Sentimental Analysis of Restaurant Reviews

By:

Sayali Ambre & Anisha Menezes

Team 5

Professor:

Jianmin Chen

1.ABSTRACT

Sentimental Analysis is a huge volume increasing at a humogonous rate everyday which has made it almost impossible to evaluate the data manually.In social media,twitter,restaurant site people share their opinion as in a huge number of their prevalence.In order to make the process of analyzing the text automatic there are various machine learning techniques that could be applied.The data set is for those enthusiastics who are willing to play with text data and perform sentiment analysis or text classification.The huge quantity of data in textual is generated every day has no value unless processed.This data set consists of actual reviews from real people.So this data set will give a real time experience as to how deal with textual data.

2. INTRODUCTION

Recently there has been number of restaurants when you are on the lookout for a new place to eat,what is the best way to find a great restaurant.Ask someone who's been there,ofcourse.If you don't have someonne to personally ask,then you can always turn to online reviews.Customers take many factors into consideration that when deciding where to eat.It's not just about how great the food tastes but how good the service is,how polite the employees are,and how well maintained the facilities are.The truth is,consumers are trusting advertising less and less and turning to reviews to find out what dining at a restaurant is really like.Customers having testimonials will give the potential and also thee customers assurance that they may have a great experience. Customers want to know what to expect when trying a new restaurant.And who better to tell them than a previous Customer .The more individuals hear about your restaurant,the more inclined they will be to dine here.Now it is reviews first than to decide where to eat.Dont let a lack of reviews for your restaurants prevent you from standing out.Collecting the recommendations by making the customers easy to talk about how great their experience was in choosing.

Dataset sourced from online restaurant review platforms (e.g., Yelp, TripAdvisor), with a size of 61,332 Bytes.

Labels,the target variable or labels in the dataset are sentiment scores - positive or negative.There views are labeled as either positive (1) or negative (0).This is a binary classification task predicting sentiment.

Natural Language Processing (or NLP) is applying Machine Learning models to text and language. Teaching machines to understand what is said in spoken and written word is the focus of Natural Language Processing. NLP can be used on a text review to predict if the review is a

good one or a bad one. NLP can be used on an article to predict some categories of the articles you are trying to segment. Also can be used on a book to predict the genre of the book. And it can go further, NLP can be used to build a machine translator or a speech recognition system, and in that last example you use classification algorithms to classify language. A very well-known model in NLP is the Bag of Words model. It is a model used to preprocess the texts to classify before fitting the classification algorithms on the observations containing the texts.Here we learn things like,Clean texts to prepare them for the Machine Learning models,Create a Bag of Words model,Apply Machine Learning models onto this Bag of Worlds model. The history of natural language processing (NLP) generally started in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence. Up to the 1980s, most natural language processing systems were based on complex sets of handwritten rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing. Identifying the mood or subjective opinions within large amounts of text, including average sentiment and opinion mining,Use it to predict the genre of book,Question answering,Use NLP to build a machine translator or speech recognition system,Document summarization Main toolkits of NLP are **Natural Language Toolkit – NLTK,Spacy,Stanford NLP,OpenNLP.**

Various steps involved in text processing and the flow of NLP are **importing data set and libraries, Text Cleaning or Preprocessing, splitting sentences and words from the body of the text, Making the bag of words via sparse matrix,splitting into Training and Test set, Fitting a Predictive Model, Predicting Final Results, To know the accuracy, confusion matrix is needed**.

Data collection ,in this step data is taken out in a recognized format.Missing fields are evacuated in this process and thus the data is transformed.

Sentimental analysis can be considered a classification process.There are three main classification levels in sentimental analysis document-level,sentence-level and aspect-level sentimental analysis.Level of document it aims to classify an opinion expression.It considers the full document as a basic information unit.

Data preprocessing,the collected raw data of restaurants reviews consist of large number attributes and also there will be missing values.Reducing the attributes is required,extracting

the required attributes is also much essential.In data cleaning once attributes are removed,filling the missing values,removing inconsistent datameasuring the central tendency for the attribute is done.In data process the data is cleaned and the extracted data before analysis.Non textual contents and contents that are irrevalent for the analysis are ientified and eliminated.

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. The process of classifying whether a block of text is positive, negative, or, neutral. Sentiment analysis is contextual mining of words which indicates the social sentiment of a brand and also helps the business to determine whether the product which they are manufacturing is going to make a demand in the market or not. The goal which Sentiment analysis tries to gain is to analyze people's opinion in a way that it can help the businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

Classification is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – speech recognition, face detection

handwriting recognition, document classification

Software-libraries

Pandas

It's an open source data analysis library for providing easy-to-use data structures and data analysis tools. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Numpy

NumPy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

Unicode

Unicode is a specification that aims to list every character used by human languages and give each character its own unique code. The Unicode specifications are continually revised and updated to add new languages and symbols.

re-Regular expression

Regular expressions use the backslash character ('\') to indicate special forms or to allow special characters to be used without invoking their special meaning. This collides with Python's usage of the same character for the same purpose in string literals; for example, to match a literal backslash, one might have to write '\\\\' as the pattern string, because the regular expression must be \\, and each backslash must be expressed as \\ inside a regular Python string literal.

String

The string module contains a number of useful constants and classes, as well as some deprecated legacy functions that are also available as methods on strings. In addition, Python's built-in string classes support the sequence type methods described in the Sequence Types — str, unicode, list, tuple, bytearray, buffer, xrange section, and also the string-specific methods described in the String Methods section. To output formatted strings use template strings or the % operator described in the String Formatting Operations section. Also, see the re module for string functions based on regular expressions.

NLTK

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Tokenization, Stemming, Lemmatization, Punctuation, Character count, word count are some of these packages.

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

Algorithms

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

K-Nearest Neighbors

The KNN algorithm assumes that similar things exist in close proximity. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables PCA is a most widely used tool in exploratory data analysis and in machine learning for predictive models.

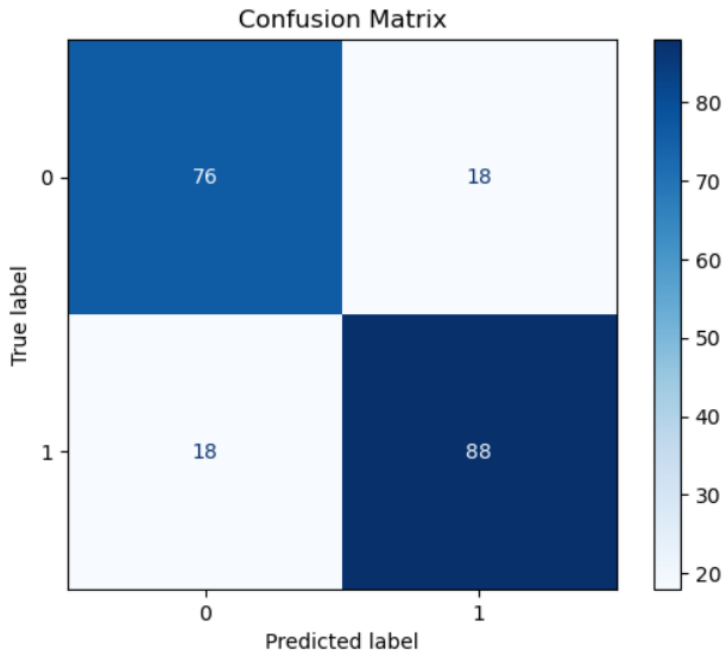
Recurrent Neural Networks (RNNs) are a type of artificial neural network designed for sequential data processing. Unlike traditional feedforward neural networks, RNNs have internal memory, allowing them to capture temporal dependencies in sequences. They excel in tasks like natural language processing, time series analysis, and speech recognition, as they can maintain context and information from previous steps in the sequence. RNNs process input sequentially, updating their hidden state at each step, making them well-suited for dynamic and sequential data modeling.

RESULTS:

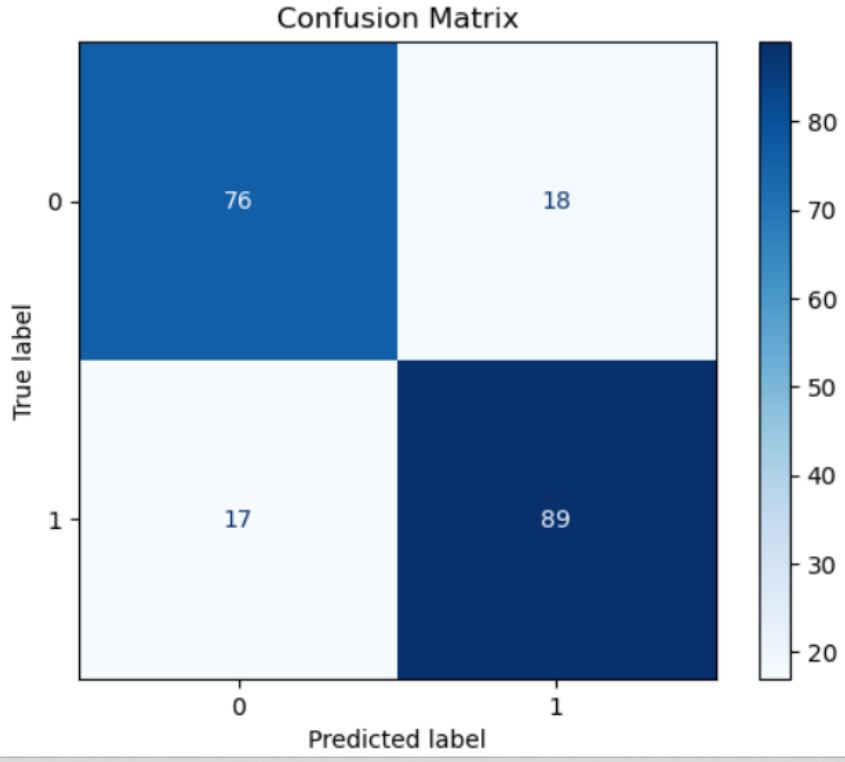
	Accuracy	**Precision**	**Recall**	**F1 score**
KNN	**82**	**83**	**83**	**83**
RNN	**79**	**76**	**81**	**81**
SVM	**85**	**86**	**85**	**85**
SVM with PCA	**85**	**86**	**85**	**85**

Confusion Matrix :

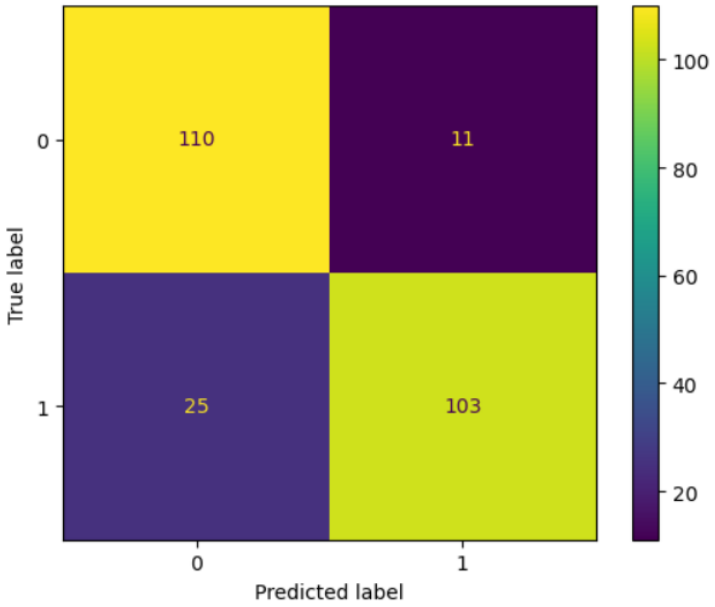
KNN



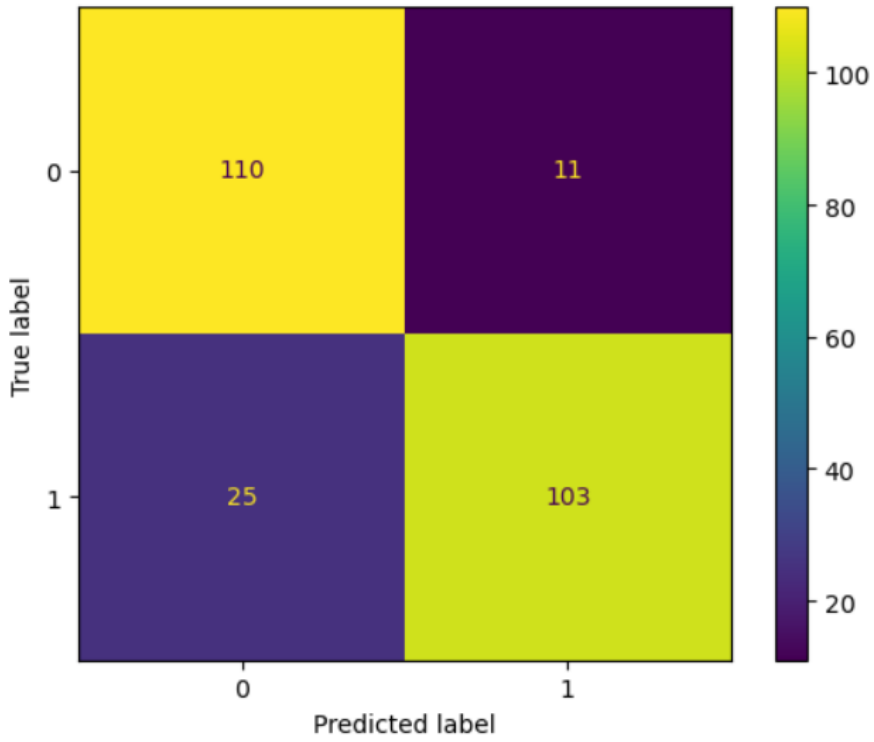
RNN



SVM



SVM with PCA



In the sentiment analysis of restaurant reviews, four distinct models were employed to gauge their performance in classifying sentiments as positive or negative. The K-Nearest Neighbors (KNN) model achieved a testing accuracy of 82%, with precision, recall, and F1 Score all standing at 83%, showcasing a well-balanced performance. The Recurrent Neural Network (RNN) displayed a testing accuracy of 79.5%, with a noteworthy precision of 76.9%, recall of 87.7%, and an F1 Score of 81.9%, indicating its effectiveness in capturing sentiment intricacies. The Support Vector Machine (SVM) model demonstrated superior results, boasting an 85.5% testing accuracy, 90.4% precision, and an 85.5% recall and F1 Score. Interestingly, applying Principal Component Analysis (PCA) to SVM did not significantly alter its performance, maintaining an 85.5% accuracy and balanced precision, recall, and F1 Score. Overall, these models provide valuable insights into the sentiment of restaurant reviews, with SVM standing out for its robust performance, while KNN and RNN also showcase competitive results in sentiment classification.