Original Article

App Review Prediction Using Machine Learning

S. Shiva Prakash[1], Dr.C. Siva Kumar[2]

[1]Assistant Professor, Department of Computer Science and Systems Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhrapradesh, India. E-mail: shivasthaneekam@gmail.com

[2]Assistant Professor, Department of Computer Science and Systems Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhrapradesh, India. E-mail: sivakumar.c@vidyanikethan.edu

Abstract

Millions of users utilize Google Play Store, the company's official android market. It can be utilized to access magazine, gaming, music, movie, and television material, among other things. Users can rate and review apps in the android market after installing them to express their own experiences with them, and this goes both ways in that other user can be inspired by other user's reviews. Users' experiences usually explain the app's usability, performance, and, in some cases, issues that they have encountered while using it. The goal is to categorise Google's app evaluations using machine learning approach. The Machine learning approaches are employed to gather element identify things such as emotions, element polarities, and to build a sample utilizing these characteristics, that could then go into a series of procedures including data validation or cleaning, visualization, and finally classification into positive, neutral, and negative sentiments.

Keywords: Algorithm Comparison, Classification Problems, Machine Learning (ML).

DOI:10.47750/pnr.2022.13.04.174

INTRODUCTION

Machine learning(ML) is a approach which predicts a Computers can find out about a specific collection of events or data without needing to be hard coded or manually designed [2]. Because of the principles of ML, including the execution of machine learning techniques in Python, machine learning is focused only with the development of algorithms that will adjust to new inputs.

As part of training and prediction phase, specialised algorithms are used. The training data is fed into an approach, which makes use of this to create forecasts supported by recent test results. Machine learning can be divided into the three they are: unsupervised learning, supervised learning and reinforcement learning. To gain access to data that should first be categorized by a human, a SL system is provided between two raw input and its associated labels.

Unsupervised learning doesn't have any labels. The algorithm for learning was given access to it. The input data must be grouped, hence this algorithm must figure out how to do so. Finally, reinforcement learning has a tangled relationship with its environment, delivering both criticism and praise to increase output. Data scientists uses a variety of Machine Learning methods to find trends in python that help them making meaningful observations.

These methods have been categorised as either supervised or unsupervised learning, depending on how they "read" data so on predict. Classification is that method of assuming the category of provided data points. This research seeks to develop a paradigm for learning algorithms is categorising android market [3] comment it might potentially restore adaptable SML learning categorization methods via predicting events with the simplest precision possible via using supervised algorithm comparison.

When learning is done under supervision, we can apply a method to identify the transformation matrix first from source to the destination, which is y = f. (X). The purpose is just to get close enough to the mapping works so effectively that after we already got new csv file (X) from this we just could forecast the independent variable (y) from this input. decision trees, Logistic regression, multi-class classification and other approaches fall within the category of supervised machine learning. SL demands that as the habituated to data for the algorithm's training is must already be labeled with correct answer. SL problems are subsets of Classification problems.

This purpose of this task is to development of something like a simple analysis that could predictive the amount of the such specified dependent characteristics based on other attribute variables. The two activities are identical save from one thing is that a proven the dependent's fact characteristic in categorical identification is numerical. A classification method seeks to derive a result using variables that have been evaluated. A classification method will attempt to assess the cost of one or even more outcomes provided one or even more input. The classification problem is when the outcome variable could be a category, such as "red" or "blue".

RELATED WORK

Scientific publications are complicated, and grasping their

utility necessitates prior knowledge. Peer reviews are expert opinions on a manuscript offered by specialists inside that area and contain a significant quantity of information, for not only the publisher's and chairperson to make that decision final, but also to appraise the paper's prospective effect. Under this research, we suggest that to extract meaningful information from scientific reviews using aspect-based sentiment analysis, which corresponds well for the admit choice.

Sentiment Analysis is a text analysis technique for rapidly recognising, evaluating, and studying a wide range with emotional states through the uses of the natural language. Applications that employ sentiment classification on the web and web-based online social networks, social service resources and inspections, and researcher response are examples of applications that apply marketing to customer management to clinical medication. Many websites, such as Amazon, encouraged consumers to leave product reviews on their sites. However, Amazon has a content limit for posting reviews.

With the rise of social media on the internet, sentiment analysis has become one of the most important research fields. Social media services such as Facebook and Twitter are used by millions of individuals to share their thoughts, ideas, expressions, feelings, and opinions. Sentiment analysis, often known as opinion mining, is concerned with the classification and prediction of people's attitudes about a certain topic. It entails categorising text documents or words according to whether the offered opinion is positive or negative regarding a particular issue. Although sentiment analysis appears to be similar to text categorization, it faces numerous problems that have prompted extensive research in this area. Various machine learning and lexicon-based algorithms have been developed in the literature to automate the sentiment analysis task. Despite their widespread use for sentiment categorization, these strategies have failed to deliver the best outcomes in terms of accuracy and resolution of all issues. As a result, new automated procedures must be developed to address all problems and provide the best results.

BACKGROUND

A. Machine Learning

Machine learning is being used to predict an outcome using past data. Machine learning (ML) is an AI technology that enables machines to understand without being explicitly programmed. Machine learning encompasses both the production of data-adaptive computer programmes, and the principles of algorithms, such as the development of something like a simple machine learning model in python. Sophisticated algorithm is employed as in learning and forecasting phase. It provides training information to an algorithms, it then utilized the trained data to forecast on new test data.

B. NLP

Machines can read and comprehend human language due to natural language processing (NLP). Natural-language interfaces design and direct effective learning of knowledge directly from sentient sources, such as newswire texts, it might possibly be doable with a sophisticated natural language processing system. Information extraction, text categorization, knowledge discovery, and translation software are some simple uses of natural language processing. To generate syntactic representations of text, many contemporary techniques use word co-occurrence frequencies. "Keyword spotting" search algorithms are popular and scalable, but they're also stupid; a search query for "dog" might only return pages that contain the literal word "dog" and ignore papers that contain the term "poodle." Lexical affinity techniques measure the emotion of a document by looking for words like "accident".

THE PROPOSED SCHEMES

The proposed plans and the system model will be presented within this part.

A. Analyse Exploratory Data

supervised classification using ml algorithms methods would be utilised the analyze a dataset and identify trends that will aid in categorizing comments, allowing apps to develop decent feature judgments going forward.

B. Data Manipulation

In this first import the file, verify to ensure cleanliness, then reduce and cleanse the catalog of inspection within the report's section.

C. Data Collection

There are two components in the data collection to be able to classify information provided: testing and training. In most cases, data is split into train and test sets using a 70:30 ratio. The train set is then loaded with the SMLT created Data Model, and validation set forecast is performed based on the test result precision. The comments are examined, demonstrating the utility of machine learning to separate the comments.

MODULE DESCRIPTION

The basic functionalities of the employed modules are described in this section. The following is a list of modules.

- Process of Data Validation
- Exploration of data analysis and visualisation
- Comparing Algorithms to Predictions in the Form of the Highest Accuracy
- The best accuracy result is obtained by combining RNN and LSTM.

- Sentiment prediction output using input sentence.

a) Data Validation Process

Validation approaches for machine learning are used to determine the failure the ML algorithm's rate, which is as follows feasible to the specific sample's error rate, replicate the value, resulting in the data form description to get whether there is a missing value that's a float variables or an integer variables. To offer a neutral reference point for tuning model hyper parameters, sample data is employed estimate on the system to suit on the actual training set. The verification sample is employed to test a model, although it isn't On a regular basis, it's used. Engineers working on machine learning used this data to control the modeling extraordinary parameters.

The processing of data can understand it, and it manages the data correctness, structure, analyze information take a long time. During the data recognition process, it is beneficial to understand the data and its qualities; the knowledge can assist you in deciding which algorithm to use to build your design. for ex, to show a dataset's data type format. The Pandas library is used for a variety of data cleaning jobs, with an emphasis on the most common data cleansing tasks, such as the ability to clean data more quickly, and missing values.

b) Exploration of Data Analysis and Visualisation

Data visualisation is a critical capability in applied analytics and machine learning. Objective data interpretations and estimations are the focus of statistics. Data visualization is a set of approaches for providing a qualitative analysis of the information. While exploring and getting to know a dataset, this might be helpful for finding new technologies, fake outliers, results, and other tidbits. Data visualizations can be utilised to express and exhibit crucial relationships in graphs and maps which are more intuitive and relevant to companies than signs of association or relevance based on restricted domain data.

Data cannot be understood until it is presented graphically, like those in graphs and charts. For both applied statistics and related machine learning the capacity to visualise data samples and other objects quickly is a valuable asset significant talent. It will show you how to better understand your data by using the many sorts of diagrams accessible when displaying data in Python.

Raw data is converted to clean data after preprocessing. To put it another way, data is gathered from variety of sources and processed in raw format, making interpretation difficult. The data must be appropriately organised in order to produce better results from the implemented Machine Learning model. Any machine learning a model is a collection of data in a specific format of set procedure that does not allow null values. As a result, in order to execute the random forest algorithm, all missing values from the initial input data collection are processed.

FALSE POSITIVE (FP): A defaulter is someone who has failed to pay a debt. If the intended class is yes, but the actual class is no. The actual class, for example, suggests that the passenger will survive, but the true class implies that an individual did not.

FALSE NEGATIVE (FN): A defaulter is more than likely a person who pays. When the class is really held higher than the projected class. If the passengers true class value indicates that they survived, yet the predicted class value predicts that this passenger would perish.

TRUE POSITIVE (TP): A defaulter is someone who just doesn't pay their debts. Both successfully forecast positive value, displayed the real class value as well as the coming class value. For example, if both the actual and predicited class values suggest that the passenger made it, then this passenger is alive.

TRUE NEGATIVE (TN): A defaulter is more than likely a payer. It's both precisely calculated negative number's, showing the real and projected classes have the same value. For instance, if the traveller did not survive in both the real and expected classes.

c) Comparing Algorithm with Prediction in the form of Best Accuracy Result

The following five algorithms are compared:

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest neighbour

The K-fold training and testing procedure is employed to judged all algorithms, this assures to that of identical divides into training dataset/example is made and also checks that every algorithms had been accessed with similar way. Separate the testing and learning set as well. By evaluating accuracy, it's feasible to predict the conclusion.

A linear model with potential predictors is frequently utilized in the logistic regression process to estimate a value. The anticipated value ranges from negative and positive infinity. By comparing the best accuracy, the logistic regression model produces a higher accuracy prediction result.

TPR= TruePositive / (TruePositive + FalseNegative)
FPR= FalsePositive / (FalsePositive + TrueNegative)

Accuracy Calculation

Accuracy = (TruePositive + TrueNegative) / (TruePositive + TrueNegative + FalsePositive + FalseNegative)

The vast straightforward and efficient statistic is precision, that is defined as the proportion of accurately outcome expectations to all occurrences. If the model's accuracy is high, one might believe it is the best. Only when the records are symmetrical as well as the rates of false positive and negative is mostly now equal is the accuracy metric relevant.

precision = TruePositive / (TruePositive + FalsePositive)
Recall = TruePositive / (TruePositive + FalseNegative)
Fmeasure = 2*TP / (2*TP + FP + FN)
F1score = 2*(Recall * Precision) / (Recall + Precision)

ALGORITHMS AND TECHNIQUES

Categorization as an ensemble learning strategy in which the computing algorithm studies from input data and after it apply what it has learnt to categorise new finds.

The dataset can have two or more classes. Recognition of handwriting and speech, biometric authentication, and other categorization issues are some examples. In Supervised algorithms educated from labelled results. After knowing the information, the algorithm evaluates whether a mark must be given to the newly information premised on the patterns and connecting the pattern onto the unlabeled new information.

Logistic Regression

It just the method of reviewing the dataset into the which a single or multiple distinct factors have impact on the outcome. The outcome is evaluated using a variable that is either true or false. The goal of the logistic regression is the to select the top model for explaining the association between such a number of autonomous factors and an interesting dualistic traits of interests.

Decision tree Algorithm

It just the well-known and highly effective method. A supervised learning method, the decision tree algorithm is classified as such.It can be used to create output variables that are both continuous and categorical. Hypotheses for decision trees we presume that the entire training data is the base at first. Characteristics are expected to just be categorical in regard to knowledge gain, however continuous characteristics are presumptions. Based on attribute values, entries are repeatedly allocated. It dismantles a large dataset into smaller subgroups over time while also building a decision tree.

K-Nearest Neighbour

This was the supervised computer learning approach this tracks every occurrences in multi-dimensional spaces that correspond to experimental data sets. Whenever unspecified separate data is collected, this examines the closest k preserved occurences and gives its most common in the estimation category, whereas legitimate data has to be returned the the k closest average neighbours. This uses the set of the designated areas as input andmakes use of them to instruct on its own to identify further. It polls the specified spots closest to it and polls its neighbours to mark a new point.

Random Forest Algorithm

It is just a regression approach or supervised classification that could develop a lot of decision trees and produce an outcome class based on the weighted average of all the trees. The algorithm is built on a system of ensembles training. Ensembles training is a method of creating a more efficient predictive model by integrating different versions of the same or comparable algorithms.

Support Vector Machine

It's just a classification system that sorts data into categories by finding the best hyperplane among the key points. This classification was selected because it is extremely versatile with relation to the number of various functions of kernelling that can be applied, and that also has a high prediction score. When this was initially invented in the 1990s, they were a huge hit, and they're still a must have tool for such a high-performing algorithms with no adjustment today.

RESULTS

The simulation results are basis on chi-square test of independence [1]. The findings show that the proposed classification method is accurate.

```
#Checking datatype and information about dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 37427 entries, 0 to 64230
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   App                     37427 non-null  object
 1   Translated_Review       37427 non-null  object
 2   Sentiment               37427 non-null  object
 3   Sentiment_Polarity      37427 non-null  float64
 4   Sentiment_Subjectivity  37427 non-null  float64
dtypes: float64(2), object(3)
memory usage: 1.7+ MB
```

Checking duplicate values of dataframe:

Figure 1: data information type

```
#Checking sum of missing values
df.isnull().sum()
```

```
App                     0
Translated_Review       0
Sentiment               0
Sentiment_Polarity      0
Sentiment_Subjectivity  0
dtype: int64
```

Spliting Train/Test:

Figure 2: Missing values

```
# plotting graph by length.
pasitive =df[df['Sentiment'] == 0]['Translated_Review'].str.len()
sns.distplot(pasitive, label='negative')
negative = df[df['Sentiment'] == 1]['Translated_Review'].str.len()
sns.distplot(negative, label='Neutral')
negative = df[df['Sentiment'] == 2]['Translated_Review'].str.len()
sns.distplot(negative, label='Positive')
plt.title('Distribution by Length')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x13fdbb4ba30>
```

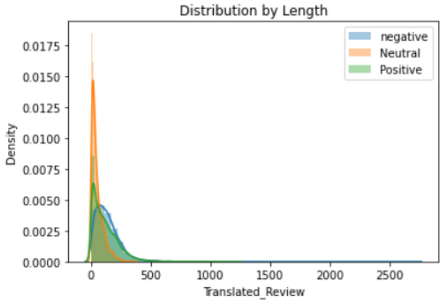

Figure 3: Density plot

```
# constructs the model with 128 LSTM units
model = get_model(tokenizer=tokenizer, lstm_units=128)
```

```
Reading GloVe: 400000it [00:52, 7629.91it/s]
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 100)          2208500

 lstm (LSTM)                 (None, 128)               117248

 dropout (Dropout)           (None, 128)               0

 dense (Dense)               (None, 3)                 387

=================================================================
Total params: 2,326,135
Trainable params: 117,635
Non-trainable params: 2,208,500
_________________________________________________________________
```

Figure 4: Neural networks building

```
In [20]: #plotting graph for distribution
         import matplotlib.pyplot as plt
         import seaborn as sns
         sns.countplot(x = "Sentiment", data = df)
         df.loc[:, 'Sentiment'].value_counts()
         plt.title('Distribution of  Sentiments')
```

```
Out[20]: Text(0.5, 1.0, 'Distribution of  Sentiments')
```

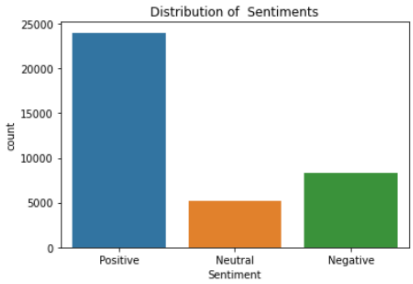

Figure 5: Visualize sentiment types on plot

```
Accuracy result of K-Nearest Neighbor is: 98.95805503606732


Classification report of K-Nearest Neighbor Results:

              precision    recall  f1-score   support

           0       0.99      0.97      0.98      2481
           1       0.99      1.00      0.99      8748

    accuracy                           0.99     11229
   macro avg       0.99      0.98      0.98     11229
weighted avg       0.99      0.99      0.99     11229

Confusion Matrix result of K-Nearest Neighbor is:
 [[2396   85]
 [  32 8716]]

Sensitivity :  0.9657396211205159

Specificity :  0.9963420210333791
```

Figure 6: KNN accuracy

```
Accuracy result of Naive bayes is: 97.80924392198771


Classification report of Naive bayes: Results:

              precision    recall  f1-score   support

           0       1.00      0.90      0.95      2481
           1       0.97      1.00      0.99      8748

    accuracy                           0.98     11229
   macro avg       0.99      0.95      0.97     11229
weighted avg       0.98      0.98      0.98     11229

Confusion Matrix result of Naive bayes: is:
 [[2235  246]
 [   0 8748]]

Sensitivity :  0.9008464328899637

Specificity :  1.0
```

Figure 7: Naive bayes accuracy

```
[20]: graph()
```

Figure 8: Reviews prediction

```
In [66]: text=str(input("enter the statement: "))
```

```
enter the statement: This help eating healthy exercise regular basis
```

```
In [67]: #text = "We stayed for a one night getaway with family on a thursday. Triple AAA "
         print(get_predictions(text))
```

```
Positive
```

```
In [68]: text = "Language barrier. I understand Korea. I want English."
         print(get_predictions(text))
```

```
Neutral
```

Figure 9: Input & output

Conclusion and Future Work

The analytical procedure starts from data cleaning and preparation, missing value, exploratory analysis and finally model creation and testing. The best accuracy on public

testing set is higher overall accuracy score will be find out. This tool can assist you in determining the Prediction of Google play store review classification prediction.

Future work of this app review prediction using machine learning is to review classification prediction to connect with cloud and to optimise the work for Artificial Intelligence implementation.

REFERENCES

Chisquared test of the independence. http://www.rtutor.com

Safvat Hussan, Cor Paul Bezemer, and Ahmed E. Hu ssan, "Studying Bad Updates of Top Free-To-Download Apps in the Google Play Store".

Pan Li, and Alexander Tuzhilin, "Learning Latent Multi Criteria Rating from User Reviews for Recommendations", Journal of latex class file, no. 8, volume. 14, Dec 2019.

T. Donker, B. Loep, and J. Ziegler, "Explaining recommendation by means of user reviews," in Proc. 1st Workshop Explainable Smart Sys (ExSS), 2018. [Online]. Available: https://ceur-ws.org/Vol2068/exss8.pdf

K. Moran, B. Li, C. Bemal-C´ardena, D. Jelf, and D. Poshyvank. Automated reporting of GUI design violations for mobile apps. In Proceedings of the 40th International Conference on Software Engineering, ICSE 18, 2018.

"The Mobile Marketer Guide to Apps Store Rating & Review," https://www.apptentive.com/blog/2015/05/05/

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

M. Ramu, Dr. Nazeer Shaik, "Study on Potential AI Applications in Childhood Education", International Journal of Early Childhood Special Education (INT-JECS), ISSN: 1308-5581 Vol 14, Issue 03 2022.

S. Shiva Prakash, Smt. Bingi Manorama Devi, Dr.P. Arulprakash, "Educating and communicating with deaf learner's using CNN based Sign Language Prediction System", International Journal of Early Childhood Special Education, ISSN: 1308-5581 Vol. 14, Issue 02, 2022.

K. Pujitha, R. Vidya," Semantic based search engine for real images and web url's using hypergraph distance measure algorithm", International journal of pure and applied mathematics,ISSN:1311-8080 Vol 115,issue no 6, 2017.

P. Yogendra Prasad, Dr. Dumpa Prasad, Dr.D Naga Malleswari,"Implementation of Machine Learning Based Google Teachable Machine in Early Childhood Education", International Journal of Early Childhood Special Education (INT-JECSE), ISSN: 1308-5581 Vol. 14, Issue 03 2022.