Women’s Clothing E-Commerce Reviews Analysis

Introduction

The project focuses on analysing customer reviews from a Women’s Clothing E-Commerce dataset. It aims to derive insights from review texts, classify them as recommendations or not, and understand customer preferences and sentiments. The dataset comprises 23,486 reviews, each with several features like age, review text, rating, and recommendation status.

Data Preparation

The initial phase involved cleaning and preprocessing the dataset. Unnecessary columns like 'Age', 'Unnamed: 0', 'Clothing ID', 'Rating', 'Positive Feedback Count', 'Title', and 'Unnamed: 0' were dropped to focus on the textual review and recommendation indication. Missing values in 'Class Name', 'Division Name', and 'Department Name' were handled by dropping such entries.

Feature Engineering

A new feature, 'Review_length', was created to analyse the length of each review. The recommendation indicator was converted from binary to a string format for better interpretability ('Positive' for 1 and 'Negative' for 0). The text was further processed by removing non-alphanumeric characters, converting to lowercase, and removing stop words.

Exploratory Data Analysis

Descriptive statistics were used to understand the review lengths. The longest and shortest reviews were identified for insights. A word cloud was generated to visualize the most frequent words in the reviews, providing a graphical representation of the data's textual elements.

Model Building and Evaluation

Two models, Logistic Regression and Random Forest, were trained on the processed text. The text data was vectorized using CountVectorizer, transforming it into a format suitable for model input. The models were evaluated based on accuracy, with Logistic Regression achieving an accuracy of approximately 88% and Random Forest around 86%.

LSTM Neural Network

An LSTM (Long Short-Term Memory) model was also developed to handle the sequence nature of text data. The model consisted of an embedding layer, LSTM layer, and a dense layer with sigmoid activation. It was trained on tokenized and padded sequences of the reviews. The model achieved an accuracy of around 89% on the test set.

Predictions on Unseen Data

The models were tested on new textual data to predict their recommendations. Different outcomes from the models highlighted the varied interpretations and responses to the unseen data.

Challenges and Learnings

Data Cleaning: The initial challenge was cleaning and preprocessing the dataset, requiring decisions on which features to keep or drop.

Text Processing: Implementing text processing techniques like tokenization and stop words removal was critical to prepare the data for modelling.

Model Selection and Tuning: Choosing and tuning the right models to handle textual data was a significant learning curve.

Interpreting Results: Understanding and interpreting the model outputs, especially the differences in predictions between models, provided insights into their strengths and weaknesses.

Future Scope

Future enhancements can include integrating more sophisticated NLP techniques, experimenting with other deep learning architectures, and exploring sentiment analysis for a more nuanced understanding of customer reviews.

Industry Application

This project is highly relevant in the e-commerce industry, particularly for online clothing retailers. The insights from this analysis can help in:

1. Understanding customer sentiments towards products.
2. Enhancing product recommendations based on customer reviews.
3. Improving inventory and marketing strategies based on popular items and customer feedback.

Providing valuable insights for customer service to address common complaints or praises mentioned in reviews.

Overall, this project demonstrates the application of machine learning and deep learning techniques in analysing customer reviews, which is an asset in customer relationship management and business strategy in the e-commerce sector.