Jairaj Tikam jpt1342
Jyotsna Nakte jnn2078

MILESTONE 01: PROJECT PROPOSAL

Title: Sentiment Classification and Prediction of Women's Apparel Reviews

Team members:
Jairaj Tikam (jpt1342)
Jyotsna Namdeo Nakte (jnn2078)

1.] Description of the problem:

Today, while shopping online in any department for grocery, apparel, electronics from sites like Amazon, eBay, Flipkart, 90% of the customers first look for the ratings and reviews of the product and then make an informed decision. More often than not in the E-Commerce business, the popularity of the product depends on the recommendations and reviews given by other users. The social networking sites play an essential role amongst the customers to put forward their opinions, share perspectives, and reviews about these products. Clothing is prominent sector where user opinion matters the most to increase the revenue of sellers and provide the customers with a quality product. User sentiments towards the product help improve the product alongside seller revenue. The data available on social media by legit users can be used for information retrieval. Extracting information from raw data about the clothes can prove beneficial to the seller in many ways. While analyzing and studying the data, we could find variables/features in the dataset, which could help for better customer experience. The massive amount of data available poses a problem for humans to sort clothes types according to customer likings, tagging comments manually, recommending to various customers, analyzing the reviews by users for seller and users benefits. The whole process is cumbersome because of the amount of data being posted by individual users. The project aims to build an automated system which classifies the user rating using sentiment analysis which helps to predict the likelihood of a product being recommended by the customer to someone else based on the product review.

2.] Function of the system:

The system proposed works towards tagging the reviews based on user rating and perform recommendation and sentiment classification of the review. The sentiment classification on the data based on reviews helps for the prediction of recommendation. Multivariate, univariate analysis, and descriptive statistics would be performed on the data for exploration. The system even performs text classification and analysis for natural language processing by performing word frequency distribution for the recommendation, information retrieval Using LSA(Latent Semantic Analysis), finding most commonly used words for the positive and high rating, tf-idf matrix for text analysis. The system would use one core algorithm Naive Bayes classification to build the model on training data. Other Algorithms like Random Forest Classifier, Logistic Regression, Linear SVC, Neural Network, would be used on testing data for comparative analysis of results based on accuracy, precision, recall, f1-score, and support.

Jairaj Tikam jpt1342
Jyotsna Nakte jnn2078

3.] Need of system, use and benefits:

The system is utilized for sentiment classification and prediction of recommendation. The former system would be helpful for both the sellers and customers. It would help improve the type of clothes to produce for selling, customers to buying, and revenue of the sellers. It is useful to discover the connection between the age group of reviewers according to the type of clothes sold for better business strategy. The system assists customers to discover clothes based on recommendations by other user reviews. The importance of various factors like age group, type of clothes, clothing department in the data is realized for prediction or recommendation for revenue gains. It benefits to find relations between the words used for reviews and ratings. Based on reviews, the sellers have better knowledge about their product as well as users.

4.] About the data and candidate core algorithms:

The dataset chosen to perform analysis is Women's E-Commerce Clothing Reviews from Kaggle. The dataset contains 10 feature variables with 23486 records of different customer ratings. The features are Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, Class Name.

Core Algorithm: Naive Bayes Classification

Other Algorithm for Comparative Analysis:

Random Forest Classifier
Logistic Regression
Linear SVC
Neural Network

Dataset Link: https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews