Natural Language Processing Project

III-II B.Tech (Artificial Intelligence & Machine Learning)

Malla Reddy University

Sentiment Analysis on Customer Reviews

Abstract

Sentiment analysis is a Natural Language Processing (NLP) task aimed at identifying and classifying the sentiment expressed in textual data. This project focuses on analyzing customer reviews to determine whether they convey positive, negative, or neutral sentiments. The process involves text preprocessing techniques such as cleaning, tokenization, lemmatization, and vectorization, followed by the development and evaluation of machine learning models for classification. The outcome of this project provides insights into customer opinions, aiding businesses in decision-making and improving customer satisfaction.

Methodology

1. Data Collection

Source: Use publicly available datasets such as Kaggle's customer review datasets or scrape data from e-commerce websites.

2. Data Preprocessing

a. Cleaning the Text:

Remove special characters, numbers, and punctuation.

Convert text to lowercase.

Remove stopwords (e.g., "is," "the," "and").

b. Tokenization:

Split the text into individual words or tokens.

c. Lemmatization:

- Reduce words to their base or root form (e.g., "running" -> "run").

d. Vectorization:

- Convert textual data into numerical form using methods like:
    - Bag of Words (BoW)
    - Term Frequency-Inverse Document Frequency (TF-IDF)
    - Word Embeddings (e.g., Word2Vec, GloVe, or BERT).

3. Model Development

a. Train-Test Split:

- Split the dataset into training and testing sets (e.g., 80% training, 20% testing).

b. Model Selection:

- Use classification algorithms like:
    - Logistic Regression
    - Support Vector Machines (SVM)
    - Naïve Bayes
    - Random Forest
    - Neural Networks (if using deep learning).

4. Model Training

- Train the model using the preprocessed data.

5. Model Evaluation

- Evaluate the model's performance using metrics such as:
    - Accuracy
    - Precision
    - Recall
    - F1-score
- Visualize performance using a confusion matrix.