CUSTOMER REVIEW CLASSIFICATION USING NATURAL LANGUAGE PROCESSING

DATA 606 – DATA CAPSTONE IN DATA SCIENCE

Project Members:

- BHANU HARISH SURISETTI (IO02971)
- JESSIE CAROLINE MERUGU (UW01251)
- SREEJA PENDOTA (XU32613)

Project Idea:

The objective of the project is to use NLP and machine learning techniques to predict customer satisfaction with products based on their reviews. The dataset contains 71,044 rows and 25 columns, including the product ID, brand, category, review text, and whether the reviewer purchased the product or not. The project will use various classifiers such as Logistic Regression, Decision Tree, K Nearest Neighbor, and Support Vector Machine to predict the outcome, with the maximum accuracy obtained so far being 70%. The project also involves data preprocessing to remove null and duplicate rows, and exploratory data analysis (EDA) to analyze the correlation between review rating and the number of words used in the review text, as well as to visualize which words are used most and check if there is any regional impact on the reviews provided. The project also plans to use future classifiers such as Naïve Bayes Classifier, KNN, and XLNet to improve accuracy. Overall, the project seeks to use NLP and machine learning to better understand customer satisfaction and to identify factors that influence it.

Data Source: https://www.kaggle.com/code/duttadebadri/detailed-nlp-project-prediction-visualization/data

Size of Dataset: 99.44MB

Data Set Columns:

id	ID of the product
brand	Product Brand
categories	category group, in which the product can be used
dateAdded	Date which the review is written
dateUpdated	Final review updated date
ean	Produce unique bar code
keys	Unique key of the review
manufacturer	Manufacturer of the product
manufacturerNumber	Manufacturer number
name	Name of the product
reviews.date	Date which review is given
reviews.dateAdded	Date added
reviews.dateSeen	Review seen date

reviews.didPurchase	Did the reviewer really purchased the product
reviews.doRecommend	Is the reviewer recommending the product
reviews.id	Reviewers ID
reviews.numHelpful	Is the reviewers review helpful
reviews.rating	Reviewers rating for the product
reviews.sourceURLs	Product review URL
reviews.text	Product complete review
reviews.title	Review Title
reviews.userCity	Reviewers City Place
reviews.userProvince	Reviewers State
reviews.username	Reviewer Name
upc	Product number which is nothing but product barcode

In this project, the main feature columns would be review.text. And the target variable is review.didpurchase.

Existing Regressions Used: Random Forest, XGBoost

Reference: https://www.projectpro.io/article/machine-learning-nlp-text-classification-algorithms-and-models/523