

DATA MINING STUDY OF RESTAURANTS IN US



BY,
KRITHIKA RAGHAVAN
AMOL KUMTAKAR
GUNJAN BATRA

CONTENTS

INTRODUCTION
TOOLS USED
ZOMATO API	
Data Collection
MS - ACCESS	
Data Collection
Scripting
WEKA	
Pre-Processing
Classification
Clustering
TABLEAU	
Visualization

INTRODUCTION

The objective of this project is to perform a study on the Restaurant Industry in US.

Using Tableau, we have analyzed the restaurant data set of 5 cities of US based on the parameters of Cost of restaurant, Cuisines, User review rating, Locality, Health Inspection Grade and Text in the Reviews. Using WEKA, we use different classification algorithms to build a model for Text Data of User reviews to be classified as Positive/ Negative

Further the project, discusses text mining results from different Data Mining algorithms. We have started the project by collecting data from Zomato (popular site for checking reviews about each restaurant) on hotels available in New York.

The following information were collected for 5 cities of US – New York, San Francisco, Chicago, Detroit, Houston

Restaurant_ID, Restaurant_Name, Restaurant_URL, Restaurant_Address, Restaurant_Locality, Restaurant_City, Restaurant_Zipcode, Restaurant_Cuisines, Restaurant_Avg_cost_for_two, Restaurant_UserRating, Restaurant_UserRatingText, Restaurant_UserRatingVotes

Further for New York City, Health Inspection Rating and the restaurant reviews from Zomato were collected manually from <http://www.nyc.gov/html/doh/html/environmental/food-service-inspection.shtml> and www.zomato.com respectively.

Each hotel has both positive as well as negative reviews. In this project we apply multiple algorithms on this data set, analyze the resulting efficiency and the ROC graphs (Weka) and conclude which has been the best classification Algorithm for our data set.

Below is a snapshot of Zomato website for a specific hotel; Supper

The screenshot shows the Zomato website interface for searching restaurants in New York City. The search term 'supper' is entered in the keyword field. The results page features a map of New York City with a yellow marker pointing to the location of 'Supper' in Alphabet City. The restaurant's details are displayed on the right, including its name, address, cuisine type (Italian), and a user rating of 3.7 from 394 votes. The sidebar on the right includes a promotional message for the 'Brooklyn Crush' event and a call to action to download the Zomato app for exploring more restaurants.

DATA MINING PROJECT

Each hotel has three tabs namely reviews, blogs and popular. Zomato rates each hotel based on the data in all the three tables. Below snapshot shows how for hotel **Supper** we have three tabs:

The screenshot shows a web browser displaying the Zomato website at <https://www.zomato.com/new-york-city/supper-alphabet-city/reviews#tabtop>. The page is titled "Popular 10" and "Reviews 11" and "Blogs 9". There are two reviews listed:

- Stacy Landers** (15 Reviews, 30 Followers) posted 12 days ago via Zomato for iOS. RATED 4.5. Review: "I am a big fan of Frank Prisinzano. The ambiance is very reminiscent of Frank and Lil' Frankie's. Everything was good, but Lil' Frankie's is where my heart lies." Buttons for Like (0), Dislike (0), and Share are shown.
- Dallas Trends** (133 Reviews, 106 Followers) posted 3 months ago. RATED 4.5. Review: "Good food. Great atmosphere. I was with someone who never dined at a restaurant where they seat others at your table, and it was exciting. I had the privilege of dining for dinner. Somewhat limited seating. Friendly service. We decided to try this place after a local recommended it. I'd love to return." Buttons for Like (0), Dislike (0), and Share are shown.

The data is taken only from the reviews tab and not from any other tab.

TOOLS USED -

We have used Zomato API, Ms-Access, Weka and Tableau to implement various data mining algorithms and techniques and also for data visualization.

Zomato API → This a tool for developers to obtain real time data from Zomato by writing web scripts

MS-Access → It is a database used to store a huge data set. It makes the job easy to create forms or load data when the data set are huge. MS-Access is SQL database and we have **FORMS** option to generate a form to accept the inputs and store them in a tabular format.

Weka → Weka is a data mining tool which has in built machine learning algorithms. This tool is extremely useful when we have to work on mining of data with various algorithms. This tool lets us compare the output of various algorithms on each data set and lets us decide which algorithm worked best with the given data set.

Tableau → Tableau is a data visualization tool. Even though we can do visualization with Weka, Tableau is the right tool that can show each point with perfect plotting and gives clarity in the data visualization. It produces a family of interactive data visualization products focused on business intelligence.

DATA MINING PROJECT

DATA COLLECTION

We collected the details of 495 restaurants in 5 cities of US from the Zomato API by writing scripts in the Zomato Developer API.

Restaurant ID	Restaurant Name	Restaurant URL	Restaurant Address	Restaurant Locality	Restaurant City	Restaurant Zipcode	Restaurant Cuisines	Restaurant cost_for_two	Restaurant_Avg	Restaurant_UserRating	Restaurant_UserRatin	Restaurant_RatingVotes	Health Inspection
16781904	Momofuku	https://www.zomato.com/us/momofuku	171 1st Ave	East Village	New York City	10003	Asian, Ramen	60	4.1	Excellent	1530	1530	A
16767139	Gramercy Tavern	https://www.zomato.com/us/gramercy-tavern	42 E 20th St	Union Square	New York City	10003	American	160	3.7	Very Good	1754	1754	A
16760100	Balthazar	https://www.zomato.com/us/balthazar	80 Spring St	Soho	New York City	10012	French, Cafe	140	3.7	Very Good	4005	4005	A
16775039	Peter Luger	https://www.zomato.com/us/peter-luger	178 Broadway	Williamsburg	New York City	11211	Steakhouse	150	3.8	Very Good	2091	2091	A
16783153	Shake Shack	https://www.zomato.com/us/shake-shack	366 Columbus	Upper West Side	New York City	10024	American, Burgers	30	4.2	Excellent	1524	1524	A
16783998	The Halal Guys	https://www.zomato.com/us/the-halal-guys	6th Avenue	Theater District	New York City	10019	Middle Eastern	25	4.4	Excellent	493	493	A
16761344	Buddakan	https://www.zomato.com/us/buddakan	75 9th Avenue	Meatpacking	New York City	10011	Chinese, Fusion	150	3.9	Very Good	1483	1483	A
16761402	Burger Joint	https://www.zomato.com/us/burger-joint	Le Parker Meridien	Theater District	New York City	10019	American, Burgers	25	4.1	Excellent	1381	1381	A
16785398	Shake Shack	https://www.zomato.com/us/shake-shack	691 8th Avenue	Hell's Kitchen	New York City	10036	American, Burgers	30	4.4	Excellent	856	856	A

Further, for New York City collected the Health Inspection Rating for restaurants, from <http://www.nyc.gov/html/doh/html/environmental/food-service-inspection.shtml>

Also, we have collected 1545 reviews (Weka – instances) across 90 hotels in New York City area. We created a Form in MS-Access where we loaded all the reviews to build the data set. Below is the snapshot of the form and data set.

The screenshot shows a Microsoft Access application window. The ribbon at the top has tabs for File, Home, Create, External Data, Database Tools, and a search bar. The 'Home' tab is selected. On the left, the navigation pane shows 'All Access Objects' with sections for Tables, Queries, and Forms. Under 'Forms', there is a list with 'hotel_nyc' highlighted. The main workspace displays a form titled 'hotel_nyc'. The form contains the following fields:

- Id
- Hotel_name: Supper
- Address: 156 E 2nd Street,
- Location: New York
- Zip: 10009
- cuisine: Italian
- Review1:
I am a big fan of Frank Prisinzano. The ambience is very reminiscent of Frank and Lil' Frankie's. Everything was good, but

At the bottom of the form, there is a status bar with 'Record: 14 5 of 99' and a 'Search' button.

DATA MINING PROJECT

Screenshot of Microsoft Access showing a query results grid.

The ribbon tabs visible are: File, Home, Create, External Data, Database Tools, and a search bar.

The left pane shows the "All Access Objects" navigation pane with sections for Tables, Queries, and Forms. The "hotel_nyc" table is selected.

The main area displays the results of the "Final_Query" query:

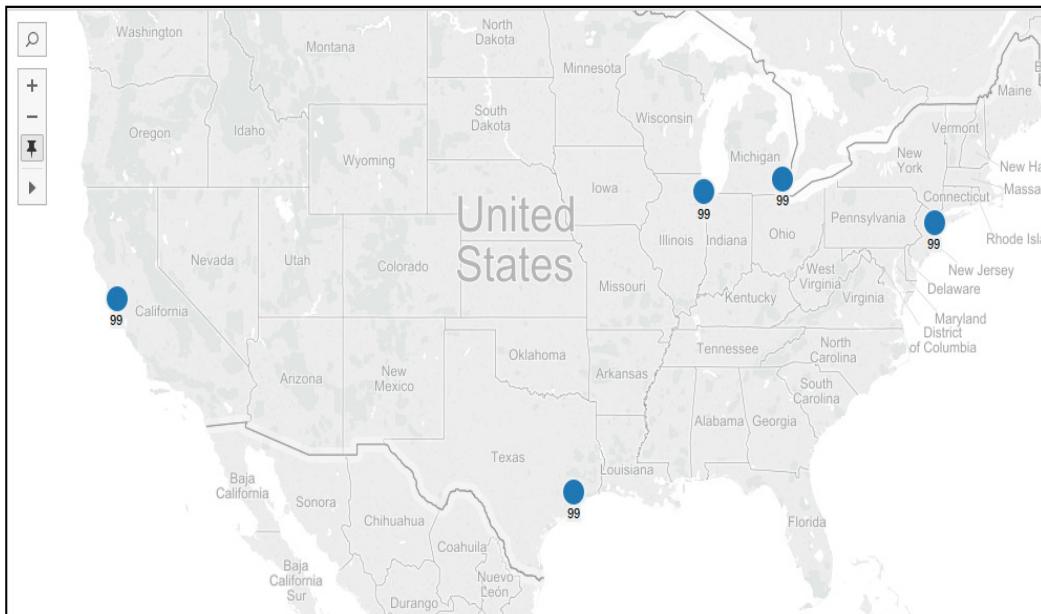
	id	hotel_name	address	location	zip	cuisine	Expr1006	Expr1007	Expr1008	Expr1009
1	1	Supper	156 E 2nd Street	New York	10009	Italian	I am a big fan of good food, great supper? had it perfect. I could			
2	2	Root & Bone	200 E 3rd Street	New York	10009	Southern, Café	great food, love if you want a cute atmosphere	I saw a picture		
3	3	Zum Schneider	107 Avenue C,	New York	10009	German, Bar	Food good food, large amazing food	zum schneider ja bier. this place		
4	4	Matcha Cafe W	233 E 4th Street	New York	10009	Cafe, Ice Cream	japanese owned	na	na	
5	5	Tuome	536 E 5th Street	New York	10009	Asian	the restaurant small, big, side	after seeing me	the restaurant	
6	6	Esperanto	145 Avenue C,	New York	10009	Cuban, Latin American	I think that by probably the best	went for dessert	good brailliant	
7	7	Katz's Delicatessen	205 E Houston Street	New York	10002	Sandwich	If you've been send your boy	I personally	one of the best	
8	8	Poco	33 Avenue B,	New York	10009	Spanish, Tapas	Ming this is the brunch is awesome	terrible, please	tapas and drink	
9	9	Minca	536 East 5th Street	New York	10009	Ramen	Pretty good rare had the chicken	great place for	I had the chicken	
10	10	Yerba Buena	23 Avenue A,	New York	10009	Latin American	great NY latin food	great place for	na	
11	11	Buenos Aires	513 E 6th Street	New York	10009	Argentine, Steakhouse	great authentic dining	planning to re	I've dined here	
12	12	Gnocco	337 E 10th Street	New York	10009	Italian	Cute little place gotta get the a	good vibe/good	na	
13	13	Black Iron Burg	540 E 5th Street	New York	10009	American, Burgers	Awesome burger best burger job	solid, great burger		
14	14	Balthazar	80 Spring Street	New York	10012	French, Cafe	B had one of my favorite bras	yes yes yes!	love the strawberry	
15	15	Momofuku Noodle Bar	171 1st Avenue	New York	10003	Asian, Ramen	The queue at	went on a very long	noodles and good	not all that, and
16	16	Buddakan	75 9th Avenue	New York	10011	Chinese, Fusion	Wow, from the	I always told	our hotel, so	one of the best
17	17	Per Se	Time Warner Center	New York	10019	French	Exceptionally perfection. has	I was really excited	extremely delicious	world famous
18	18	Peter Luger Steakhouse	178 Broadway	New York	11211	Steakhouse, American	Definitely live	extremely delicious	the best place	

TABLEAU

Using Tableau, we analyze the restaurant data set of 5 cities of US based on following parameters

- Cost of restaurant
- Cuisines
- User review rating
- Locality
- Health Inspection Grade
- Text in the Reviews

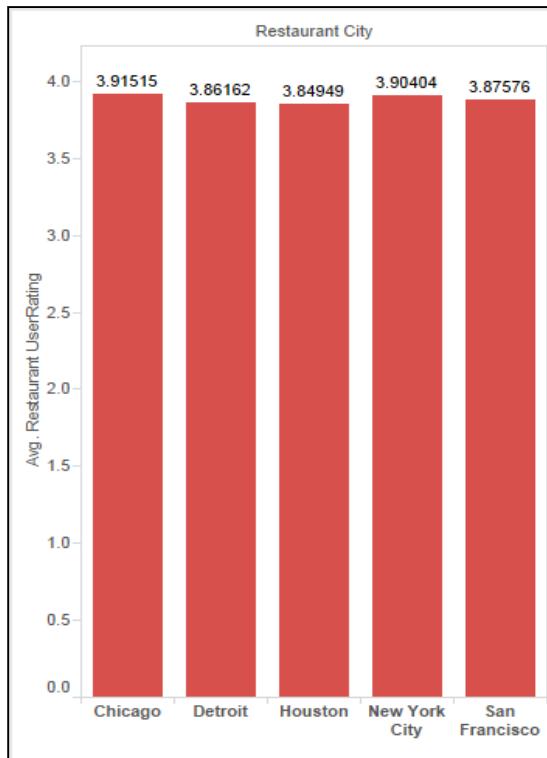
Description of Data



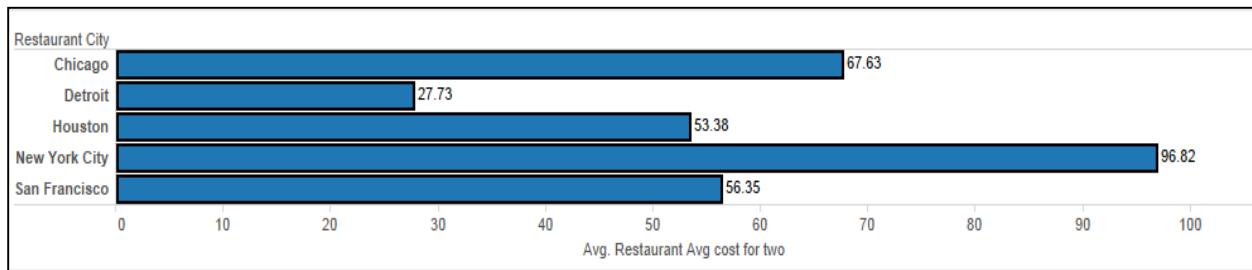
We took 99 Restaurant records for each city – New York, Houston, Detroit, San Francisco, and Chicago. 20 data records at each time were obtained for each city at a time in the form of json files which was converted to csv we combined the data set to obtain a collective data of 496 restaurants for exploration in Tableau and Weka. After exploration in Weka and Tableau the attributes were reduces from 31 to 13

DATA MINING PROJECT

Data Analysis

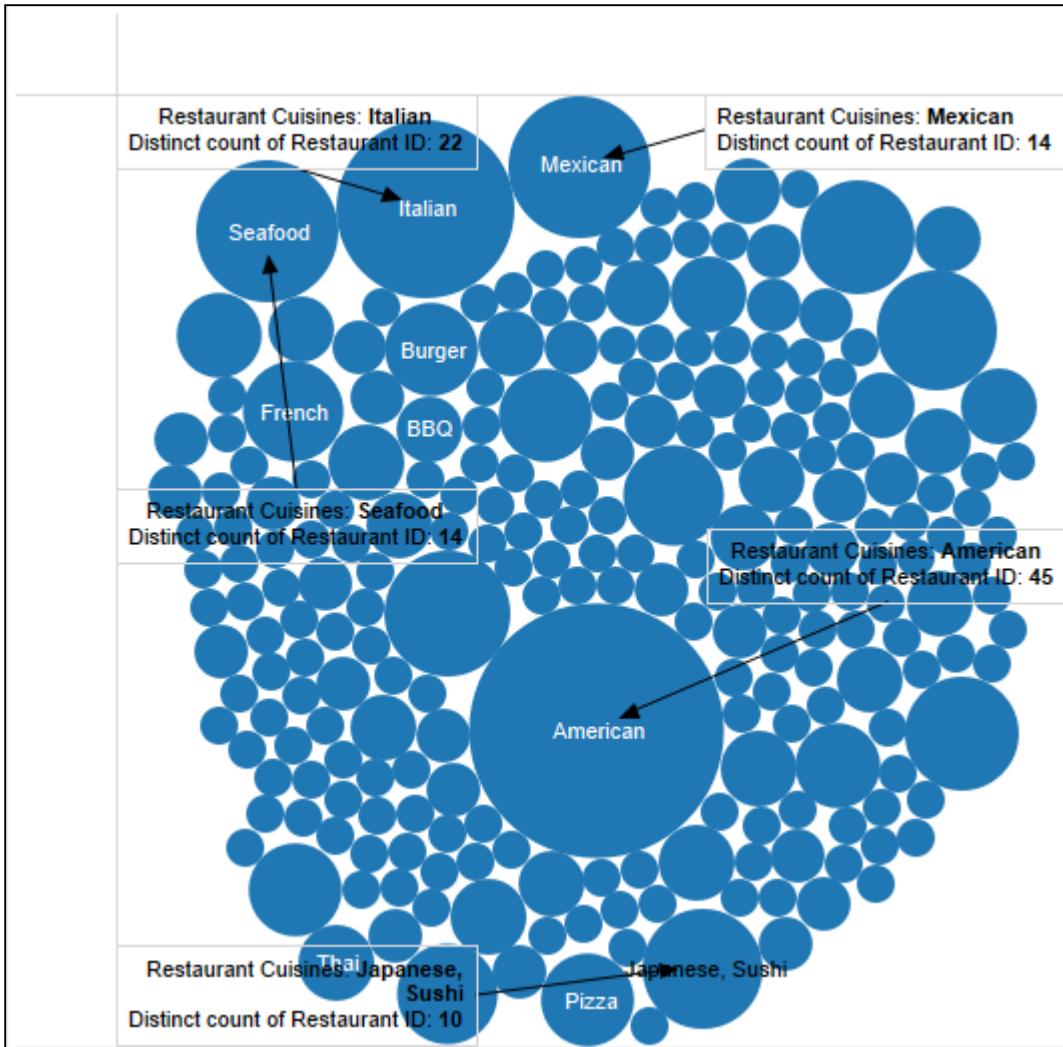


The graph between average user rating and the 5 cities shows that Chicago restaurants have the best user ratings



This graph between the Restaurant cities and Average of Average cost for two people shows that New York City has most expensive restaurants

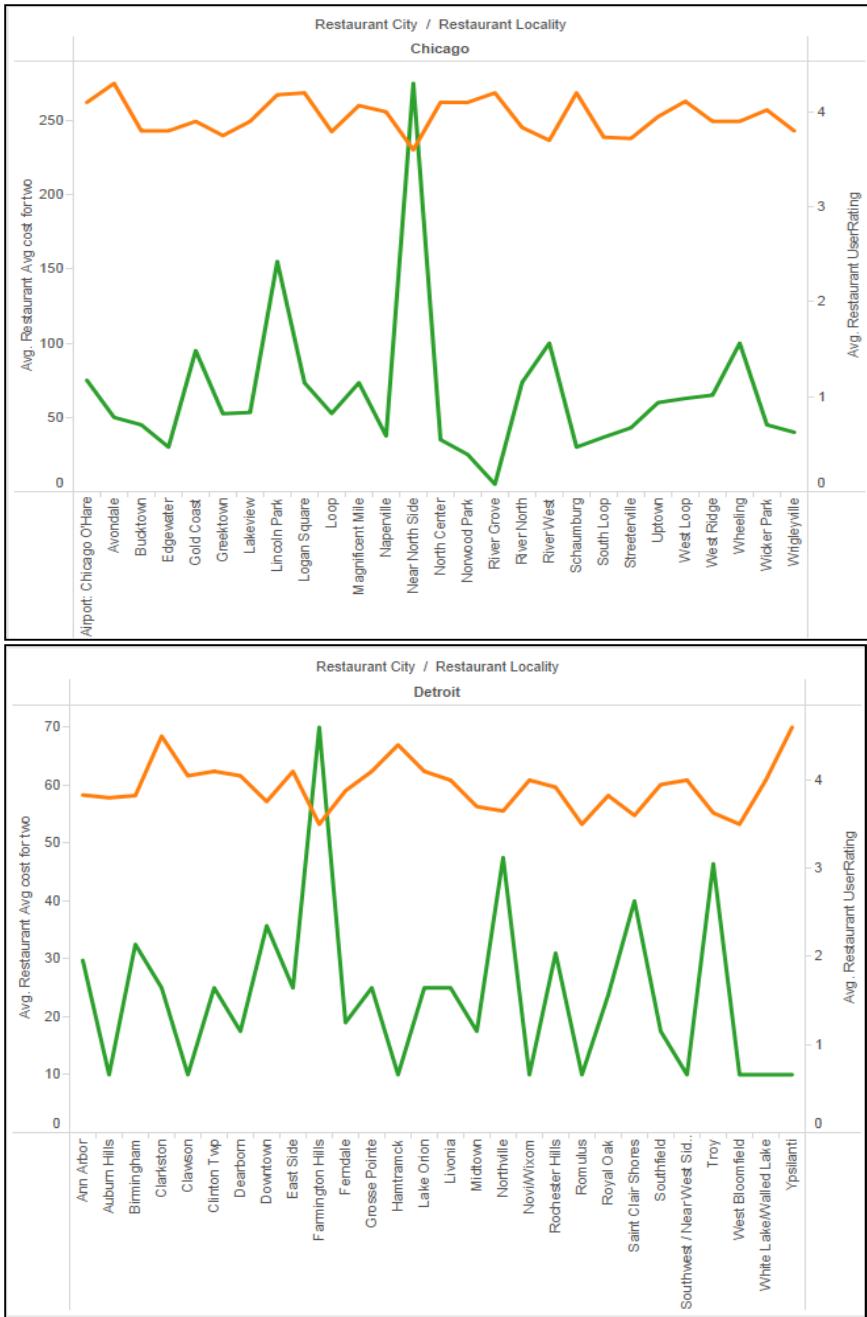
DATA MINING PROJECT



We took the count of restaurant for every cuisine to get the most common cuisine. The most popular cuisine in our data set is American which is in 45 restaurants, followed by Italian, Mexican, Seafood and Japanese

We did a study **of relationship between Average Cost for two and Average User Rating of a Restaurant** for the five different cities

DATA MINING PROJECT



DATA MINING PROJECT



DATA MINING PROJECT

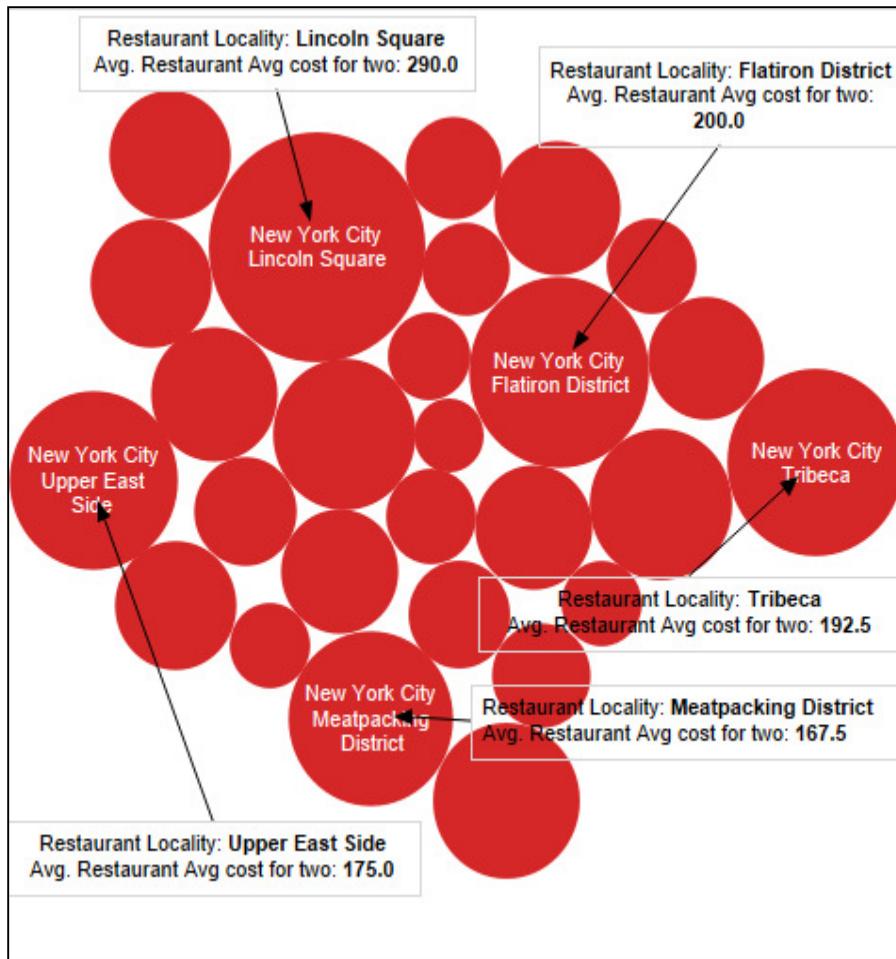


We couldn't see much correlation between the two variables. We also calculated Pearson Correlation Coefficient between the entire data set of Average Cost for two and User review rating which came out as -0.09466647716057779 which mean that they are probably not correlated.

DATA MINING PROJECT

New York City

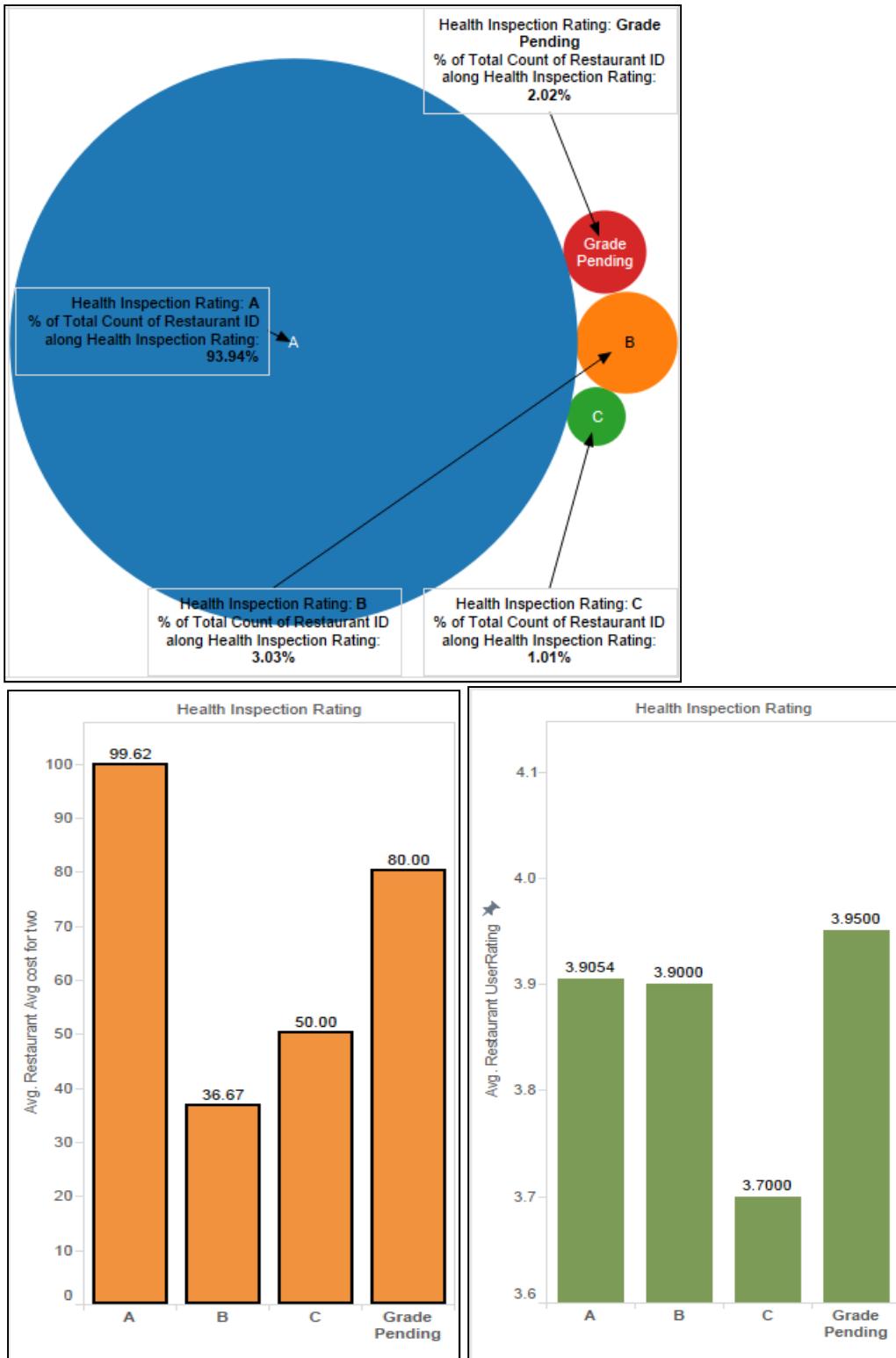
Locality wise – Average cost of two people



We found the average cost per locality in New York City and found that the most expensive restaurants in New York City are in Lincoln Square, followed by Flatiron District, Upper East Side, Tribeca and Meat Packing District

DATA MINING PROJECT

Results of Health Inspection Grades

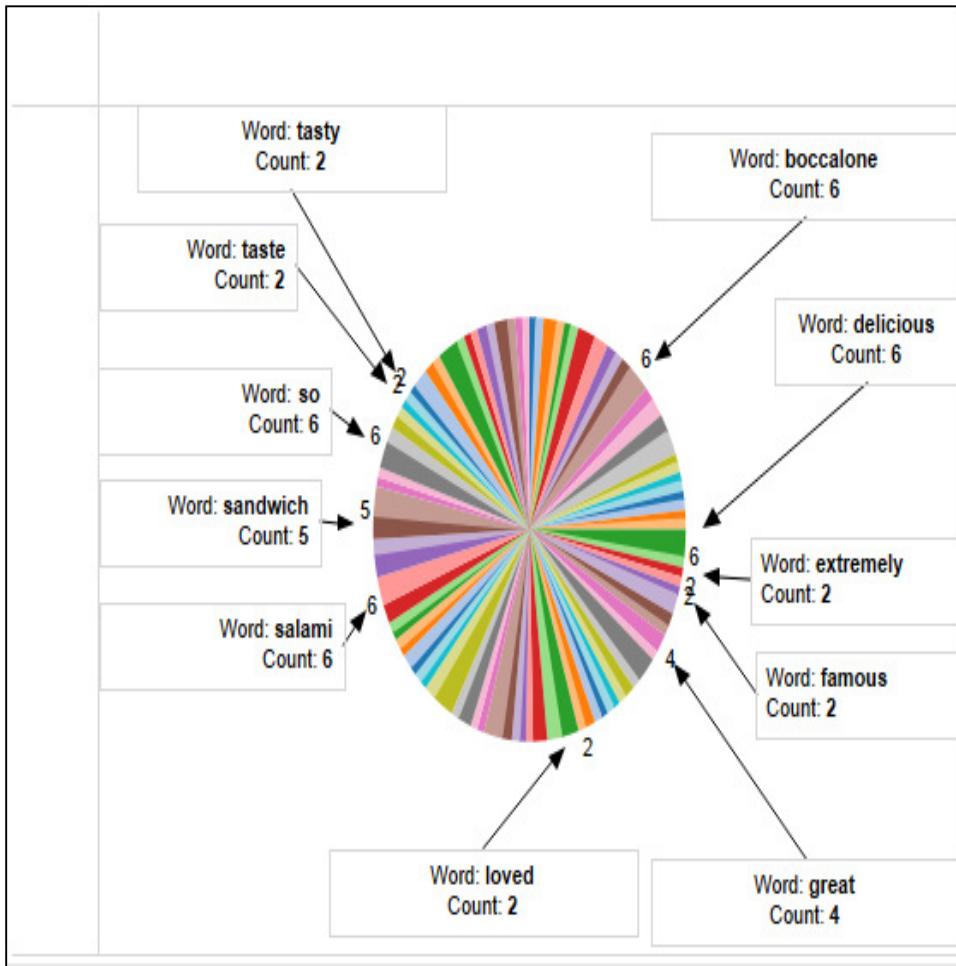


DATA MINING PROJECT

We took the percentage of Restaurants in each grade A, B, C for New York City and also found the average User rating of restaurants in each grade. We also found the Average Health Inspection rating of restaurants in each grade

The average user review rating for A grade restaurants is much higher than C grade restaurants

The average cost for two in A grade restaurants is much higher than in B and C grade restaurants



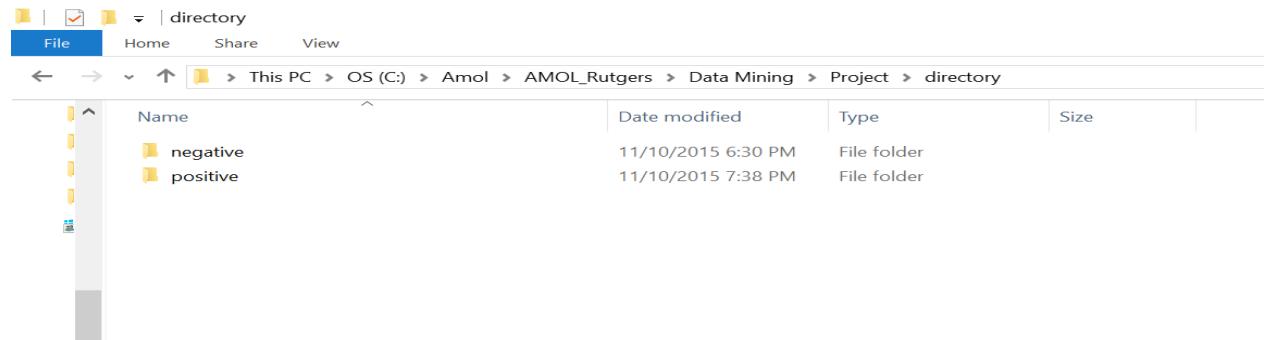
We did a study on Buccalone the restaurant with highest rating and found that the words adding to positive review of Buccalone: delicious, tasty, taste, famous, great, loved...

WEKA

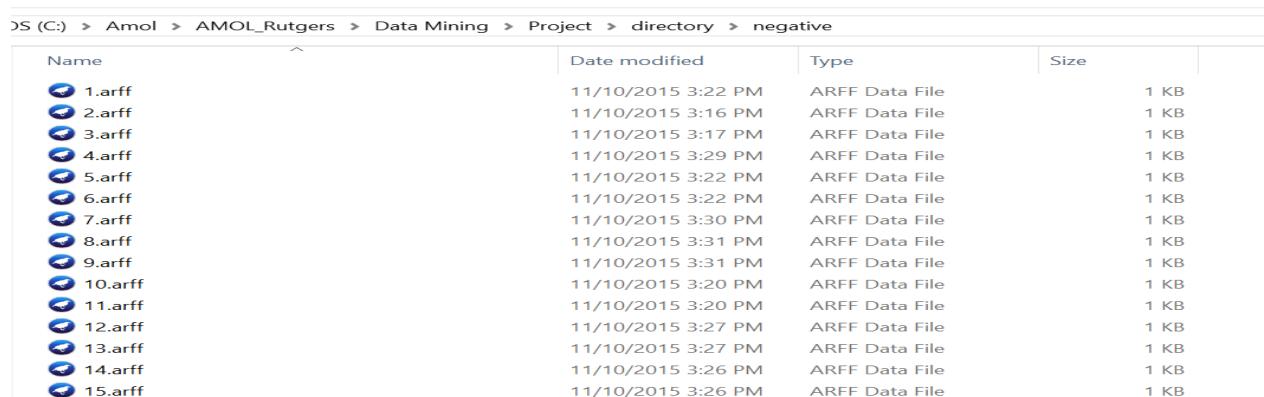
SCRIPTING

The following snapshots will show the positive and negative arff files for a specific hotel. Below we have shown the directory of both positive and negative reviews with few instances of arff files for both positive reviews and negative reviews.

All reviews are classified under two folders under one directory. One positive folder which contains **67 files of positive reviews** for each hotel and the second is **negative folder which contains 70 files** of negative reviews for each hotel. The 70 files in each folder are arff files which were written manually for both positive and negative reviews.



Name	Date modified	Type	Size
negative	11/10/2015 6:30 PM	File folder	
positive	11/10/2015 7:38 PM	File folder	



Name	Date modified	Type	Size
1.arff	11/10/2015 3:22 PM	ARFF Data File	1 KB
2.arff	11/10/2015 3:16 PM	ARFF Data File	1 KB
3.arff	11/10/2015 3:17 PM	ARFF Data File	1 KB
4.arff	11/10/2015 3:29 PM	ARFF Data File	1 KB
5.arff	11/10/2015 3:22 PM	ARFF Data File	1 KB
6.arff	11/10/2015 3:22 PM	ARFF Data File	1 KB
7.arff	11/10/2015 3:30 PM	ARFF Data File	1 KB
8.arff	11/10/2015 3:31 PM	ARFF Data File	1 KB
9.arff	11/10/2015 3:31 PM	ARFF Data File	1 KB
10.arff	11/10/2015 3:20 PM	ARFF Data File	1 KB
11.arff	11/10/2015 3:20 PM	ARFF Data File	1 KB
12.arff	11/10/2015 3:27 PM	ARFF Data File	1 KB
13.arff	11/10/2015 3:27 PM	ARFF Data File	1 KB
14.arff	11/10/2015 3:26 PM	ARFF Data File	1 KB
15.arff	11/10/2015 3:26 PM	ARFF Data File	1 KB

DATA MINING PROJECT

Name	Date modified	Type	Size
hotel_1.arff	11/10/2015 4:56 PM	ARFF Data File	2 KB
hotel_2.arff	11/10/2015 3:38 PM	ARFF Data File	1 KB
hotel_3.arff	11/10/2015 3:42 PM	ARFF Data File	1 KB
hotel_4.arff	11/10/2015 3:44 PM	ARFF Data File	1 KB
hotel_5.arff	11/10/2015 3:46 PM	ARFF Data File	1 KB
hotel_6.arff	11/10/2015 3:59 PM	ARFF Data File	1 KB
hotel_7.arff	11/10/2015 4:04 PM	ARFF Data File	1 KB
hotel_8.arff	11/10/2015 4:18 PM	ARFF Data File	1 KB
hotel_9.arff	11/10/2015 4:54 PM	ARFF Data File	1 KB
hotel_10.arff	11/10/2015 4:51 PM	ARFF Data File	1 KB
hotel_11.arff	11/10/2015 4:54 PM	ARFF Data File	1 KB
hotel_12.arff	11/10/2015 5:00 PM	ARFF Data File	1 KB
hotel_13.arff	11/10/2015 5:04 PM	ARFF Data File	1 KB
hotel_14.arff	11/10/2015 5:06 PM	ARFF Data File	1 KB
hotel_15.arff	11/10/2015 5:09 PM	ARFF Data File	1 KB

We have shown it accordingly with the Zomato site for better reference.

Positive review arff →

Cash only

Stacy Landers
15 Reviews, 30 Followers
12 days ago via Zomato for iOS

RATED 4.5 I am a big fan of Frank Prisinzano. The ambience is very reminiscent of Frank and Lil' Frankie's. Everything was good, but Lil' Frankie's is where my heart lies.

Like 0 | Comment 0 | Share

Dallas Trends
133 Reviews, 106 Followers
3 months ago

RATED 4.5 Good food. Great atmosphere. I was with someone who never dined at a restaurant where they seat others at your table, and it was exciting. I had the privilege of dining for dinner. Somewhat limited seating. Friendly service. We decided to try this place after a local recommended it. I'd love to return.

trendydallas.com

Like 0 | Comment 0 | Share

EXPLORING!
Download the Zomato app and discover great restaurants around on-the-go!

AVAILABLE ON 

OR LET US TEXT YOU A [DOWNLOAD LINK](#)

ZOMATO SPOONBACK 

C:\Amo\directory\positive\hotel_1.arff - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
hotel_1_arff
1 @Relation Hotel
2
3 @Attribute Hotel_name STRING
4 @Attribute Review STRING
5 @Attribute Rating {1,2,3,4,5}
6
7 @Data
8 "Supper", "i am a big fan of frank prisinzano. the ambience is very remini
9 "Supper", "good food great atmosphere. i was with someone who never dined
10 "Supper", "perfect. i couldn't have been happier with my dinner at supper
11 "Supper", "mostly buon appetito. the food was really good, it's a little pricier
12 "Supper", "best italian in the village!. amazing pasta, rustic atmosphere.
13 "Supper", "mostly buon appetito. i've only been here twice, the first time
14 "Supper", "the best!. cool place, grrrrrrreat service and the food is ex
15 "Supper", "snug but fantastic!. the old-worldish, faded glory chic lends a
16
17
18
19
20

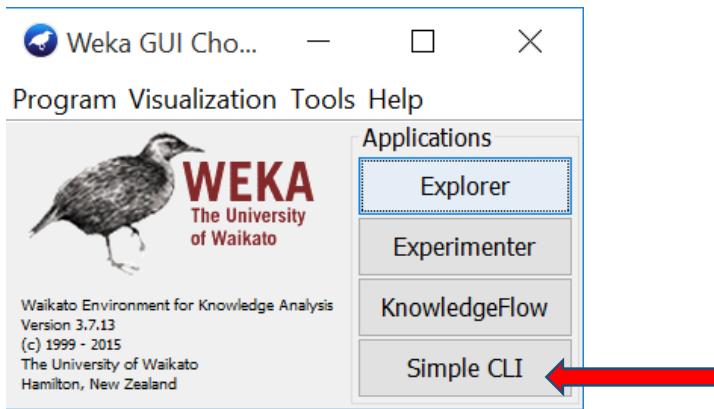
Negative Review arff →

DATA MINING PROJECT

The screenshot shows a Zomato review page for a restaurant named "Supper?". The review, posted by user "6968tim" on October 14, 2013, is labeled "NEGATIVE" and describes a disappointing meal experience. To the right of the review, a portion of an ARFF (Attribute-Relationship File Format) file is displayed. The ARFF file defines a relation "Hotel" with attributes "Hotel_name" (STRING), "Review" (STRING), and "Rating" (an integer from 1 to 5). It also defines a data set with two entries: "Supper?", which has a rating of 1, and "Overpriced...with terrible service", which has a rating of 2.

```
1 @Relation Hotel
2
3 @Attribute Hotel_name STRING
4 @Attribute Review STRING
5 @Attribute Rating {1,2,3,4,5}
6
7 @Data
8 "Supper?","Supper?. Had meal at Supper more than once
9 "Supper?","Overpriced...with terrible service ",2
10
11
12
13
14
15
16
17
18
19
20
21
22
```

After the arff files are created, using Weka CLI (command line interface) script was written to convert all the arff files into single data. Below is the snapshot of the CLI command used to convert the arff files into single data:



DATA MINING PROJECT



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.' or './'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
  java <classname> <args> [ > file]
  break
  kill
  capabilities <classname> <args>
  cls
  history
  exit
  help <command>
```

```
java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:\Amol\directory\hotel_nyc.arff
```

The command tells from where the data of positive reviews and negative reviews are to be referred from. We have also given the path where the single arff file indicating the path where the complete data should be stored.



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.' or './'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
  java <classname> <args> [ > file]
  break
  kill
  capabilities <classname> <args>
  cls
  history
  exit
  help <command>

> java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:\Amol\directory\hotel_nyc.arff
Finished redirecting output to 'C:\Amol\directory\hotel_nyc.arff'.
```

DATA MINING PROJECT

Once we merge all the positive and negative arff files to build one Master arff files, we see the data organised as below

```
@relation C_Users_krith/Desktop/DataMining_Project_fwddataminingdataset_directory

@attribute text string
@attribute @@class@ {positive,negative}

@data
'@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Supper\",\\\"last meal was\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Yerba Buena\",\\\"not up to\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Buenos Aires\",\\\"overpriced\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Gnocco\",\\\"Service was\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Black Iron Burger\",\\\"quiet\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Balthazar\",\\\"disappointing\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Momofuku Noodle Bar\",\\\"the\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Buddakan\",\\\"food sucks\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Per Se\",\\\"over priced\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Peter Luger Steak House\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Gramercy Tavern\",\\\"meals\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Root & Bone\",\\\"the place\n@Relation Hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"The Spotted Pig\",\\\"we\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"The Halal Guys\",\\\"it's\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Burger Joint\",\\\"terrible\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Momofuku Ssäm Bar\",\\\"very\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Max Brenner\",\\\"disappointing\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Max Brenner\",\\\"very bland\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Joe's Shanghai\",\\\"meals\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Nobu\",\\\"I'm so disappointed\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Babbo\",\\\"it took 18 months\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Les Halles Park Avenue\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Zum Schneider\",\\\"terrible\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Del Posto\",\\\"not worth it\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Cookshop\",\\\"skip it.\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Chinatown Ice Cream Factory\",\\\"ice cream\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Havana Central\",\\\"I ordered\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"The Cupping Room Café\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Mercer Kitchen\",\\\"I am\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Hard Rock Cafe\",\\\"Bad food\n@Relation hotel\r\n\r\n@Attribute Hotel_name\tSTRING\r\n@Attribute Review\t\tSTRING\r\n@Attribute Rating\tt{1,2,3,4,5}\r\n\r\n@Data\r\n\"Normal's\",\\\"sucks. food is terrible\n'
```

The Master arff file contains the metadata of the data in the file. The Meta data says the datatype of the attributes we have in the file along with the classes present in the file. The file lists the two classes namely positive and negative from the data set.

It can be noted that the master data details are contained inside the **data** (@data) attribute. The same data can be seen for each review of a hotel as shown in the below snapshot:

DATA MINING PROJECT

C:\Users\krith\Desktop\DataMining_Project\fwddataminingdataset\directory\data.aff - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

change.log weather.csv weather.aff Test1.aff Test2.aff data.aff

```
107 ig\t\t(1,2,3,4,5)\r\n\r\n\r\nData\r\n"The Halal Guys",\"great food. quick service.\",\r\nwell the rating speaks for it self\",4\r\n"The Halal Guy\r\nBurger Joint",\"i love this place! not just the delicious burgers\",5\r\n"Burger Joint",\"the best thing about the burger j\r\nMomoFuku Ssäm Bar",\"a really great job!\",3\r\n"MomoFuku Ssäm Bar",\"i loved at atmosphere at this place.\",4\r\n"Momofuku\r\nShake Shack",\"the best burger i tried in new york\",5\r\n"Shake Shack",\"great burger and shakes, can get very busy\",4\r\nMax Brenner",\"nice place, great staff & good food. i liked their mac & cheese, it's very hearty & heavy\",4\r\n"Max Brenne\r\n's Shanghai",\"i recommend joe's shanghai but don't expect your own table, any sort of privacy, or a friendly server\",5\r\nNobu",\"black cod with miso. delicious!!\",4\r\n"Nobu",\"great food and great service and i would definitely come back!\",5\r\nBabbo",\"memorable experience\",4\r\n"Babbo",\" food was good but not great and service varied from good to rude!\",3\r\nLes Halles Park Avenue",\"the best service i have seen! the atmosfer was perfect!\",5\r\n"Les Halles Park Avenue",\"amazing\r\nZum Schneider",\"good food, large portions. the size of the schnitzel plate here is insane.\",4\r\n"Zum Schneider",\"amazing\r\nDel Posto",\"have been to this place more than 5 times . it never fails to surprise me.\",5\r\n"Del Posto",\"party of four.\r\nRoberta's",\"great pizza, really friendly staff\",5\r\n"Roberta's",\"one of my newyorker friends took me there. it's ver\r\nCookshop",\"wonderful place to stop in for breakfast before taking a stroll on the high line\",4\r\n"Cookshop",\"food was g\r\nChinatown Ice Cream Factory",\"it was creamy, flavourful, and not too sweet!\",3\r\n"Chinatown Ice Cream Factory",\"their\r\nNathan's Famous",\"impeccable. this hot dog tops the one from crif dogs and is now my #1\",5\r\n"Nathan's Famous",\"so i f\r\nHavana Central",\"first time here and iffy to try the cuisine but it was deliciou\",4\r\n"Havana Central",\"lovely place. l\r\nThe Cupping Room Café",\"never get bored of getting in cafes to gave amazing desserts\",5\r\n"The Cupping Room Café",\"the\r\nMercer Kitchen",\"good address for dinner with nice atmosphere.\",5\r\n"Mercer Kitchen",\"wow, the mercer burger was insane.\r\nHard Rock Café",\"one of my life goals is to visit every single hard rock cafe in the entire world\",5\r\n"Hard Rock Café",\"Norma's",\"loved it so much that i actually went back another time\",5\r\n"Norma's",\"this was highly recommended as one\r\nMatcha Cafe Wabi",\"japanese owned and operated everything we had was delightful.\",4\r\n"Matcha Cafe Wabi",\"positive\r\nCarmine's",\"nice family style restaurant in the upper west , a must visit for me every time i am in manhattan\",5\r\n"Carmine's",\"fusilli with octopus and bone marrow. this dish is perfection\r\nMarta",\"salted caramel semifreddo. delicious\",4\r\n"Marta",\"not sure what all the fuss is about. g\r\nBrooklyn Diner",\"if i lived in nyc it would surely be a go-to place for me.\",4\r\n"Brooklyn Diner",\"i would\r\nLes Halles",\"stopped in for brunch and everything was great.\",4\r\n"Les Halles",\"probably the best dinner i had so far in\r\nSmith & Wollensky",\"stopped in for a drink and dessert late in the evening\",5\r\n"Smith & Wollensky",\"the meat lovers pi\r\nGlass House Tavern",\"oh at best; stay on top of pace.\",3\r\n"Glass House Tavern",\"awesome corn/cous-cous risotto, great\r\nBlue Bottle Coffee",\"the coffee is clearly great and the staff was sweet enough\",5\r\n"Blue Bottle Coffee",\"great 3rd wa\r\nRadegast Hall & Beer Garden",\"despite being from san diego - arguably the craft beer capital of america\",3\r\n"Radegast Ha\r\nTortilla Flats",\"awesome. great atmosphere and people, great food and a lot of it.\",5\r\n"Tortilla Flats",\"nice food and\r\nTuome",\"this was perfectly charismatic\",5\r\n"Tuome",\"the flavors your taste buds will soon experience are anything but.\r\nThe Fat Radish",\"delicious!\",5\r\n"The Fat Radish",\"i would recommend \",5\r\n"The Fat Radish",\"great place for laten\r\nEsperanto",\"one of my top favorites\",5\r\n"Esperanto",\"probably the best french toast i've ever had\",4\r\n"Esperanto'\r\nKatz's Delicatessen",\"this joint has a rich history and some awesome grub.\",3\r\n"Katz's Delicatessen",\"this is an ol\r\nPoco",\"mg this is the cutest little spot in alphabet city\",3\r\n"Poco",\"brunch is awesome!. they have unlimited mimosas\r\nMinca",\"pretty good ramen. but a five or four rating? meh. more like 3 to me\".3\r\n"Minca",\"great place for ramen. out o
```

< >

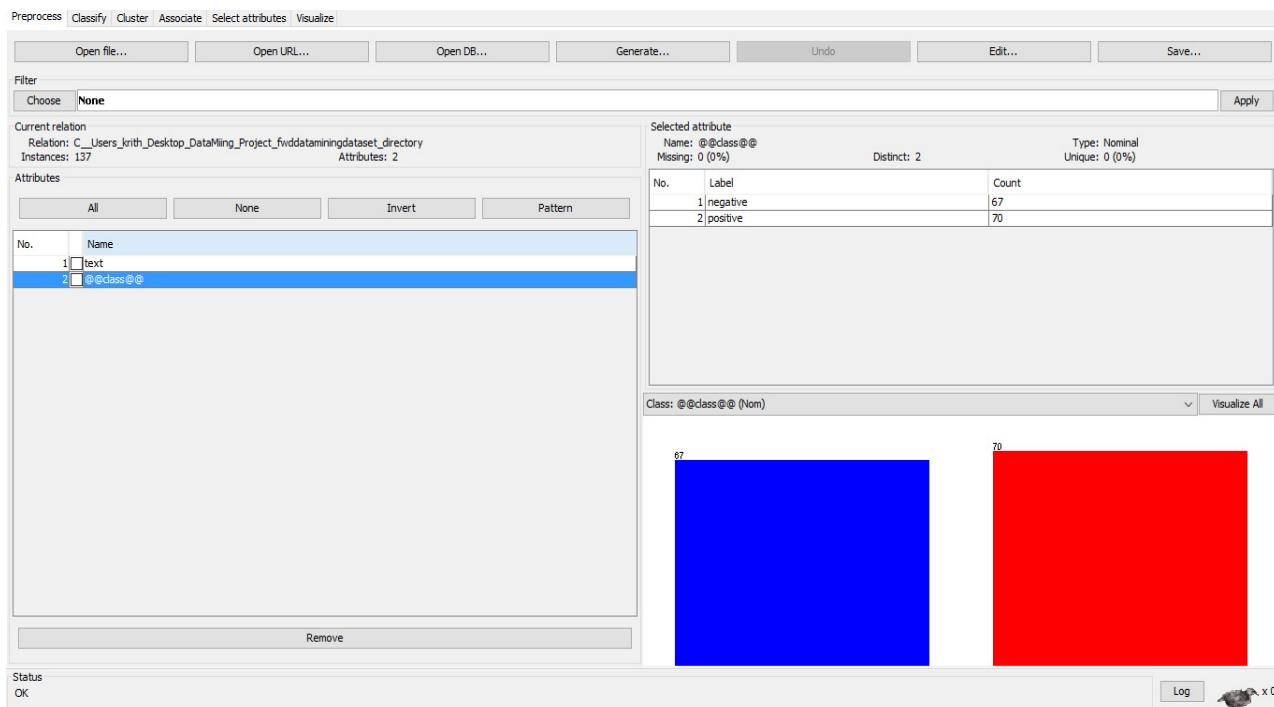
Normal text file

length:61580 lines:144 Ln:142 Col:570 Sel:0|0 UNIX UTF-8 IN5

Search the web and Windows 12:03 PM 11/11/2015

PRE-PROCESSING

The first step in Weka is Pre-processing. The data once loaded generates a histogram based on the class (positive and negative) as below.



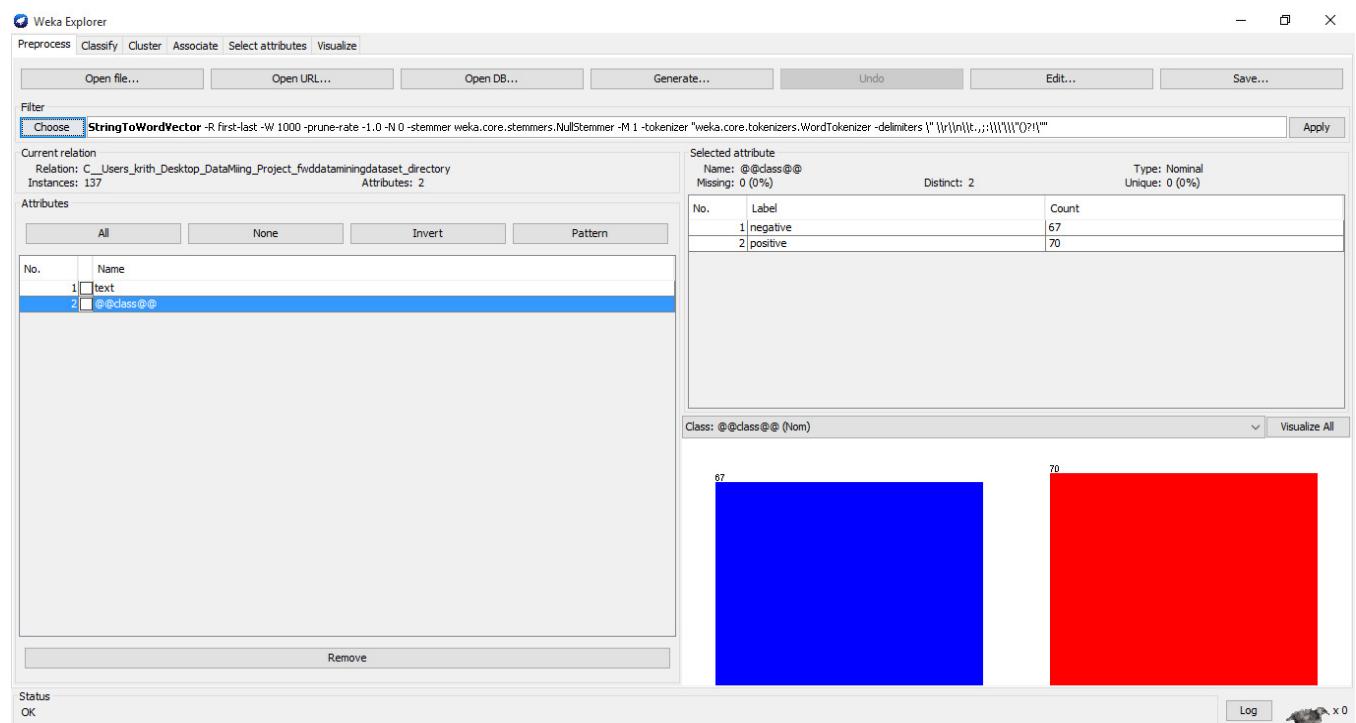
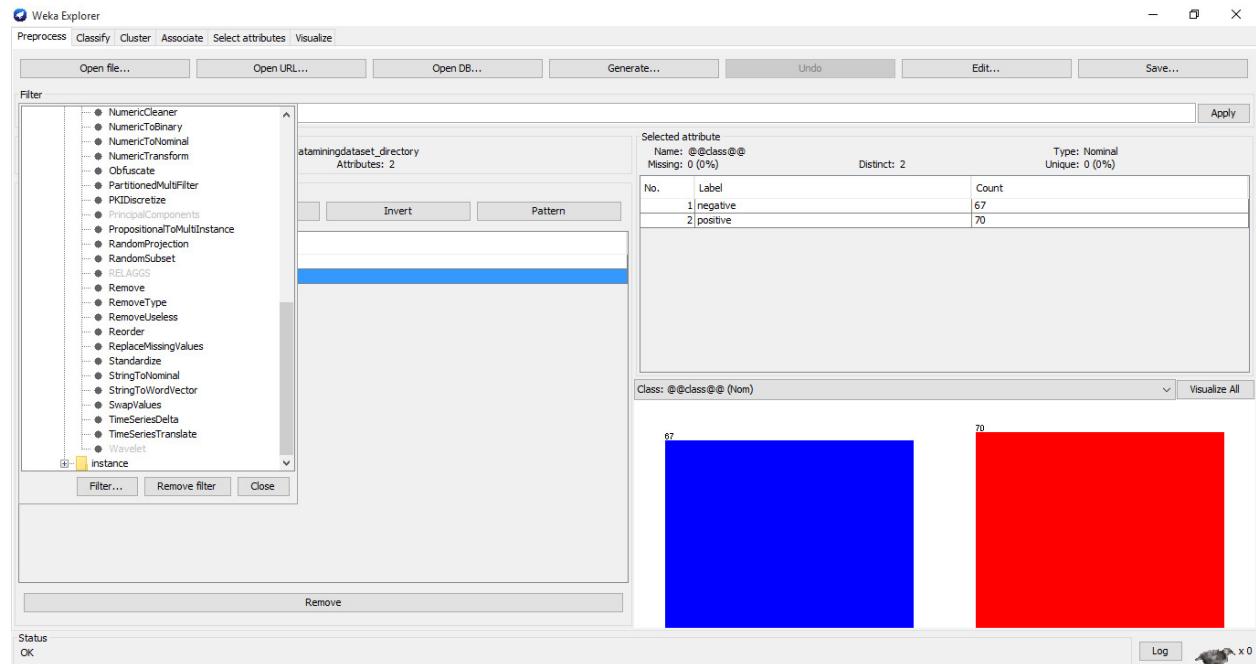
The count on each graph shows the number of instances we have in each class. The reviews are split into each word and the words are compared and classified to check if each word falls either in positive or negative class. We can choose the filters on which data pre-processing can be done. Weka gives us three options to filter the data on. Firstly, data can be filtered on all available filters or multiple filters can be selected. If not, Weka lets us filter data on supervised or unsupervised filters.

Each supervised and unsupervised filters have different built-in functions to filter the data on. For instance, supervised filter has functions called **Attribute classification** based on which we classify the attribute for each class.

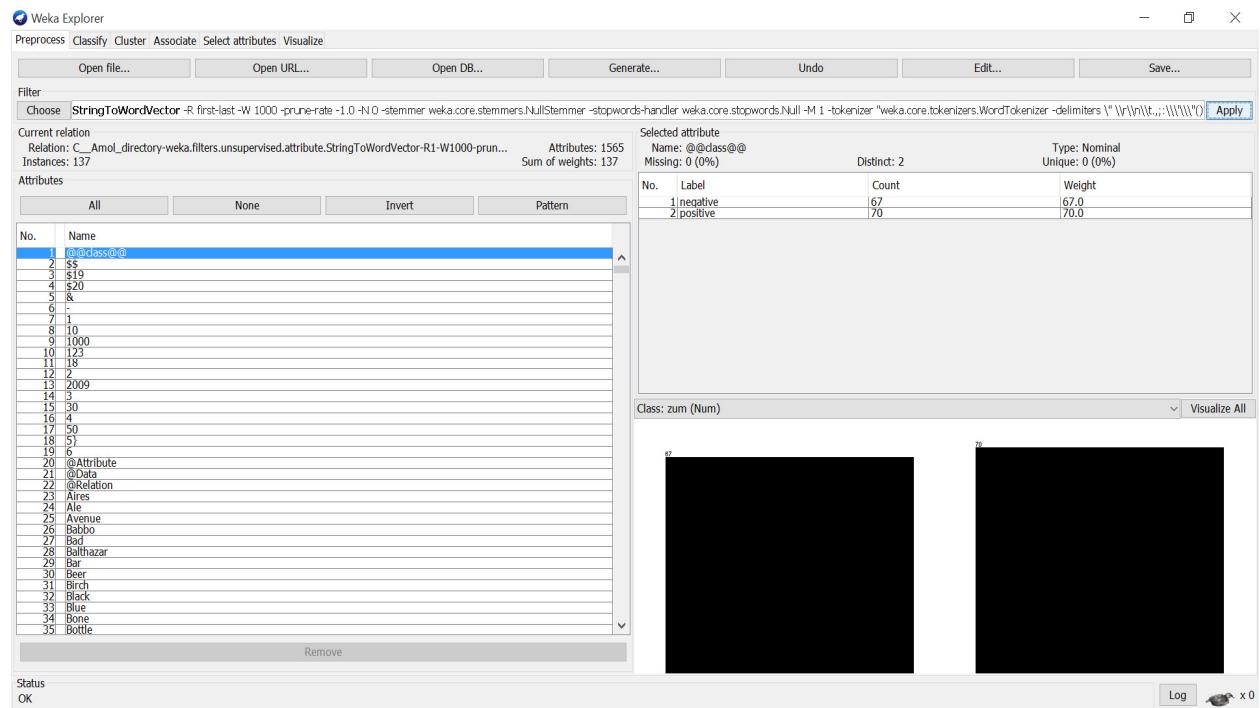
Here, we have used unsupervised **StringToWordVector** filter to split each word in the reviews to form a vector of individual words. **StringToWordVector** is used to Converts String attributes into a set of attributes representing word occurrence information from the text contained in the strings.

Following are the steps to select and apply the unsupervised filter **StringToWordVector**.

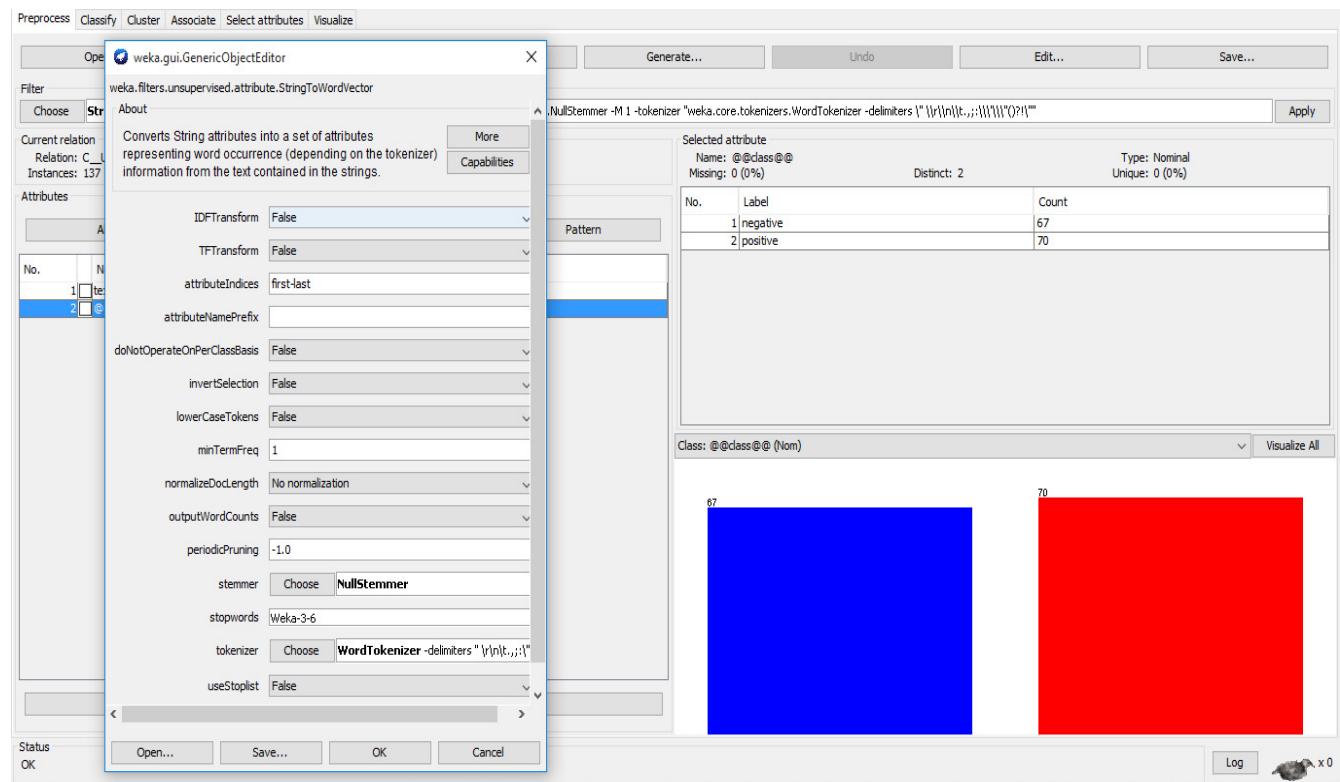
DATA MINING PROJECT



DATA MINING PROJECT



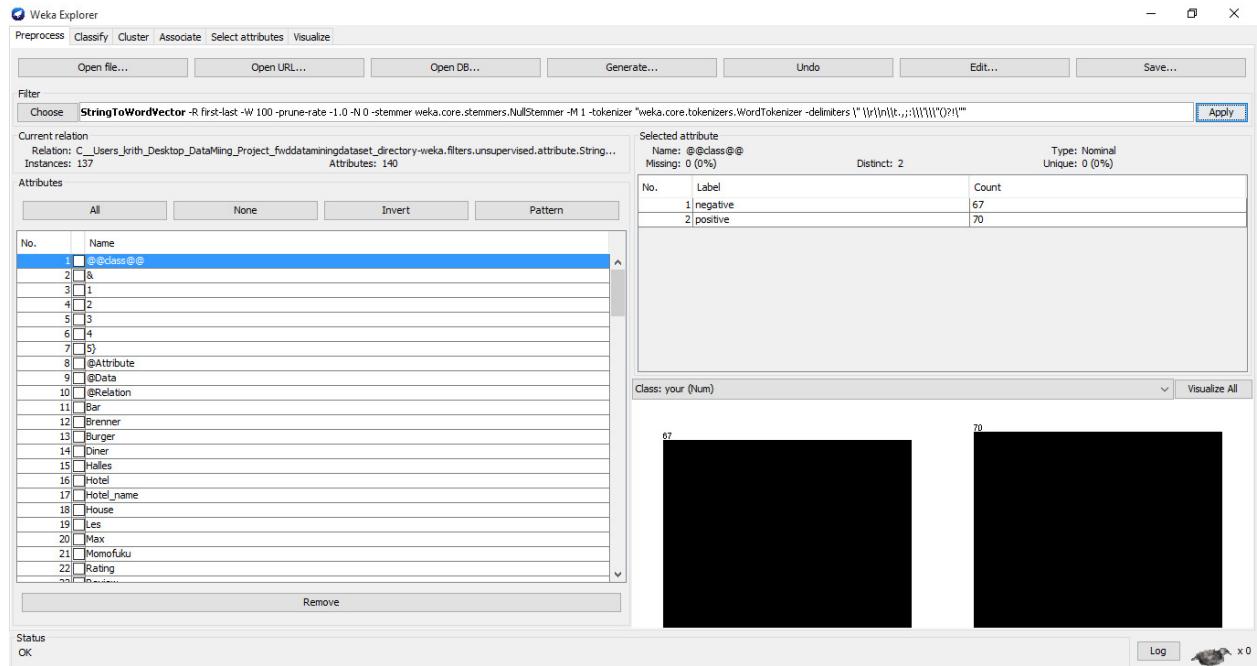
We can modify the filter options by clicking the field that shows the selected filter name to open the edit menu of the filter.



DATA MINING PROJECT

The first few options provided in edit option are briefed below:

1. IDFTransform → Inverse document frequency gives the count of the words in the documents.
2. TFT → Term frequency transform gives the count of each term in all the documents.
3. Minimum term frequency → when set, gives those terms which has atleast the number of times set for the MTF.
4. Stemmer → Used to combine words together like **Works**, **Working** and so on. It gives us option to choose between four different stemmers.
5. StopwordsHandler → We can provide the word list to filter the document with those words. By default Weka has English language stop words. We can also select different language stop words.
6. WordToKeep → If set lets us to retain only those number of words



After the filter is applied, we can see the number of attributes and number of instances depending on the filer options you set. We can see the change in the count of attributes before and after applying the filers.

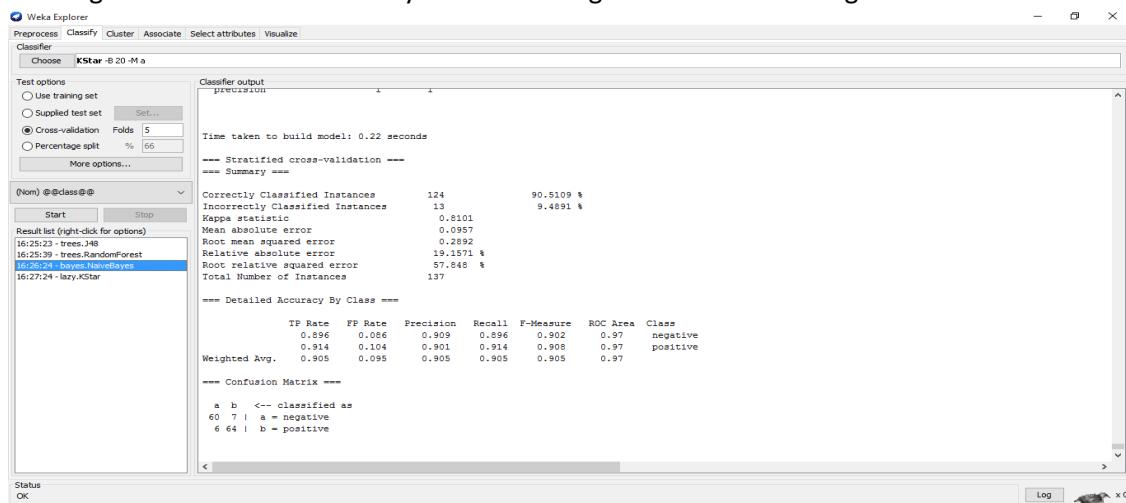
CLASSIFICATION

After pre-processing the data, data needs to be classified using the data mining algorithms to check if the words that are split fall in the respective class. We have used the below four algorithms to classify the texts. We have plotted their respective ROC curves.

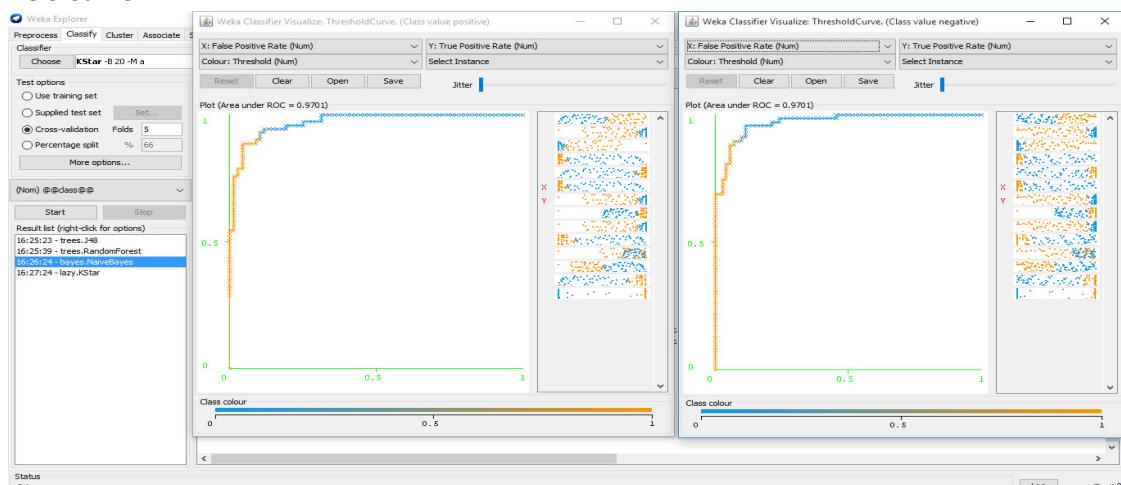
ROC (Receiver Operating Characteristics) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. This is plotted against “False Positive Rate” (X-axis) and “True Positive Rate” (Y-axis). True Positive Rate is also called as the **Recall**.

1. Naïve Bayes → In machine learning, **naïve Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. Naïve bayes are highly scalable requiring a number of parameters linear in the number of variables. This comes under bayes inside the classifiers.

The below snapshot shows the naïve bayes algorithm applied on the input data set. We can see the confusion matrix at the end that shows how many data were correctly classified as positive and negative. Shows the accuracy and the error generated with the algorithm on the data set.



ROC Curve -



DATA MINING PROJECT

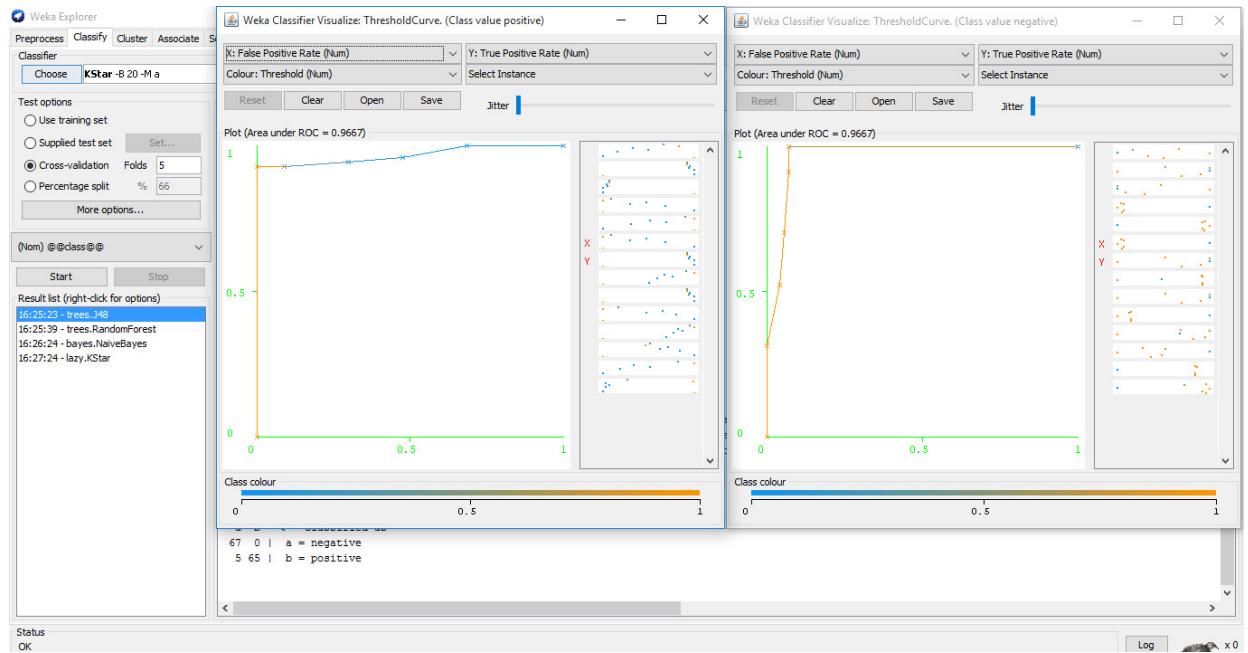
2. J48 → This is inside the trees within classifier. The J48 is an open source java implementation inside C4.5 algorithm. The C4.5 algorithm creates a decision tree based on a set of labeled data. For instance here we have two classes positive and negative. The below snapshot shows the amount of data correctly classified. Total of 132 instances were correctly classified and 5 were incorrectly classified. We can also note the difference in the confusion matrix where there are 2 misclassified data in negative and 3 misclassified data in positive.

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose J48 -C 0.25 -M 2
Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds 10 %
 Percentage split % 66
More options...
(Nom) @class@ @
Start Stop
Result list (right-click for options)
17:54:38 - bayes.NaiveBayes
17:58:39 - trees.J48
Number of Leaves : 4
Size of the tree : 7
Time taken to build model: 0.33 seconds
--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances 132 94.3504 %
Incorrectly Classified Instances 5 3.6496 %
Kappa statistic 0.927
Mean absolute error 0.0472
Root mean square error 0.1234
Relative absolute error 9.4515 %
Root relative square error 36.4669 %
Coverage error 0.0000 %
Mean absolute error (0.95 level) 0.0453
Mean rel. region size (0.95 level) 54.0146 %
Total Number of Instances 137
--- Detailed Accuracy By Class ---
           TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
negative 0.970 0.043 0.950 0.970 0.943 0.927 0.979 0.979 negative
positive 0.964 0.036 0.964 0.964 0.964 0.927 0.979 0.973 positive
Weighted Avg. 0.964 0.036 0.964 0.964 0.964 0.927 0.979 0.973
--- Confusion Matrix ---
a b <-- classified as
65 2 | a = negative
3 67 | b = positive

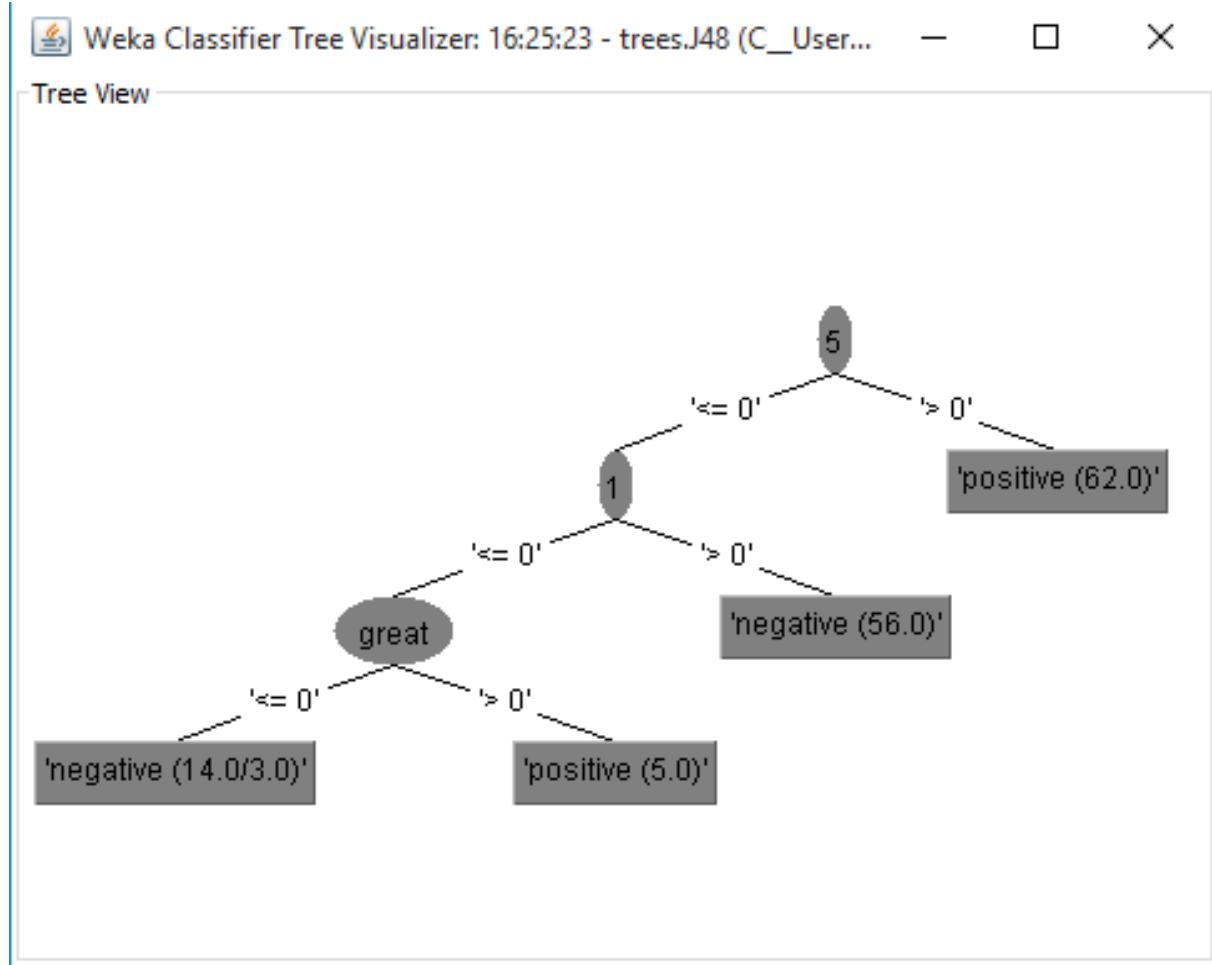
```

ROC Curve:



DATA MINING PROJECT

The J48 builds a decision tree based on the classes and the instances.



DATA MINING PROJECT

3. Random Forest → This is again available inside the trees within classifier. This is an ensemble model where we create a decision tree by modifying the input features. In this approach a subset of input features is chosen to form each training set. The subset can be chosen randomly or based on the recommendation of domain experts.

The below snapshot shows the amount of incorrectly classified to correctly classified. We can see that Random forests have incorrectly classified 14 instances out of the total 137 instances where 2 are from negative class and 12 are from positive class. The confusion matrix can be found below with the difference.

```

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier Choose RandomForest -I 100 -K 0 -S 1 -num-slots 1
Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
More options...
(Nom) @@@class@@@ Start Stop
Result list (right-click for options)
17:54:38 - bayes.NaiveBayes
17:58:39 - trees.J48
18:02:50 - trees.RandomForest
18:02:50 - trees.RandomForest

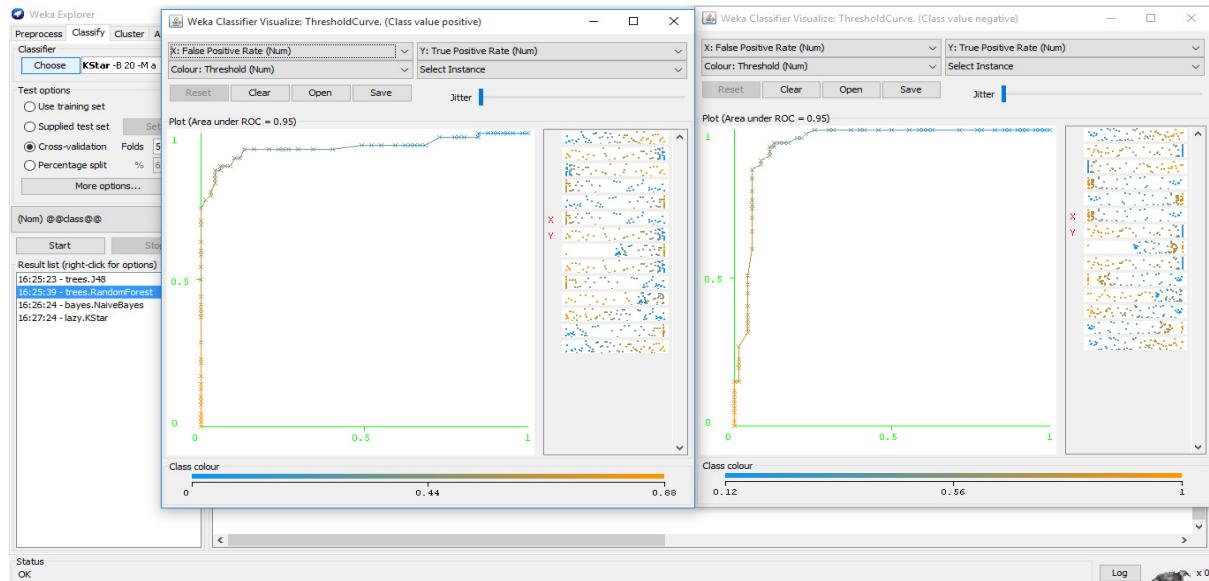
Time taken to build model: 0.77 seconds
Stratified cross-validation ===
Summary ===
Closely Classified Instances 128 89.74%
Incorrectly Classified Instances 14 10.21%
Mean absolute error 0.3985
Root mean squared error 0.3981
Relative absolute error 61.1433 %
Root relative squared error 69.5719 %
Coverage of cases (0.95 level) 100 %
Mean deviance (0.95 level) 36.7553 %
Total Number of Instances 137

Detailed Accuracy By Class ===
    TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Auc Class
    0.970 0.171 0.944 0.970 0.903 0.805 0.959 0.940 negative
    0.829 0.030 0.967 0.829 0.892 0.805 0.959 0.968 positive
Weighted Avg. 0.898 0.099 0.907 0.898 0.897 0.805 0.959 0.954

Confusion Matrix ===
    a b <-- classified as
    65 21 | a = negative
    12 55 | b = positive

```

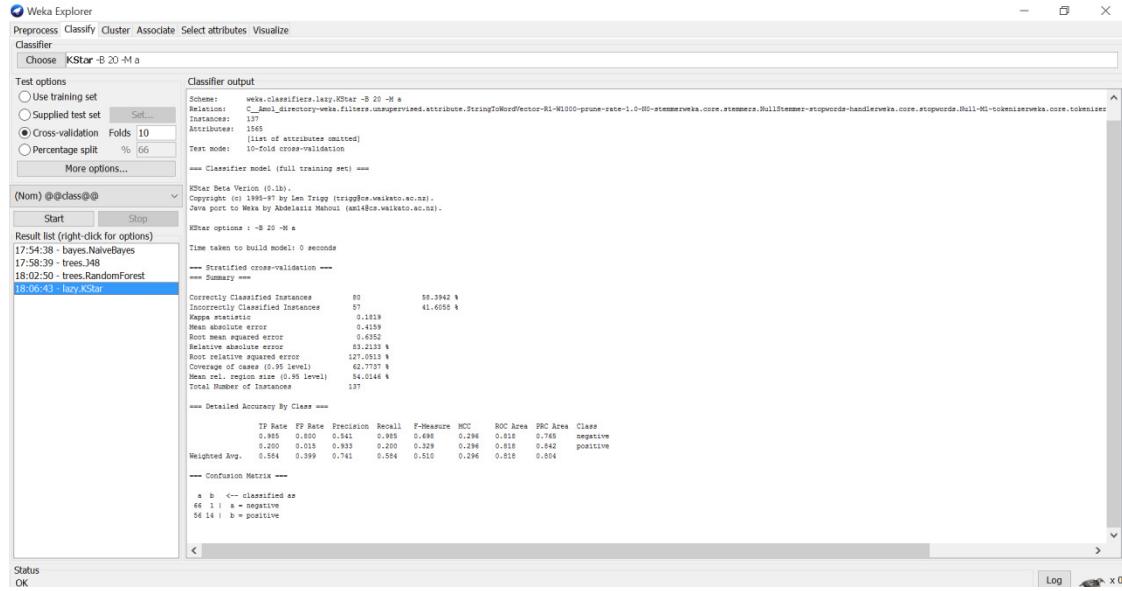
ROC Curve:



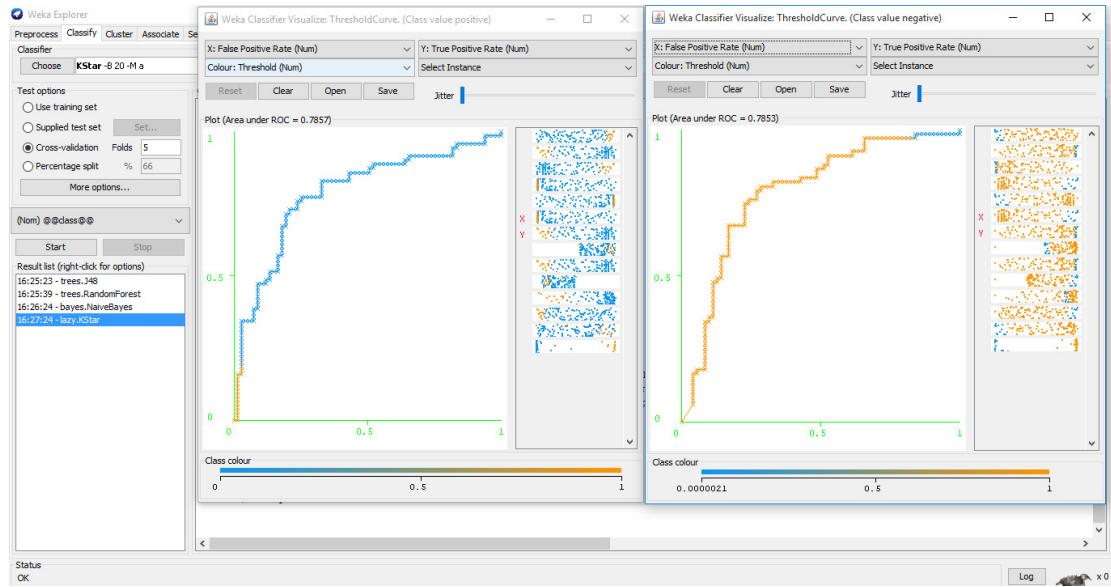
DATA MINING PROJECT

4. KStar → Kstar is a lazy classifier. **Lazy learning** is a learning method in which generalization beyond the training data is delayed until a query is made to the system, as opposed to in eager learning, where the system tries to generalize the training data before receiving queries.

KStar uses entropic distance measure to evaluate the data set. Kstar is an instance based classifier that is the class of a test instance is based upon the class of those training instances similar to it.



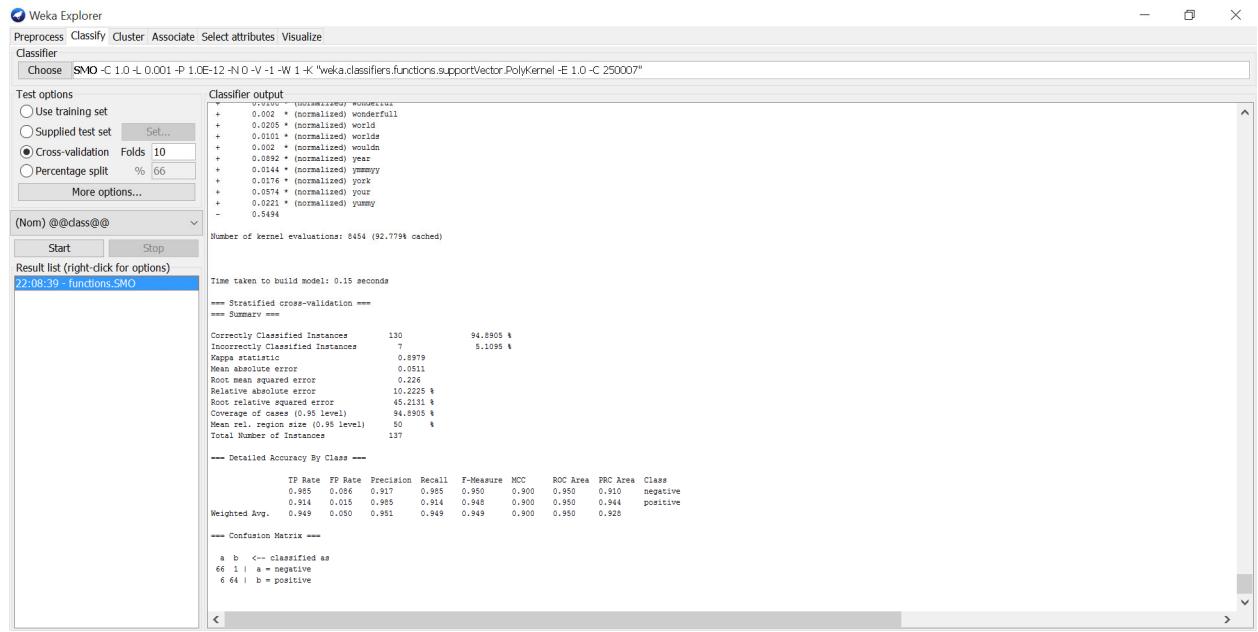
ROC Curve:



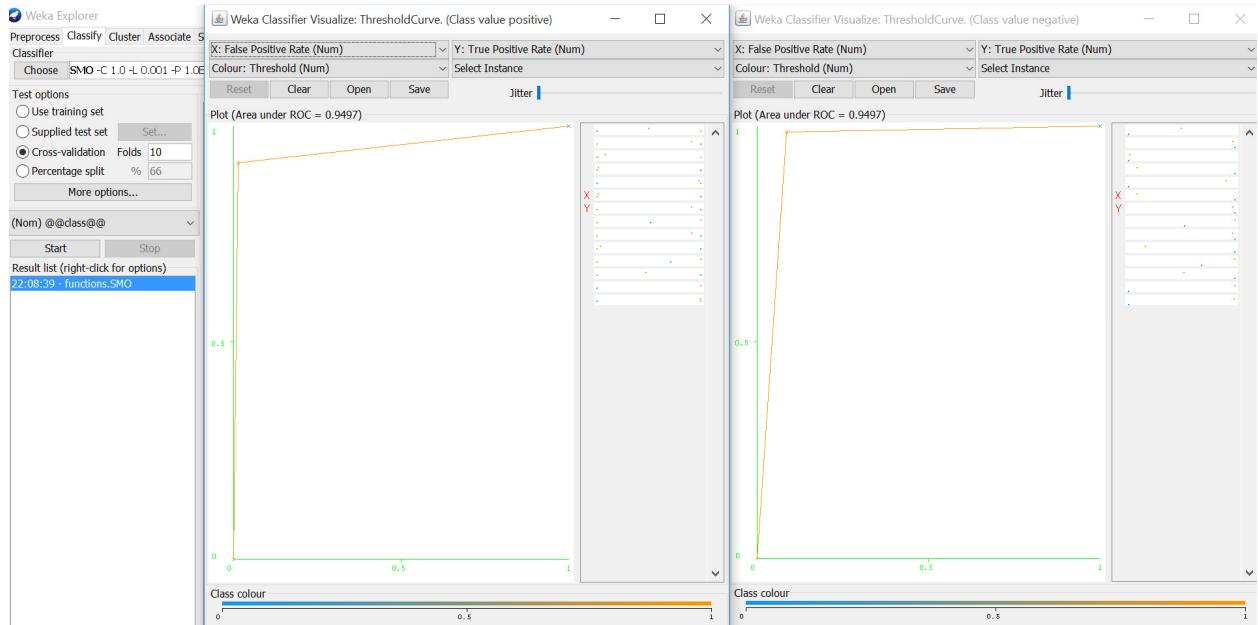
DATA MINING PROJECT

5) Support Vector Machine →

SVM also works very well with high-dimensional data and avoids the curse of dimensionality problem. Another unique aspect of this approach is that it represents the decision boundary using a subset of the training examples, known as the support vectors.



ROC Curve:



COMPARISON OF CLASSIFICATION MODELS ➔

Sr.No.	Naïve Bayes	J 48	Random Forest	K - Star	Support Vector Machine
Area under ROC	0.9701	0.966	0.95	0.78	0.94
Correctly classified Instance (%)	90.5	96.35	89.7	58.3	94.8
Time taken to build model (secs)	0.22	0.33	0.77	0	0.15

Conclusion ➔

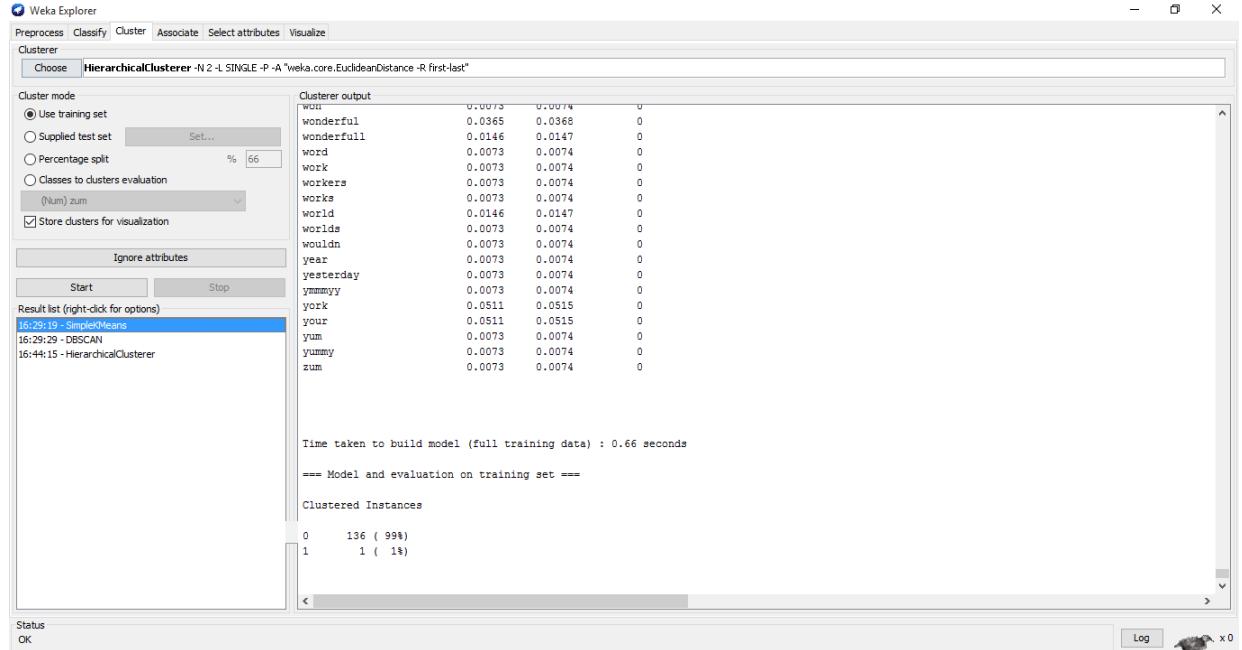
J48 tree and **Support Vector Machine** have done the best classification for our textual data set.

DATA MINING PROJECT

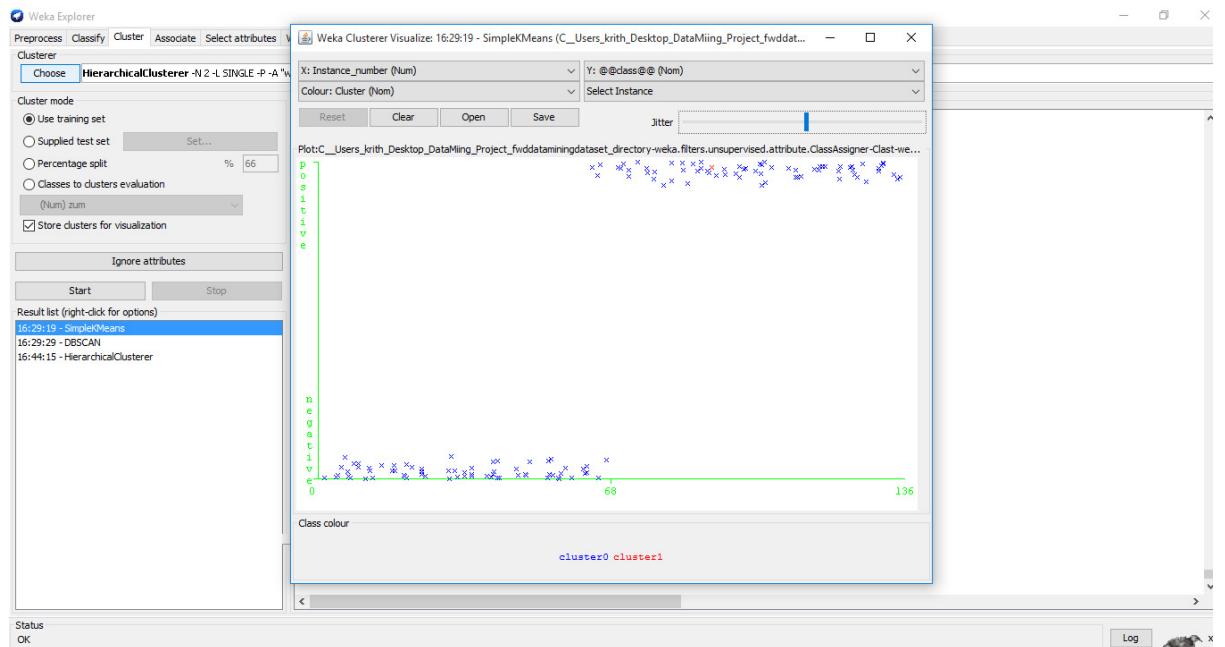
CLUSTERING

1) SimpleKMeans →

This is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids.



Simple Kmeans visualization of cluster -

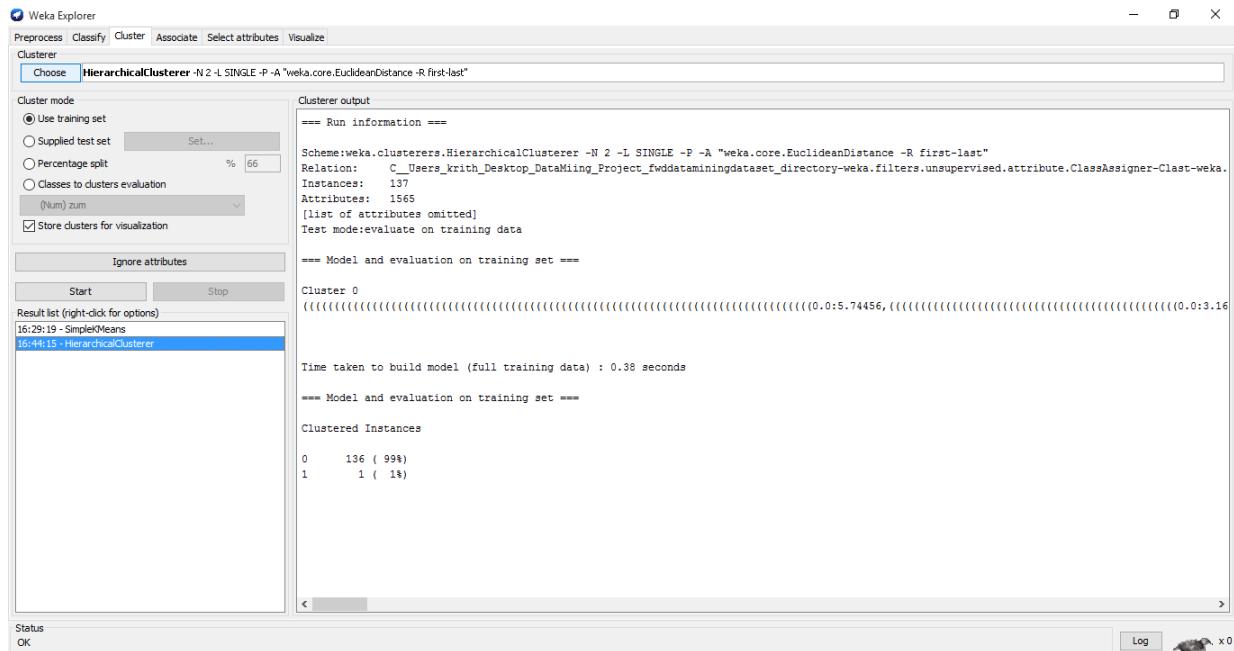


Here we see one instance has been wrongly clustered.

DATA MINING PROJECT

2) Hierarchical Cluster →

This clustering approach refers to a collection of closely related clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all encompassing cluster remains.



Dendrogram-

