KENNESAW STATE
UNIVERSITY

Data WH and Data Mining

SENTIMENT ANALYSIS

Grishma Saparia

Kashish Patel

1. PROBLEM DEFINITION

In this project, we are going to implement Sentiment Analysis also known as opinion mining. Sentiment analysis is the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

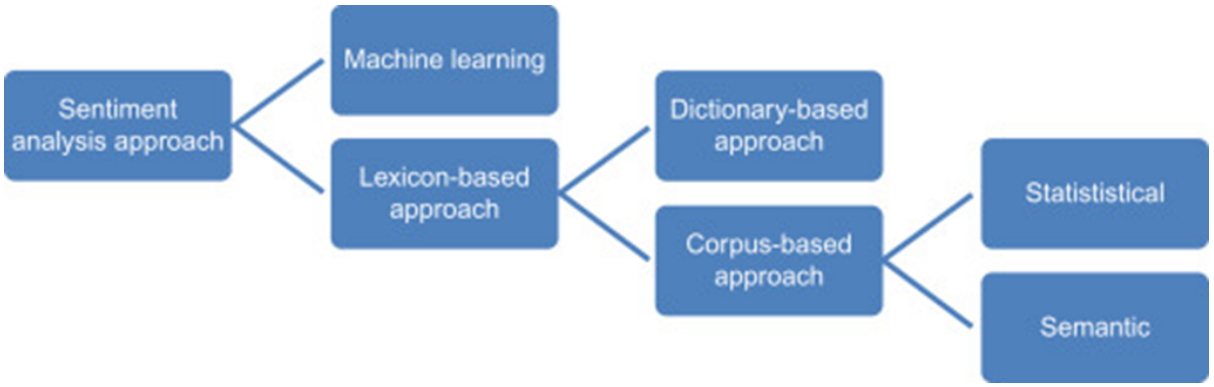
2. DATA SET USED

We have used publicly available data from the website called Booking.com. The entire content of the file is already freely accessible to everyone. This dataset includes 1493 premium hotels from across Europe and 515,000 customer evaluations and ratings.

3. EXISTING ALGORITHM

Lexicon based approach.

They use only general domain. They cannot handle different dialects and informal or slang words. They don't contain acronyms and shorthand.



4. ALGORITHM USED IN OUR MODEL

Preprocessing of Dataset :
As part of the model development, we labelled the data as ‘positive review’ and ‘negative review’. We used the natural language toolkit to capture the adjective, verb, noun and adverb from the data set and removed those from prediction consideration. We also did a few other things to do the data cleanings like removing white spaces, changing the cases of texts to lower, removing words having digits from the dataset. After that, we added a sentiment analysis column to the program and created vector columns to hold the values. Also our focus is to capture the most frequently used words in the review dataset and visualise those based on the weightage.

Random Forest Classifier :
Multiple decision trees are used instead of a single tree for prediction. The class with the most votes becomes the class predicted by our model, which is generated by the individual trees.

5. TEST RESULTS

Test Result Screenshots from the program :

	review	is_bad_review
0	I am so angry that i made this post available...	1
1	No Negative No real complaints the hotel was g...	0
2	Rooms are nice but for elderly a bit difficul...	0
3	My room was dirty and I was afraid to walk ba...	1
4	You When I booked with your company on line y...	0

Figure 5.1(a) Bad Review

	review	is_bad_review	review_clean	neg	neu	pos	compound	nb_chars	nb_words	doc2vec_vector_0	...	word_yoghurt	word_yogurt
333976	Nothing really Great location staff rooms br...	0	nothing really great location staff room break...	0.213	0.485	0.303	0.1253	76	14	-0.006474	...	0.0	0.0
464140	The beds are comfy the location is superb the...	0	bed comfy location superb staff helpful	0.000	0.605	0.395	0.8016	69	15	-0.052489	...	0.0	0.0
293833	Well kept hotel Excellent location and staff ...	0	well kept hotel excellent location staff break...	0.000	0.370	0.630	0.8979	72	13	-0.011393	...	0.0	0.0
424002	It is perfect if you are ar a conference at R...	0	perfect ar conference rai walk nice garden hes...	0.033	0.689	0.279	0.9403	255	58	0.090006	...	0.0	0.0

Figure 5.1(b)

	review	pos
43101	A perfect location comfortable great value	0.931
396140	Noting Best best best	0.926
265458	Friendly Smiling Efficient Helpful staff	0.922
406627	Excellent service super comfortable	0.916
127959	Great location excellent value	0.911
177643	Friendly welcome Comfortable room	0.905
169970	Very comfortable Good value	0.904
400252	Clean comfortable good facilities	0.899
270616	Good friendly helpful staff	0.899
152433	Very clean good value	0.898

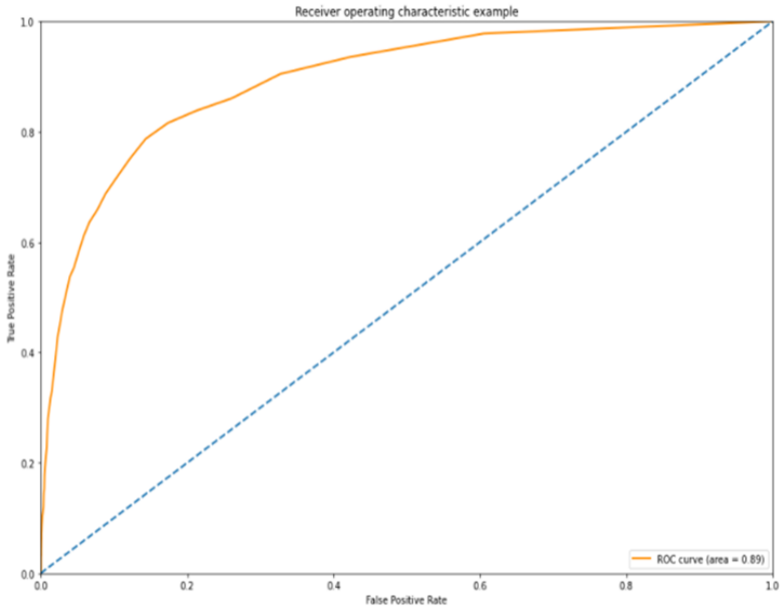
Figure 5.1 (c) Highest Positive Sentiment Review

	review	neg
179987	No A C	1.000
321288	Nothing Classy calm perfect	0.880
29666	A bit noisy No	0.796
257753	Nothing Friendly helpful welcoming staff	0.786
271741	bad wifi very bad	0.778
320301	Terrible and disgusting wellnes	0.765
458697	Nothing Great friendly staff	0.748
50226	Nothing Lovely friendly hotel	0.740
310337	Poor decor uncomfortable bed	0.740
235423	dirty and tired breakfast not good	0.732

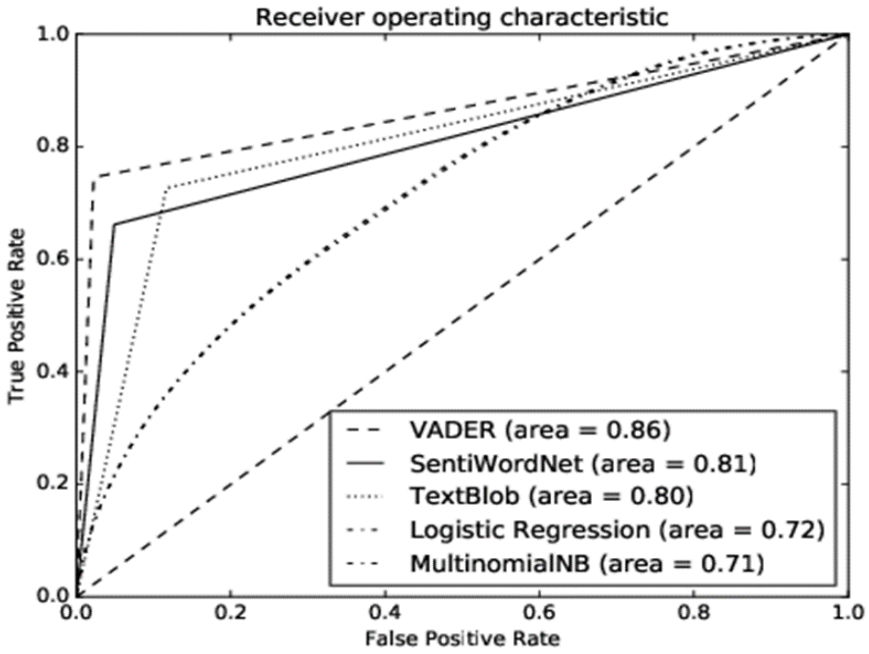
Figure 5.1 (d) Lowest Negative Sentiment Review

6. RESULTS

We compare our results with the lexicon based approach. Plot distribution between positive and negative review, ROC curve to evaluate our results.



ROC CURVE for Random Forest Tree Model



ROC CURVE for Lexicon Based Model.

So far we made this project work more readable and understandable, we also captured the highest positive sentiment review, lowest negative sentiment review to make the analysis on the dataset. We have compared ROC curves for both Lexicon based model and Random Forest model where we found that our model worked best with 0.89 ROC curve area compared to Lexicon based model with 0.86 ROC curve area.