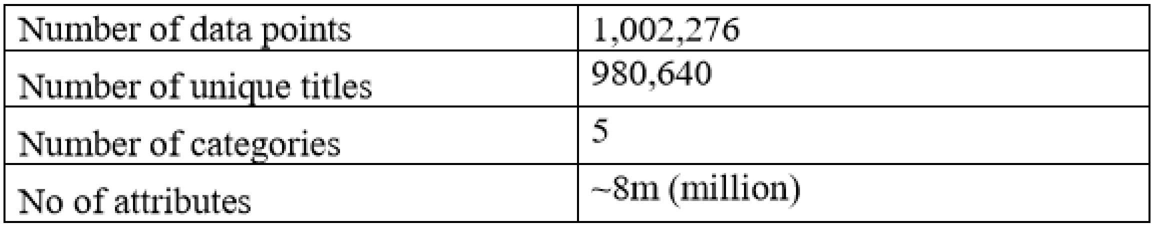
OBJECTIVE:

1. To cluster data from e-commerce website and give the corresponding measures to validate the clustering. (Clustering Analysis)
2. To classify the reviews which have been posted on hotel websites are true or deceptive. (Classification)

DATA ANALYSIS AND STATISTICS:

For Clustering:

Taking a look inside the data:



Number of data points	1,002,276
Number of unique titles	980,640
Number of categories	5
No of attributes	~8m (million)

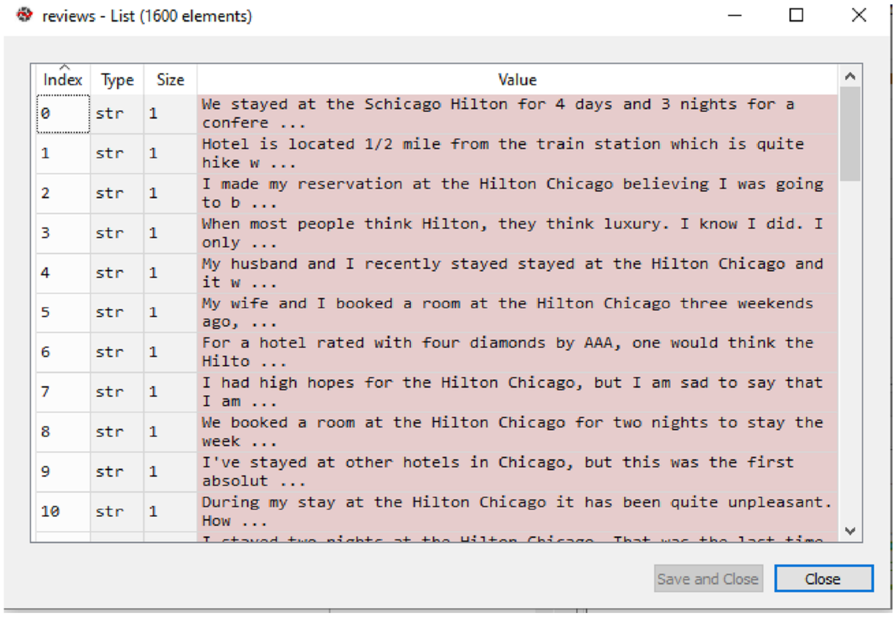
The dataset provided by ecommerce site contains over 1 Million data points, which have about 1 million titles and over 8 million corresponding attributes. The entire dataset has data of five different categories. This dataset includes following columns:

Columns

- category – A number representing the category to which the listing belongs
- title – The title of the listing
- subtitle – The subtitle of the listing
- gallery_url – The primary image URL for the listing
- picture_url – Semicolon ";" separated list of all associated image URLs for the listing
- attributes – Seller-provided attributes for the listing as a list of key-value pairs, see below for details
- description_html_base64 – Seller-provided listing description HTML which is base64 encoded
- index – A numeric identifier for each listing

For Classification:

The dataset contains 1600 reviews per polarity which is positive and negative polarity. The Polarity folders are taken with data from Web as well as MTruck in each of them.



Index	Type	Size	Value
0	str	1	We stayed at the Schicago Hilton for 4 days and 3 nights for a confere ...
1	str	1	Hotel is located 1/2 mile from the train station which is quite hike w ...
2	str	1	I made my reservation at the Hilton Chicago believing I was going to b ...
3	str	1	When most people think Hilton, they think luxury. I know I did. I only ...
4	str	1	My husband and I recently stayed stayed at the Hilton Chicago and it w ...
5	str	1	My wife and I booked a room at the Hilton Chicago three weekends ago, ...
6	str	1	For a hotel rated with four diamonds by AAA, one would think the Hilto ...
7	str	1	I had high hopes for the Hilton Chicago, but I am sad to say that I am ...
8	str	1	We booked a room at the Hilton Chicago for two nights to stay the week ...
9	str	1	I've stayed at other hotels in Chicago, but this was the first absolut ...
10	str	1	During my stay at the Hilton Chicago it has been quite unpleasant. How ...
			I stayed two nights at the Hilton Chicago. That was the last time

METHOD USED:

For Clustering:

1. Data preprocessing:

   The pre-processing is divided into two parts:

   a) Pre-process data frame entirely A function preProcess uses the input data which is the mlchallenge_set.csv file and outputs panda data frame. Currently it processes 'Titles' and 'attributes' only. The following steps are used at this initial phase of preprocessing:

   i. Remove all the special characters present, except comma (,) and semicolon (:).

   ii. Discard all single characters.

   iii. Carry out multiple to single mapping of comma's (,) and semicolon's (:).

   iv. Remove all the leading, sandwiched and trailing additional spaces.

   v. Make the data case insensitive.

   vi. Stem the data frame.

   b) Collect attributes and values and pre-process individually. The first thing that is done at this point is to section the data based on the categories as mentioned there are five categories, for each section we have to get all attributes, process them, find the unique ones among them and create a master dictionary

2. Tokenization

   Once the attribute list is formed, a specific value basically tokens are needed to be assigned for each attribute. As the data we have is a collection of strings and it becomes difficult for the clustering algorithm to handle string data, also performance wise the model becomes slow. Thus tokenization converts the strings into numerical values. Using these values a data-attribute matrix can be formed for clustering. The following steps are performed for tokenization:

i. Loop through each category.

ii. Look up master dictionary for each attribute, determine tokens for each value and Update Data Frame for the category

3. Apply Clustering Algorithm

After going through a series of algorithms K Means and EM clustering algorithms though being good for clustering categorical attributes have a prior requirement, for both these algorithms we need to specify the number of clusters that we want to be formed, in this case we don't know the exact number of clusters and therefore they are of no use here. Agglomerative clustering is one of the good choices as the data provided forms a hierarchy, but this clustering algorithm has a running time complexity of O(n3). Another better algorithm for this clustering is DBSCAN algorithm which has a complexity of O(n2). Agglomerative clustering and DBSCAN gives the similar results during initial testing with conservative values of tuning parameter. However, the run time was significantly different for the two algorithms as expected. just agglomerative clustering takes more time in comparison to DBSCAN.

4. Validation.

Two types of validations are conducted here:

i. Base Case with all values belonging to one cluster o 100 data points, same attributes ▪ 1 cluster containing all the data points is expected o 101 data points (100 pts with same attributes, 1 with different) ▪ 2 clusters with formed as expected

ii. Base case with no data points having PLE o Expected output: ▪ Expected 0 clusters, 100 separate pts.

For Classification:

1. Data Preprocessing

   The data which is collected is not consistent using various machine learning algorithms the data is trained and presented in a particular format. The pre-processing phase includes two preliminary operations, which help in transforming the data before the actual task. Data pre-processing plays a significant role in many supervised learning algorithms. To prepare the dataset for learning involves transforming the data by using the String To Word Vector filter, which is the main tool for text analysis in Weka. The String To Word Vector filter makes the attribute value in the transformed datasets Positive or Negative for all single words, depending on whether the word appears in the document or not. Removing the poorly describing attributes can significantly increase the classification accuracy, in order to maintain a better classification accuracy, because not all attributes are relevant to the classification work, and the irrelevant attributes can decrease the performance of the used analysis algorithms, an attribute selection scheme was used for training the classifier.

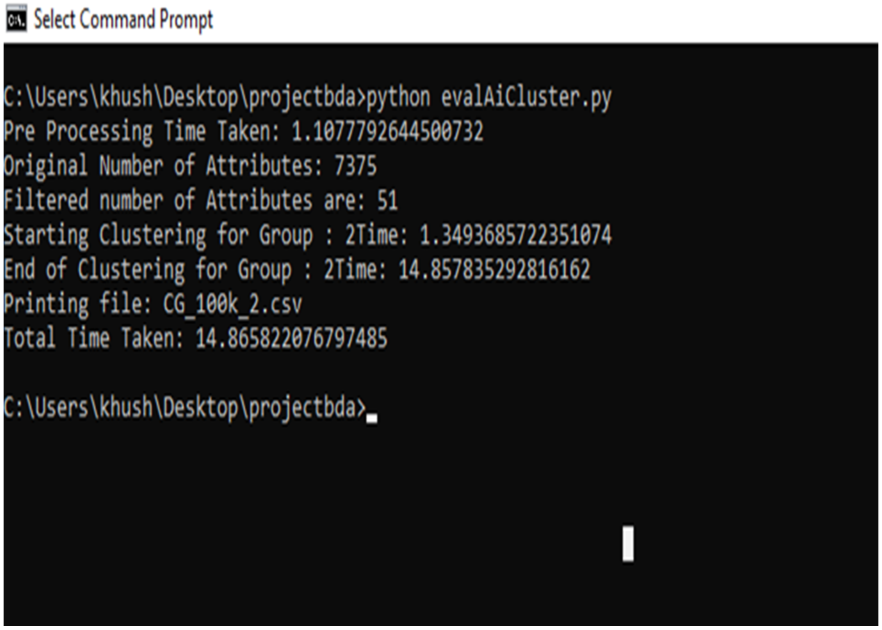
2. Using Naive Bayes Multinomial method as the classifier and training the data.

   This module deals with the various feature sets under consideration and analyse them in order to obtain insights from it. Feature selection is an approach which is used to identify a subset of features which are mostly related to the target model, and the goal of feature selection is to increase the level of accuracy. The NB classifier is a basic probabilistic classifier based on applying Bayes' theorem. The NB calculates a set of probabilities by combinations of values in a given dataset. Also, the NB classifier has fast decision-making process. In this case the training set is 0.8 and the test set is 0.2.

3. Display the accuracy, precision score, recall score , confusion matrix for validation.

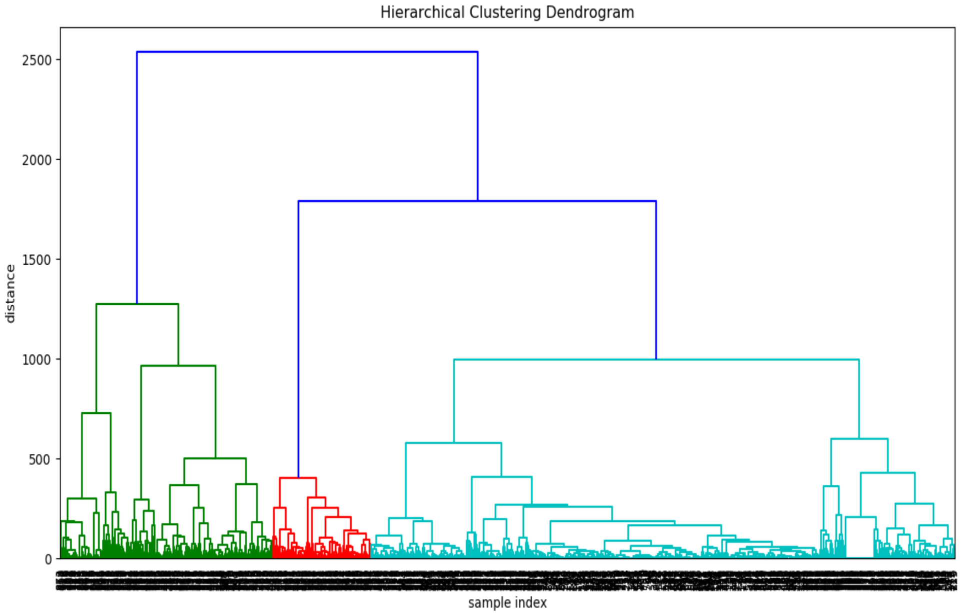
RESULTS:

Clustering :

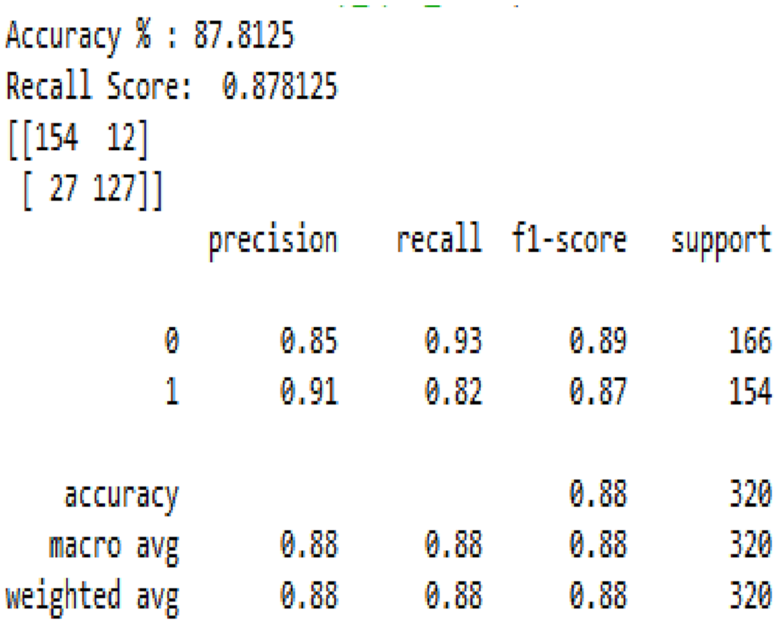


```
C:\Users\khush\Desktop\projectbda>python evalAiCluster.py
Pre Processing Time Taken: 1.1077792644500732
Original Number of Attributes: 7375
Filtered number of Attributes are: 51
Starting Clustering for Group : 2Time: 1.3493685722351074
End of Clustering for Group : 2Time: 14.857835292816162
Printing file: CG_100k_2.csv
Total Time Taken: 14.865822076797485

C:\Users\khush\Desktop\projectbda>_
```



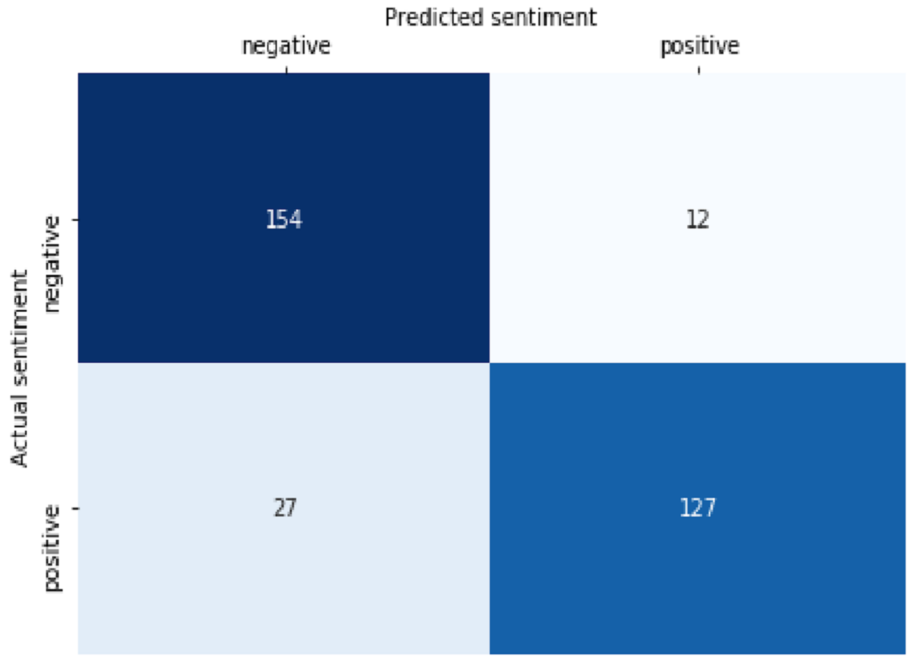
Classification:



```
Accuracy % : 87.8125
Recall Score:  0.878125
[[154  12]
 [ 27 127]]
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       166
           1       0.91      0.82      0.87       154

    accuracy                           0.88       320
   macro avg       0.88      0.88      0.88       320
weighted avg       0.88      0.88      0.88       320
```



CONTRIBUTION:

TASK	PEOPLE
Gathering Clustering Dataset	Samiksha Gaikwad
Gathering Classification Dataset	Khushboo Mantri
Preprocessing Clustering Set	Samiksha Gaikwad
Preprocessing Classification Set	Khushboo Mantri
Clustering Algorithm and Validation	Khushboo Mantri
Classification Algorithm and Validation	Samiksha Gaikwad
Report	Samiksha Gaikwad, Khushboo Mantri
Presentation Slides	Samiksha Gaikwad, Khushboo Mantri