GOOGLE PLAYSTORE REVIEWS (NLP) PREDICTION USING ML

MsAbinaya C[1],Pavithra S[2],Elakiya A[3]

[1]Assistant Professor, CSE, Agni College of Technology, Chennai, Tamil Nadu, India.

[2]Student,B.E(CSE), Agni College of Technology, Chennai, Tamil Nadu, India.

[3]Student,B.E(CSE), Agni College of Technology, Chennai, Tamil Nadu, India.

Abstract

Google Play Store, Google's app store, is being used by millions of people on a daily basis. It is intended to access, among other things, publications, games, music, movies, and even television.After installing apps, users can submit reviews in the app store to share their personal experiences with others, and this works both ways, with one user being motivated by the ratings of others.The app's usability, performance, and, in rare cases, faults that users have faced while using it are usually explained by users' experiences. The goal is to use Supervised Machine Learning Techniques to classify Google app reviews (SMLT).The SMLT approaches are used to collect variable identification such as sentiments and sentiment polarity, as well as to create a dataset with these variables. It will then go through a series of stages, including data validation and cleaning, visualization, and classification into good, neutral, and negative sensations.

1.INTRODUCTION:

1.1DATA SCIENCE :

Data science is an interdisciplinary field that use scientific methods, procedures, algorithms and systems for extracting knowledge and information from structured data, as well as actionable insights, and using that knowledge and actionable insights across a variety of applications areas.It has become one of the most popular and stylish positions in the industry in less than a decade.Data science is a subject that combines domain knowledge, computational skills, and math and statistics knowledge to extract useful insights from data.Data science is a discipline that combines mathematics, business expertise, tools, algorithms, and machine learning approaches to assisting in the discovery of hidden patterns or insights in raw data that can be used to make critical business decisionsin finding hidden patterns or insights in raw data that can be used to make key business alternatives.

1.2 ARTIFICIAL INTELLIGENCE:

Artificial intelligence (AI) is the emulation of human intelligence in machines that have been programmed to think and act in human-like ways. The phrase can also refer to any machine that demonstrates human-like traits like learning and problem-solving.Machine intelligence, as opposed to natural intelligence shown by humans or animals, is referred to as artificial intelligence (AI).The simulation of human intelligence processes by technology, particularly computer systems, is known as artificial intelligence. AI applications include expert systems, natural language processing, speech recognition, and machine vision.Advanced web search engines, recommendation systems (like those used by YouTube, Amazon, and Netflix), understanding human speech (like Siri or Alexa), self-driving Vehicles (such as Tesla) and competing at the highest levels in strategic gaming systems (such as chess and Go)are examples of AI.Optical character recognition, for example, is frequently overlooked in AI debates, despite the fact that it is a widely used technique.The various sub-fields of AI research are based on different objectives and techniques.Reasoning,knowledge representation, planning, learning, natural language processing, sensing, and the ability to move and manipulate objects are all traditional AI research aims. One of the field's long-term goals is general intelligence (the capacity to solve any problem).AI incorporates computer science, psychology, linguistics, philosophy, and a range of other fields,because of its enormous potential and power, AI has been presented as an existential threat to humanity in science fiction and futurology.Although no single computer language is synonymous with AI, a handful stand out, including Python, R, and Java.AI systems, in general, work by consuming vast volumes of labelled training data, analyzing the data for correlations and patterns, and then using these patterns to forecast future states. learning, reasoning, and self-correction are the three cognitive functions that AI programming focuses on.Processes of learning this element of AI programming is concerned with gathering data and formulating rules for turning it into useful information. this field of AI programming tries to constantly fine-tune algorithms in order to deliver the best accurate results. AI is significant because it has the potential to provide organizations with previously unobtainable insights intotheir operations.In some situations, AI can execute tasks better than humans. When it comes to repetitive, detail-oriented processes like analyzing massive numbers of legal paperwork to AI systems frequently perform processes quickly and with few errors, so make sure all necessary fields are filled up correctly.

1.3NATURAL LANGUAGE PROCESSING (NLP) :

Using natural-language user interfaces and learning straight from human-written materials such as books a sufficiently capable natural language processing system, such as newswire texts, it would be fair. Information retrieval, text mining, question answering, and machine translation are all applications of natural language processing.

To generate syntactic representations of text, many contemporary techniques use word co-occurrence frequencies."Lexical affinity approaches look for words like "accident" to determine the sentiment of a document. Modern statistical NLP systems can incorporate all of these strategies, as well as others, to attain acceptable accuracy at the page or paragraph level.In addition to semantic NLP, the ultimate goal of "narrative" NLP is to have a complete understanding of everyday cognition.

1.4 MACHINE LEARNING:

Machine learning is the process of predicting the future based on historical data. Machine learning (ML) is an artificial intelligence (AI) technique that allows computers to learn without having to be explicitly programmed. The process of creating computer programthat can adapt to new data is known as machine learning.In the training and prediction phase, specialized algorithms are used. It gives the training data to an algorithm, which then uses the training data to create predictions using new test data.For a great user experience and high-quality program, identifying developing concerns in a timely and correct manner is critical.

Machine learning can be divided into three distinct areas. Learning can be classified into three categories: supervised, unsupervised, and reinforced.

1.Supervised Learning: To learn data that must first be tagged by a person, a supervised learning algorithm is provided both the input data and the accompanying labelling.

2.Unsupervised learning: In unsupervised learning, there are no labels. It was made available to the learning algorithm. This method must determine the clustering of the incoming data.

3.Reinforcement learning: reinforcement learning interacts with its environment in a dynamic manner and receives positive or negative feedback in order to enhance its performance.

The task of estimating a mapping function from input variables (X) to discrete output variables is known as classification predictive modelling (y). In machine learning and statistics, classification is a supervised learning method in which the software learns from the data it receives and then uses what it has learnt to classify new data.It's possible that this data collection is bi-class or multi-class. Examples include speech recognition, handwriting recognition, biometric identity, document classification, and other classification issues. In the great majority of real-world machine learning applications, supervised machine learning is used.You are using a technique to acquire the mapping function from the inlet to the outlet, which would be y = f, when you have variables (X) and output variables (y) (X). The goal is to estimate the mapping function to the point that you can forecast the output variables (y) for new input data (X). Logistic regression,

multi-class classification, Decision Trees, and support vector machines are examples of supervised machine learning techniques. As supervised learning requires, the data needed to train the system must already be labelled with correct responses.Classification problems are a sub-set of supervised learning problems.The purpose of this task is to develop a simple model that uses only attribute variables to predict the value of the dependent attribute. The only difference between the two tasks is that the dependent attribute in categorical classification is numerical. A classification model tries to infer something from data that has been observed.A classification model will attempt to predict the value of one or more outputs given one or more inputs. When the output variable is a category, such as red or "blue," the problem is known as a classification problem.

Figure 1:Process of machine learning

2.EXISTING SYSTEM :

A successful app relies on a positive user experience and well-designed functions. Popular apps commonly schedule their updates on a regular basis to achieve this. Developers can make fast updates and maintain a positive user experience if we can record significant app issues faced by users in a timely and accurate manner.It is critical to identify emerging difficulties in a timely and accurate manner in order to provide a favourable user experience and maintain high-qualityprograms. We present MERIT, an unique topicmodeling-based approach for spotting emergent concerns by evaluating online app reviews, in this research. By better modelling brief review texts, jointly modelling subjects and sentiment, and employing word embeddings to better interpret topics, MERIT improves on the current state-of-the-art strategies.

2.1 DOWNSIDES OF EXISTING SYSTEM:

Because they did not employ machine learning techniques, their accuracy is low.

3.PROPOSED SYSTEM:

3.1 Exploratory Data Analysis:

Machine learning supervised classification algorithms will be utilized to take a given dataset and identify patternsthis will help categorize the feedback, allowing apps to make better feature decisions in the future.

3.2 Data Wrangling:

You will load in the data, check for cleanliness, and then trim and clean the dataset for analysis in this phase of the report.

3.3Data collection:

The data set gathered for classifying the given data is divided into two parts: training and testing. In most cases, a 70:30 split is used to divide the Training and Test sets. The SMLT-created Data Model is applied to the Training set, and Test set prediction is made based on the test result accuracy.

3.4Building the classification model:

The likelihood of a positive Google Play Store review is high due to the following factors in machine learning algorithm prediction model and is effective: In a classification problem, it produces superior occurrences.

1.During preprocessing, it excels at removing outliers, irrelevant variables, and a mix of continuous, categorical, and discrete data.

2.It yields out-of-bag estimate errors, which has been demonstrated to be impartial in numerous studies and is quite easy to alter.

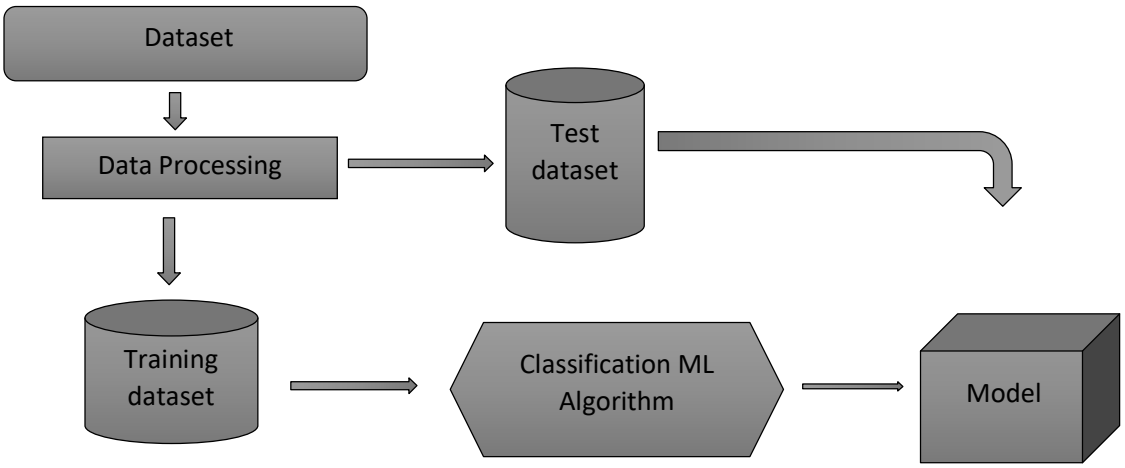


Figure 2:Architecture of Proposed model

3.5 ADVANTAGES OF PROPOSED SYSTEM:

•Improving the performance of the prediction.

• Different algorithms' performance measures are evaluated, and the best prediction is made.

4. REVIEW OF LITERATURE SURVEY

Title : Aspect-based Sentiment Analysis of Scientific Reviews
Author: Souvic Chakraborty
Year : 2020

Scientific publications are complicated, and grasping their utility necessitates prior knowledge. Peer reviews are expert comments on a manuscript that provide a large quantity of information, not only for the editors and chairs to make the final choice, but also to appraise the paper's prospective effectsOne of the study's most noteworthy conclusions is that assessments with a constant reviewer score and aspect are more reliable. In this study, we suggest employing aspect-based sentiment analysis to extract significant information from scientific reviews, which aligns well with the accept/reject decision.We employ an active learning framework to generate a training dataset for aspect prediction, which is then used to acquire the aspects and sentiments while working on a dataset of roughly 8,000 reviews from ICLR, one of the main conferences in the field of machine learning, for the entire dataset. We show that for accepted and rejected papers, the distribution of aspect-based sentiments obtained from a review is significantly different. We make an important conclusion using the aspect sentiments from these reviews: certain elements present in a paper and highlighted in the review heavily influence the ultimate recommendation. A second goal is to assess the degree of disagreement among the reviewers who are evaluating a work.We also look into the level of disagreement between the reviewers and the chair, and discover that there is a link between inter-reviewer conflict and disagreement with the chair. One of the study's most noteworthy conclusions is that assessments with a constant reviewer score and aspect are more reliable. sentiments retrieved from the review text written by the reviewer are more likely to be congruent with the chair's decision.

Title : ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction
Author: Jiahao Bu1 , Lei Ren , Shuang Zheng , Yang Yang , Jingang Wang , Fuzheng Zhang , Wei Wu
Year : 2021

Sentiment analysis is gaining traction in the e-commerce world. For business information, the sentiment polarity that underpin user reviews are extremely valuable. Aspect category sentiment analysis (ACSA) and review rating prediction are two key tasks for recognizing fine-to-coarse sentiment polarities (RP). In real-world e-commerce scenarios, ACSA and RP are often used together because they are closely connected. While most public datasets for ACSA and RP are created independently, this may limit future usage of both program.We will give a large-scale Chinese restaurant review dataset with 46, 730 real reviews as soon as possible. from a top online-to-offline (O2O)

e-commerce platform in China, to solve the problem and advance related research. Each review is manually marked according to its sentiment polarities toward 18 pre-defined aspect categories, in addition to a 5-star scale rating.We believe that the release of the datasets would shed light on the field of sentiment analysis. Furthermore, we provide an evident yet effective combined model for ACSA and RP. On both tasks, the joint model outperforms state-of-the-art baselines, according to experimental results. ASAP, a large-scale Chinese restaurant review dataset for aspect category sentiment analysis (ACSA) and rate prediction, is presented in this study (RP). ASAP is a collection of 46, 730 restaurant user reviews with star ratings collected from a big Chinese e-commerce platform. Based on its emotion polarities, each review is carefully annotated on 18 fine-grained aspect categories.We offer a combination model to address ACSA and RP synthetically, which outperforms previous In addition to individual evaluations of ACSA and RP models on ASAP, state-of-the-art baselines were greatly improved. We believe that the release of ASAP will help to further related research and applications.

Title : Sentiment Analysis for Amazon Products using Isolation Forest

Author: S. Salmiah, DadangSudrajat, N. Nasrul, Tuti Agustin, Nisa Hanum Harani, Phong Thanh Nguyen

Year : 2019

Sentiment Analysis is a text analysis technique for rapidly recognizing, evaluating, and studying a wide range of affective states through the use of natural language. Applications that use advertising to client administration to clinical medication administration include sentiment analysis on the web and web-based social networking, human services materials and audits, and research reactions. Many websites, such as Amazon, encouraged consumers to leave product reviews on their sites. Amazon, on the other hand, imposes a content limit for writing reviews.The review aids in the analysis of the product for various uses, albeit the evaluation for numerous products will differ. Using data from Amazon, this study applies and expands on earlier work in the disciplines of sentiment analysis and natural language processing. The technique employs Machine Learning algorithms to categorize surveys as favourable or negative. Sentiment Analysis is a content evaluation method for successfully perceiving, assessing, and studying full of feeling states without the need of traditional language preparation.Apps that use online presumption assessment and online person-to-person communication, as well as human administrations materials and services reviews, and study responses are examples of applications that use promoting to customer organizations to clinical medication. Many websites, such as Amazon, encouraged customers to publish product reviews on their sites. Amazon, on the other hand, is the furthest point of substance from which to post the questionnaires. The audit breaks down the item for numerous uses, despite the fact that the survey for a few products will be exceptional.This project examines Amazon data and builds on past work in the fields of opinion research and

common language preparation. Machine Learning approaches are used to classify studies as positive or bad.

Title : A Review on Sentiment Analysis Approaches
Author: Ashwini Patil, Shiwani Gupta
Year : 2021

With the rise of social media on the internet, sentiment analysis has become one of the most important research fields. Millions of people use social media sites like Twitter and Facebook to share their thoughts, ideas, expressions, feelings, and opinions.Sentiment analysis, often known as opinion mining, is focused with categorizing and predicting people's sentiments toward a specific topic. It entails categorizing written documents or sentences into positive or negative categories based on the expressed perspective on a certain issue. Although sentiment analysis appears to be comparable to text classification, it confronts a number of obstacles that have prompted a lot of research in this area. To automate the sentiment analysis process, various machine learning and lexicon-based algorithms have been developed in the literature. Despite the fact that these strategies have been widely employed for sentiment categorization, they have not been able to achieve the best results in terms of accuracy and settlement of all issues. difficulties. As a result, new automated techniques must be devised to overcome all obstacles and deliver the finest results.

Title :Sentiment analysis using product review data
Author: Xing Fang and Justin Zhan
Year :2015

One of the most important NLP jobs is sentiment analysis, also known as opinion mining (Natural Language Processing). The purpose of this study is to address one of the most basic challenges in sentiment analysis: categorizing sentiment polarity.A generic process for categorizing sentiment polarity is proposed, along with comprehensive process descriptions. The data for this study came from Amazononline product reviews. Experiments on both sentence-level and review-level categorization have generated promising results. Sentiment analysis, often known as opinion mining, is a branch of research that examines people's feelings, attitudes, and emotions about specific entities. Sentiment polarity categorization is a key challenge in sentiment analysis that is addressed in this study. Amazon.comthe data for this study was gathered from internet product reviews. A process for categorizing sentiment polarity has been suggested, with thorough descriptions of each stage. Experiments on sentence-level categorization as well as review-level categorization were conducted.

5. SCOPE OF THE PROJECT :

The major focus is on the Google Play Store review, which is a classic text classification problem that can be solved with the help of natural language processing and machine learning algorithms. It's required to create a model that can tell the difference between the two.

Feasibility study:

5.1 Data Wrangling

The data will be loaded into this area of the report, and it will be checked for cleanliness before being trimmed and cleaned for research. Ensure that the document steps are followed carefully and that cleaning decisions are justified.

5.2 Data collection

There are two components to the data set for predicting provided data: training and testing.The Training and Test sets are usually divided into 7:3 ratios. The Data Model which was created using Random Forest, logistic, and are applied on the Training set and based on the test result accuracy, Test set prediction is done.

5.3 Preprocessing

There is a chance that the data obtained contains missing values, which could lead to inconsistencies. To get better results, data must be preprocessed to boost the algorithm's performance.The outliers should be eliminated and the parameters are converted.

5.4 Building the classification model

According to the Google Play Store review, A high-accuracy prediction model is effective due to the following factors: It produces superior results in a classification task.During preprocessing, it excels at removing outliers, irrelevant variables, and a mix of continuous, categorical, and discrete data.It generates out-of-bag estimate error, which has been shown to be unbiased in numerous experiments and is quite simple to tweak.

6.SYSTEM ARCHITECTURE :

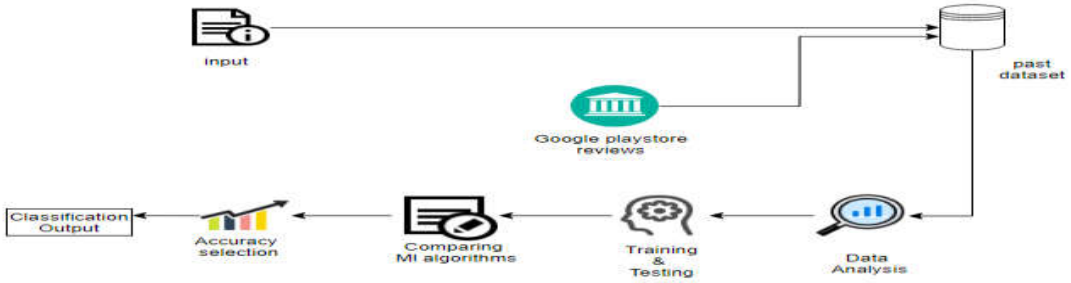


Figure 3:System Architecture

7.WORKFLOW DIAGRAM :

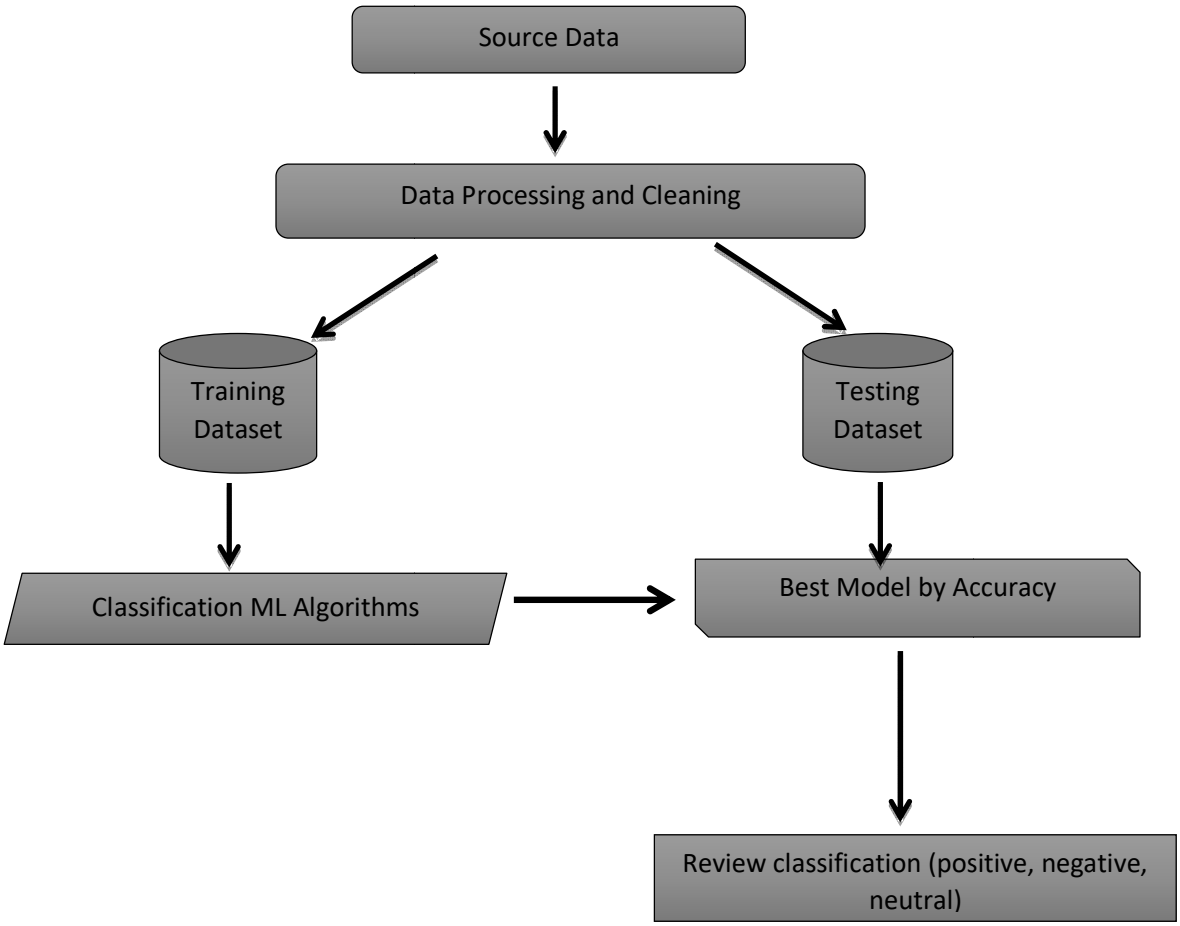


Figure 4: Workflow Diagram

8.MODULES:

1. Data Pre-processing
2. Data Analysis of Visualization
3. Comparing Algorithm with prediction in the form of best accuracy result
4. Deep Learning RNN with LSTM get best accuracy result
5. Output for prediction of sentiment by giving input sentences.

8.1 MODULE DESCRIPTION:

8.1.1 DATA PRE-PROCESSING:

Machine learning validation approaches are used to calculate the error rate of the Machine Learning (ML) model that is as near to the dataset's true error rate as possible. Validation processes may not be required if the data volume is large enough to be representative of the population. In real-world circumstances, however, working with data samples that are not always representative of the dataset's population is necessary. If the variable is a float or an integer variable, duplicate the value and the data type description to find the missing value. While tuning model hyper parameters, a sample of data is employed to offer an unbiased evaluation of a model fit on the training data source.As competence on the validation dataset is incorporated into the model setup, the evaluation becomes increasingly biased. The validation set is used to test a model, although it is only used on a regular basis. This data is used by machine learning specialists to fine-tune the model hyper parameters. Data collection, analysis, and the process of addressing data content, quality, and organization can be time-consuming. Understanding your data and its properties is helpful during the data identification phase; this knowledge will assist you choose which algorithm to employ to build your model.A collection of data cleaning tasks using Python's Pandas module, with an emphasis on the most common data cleaning tasks, missing values, and the ability to clean data more quickly. It prefers to spend less time cleaning data and more time analyzing and modelling it.Some of these sources are simply unintentional errors. Other times, there may be a more serious reason for the lack of data. From a statistical standpoint, it's critical to comprehend the various sorts of missing data. The type of missing data will determine how missing values are filled in, how missing values issuing data is dealt with using simple imputation and detailed statistical procedures, as well as how missing data is found. It's crucial to understand the sources of missing data before putting it into code. Here are some common explanations for missing data:

• A field was left blank by the user.

• Data was lost during a manual transfer from a legacy database.

• A programming error occurred.

• Users declined to fill out a field related to how the results would be utilised or interpreted based on their opinions.

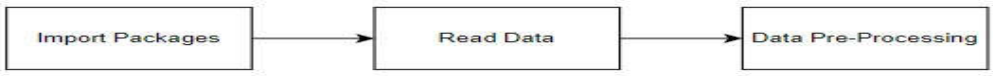
Variables are identified using univariate, bivariate, and multivariate analysis:

Import libraries for access and functionality, as well as the dataset you've been given. Analyzing the General Properties of a Dataset As a data frame, display the specified data set. Columns to show.

To give a description of the data frame

Checking for duplicate data,checking for missing values in a data frame checking for unique values in a data frame Checking count values in a data frame Rename and drop the given data frame to specify the type of values To construct extra columns

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data
output : removing noisy data

DATA VALIDATION/CLEANING/PREPARING

Importing library packages and loading the specified dataset. Identifying variables based on data shape, data type, and evaluating missing values and duplicate values. A validation dataset is a set of data that has been stored after your model has been trained and used to assess its performancemodel skill when tweaking models and processes for making the greatest use of validation and test datasets while evaluating your models. To analyze the uni-variate, bi-variate, and multi-variate processes, data cleaning / preparation is achieved by renaming the given dataset and eliminating the columns, among other things. The methods and techniques for cleaning data will differ depending on the dataset. The fundamental purpose of data cleaning is to find and fix mistakes and abnormalities so that data may be used for analytics and decision-making.

```
In [10]: #To describe the dataframe
         df.describe()
```

Out[10]:

	Sentiment_Polarity	Sentiment_Subjectivity
count	37427.000000	37427.000000
mean	0.182171	0.492770
std	0.351318	0.259904
min	-1.000000	0.000000
25%	0.000000	0.357143
50%	0.150000	0.514286
75%	0.400000	0.650000
max	1.000000	1.000000

Out[22]:

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	0	8865	2	6194	2521
1	0	22964	2	4396	311
3	0	26302	2	5342	4478
4	0	1986	2	6194	358
5	0	2066	2	6194	358

8.1.2 EXPLORATION DATA ANALYSIS OF VISUALIZATION

Data visualization is a critical skill in applied statistics and machine learning. Statistics is concerned with quantitative data descriptions and estimations. Data visualization is a valuable set of tools for acquiring a qualitative understanding of data. This might be useful for spotting patterns, faulty data, outliers, and other things when exploring and getting to know a dataset.For stakeholders, measurements of association or relevance with a small issue are more emotional and meaningful knowledge, and data visualizations can be used to express and demonstrate crucial relationships in plots and charts. Data visualization and exploratory data analysis are fields in and of themselves, and for more information, it is suggested that you read some of the works listed at the conclusion.

Data may not make sense until it is presented in a visual format, such as charts and graphs. The ability to visualize data samples and other objects instantly is a crucial talent in both applied statistics and applied machine learning. It will show you how to use the many sorts of plots available when visualizing data in Python to better understand your own data.

How to visualize time series data with line plots and categorical data with bar charts.

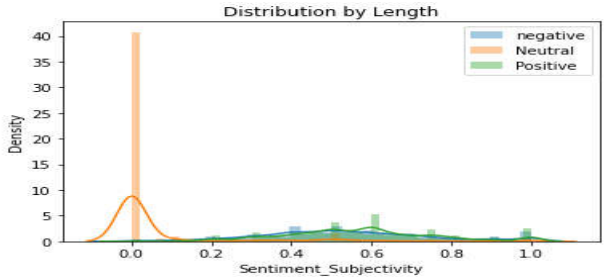
How to summarize data distributions using histograms and box graphs.

MODULE DIAGRAM

GIVEN INPUT EXPECTED OUTPUT

input : data

output : visualized data



Pre-processing refers to the changes we make to our data before feeding it to the algorithm. Data Preprocessing is a technique for converting raw data into a tidy set of data. In other words, data is collected in a raw format whenever it is received from numerous sources, making analysis impossible.To get better results from the used model in Machine learning, the data in a specific format, such as the Random Forest method,which does not accept null values. As a result, in order to perform the random forest algorithm,null values from the original raw data set must be handled.

False Positives (FP):

A individual who is expected to pay is referred to as a defaulter. When the expected class is yes and the actual class is no. For example, if the actual class indicates that this passenger did not survive, but the forecast class indicates that this passenger will.

FalseNegatives (FN):

In the future, a defaulter is more likely to become a payer.When the actual class is good, but the predicted class is bad. For example, if the passenger's actual class value indicates that he or she survived, while the predicted class value implies that the person would die.

True Positives (TP):

Defaulter is a term used to describe someone who does not pay their bills. These are successfully predicted positive values, indicating that the value of the real class is yes, as well as the value of the anticipated class. For example, if the actual class value indicates that this passenger survived and the anticipated class also suggests that this passenger survived.

True Negatives (TN):

A payer has a higher chance of defaulting. These are accurately predicted negative values, indicating that the value of the real class is zero and the value of the projected class is zero as well. For example, if the real class states the passenger did not survive and the predicted class says the same.

8.1.3 COMPARING ALGORITHM WITH PREDICTION IN THE FORM OF BEST ACCURACY RESULT

It is critical to compare the performance of various different machine learning algorithms consistently, and this tutorial will demonstrate how to create a test harness in Python using scikit-learn. This test harness can be used as a starting point for your own machine learning projects, with several algorithms to compare each model's performance characteristics will differ. Using resampling techniques like cross validation, you can get an indication of how trustworthy each model is on unknown data. It should be able to use these estimates to pick one or two of the best models from the collection you've created. When you have a fresh dataset, it's a good idea to visualize it using a variety of ways so you can see it from multiple angles. When it comes to model selection, the same logic applies, you should consider the estimated accuracy of your machine learning algorithms in a variety of ways when deciding which one or two to complete. Using various visualization approaches to display the average accuracy, variance, and other features of the distribution of model accuracies is one way to accomplish this.

The key to a fair comparison of machine learning algorithms is to ensure that each method is evaluated in the same way on the same data, which can be accomplished by requiring each algorithm to be evaluated on the same test harness.

Two distinct algorithms are compared in the example below:

→ Logistic Regression is a statistical technique for predicting the outcome of a situation.

→DT The K-fold cross validation technique, which is designed with the same data, is Each The algorithm is put to the test, as well as the fact that each technique is evaluated

equally. Prior to the comparison algorithm, Using the Scikit-Learn packages, create a machine learning model. Preprocessing, linear model with logistic regression method, cross validation with K-Fold method, ensemble with random forest method, and tree with decision tree classifier are all included in this library package. Additionally, the train and test sets should be separated. By comparing accuracy, it is possible to forecast the outcome.

PREDICTION ACCURACY BY RESULT:

A linear equation with independent predictors is also used in the logistic regression process to predict a value. The anticipated value ranges from negative infinity to positive infinity. The algorithm's output must be classed as variable data. When compared to the best accuracy, the logistic regression model has a higher accuracy in predicting the outcome.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

ACCURACY :

The percentage of total predictions that are correct;alternatively,how often the model properly predicts defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

The simplest intuitive performance metric is accuracy, which is just the ratio of properly predicted observations to all observations. One would believe that if our model is accurate, it is the best. Yes, accuracy is useful, but only when the datasets are symmetric and the number of false positives and false negatives is almost equal.

PRECISION :

The percentage of optimistic predictions that are accurate.

The ratio of accurately anticipated positive observations to total predicted positive observations is known as precision. The question that this measure answers is how many of the passengers who were identified as having survived actually did. The low false positive rate is related to high precision. We have a precision of 0.788, which is rather good.

RECALL :

The proportion of observed positive values that were anticipated correctly.

Recall = TP / (TP + FN)

The F1 Score is the weighted average of Precision and Recall. As a result, both false positives and false negatives are taken into account in this score. Although it is not as intuitive as accuracy, F1 is frequently more useful than accuracy, especially if the class distribution is unequal.When the costs of false positives and false negatives are equal, accuracy works well. It's best to look at both Precision and Recall if the cost of false positives and false negatives is considerably different.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

LOGISTIC REGRESSION :

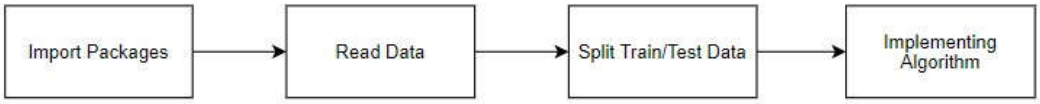
It's a statistical technique for analyzing a data set with one or more independent factors that influence the outcome. A dichotomous variable is used to assess the outcome (in which there are only two possible outcomes). The goal of logistic regression is to find the best-fitting model to capture the relationship between a group of individuals (primary indicator as a group of individuals).well as explanation) attributes and a dichotomous feature of interest (dependent variable = response or dependent variable = response, or dependent variable = response, or dependent variable = response, or dependent variable = response, or dependent variable = response, or dependent variable = response, or dependent variable = response, or dependent variable = response or dependent variable = response or dependent variable = response or dependent variable= dependent variable or response = dependent variable or response =response variable, also known as a dependent variable = dependent variable or response = dependent variable or response = response or dependent variable = response or dependent variable = response or dependent variable = response or dependent variable = dependent variable or response = dependent variable or response = dependent variable or response (response or outcome variable) = response or dependent(response or outcome variable)= dependent variable).To forecast the likelihood of a categorical dependent variable, a Machine Learning classification approach called logistic regression is utilised.The dependent variable in logistic regression is a binary variable that comprises data coded as 1 or 0.(Success, yes, and so on.) . or 0 (no) (no, failure, etc.).

In other words, this logistic regression model predictions P(Y=1) as a function of X. Logistic regression is a technique for predicting the outcome of Assumptions: Binary logistic regression necessitates a binary dependent variable.

The dependent variable's factor level 1 should represent the desired outcome in a binary regression.Only the most important factors should be included.The independent variables must not be connected in any way. To put it another way, the model should have very little.

The log chances are linearly connected to the independent variables.

Logistic regression necessitates a high sample size.



GIVEN INPUT EXPECTED OUTPUT

input : data

output : getting accuracy

```
Accuracy result of Logistic Regression is: 99.99109448748776

Classification report of Logistic Regression : Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2481
           1       1.00      1.00      1.00      8748

    accuracy                           1.00     11229
   macro avg       1.00      1.00      1.00     11229
weighted avg       1.00      1.00      1.00     11229

Confusion Matrix result of Logistic Regression : is:
 [[2480    1]
 [   0 8748]]

Sensitivity :  0.9995969367190649

Specificity :  1.0
```

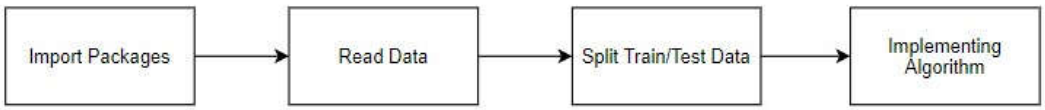
DECISION TREE CLASSIFIER :

Using Decision Trees as a Classifier As a Classifier, Decision Trees are Used It's a well-known and highly effective algorithm.The decision-tree algorithm is characterised as a supervised learning algorithm.It can be used with both continuous and categorical output variables. Decision tree assumptions: At first, we consider the entire training set to be the root.

Attributes are supposed to be categorical for information gain; otherwise, they are presumed to be continuous, and records are dispersed recursively based on attribute values.

As a root or internal node, we rank qualities using statistical approaches.This method is continued on the training set until a termination condition is reached.It's developed using a recursive divide-and-conquer strategy from the top down.All of the characteristics should be categorical in nature. Otherwise, they should be separated ahead of time. The information gain concept is used to identify attributes at the top of the tree that have a greater impact on classification. A decision tree can easily be over-fitted, resulting in an excessive number of branches, which can reveal anomalies due to noise or outliers.

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data
output : getting accuracy

```
Accuracy result ofDecision Tree Classifier  is: 100.0

Classification report of Decision Tree Classifier : Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2481
           1       1.00      1.00      1.00      8748

    accuracy                           1.00     11229
   macro avg       1.00      1.00      1.00     11229
weighted avg       1.00      1.00      1.00     11229

Confusion Matrix result of Decision Tree Classifier : is:
 [[2481    0]
 [   0 8748]]

Sensitivity :  1.0

Specificity :  1.0
```

8.1.4 DEEP LEARNING RNN WITH LSTM GET BEST ACCURACY RESULT :

A feedforward neural network with an internal memory is known as a recurrent neural network. RNN is recurrent in nature since it executes the same function for each data input, and the current input's outcome is dependent on the previous computation. When the output is formed, it is replicated and relayed back into the recurrent network. It evaluates the current input as well as the output it has learned from the prior input when making a decision. Unlike feedforward neural networks, RNNs can use their internal state to process sequences of inputs (memory).As a result, unsegmented, connected handwriting recognition and speech recognition are now conceivable. The inputs of other neural networks are unrelated to one another.It first extracts X(0) from the series of inputs, then outputs h(0), which, along with X(1), serves as the input for the next phase. As a result, the inputs for the next step are h(0) and X. (1). Similarly, the following step's input is h(1), and the next step's input is X(2), and so on. As a result, it remembers the context while training.

```
Train on 29941 samples, validate on 7486 samples
Epoch 1/1
29941/29941 [==============================] - 502s 17ms/step - loss: 0.5294 - accuracy: 0.7852 - precision: 0.6916 - recall:
0.4421 - val_loss: 0.4646 - val_accuracy: 0.8064 - val_precision: 0.6328 - val_recall: 0.8516

5]: <keras.callbacks.callbacks.History at 0x144d9a69e80>
```

Advantages of Recurrent Neural Network:

RNN can describe data sequences in such a way that each sample is assumed to be dependent on the ones that came before it. Even convolutional layers are employed with recurrent neural networks to broaden the effective pixel neighboorhood.

```
X_train.shape: (28070, 100)
X_test.shape: (9357, 100)
y_train.shape: (28070, 3)
y_test.shape: (9357, 3)
Train on 28070 samples, validate on 9357 samples
Epoch 1/30
28070/28070 [==============================] - 71s 3ms/step - loss: 0.6198 - accuracy: 0.7375 - precision: 0.7622 - recall: 0.8
509 - val_loss: 0.5043 - val_accuracy: 0.7919 - val_precision: 0.8319 - val_recall: 0.9259
Epoch 2/30
28070/28070 [==============================] - 68s 2ms/step - loss: 0.4733 - accuracy: 0.8078 - precision: 0.8645 - recall: 0.8
715 - val_loss: 0.4241 - val_accuracy: 0.8309 - val_precision: 0.8903 - val_recall: 0.9041
Epoch 3/30
28070/28070 [==============================] - 69s 2ms/step - loss: 0.4216 - accuracy: 0.8313 - precision: 0.8830 - recall: 0.8
859 - val_loss: 0.3906 - val_accuracy: 0.8464 - val_precision: 0.9268 - val_recall: 0.8853
Epoch 4/30
28070/28070 [==============================] - 69s 2ms/step - loss: 0.3836 - accuracy: 0.8484 - precision: 0.9003 - recall: 0.8
997 - val_loss: 0.4050 - val_accuracy: 0.8373 - val_precision: 0.9597 - val_recall: 0.8310
Epoch 5/30
28070/28070 [==============================] - 70s 2ms/step - loss: 0.3520 - accuracy: 0.8616 - precision: 0.9098 - recall: 0.9
084 - val_loss: 0.3880 - val_accuracy: 0.8477 - val_precision: 0.9414 - val_recall: 0.8790
Epoch 6/30
28070/28070 [==============================] - 70s 2ms/step - loss: 0.3217 - accuracy: 0.8745 - precision: 0.9201 - recall: 0.9
176 - val_loss: 0.3196 - val_accuracy: 0.8773 - val_precision: 0.9107 - val_recall: 0.9408
```

8.1.5 OUTPUT FOR PREDICTION OF SENTIMENT BY GIVING INPUT SENTENCES :

code2vec is a neural network that learns source code analogies. The model was developed using a Java code database, but it may be used with any codebase.
Then there's GloVe, the website's GloVe vocabulary.
We went with the bigger one because it has a better chance of detecting all of our phrases. You can save it wherever you wish, but it's best to keep it in the working directory for convenience. Now we can discover out-of-vocabulary words and count the percentage of them in the code2vec vocabulary.The code below will also work with GloVe. We tested three different word embedding algorithms for the OpenAPI specification. Despite the fact that all three perform admirably on this dataset, a closer examination of the most similar words reveals an additional pattern.

```
Reading GloVe: 400000it [00:15, 25181.09it/s]
tracking <tf.Variable 'Variable:0' shape=() dtype=int32, numpy=0> tp
tracking <tf.Variable 'Variable:0' shape=() dtype=int32, numpy=0> fp
tracking <tf.Variable 'Variable:0' shape=() dtype=int32, numpy=0> tp
tracking <tf.Variable 'Variable:0' shape=() dtype=int32, numpy=0> fn
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 100, 100)          2208500
_________________________________________________________________
lstm_1 (LSTM)                (None, 128)               117248
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 387
=================================================================
Total params: 2,326,135
Trainable params: 117,635
Non-trainable params: 2,208,500
_________________________________________________________________
```

```
|: #text = "We stayed for a one night getaway with family on a thursday. Triple AAA "
   print(get_predictions(text))

   Positive
```

9.CONCLUSION :

It is critical to accurately determine the resulting complications.In this paper,We proposed an analytical process that started from data cleaning,processing,missing

values,exploratory data analysis and finally model building and evaluation.By resulting the best accuracy on the public test set,the higher accuracy will be erected.

10.FUTURE WORK :

In the future work, We will enhance the work to implement in Artificial Intelligence environment and We'll also use the cloud to integrate the Google Playstore reviews classification prediction.

REFERENCES:

[1] "Number of apps available in leading app stores," https://www.statista.com/statistics/276623/ number-of-apps-available-in-leading-app-stores/, 2018.

[2] "The Mobile Marketer's Guide to App Store Ratings & Reviews," https://www.apptentive.com/blog/2015/05/05/ app-store-ratings-reviews-guide/.

[3] S. McIlroy, N. Ali, and A. E. Hassan, "Fresh apps: an empirical study of frequently-updated mobile apps in the google play store," Empirical Software Engineering, vol. 21, no. 3, pp. 1346–1370, 2016.

[4] S. L. Lim and P. J. Bentley, "Investigating app store ranking algorithms using a simulation of mobile app ecosystems," in Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2013, Cancun, Mexico, June 20-23, 2013, 2013, pp. 2672–2679.

[5] "App store optimization: 8 tips for higher rankings," https://searchenginewatch.com/sew/how-to/2214857/app store optimization: 8 tips for higher rankings.

[6] Q. Diao, J. Jiang, F. Zhu, and E. Lim, "Finding bursty topics from microblogs," in The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers, 2012, pp. 536–544.

[7] X. Yan, J. Guo, Y. Lan, J. Xu, and X. Cheng, "A probabilistic model for bursty topic discovery in microblogs," in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25- 30, 2015, Austin, Texas, USA., 2015, pp. 353–359.

[28] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang, "A probabilistic method for emerging topic tracking in microblog stream," World Wide Web, vol. 20, no. 2, pp. 325–350.

[9] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in Proceedings of the 22nd International Conference on Requirements Engineering (RE). IEEE, 2014, pp. 153–162

[10] "Engaging Users with App Updates," https://developer.apple. com/app-store/app-updates/, 2019. [11] "Discussion on the full screen mode of YouTube," https://www.reddit.com/r/jailbreak/comments/5tlkwu/ question full screen rotation fix/, 2019.

[12] P. J. Guo, T. Zimmermann, N. Nagappan, and B. Murphy, "Characterizing and predicting which bugs get fixed: an empirical study of microsoft windows," in Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1, ICSE 2010, Cape Town, South Africa, 1-8 May 2010, 2010, pp. 495–504.

[13] F. Thung, D. Lo, L. Jiang, Lucia, F. Rahman, and P. T. Devanbu, "When would this bug get reported?" in 28th IEEE International Conference on Software Maintenance, ICSM 2012, Trento, Italy, September 23-28, 2012, 2012, pp. 420–429.

[14] J. Cambronero, H. Li, S. Kim, K. Sen, and S. Chandra, "When deep learning met code search," in Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019, 2019, pp. 964–974.

[15] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. Murty, "On finding the natural number of topics with latent dirichletallocation: Some observations," in Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I, 2010, pp. 391–402.

[16] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, "A heuristic approach to determine an appropriate number of topics in topic modeling," in BMC bioinformatics, vol. 16, no. 13, 2015, p. S8.

[17] M. Gerlach, T. P. Peixoto, and E. G. Altmann, "A network approach to topic models," CoRR, vol. abs/1708.01677, 2017.

[18] W. J. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman, "A survey of app store analysis for software engineering," IEEE Trans. Software Eng., vol. 43, no. 9, pp. 817–847, 2017.

[19] F. Palomba, P. Salza, A. Ciurumelea, S. Panichella, H. Gall, F. Ferrucci, and A. D. Lucia, "Recommending and localizing change requests for mobile apps based on user reviews," in IEEE/ACM 39th International Conference on Software Engineering (ICSE'17), 2017, pp. 106–117.

[20] G. Grano, A. Ciurumelea, S. Panichella, F. Palomba, and H. C. Gall, “Exploring the integration of user feedback in automated testing of android applications,” in IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER’18), 2018, pp. 72–83.