PROJECT : ARTIFICIAL INTELLIGENCE

Task 2 : Sentiment Analysis of Customer Reviews

DEVELOPMENT PHASE

DLBDSEAIS02

CEM OGUZ

32008124

Project steps in brief

- Data is collected and pre-processed with sub tasks mentioned in conception phase.
- After that, exploratory data analysis phase is processed to uncover patterns & insights.
- In this stage, quantity of ratings (1 to 5 stars) and their distributions is visualized with graphs.
- Then, data is split into 80 percent training and 20 percent test, which is then used for modelling.
- After the model is created & trained with 80 percent of the data, the sentiment labels for the remaining test data is predicted .
- Next step is validation : Predicted data results are compared with the test data to check if model's predictions are accurate. This value is printed as "accuracy".
- Hyperparameter tuning techniques is used to develop the final model, and accuracy is re-calculated.

Dataset & Tools & Links

Dataset is video game reviews on amazon found on Kaggle. Kaggle Dataset : Link

Project runs on Google Colab , and notebook file is uploaded to my Github profile : Link

- pandas for data analysis
- matplotlib for creating graphs
- nltk for analyzing text data and perform sub tasks
- wordcloud for visualizing text data to see patterns
- scikit-learn for machine learning processes
- Google Colab to create and run the project file

Data Pre-processing

To improve the accuracy of our model , the text data should be pre-processed to only keep the important features ; this is done by :

- Using NLTK
- Remove null values from the review body column.
- Removing stop words
- Removing punctuation
- Stemming and lemmatization to only keep the root of the words

Feature Extraction Benefits

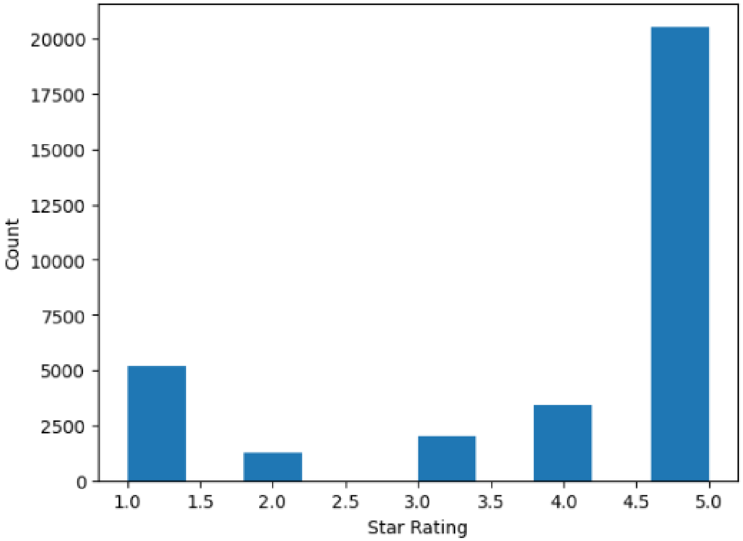
- Dimensionality reduction
- Improved model performance
- Increased interpretability
- Faster processing

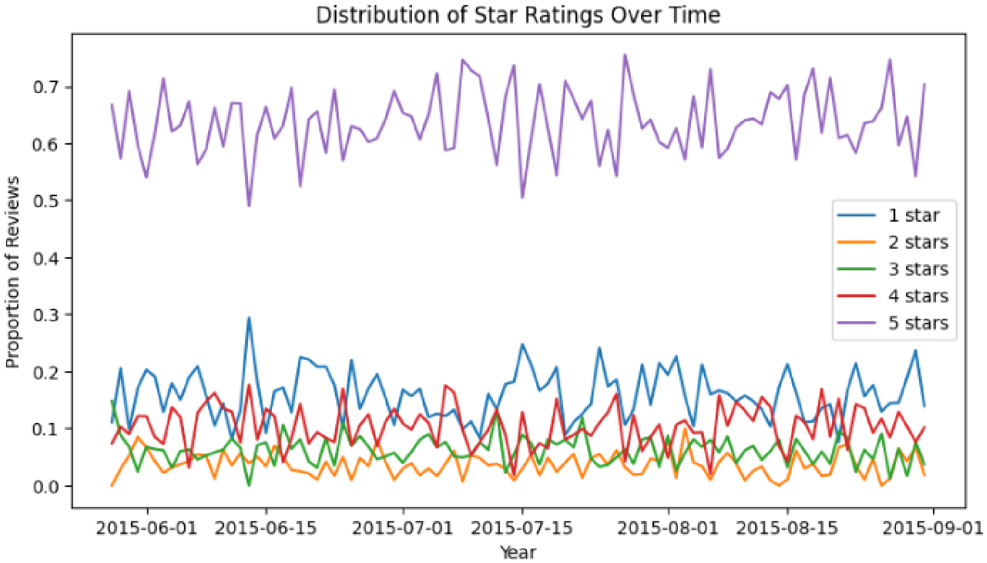
Overall, feature extraction is an important technique for data analysis and machine learning, and can provide significant benefits in terms of efficiency, accuracy, and interpretability.

Exploratory Data Analysis

- We create a word cloud to visualize the most common words used in the reviews before pre-processing.
- We plot a histogram of the star ratings to understand the distribution of ratings in the dataset.
- We convert the star ratings to sentiment labels and create a bar chart to visualize the sentiment distribution.

Results , Graphs and Visualizations





Data encoding and vectorization:

- We use LabelEncoder to convert the sentiment labels into numerical values.
- We split the data into training and testing sets.
- We use TfidfVectorizer to convert the text data into a numerical matrix for use in the machine learning algorithm.

Modelling

- Models used for sentiment analysis: Random Forest
- Random forest is one of the most popular tree-based supervised learning algorithms. It is also the most flexible and easy to use
- The algorithm can be used to solve both classification and regression problems. Random forest tends to combine hundreds of decision trees and then trains each decision tree on a different sample of the observations.
- The final predictions of the random forest are made by averaging the predictions of each individual tree.
- Why not decision trees ? : Decision trees are highly sensitive to training data which could result in high variance, so a model might fail to generalize.
- Other Advantages are : reduced overfitting, improved accuracy, easiness to tune, robustness to outliers.

Hyperparameter Tuning

- In the last step , we tune the model and expect a higher accuracy. To do that :
- We define a range of hyperparameters to search over using GridSearchCV.
- We use GridSearchCV to find the optimal hyperparameters for the Random Forest Classifier.
- We print the best hyperparameters and use them to predict sentiment labels for the test data.
- We calculate the accuracy score using the best hyperparameters.

Results and discussion:

- We achieve an accuracy score of 85,1% using the Random Forest Classifier with default hyperparameters.
- We achieve an accuracy score of 85,2% using the Random Forest Classifier with optimized hyperparameters.
- The optimized hyperparameters are: n_estimators=100, max_depth=None, min_samples_split=5, min_samples_leaf=1.
- We can conclude that the sentiment analysis model performs reasonably well in predicting the sentiment of the Amazon digital video games reviews.

Thank you.