

NAME: G.SAI SUSHANTH REDDY

REG.NO: 22MCB0005

SUBJECT: SOCIAL NETWORK ANALYTICS (MCSE618P)

ASSESSMENT-4

SENTIMENT ANALYSIS

(REPORT)

INTRODUCTION

Sentiment analysis, commonly referred to as opinion mining, is a natural language processing (NLP) technique that includes searching through text to extract introspective information and identify the sentiment or emotional tone portrayed. It aims to identify and categorise the views, opinions, and emotions that individuals have towards a particular subject, service, or event.

The primary goal of sentiment analysis is to determine whether a given text expresses a positive, negative, or neutral sentiment. To figure out the underlying sentiment or polarity, it involves evaluating the textual material, including reviews, social media posts, customer reviews, survey replies, and news items.

Sentiment analysis can be performed using various approaches, including rule-based methods, machine learning techniques and lexicon-based methods.

These approaches involve tasks such as text pre-processing, feature extraction, sentiment classification, and sentiment polarity determination.

METHODOLOGY & RESULT

I have taken the Hotel reviews dataset to perform sentiment analysis.

Review column contains both positive and negative reviews.

is_bad_review column contains 0(positive) and 1(negative) values



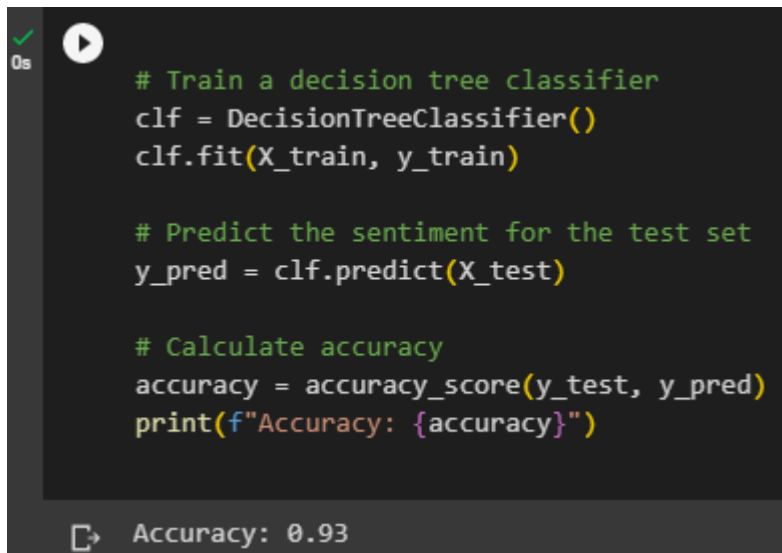
The screenshot shows a Jupyter Notebook interface with a table of hotel reviews. The table has two columns: 'review' and 'is_bad_review'. The 'is_bad_review' column contains binary values (0 for positive, 1 for negative). The table is displayed in a dark-themed environment.

	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
5	Backyard of the hotel is total mess shouldn t...	0
6	Cleaner did not change our sheet and duvet ev...	1
7	Apart from the price for the brekfast Everyth...	0
8	Even though the pictures show very clean room...	0
9	The aircondition makes so much noise and its ...	0

Decision Tree Classifier:

Decision Tree is a tree-based supervised learning algorithm that partitions the data based on a series of feature-value tests to make predictions. In sentiment analysis, Decision Trees can be used to classify text documents or sentences by learning decision rules based on various features extracted from the text.

In a Decision Tree classifier for sentiment analysis, the algorithm constructs a tree-like model where each internal node represents a feature or attribute test, and each leaf node represents a sentiment label (positive, negative, or neutral). The decision rules are learned from a labelled training dataset, where the features and their corresponding sentiment labels are known.



```
0s # Train a decision tree classifier
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)

# Predict the sentiment for the test set
y_pred = clf.predict(X_test)

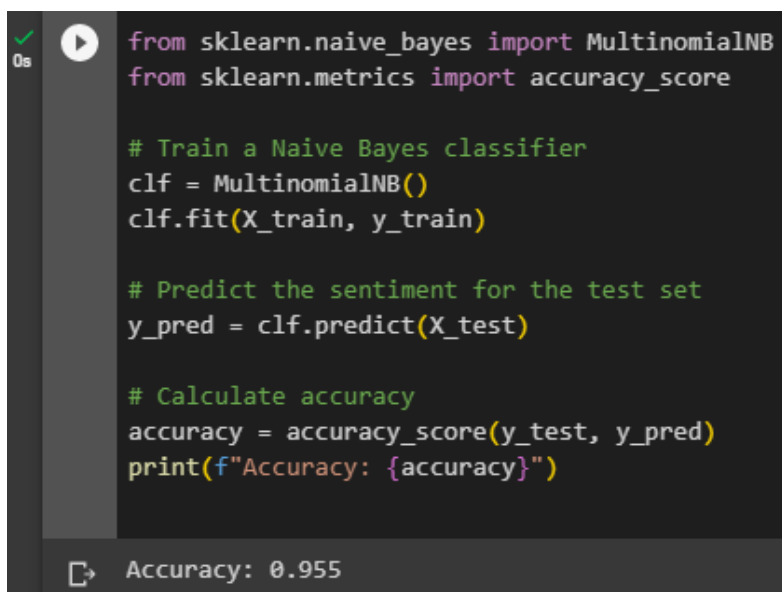
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

Accuracy: 0.93

Naïve Bayes Classifier:

In sentiment analysis, Naive Bayes can be used to classify text documents or sentences into positive, negative, or neutral sentiments based on the occurrence of specific words or features.

It builds a probabilistic model using a training dataset where the sentiment labels are known. During classification, it predicts the most likely sentiment label for a given input by computing the probability of each sentiment class and selecting the one with the highest probability.



```
0s from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score

# Train a Naive Bayes classifier
clf = MultinomialNB()
clf.fit(X_train, y_train)

# Predict the sentiment for the test set
y_pred = clf.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

Accuracy: 0.955

Sentiment Intensity Analyzer:

We have removed stop words and put into review clean.

	review	is_bad_review	review_clean	neg	neu	pos	compound
0	I am so angry that i made this post available...	1	angry make post available via possible site us...	0.083	0.859	0.058	-0.8589
1	No real complaints the hotel was great great ...	0	real complaint hotel great great location surr...	0.058	0.757	0.186	0.9494
2	Rooms are nice but for elderly a bit difficul...	0	room nice elderly bit difficult room two story...	0.111	0.671	0.218	0.8402
3	My room was dirty and I was afraid to walk ba...	1	room dirty afraid walk barefoot floor look cle...	0.099	0.754	0.147	0.9355
4	You When I booked with your company on line y...	0	book company line show picture room think get ...	0.058	0.861	0.081	0.5263
...
995	Great location room was outstanding beds were...	0	great location room outstanding bed clean comf...	0.000	0.485	0.515	0.9693
996	The cleaner knocked on our door at 8am which ...	0	cleaner knock door ideal check day need clean ...	0.037	0.633	0.330	0.9709
997	This was my second time at this hotel the fir...	0	second time hotel first time book basic room p...	0.000	0.853	0.147	0.8481
998	Going as a family with two rooms hired in the...	0	go family two room hire reservation allocate t...	0.000	1.000	0.000	0.0000
999	Some of the complimentary items of food from ...	0	complimentary item food room miss empty box mi...	0.040	0.726	0.234	0.9407

1000 rows x 7 columns

Gensim Doc2Vec Model (extension of the Word2Vec model):

The Gensim Doc2Vec model is an unsupervised learning algorithm that learns fixed-length vector representations, or embeddings, for text documents. It is widely used for learning word embeddings.

The Doc2Vec model not only captures the meanings of individual words but also the overall semantic representation of a document. The key idea behind the Doc2Vec model is to train neural networks to predict words in a document, similar to Word2Vec. However, in addition to the word vectors, Doc2Vec also learns document vectors.

TfidfVectorizer:

TF-IDF (t, d) = Term Frequency (t, d) * Inverse Document Frequency(t)

Term Frequency (TF): It measures the frequency of a term (t) in a document (d). It can be calculated as the raw count of the term or normalized by the total number of terms in the document.

Inverse Document Frequency (IDF): It measures the rarity of a term across the entire corpus. It can be calculated as the logarithm of the total number of documents divided by the number of documents containing the term.

Top 10 Positive Reviews:

Highest positive sentiment reviews with more than 5 words.

```
# highest positive sentiment reviews (with more than 5 words)
reviews_df[reviews_df["nb_words"] >= 5].sort_values(by="pos", ascending=False)[["review", "pos"]].head(10)
```

	review	pos
741	lovely helpful staff great location	0.843
936	Great location great staff comfortable rooms	0.793
965	Friendly staff great breakfast	0.785
755	Lovely comfortable beds pillows	0.780
821	N a Perfect location excellent facilitie help...	0.773
858	very clean friendly staff excellent breakfast	0.772
358	Good staff very helpful Free WIFI Nice hotel	0.756
812	Good location Comfortable room	0.756
945	Friendly staff good location	0.753
387	Great and beautiful hotel with wonderful staff	0.745

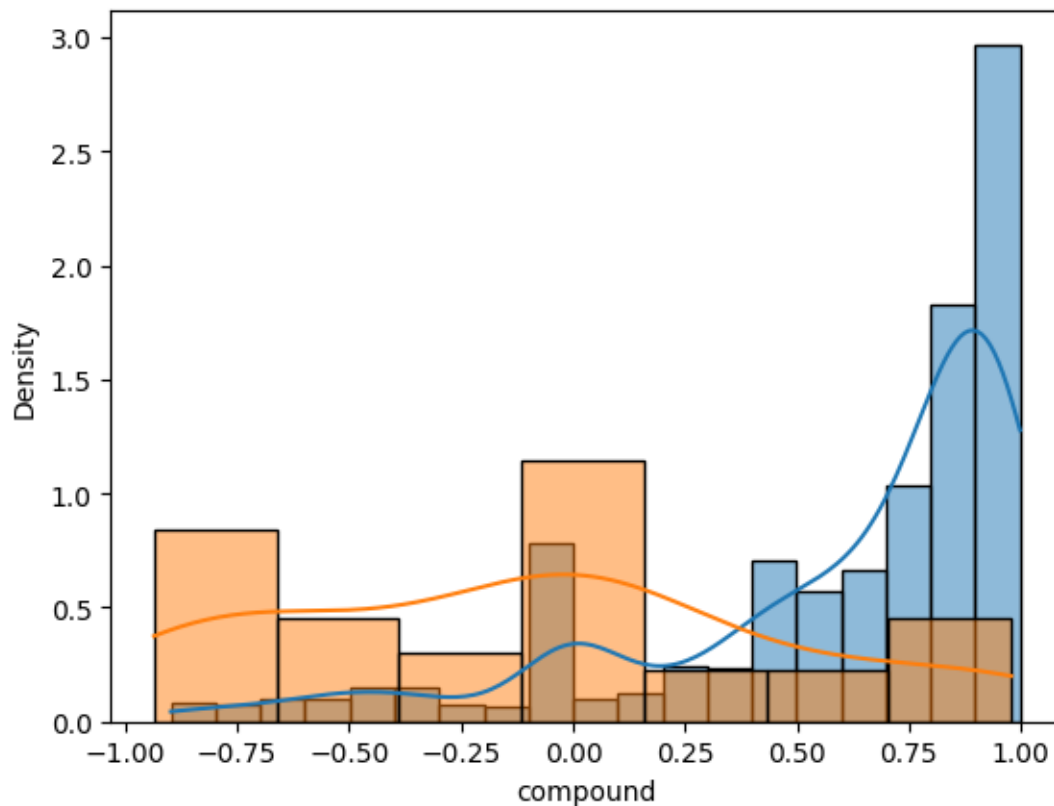
Top 10 negative reviews:

Lowest negative sentiment reviews with more than 5 words.

```
# lowest negative sentiment reviews (with more than 5 words)
reviews_df[reviews_df["nb_words"] >= 5].sort_values("neg", ascending = False)[["review", "neg"]].head(10)
```

	review	neg
710	the breakfast sucks in every aspect variety q...	0.567
708	NOTHING IOcation WAS GREAT	0.561
388	The mattress was awful	0.500
363	Plain 4 stars Hotel nothing special nothing bad	0.481
840	we didn t dislike anything	0.464
759	Small room broken tv Location	0.437
851	triple room very poor for 3 adults	0.404
697	Air conditioning very noisy	0.399
958	Nothing Location friendly staff garden	0.396
818	Air condition was poor Air condition	0.383

Final plotting of all the positive and negative reviews



The histogram plot of positive and negative reviews provided a visual representation of the sentiment distribution. From the histogram, it can be observed whether most reviews were positive or negative, and the proportion of each sentiment class. This information helps to understand the overall sentiment of the reviews and the general satisfaction level of customers.

Based on the analysis, it was found that a significant portion of the reviews were positive, indicating a positive sentiment towards the hotels. This suggests that most customers had a satisfactory experience and expressed their positive opinions in their reviews. However, it is important to note that some negative reviews were also present, indicating areas where improvements could be made to enhance customer satisfaction.

The results of the sentiment analysis revealed valuable insights into the hotel reviews dataset.

CONCLUSION:

In conclusion, this report presented an analysis of sentiment in a hotel reviews dataset using sentiment analysis techniques. The dataset consisted of customer reviews for various hotels, and the goal was to classify the reviews as positive or negative based on the sentiment expressed.

Several steps were undertaken to perform sentiment analysis, including data pre-processing, feature extraction, and the implementation of a machine learning algorithm. The text data was cleaned by removing noise, such as special characters and stop words, and then transformed into numerical features using techniques like TF-IDF or word embeddings. A machine learning classifier, such as Naive Bayes and Decision tree, was trained on the labelled data to predict the sentiment of new, unseen reviews.

Overall, sentiment analysis of hotel reviews provides valuable insights for businesses in the hotel industry. By understanding the sentiment of customer feedback, hotels can identify areas of strength and weakness, make informed decisions, and take appropriate actions to improve customer experience and satisfaction. This can ultimately lead to better customer retention, positive brand reputation, and increased business success.

Source code link:

https://github.com/SushRed10/22MCB0005/blob/main/22MCB0005_Sentiment_Analysis_Assessment4.ipynb