Project
On
Restaurant Review Using NLP

• ABSTRACT

Being able to monitor daily sales of their food and beverages is the most useful tool a restaurant can have.

At the moment, both academia and industry place a high value on recommendation systems. The management of information overload can be done with these. In this study, useful data from user reviews was analysed using machine learning algorithms that were applied to the reviews. Reviews can be helpful for both customers and business owners when making data-driven decisions.

The creation of a machine learning model using NLP approaches that extracts user opinions from user evaluations. Many businesses fail because they don't make enough money and don't take the right steps to improve. Most often, restaurant owners struggle greatly to increase their output.

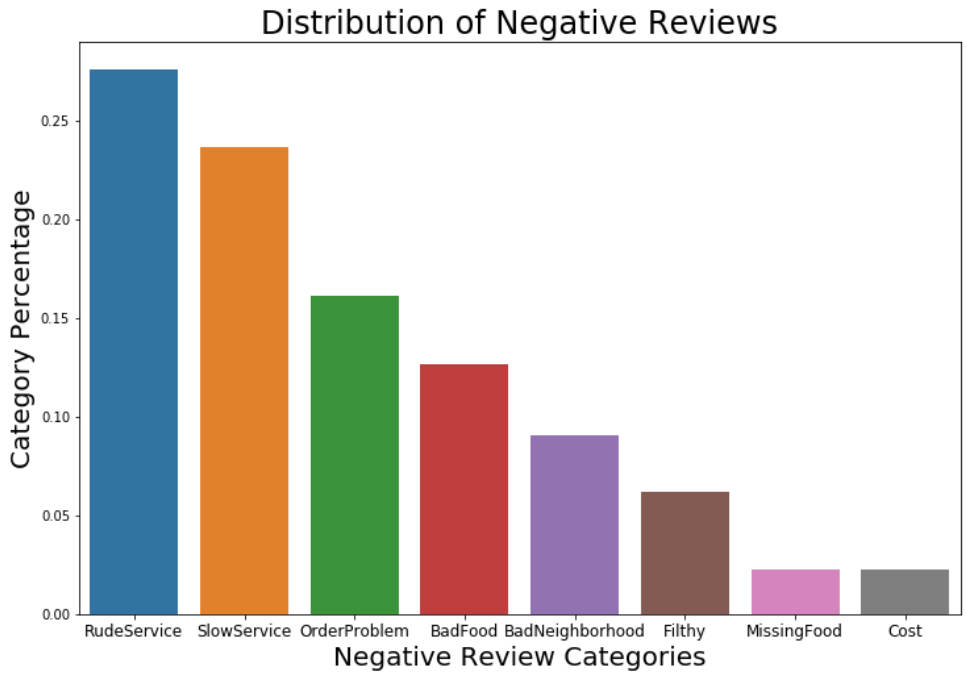
Keywords: Naive-Bayes algorithm, Python, Machine Learning, NLTK, Pandas, NLP, Survey

• OBJECTIVE

The project's objective is to identify the opinions or emotions expressed in the reviews.

For those looking to decide wisely about their eating experiences, restaurant reviews are a valuable source of information. Based on how comfortable they are, restaurant patrons rate and assess the establishment. Other customers can use these ratings and reviews to decide whether to visit a certain restaurant, and restaurant owners can use them to grow their business.

We can use NLP approaches to analyse our textual datasets. Data analysts may use NLP to put machine learning and deep learning algorithms to work on our textual datasets. Machine learning algorithms are used to categorise reviews and suggest the best restaurant. The three sorts of methodologies that are typically used in a suggested system are content-based, collaborative-based, and hybrid-based.



• INTRODUCTION

The number of guests a restaurant serves each day determines how much money it makes. There must be a typical proportion of clients who

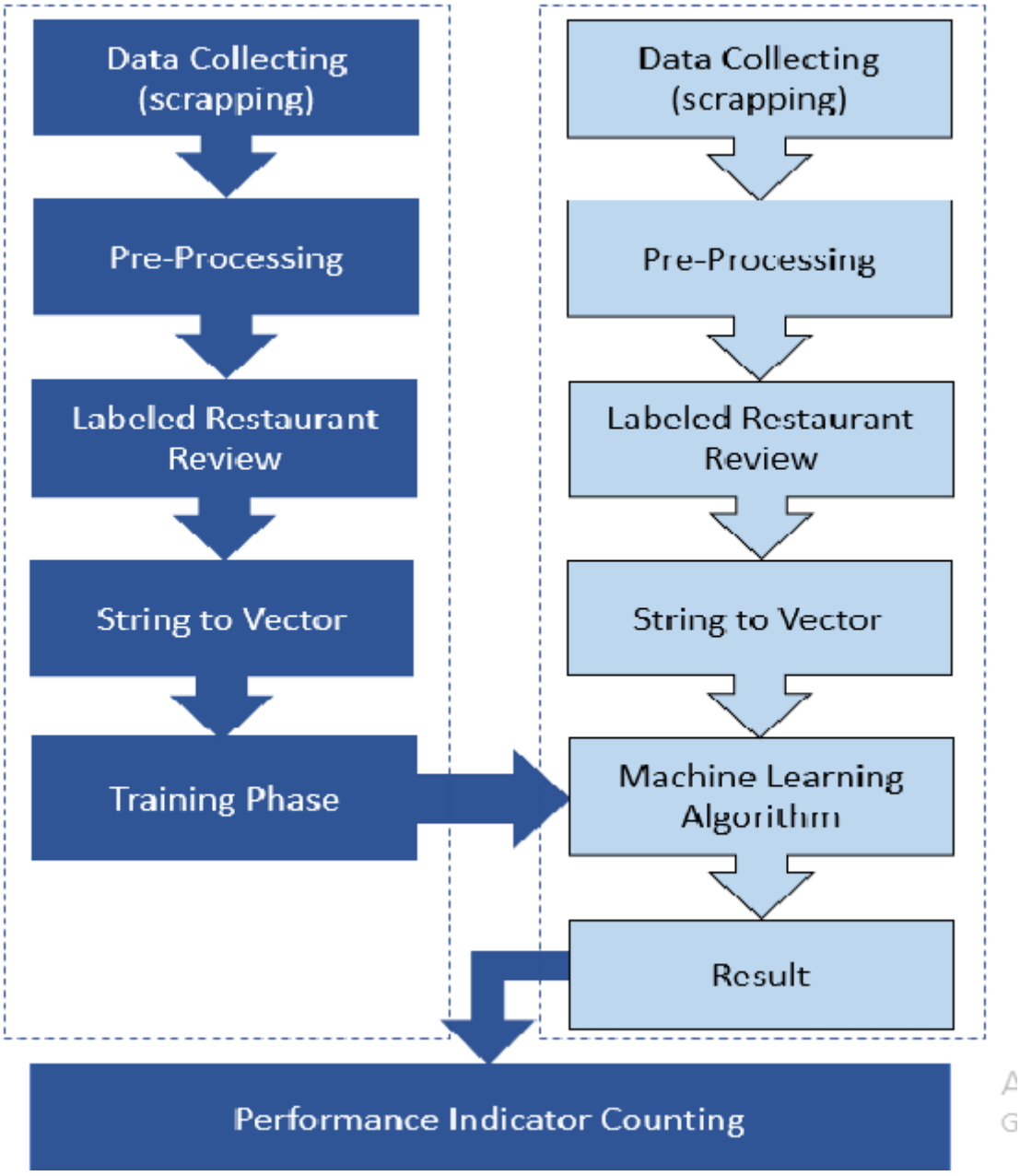
arrive at the restaurant every day. Knowing about their clientele would assist the restaurant owner realise what adjustments need to be done and how they can be accomplished if we were to truly know how they felt about the establishment.

make it. A restaurant itself can perform surveys to learn what might be lacking or what can be altered to increase profits, keep the same clients while also attracting new ones, and collect feedback from its patrons.

Pre-processing of data include removing stop words from reviews, such as "the," "and," "is," etc. The stemming method will then be used to eliminate all affixes and prefixes, allowing the model to clearly grasp the root word. It will be done to vectorize the text content. Then, Count and TFIDF vectorizers will be used. The underlying vectorizer method and ML model will then be combined using the machine learning approach.

The field of computer science and artificial intelligence known as natural language processing (NLP) is concerned with how computers interact with human (natural) languages, particularly how to programme computers to handle and analyse massive volumes of natural language data. A free machine learning library for the Python programming language is called Scikit-learn. With some basic algorithms written in Cython for efficiency, Scikit-learn is primarily developed in Python.

This algorithm can be easily applied to any other kind of text like classify a book into Romance, Friction, but for now, let’s use a restaurant review dataset to review negative or positive feedback.


Data Collecting (scrapping)
Pre-Processing
Labeled Restaurant Review
String to Vector
Training Phase
Data Collecting (scrapping)
Pre-Processing
Labeled Restaurant Review
String to Vector
Machine Learning Algorithm
Result
Performance Indicator Counting

• METHODOLOGY

To develop this system the incorporated packages used are Pandas, NLTK, Vectorization.

A. NLTK The Natural Language Toolkit (NLTK)

Python programmes that utilise data on human language and statistical natural language processing (NLP) are created using the Natural Language Toolkit (NLTK). It contains text processing packages that do tokenization, parsing, classification, stemming, tagging, and semantic reasoning. A classifier that can categorise a new supplied review as positive or negative was constructed because the data set comprises both positive and negative reviews. A supervised machine learning problem is classification. When the output has discrete and finite values, it is best employed because it identifies the classes to which the data items belong.

B. Pandas

Pandas is a software library, which is written for Python which can perform data manipulation and analysis. It provides users with data structures and operations to manipulate numerical tables and time series. We imported the restaurant review data set using the Pandas library

C. Vectorization

In NLP, a word prediction or word similarity/semantics can be found by mapping words or phrases from a lexicon to a corresponding vector of real numbers using NLP. Vectorization is the process of turning words into numbers. Vectorization was used to eliminate special characters, break up sentences, and convert all the characters to lowercase. Then we applied the TF-IDF approach, which enabled us to gauge a word's significance based on how frequently it appears in a document.

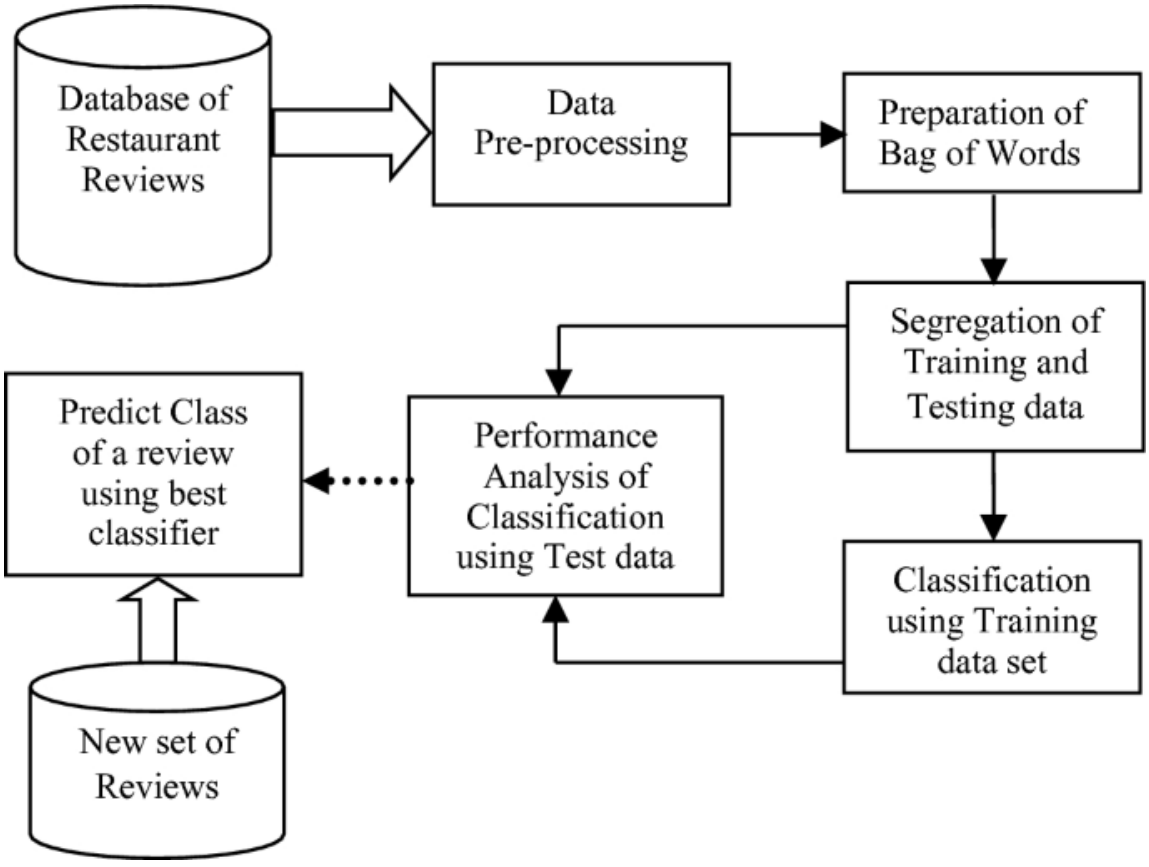
D. Process Description

We created a classifier that divides reviews into positive and negative categories. After importing a data set, stopwords were eliminated, and lemmitization was performed. We would benefit greatly from this because eliminating the stop words will allow the model to learn the categorization faster and work more effectively

and efficiently. Count vectorization was utilised to find word predictions where the phrases from reviews correspond with the vectors. After being fed these vectors, the model was 'trained'.

Now predictions using test data can also be generated. The accuracy score and classification report can thus be generated. Now if words like “bad”, “worst” and “stale” were to be present in the training data-set then the tested data will decide the review to be negative as it has been trained to do so. When the model gives out the result to be “1” then it is supposed to be a positive review and if it were to be “0” then it is supposed to be a negative review.

Multiple classification algorithms were used such as Logistic Regression, Bernoulli Naive Bayes, Multinomial Naive Bayes. We got the best results with Multinomial Naive Bayes algorithm.

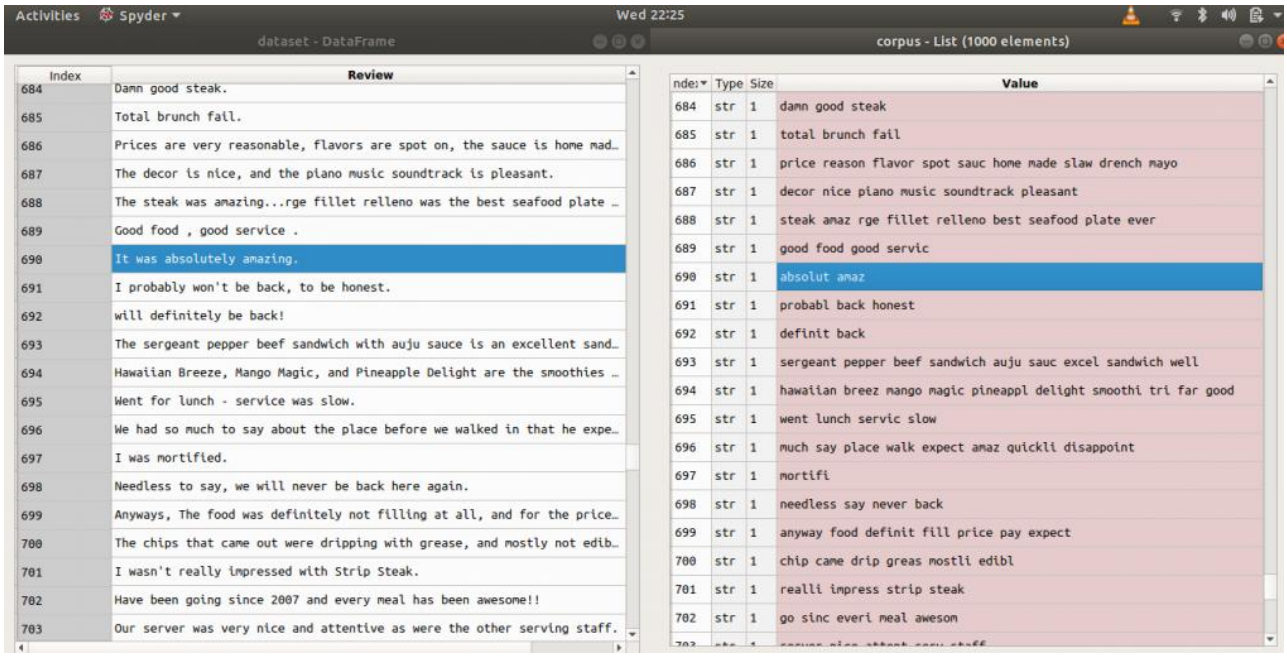


- **CODE**

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Importing dataset

# quoting = 3 is for ignoring "" for our safety.

dataset = pd.read_csv('Restaurant_Reviews.tsv', delimiter='\t',quoting = 3)

import re

import nltk

nltk.download('stopwords')

from nltk.corpus import stopwords

# Stemming means taking the root of the word eg. loved, loving, will love -> love

# This will reduce different versions of the same word and will hence reduce the sparsity of matrix

from nltk.stem.porter import PorterStemmer

corpus = []

for i in range (0, 1000):

# Removing unnecessary punctuations and numbers except letters and replacing removed words with space.

review = re.sub('[^a-zA-Z]', ' ', dataset['Review'][i])
```

```
# Converting review to lowercase

review = review.lower()

# Converting review to list(of strings)

review = review.split()

# Loop through all words and keep those which are not in stopwords list.

# set is much faster than a list and is considered when the review is very
large eg. an article,a book

ps = PorterStemmer()

review = [ps.stem(word) for word in review if not word in
set(stopwords.words('english'))]

# Joining back the review list to a string with each word seperated by a
space.

review = ' '.join(review)

corpus.append(review)

# Creating the Bag of Words Model

# Bag of Words Model is a sparse matrix where each row is the review and
each column is a unique

# Tokenization - process of taking all unique words of reviews and creating
columns for each word.

# Since this a problem of classification we have dependent and independent
variables and each

# unique word/column is like an independent variable and the
review(good/bad) depends on these words.
```

```
from sklearn.feature_extraction.text import CountVectorizer

# max_features keeps most frequent words and removes least frequent words (extra cleaning)

# max_feature reduces sparsity, increases precision, better learning and hence better prediction.

cv = CountVectorizer(max_features = 1500)

X = cv.fit_transform(corpus).toarray()

y = dataset.iloc[:, 1].values

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 0)

from sklearn.naive_bayes import GaussianNB

classifier = GaussianNB()

classifier.fit(X_train, y_train)

y_pred = classifier.predict(X_test)

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test,y_pred)
```


Activities
Spyder
Wed 22:25
dataset - DataFrame
corpus - List (1000 elements)
Index
Review
684 Damn good steak.
685 Total brunch fail.
686 Prices are very reasonable, flavors are spot on, the sauce is home mad…
687 The decor is nice, and the piano music soundtrack is pleasant.
688 The steak was amazing...rge fillet relleno was the best seafood plate …
689 Good food , good service .
690 It was absolutely amazing.
691 I probably won't be back, to be honest.
692 will definitely be back!
693 The sergeant pepper beef sandwich with auju sauce is an excellent sand…
694 Hawaiian Breeze, Mango Magic, and Pineapple Delight are the smoothies …
695 Went for lunch - service was slow.
696 We had so much to say about the place before we walked in that he expe…
697 I was mortified.
698 Needless to say, we will never be back here again.
699 Anyways, The food was definitely not filling at all, and for the price…
700 The chips that came out were dripping with grease, and mostly not edib…
701 I wasn't really impressed with Strip Steak.
702 Have been going since 2007 and every meal has been awesome!!
703 Our server was very nice and attentive as were the other serving staff.
nde: Type Size Value
684 str 1 damn good steak
685 str 1 total brunch fail
686 str 1 price reason flavor spot sauc home made slaw drench mayo
687 str 1 decor nice piano music soundtrack pleasant
688 str 1 steak amaz rge fillet relleno best seafood plate ever
689 str 1 good food good servic
690 str 1 absolut amaz
691 str 1 probabl back honest
692 str 1 definit back
693 str 1 sergeant pepper beef sandwich auju sauc excel sandwich well
694 str 1 hawaiian breez mango magic pineappl delight smoothi tri far good
695 str 1 went lunch servic slow
696 str 1 much say place walk expect amaz quickli disappoint
697 str 1 mortifi
698 str 1 needless say never back
699 str 1 anyway food definit fill price pay expect
700 str 1 chip came drip greas mostli edibl
701 str 1 realli impress strip steak
702 str 1 go sinc everi meal awesom

• CONCLUSION

Any form of business must put learning about the consumer first and foremost if it wants to succeed and have satisfied customers.

Restaurants operate in a similar manner. This initiative has a lot to offer restaurants that need to understand their clientele. We suggested machine learning natural language processing strategies for categorising restaurant reviews. We use stemming to increase productivity.

The model's goal is to identify the opinions or emotions expressed in the reviews. The proposed model accurately predicts the reviewers' feelings with a 77.67% accuracy rate. When the model was evaluated using a data set containing customer reviews, it successfully identified the attitudes.