Project Report

Of

Machine Learning

On

SENTIMENT ANALYSIS ON FOOD REVIEWS

Submitted by:

Armaan Noor Singh Brar (11901090)

Bhavnoor Kaur Gill (11901091)

Naman Jain (11901098)

Simran Garg (11901101)

Group - 3CE4

6th Sem

B. Tech CSE

Submitted to:

Mr. Lal Chand

Department of Computer Science and Engineering

Punjabi University, Patiala

2022

ACKNOWLEDGEMENT

It is our pleasure to be indebted to various people, who directly or indirectly contributed in the development of this project and who influenced our thinking, behavior and acts during the course of study. We express our sincere gratitude to Respected Raman Maini Sir, H.O.D of Computer Science and Engineering Department for providing us an opportunity to undergo through this project.

I would also like to thank my course teacher Mr. Lal Chand for teaching us the subject so easily and interestingly due to which we were easily able to understand the concepts of Machine Learning.

Thank you for encouraging us in all of our pursuits and inspiring s to follow our dreams.

Regards
Armaan Noor Singh Brar
Bhavnoor Kaur Gill
Naman Jain
Simran Garg

ABSTRACT

Evolution of the Internet in the past decade resulted in generation of voluminous data in all sectors. Due to these advents, the people have new ways of expressing their opinions about anything in the form of tweets, blog posts, online discussion forums, status updates, etc**. Sentiment analysis** deals with the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer’s attitude toward a particular topic is positive or negative. Knowing the opinion of customers is very important for any business. It is expensive to check each and every review manually and label its sentiment. So, a better way is to rely on machine learning models for that.

Hence, in this project, we analyze the reviews given by the customers of the restaurant with the help of **Machine Learning Classification algorithm**. This project mainly focuses on the implementation of various classification algorithms and their performance analysis. The simulation results showed that **Logistic Regression** resulted in the **highest accuracy of 82 %** for the given dataset.

TABLE OF CONTENTS

ABOUT THE TECHNOLOGY

Python Programming Language

Python is a high-level, general-purpose and a very popular programming language. Python is an interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. Python Programming Language is very well suited for Beginners.

The biggest strength of Python is huge collection of standard libraries which can be used for the following:

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks (like Django)
- Image processing (like OpenCV, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia
- Scientific computing (like Numpy)
- Text processing and many more.

Python libraries we have used in our project are:

1. Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source **data analysis/manipulation tool** available in any language.

The two primary data structures of pandas, **Series** (1-dimensional) and **DataFrame** (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering.

2. Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

The array object in NumPy is called **ndarray**, it provides a lot of supporting functions that make working with ndarray very easy.

3. Re – Regular expressions

The Python module **re** provides full support for Perl-like regular expressions in Python. There are various characters, which would have special meaning when they are used in regular expression. To avoid any confusion while dealing with regular expressions, we would use Raw Strings as **r'expression'**.

One of the most important **re** methods that use regular expressions is **sub**.

Syntax: re.sub(pattern, repl, string, max=0)

This method replaces all occurrences of the RE *pattern* in *string* with *repl*, substituting all occurrences unless *max* provided. This method returns modified string.

4. Tkinter

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

Creating a GUI application using Tkinter is an easy task. All you need to do is perform the following steps −

- Import the *Tkinter* module.
- Create the GUI application main window.
- Add one or more widgets to the GUI application.
- Enter the main event loop to take action against each event triggered by the user.

The tk **MessageBox** module is used to display message boxes in applications. This module provides a number of functions that we can use to display an appropriate message. Some of these functions are showinfo, showwarning, showerror, and askquestion.

5. Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by Sklearn are as follows −

- **Supervised Learning algorithms** − Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.
- **Unsupervised Learning algorithms** − On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.
- **Clustering** − This model is used for grouping unlabeled data.
- **Cross Validation** − It is used to check the accuracy of supervised models on unseen data.
- **Dimensionality Reduction** − It is used for reducing the number of attributes in data which can be further used for summarisation, visualisation and feature selection.
- **Feature extraction** − It is used to extract the features from data to define the attributes in image and text data.
- **Feature selection** − It is used to identify useful attributes to create supervised models.
- **Open Source** − It is open-source library and also commercially usable under BSD license.

Sklearn modules used in our project are:

i. **Train Test Split**

In machine learning, Train Test split activity is done to measure the performance of the machine learning algorithm when they are used to predict the new data which is not used to train the model.

We can use the train_test_split() method available in the sklearn library to split the data into train test sets.

ii. **CountVectorizer**

In order to use textual data for predictive modelling, the text must be parsed to remove certain words — this process is called tokenization. These words need to then be encoded as integers, or floating-point values, for use as inputs in machine learning algorithms. This process is called feature extraction (or vectorization).

Scikit-learn’s CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. In other words, it counts the number of times a token shows up in the document and uses this value as its weight. It also enables the pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

iii. Logistic Regression

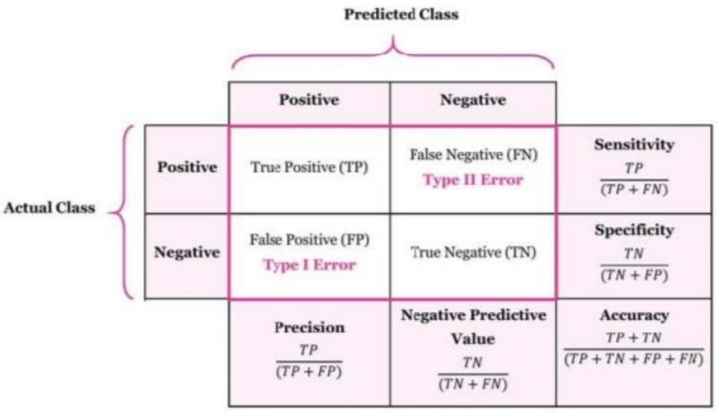
- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

iv. Metrics

Sklearn metrics are import metrics in SciKit Learn API to evaluate our machine learning algorithms. One of the most important metrics is Confusion Matrix. The confusion matrix provides a base to define and develop any of the evaluation metrics.

A **confusion matrix** is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. Some of the basic terms of confusion matrix are:

- True Positive: Actually positive (ground truth), predicted as positive (correctly classified).
- True Negative: Actually negative (ground truth), predicted as negative (correctly classified).
- False Positive: Actually negative (ground truth), predicted as positive (misclassified).
- False Negative: Actually positive (ground truth), predicted as negative (misclassified).



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

6. Pickle

The pickle module is used for implementing binary protocols for serializing and de-serializing a Python object structure.

- Pickling: It is a process where a Python object hierarchy is converted into a byte stream.
- Unpickling: It is the inverse of Pickling process where a byte stream is converted into an object hierarchy.

Module Interface:

- dumps() – This function is called to serialize an object hierarchy.
- loads() – This function is called to de-serialize a data stream.

PKL file:

A PKL file is a file created by pickle, a Python module that enables objects to be serialized to files on disk and deserialized back into the program at runtime. It contains a byte stream that represents the objects.

A PKL file is pickled to save space when being stored or transferred over a network then is unpickled and loaded back into program memory during runtime. The PKL file is created using Python pickle and the **dump()** method and is loaded using Python pickle and the **load()** method.

PKL files may also have the .PICKLE extension but more commonly have the .P extension.

ABOUT THE PROJECT

About Data set

The data set consists of 1000 reviews of fine foods. It contains 500 positive and 500 negative reviews. It is saved in a tsv (tab separated value) file. It is read using read_csv() function of Pandas.

We have the following columns:

1. Review: Review about the product
2. Liked: 0 or 1, 0 for negative review and 1 for positive review

Objective

Given a review, determine whether the review is positive (1) or negative (0).

Preprocessing text data

Text data requires some preprocessing before we go on further with analysis and making the prediction model. Hence in the preprocessing phase, we do the following in the order below:-

- Begin by removing the Html tags.
- Replace abbreviations.
- Remove contact number.
- Remove any punctuations, special characters and numbers.
- Replace more than one space with single space.
- Finally, convert data to lowercase.

Train test split

Once we are done with preprocessing, we will split our data into train and test. 80% of the data is used for training the model and 20% of the data is used for testing the model.

Vectorizing text data

After that, I have applied Count Vectorizer technique for featuring our text and saved them as separate vectors.

Machine learning Approach: Logistic regression

As our dataset is labelled, we have to use a Supervised learning approach. Our model predicts discrete output values, that's why we are using Classification algorithm. We have used Logistic Regression classification algorithm as it gives highest accuracy for the given dataset. We have also tried to train the models using SVM and Decision Trees but their accuracy was low as compared to Logistic Regression.

- Accuracy of Logistic Regression – 82%
- Accuracy of SVM – 79%
- Accuracy of Decision Trees – 16%

Save the Model

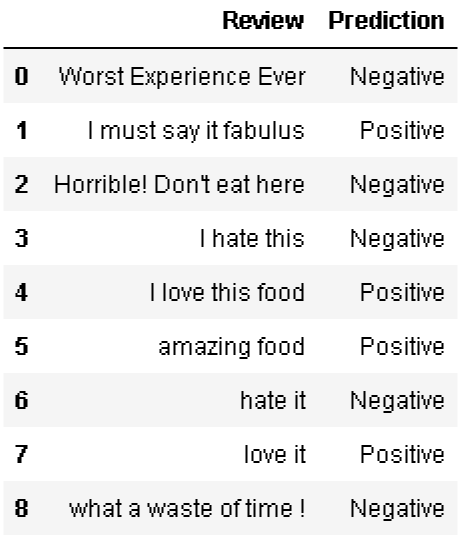
Our model is saved in a pkl file using pickle package which is then passed to our GUI.

GUI Window using Tkinter

Our project consists of a GUI window which takes review as input from the user and then drops a message box telling us whether the customer is satisfied or not.

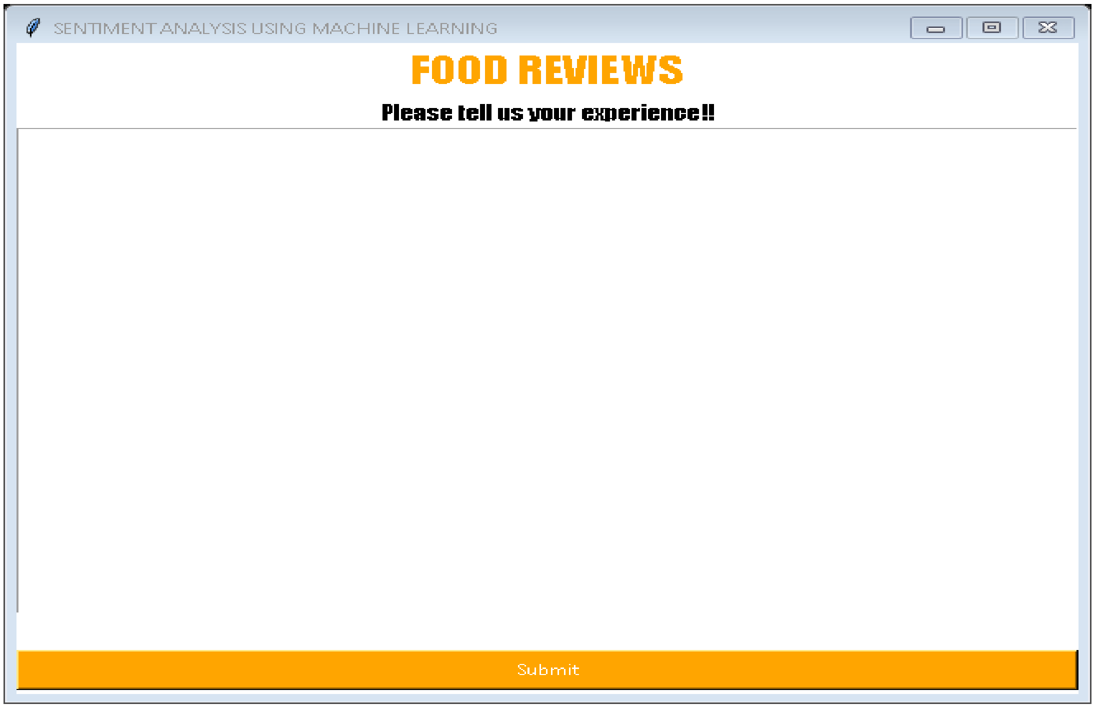
SCREENSHOTS

1. Prediction on unseen dataset



	Review	Prediction
0	Worst Experience Ever	Negative
1	I must say it fabulus	Positive
2	Horrible! Don't eat here	Negative
3	I hate this	Negative
4	I love this food	Positive
5	amazing food	Positive
6	hate it	Negative
7	love it	Positive
8	what a waste of time !	Negative

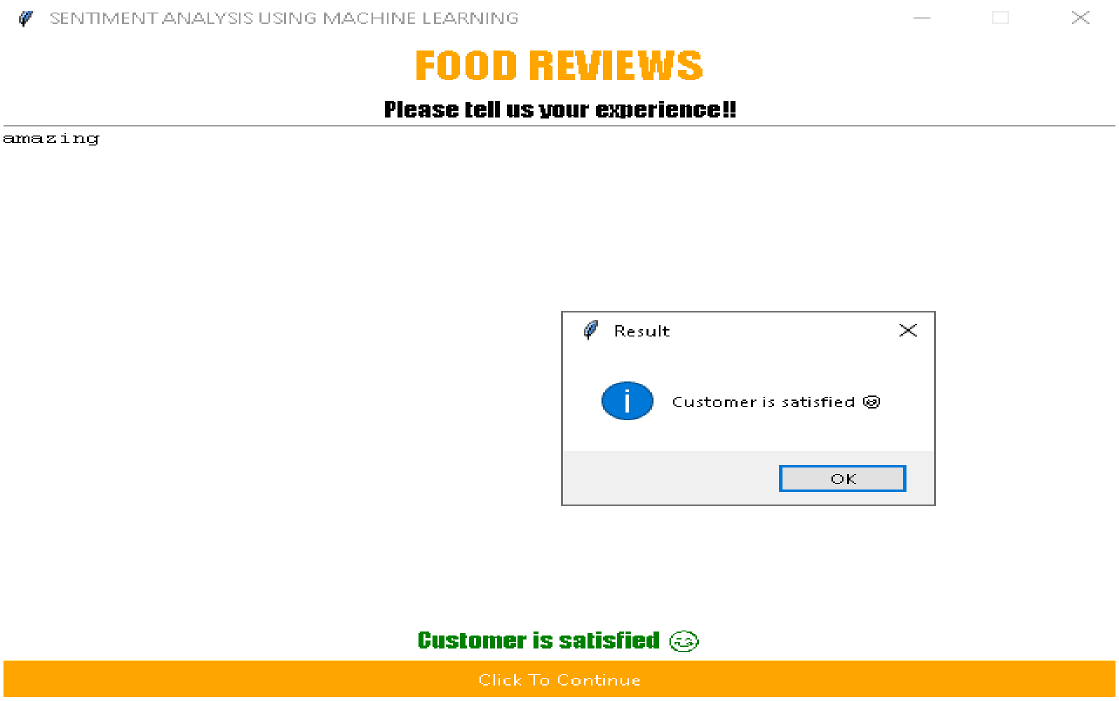
2. Tkinter window



FOOD REVIEWS

Please tell us your experience!!

3. Positive Review



SENTIMENT ANALYSIS USING MACHINE LEARNING

FOOD REVIEWS

Please tell us your experience!!

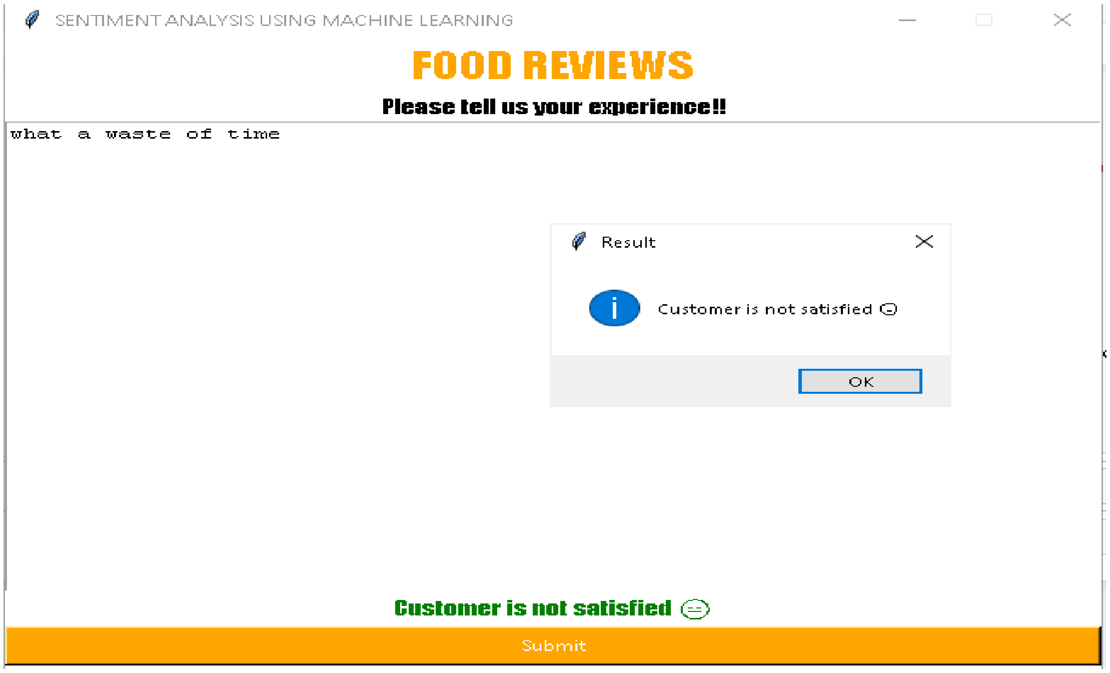
```
amazing
```

Result

Customer is satisfied 😉

Customer is satisfied 😉

4. Negative Review



SENTIMENT ANALYSIS USING MACHINE LEARNING

FOOD REVIEWS

Please tell us your experience!!

```
what a waste of time
```

Result

Customer is not satisfied 😑

Customer is not satisfied 😑

REFERENCES

1. Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques | Request PDF (researchgate.net)

2. PKL File Extension - What is a .pkl file and how do I open it? (fileinfo.com)

3. javatpoint

4. geeksforgeeks

5. tutorialspoint

6. Sentiment Analysis On Amazon Food Reviews: From EDA To Deployment | by Arun Mohan | Towards Data Science